sustainability

Knowledge Management, Innovation and Big Data

Implications for Sustainability, Policy Making and Competitiveness

Edited by

Patricia Ordóñez de Pablos and Miltiadis D. Lytras

Printed Edition of the Special Issue Published in *Sustainability*

www.mdpi.com/journal/sustainability

MDPI

Knowledge Management, Innovation and Big Data

Knowledge Management, Innovation and Big Data

Implications for Sustainability, Policy Making and Competitiveness

Special Issue Editors

Patricia Ordóñez de Pablos
Miltiadis D. Lytras

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade

Special Issue Editors

Patricia Ordóñez de Pablos
The University of Oviedo
Spain

Miltiadis D. Lytras
The American College of Greece
Greece

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Sustainability* (ISSN 2071-1050) from 2017 to 2018 (available at: https://www.mdpi.com/journal/sustainability/special_issues/Knowledge_data_sustainability).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Article Number*, Page Range.

ISBN 978-3-03928-008-7 (Pbk)
ISBN 978-3-03928-009-4 (PDF)

Contents

About the Special Issue Editors

Patricia Ordóñez de Pablos is a professor in the Department of Business Administration and Accountability in the Faculty of Economics at the University of Oviedo (Spain). Her teaching and research interests focus on the areas of strategic management, knowledge management, intellectual capital, and China. She serves as an Associate Editor for the Behaviour and Information Technology journal. Additionally, she is Editor-in-Chief of the International Journal of Learning and Intellectual Capital (IJLIC) and the International Journal of Strategic Change Management (IJSCM). She is also Editor-in-Chief of IGI Global's International Journal of Asian Business and Information Management (IJABIM), as well as editor for a number of IGI Global book publications and full book series.

Miltiadis D. Lytras is an expert in advanced computer science and management as well as an editor, lecturer, and research consultant, with extensive experience in academia and the business sector in Europe and Asia. Dr. Lytras is Research Professor at Deree College—The American College of Greece and Distinguished Scientist at the King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia. Dr. Lytras is a world-class expert in the fields of cognitive computing, information systems, technology-enabled innovation, social networks, computers in human behavior, and knowledge management. In his work, Dr. Lytras seeks to bring together and exploit synergies among scholars and experts committed to enhancing the quality of education for all.

Preface to ”Knowledge Management, Innovation and Big Data”

The evolution of knowledge management theory and the special emphasis on human and social capital sets new challenges for knowledge-driven and technology-enabled innovation. Emerging technologies including big data and analytics have significant implications for sustainability, policy making, and competitiveness. This edited volume promotes scientific research into the potential contributions knowledge management can make to the new era of innovation and social inclusive economic growth. We are grateful to all the contributors of this edition for their intellectual work. The organization of the relevant debate is aligned around three pillars:

SECTION A. DATA, KNOWLEDGE, HUMAN AND SOCIAL CAPITAL FOR INNOVATION

We elaborate on the new era of knowledge types and the emerging forms of social capital and their impact on technology-driven innovation. Topics related to social networks, sustainable smart education and the role of social capital in corporate innovation are discussed. Disruptive innovation and knowledge integration in business processes are analyzed in combination with knowledge sharing strategies for enhanced decision-making. The study of knowledge creation processes and sustainable competitive advantage is also analyzed in the terms of technological innovation capabilities. We also promote transformational training programs and quality orientation of employees. We conclude that significant investment on transformative education and training is required.

SECTION B. KNOWLEDGE MANAGEMENT & BIG DATA ENABLED INNOVATION

In this section, various advanced research related to applications and systems of knowledge management and big data enabled innovations are presented. Selected topics include:

- Crowdsourcing Analysis of Twitter Data
- Document Management through Natural Language Processing
- Data Governance Taxonomy
- Knowledge Extraction Models
- Ontology Design on Heterogeneous Architectures
- Semantic Modeling of Administrative Procedures
- Visualizing the Intellectual Structure and Hotspots of Big Data Research

SECTION C. SUSTAINABLE DEVELOPMENT

In the concluding section of this edition, the debate on the impact of knowledge management and big data research on sustainability is promoted with integrative discussion of complementary social and technological factors including:

- Impact of Big Social Network Data on Sustainable Economic Development
- Exploring the Technological Collaboration Characteristics for enhanced innovation in high-tech industries
- Business Intelligence Issues for Sustainability

- Modeling User Acceptance of Personalized Business Modes

We want to thank the professional staff of MDPI for their qualitative work that made this edition possible. We are at your disposal for any further information on this edition and we invite you to join us in our next editions on the topics of knowledge management, artificial intelligence, and sustainability.

Patricia Ordóñez de Pablos, Miltiadis D. Lytras
Special Issue Editors

sustainability

MDPI

Editorial

Knowledge Management, Innovation and Big Data: Implications for Sustainability, Policy Making and Competitiveness

Patricia Ordóñez de Pablos [1] and Miltiadis Lytras [2,3,*]

[1] Faculty of Economics and Business, The University of Oviedo, Avda del Cristo, s/n. 33071 Oviedo-Asturias, Spain; patriop@uniovi.es
[2] Deree College—The American College of Greece, 153 42 Aghia Paraskevi, Greece
[3] Effat College of Engineering, Effat University, P.O. Box 34689, Jeddah, Saudi Arabia
* Correspondence: mlytras@acg.edu; Tel.: +30-210-6009800

Received: 7 June 2018; Accepted: 16 June 2018; Published: 19 June 2018

Abstract: This Special Issue of Sustainability devoted to the topic of "Knowledge Management, Innovation and Big Data: Implications for Sustainability, Policy Making and Competitiveness" attracted exponential attention of scholars, practitioners, and policy-makers from all over the world. Locating themselves at the expanding cross-section of the uses of sophisticated information and communication technology (ICT) and insights from social science and engineering, all papers included in this Special Issue contribute to the opening of new avenues of research in the field of innovation, knowledge management, and big data. By triggering a lively debate on diverse challenges that companies are exposed to today, this Special Issue offers an in-depth, informative, well-structured, comparative insight into the most salient developments shaping the corresponding fields of research and policymaking.

Keywords: big data; competitive advantage; disruptive innovation; human capital; innovation; knowledge management; sustainability

1. Overview of the Special Issue

Knowledge Management and Big Data is a new paradigm for the integration of Internet Technology in the human and machine contexts. Now, we are able to transform raw data that are massively produced by humans and machines into knowledge and wisdom capable of supporting disruptive innovation, smart decision making, innovative services, and new business models. This Special Issue explores the role of knowledge management strategies and tools to enhance the power of big data and help decision makers in today's competitive economy. The analysis of these key issues on the interrelations among knowledge management, big data, and information technology will provide new knowledge and perspectives towards a deeper understanding of their impact on companies, economies, and societies today.

The Special Issue opens with a paper by Shan et al. titled "Internal Social Network, Absorptive Capacity and Innovation: Evidence from New Ventures in China" [1]. The authors analyze "the impact of the internal social network on new venture's innovation by building a comprehensive structural equation modeling (SEM) that integrates three streams of research: internal social network, innovation, and absorptive capacity". In their research, they use a sample of 279 new ventures from China.

Wang et al., in their paper titled "Visualizing the Academic Discipline of Knowledge Management" [2], study "the research status of knowledge management (KM) and identify the characteristics of KM in the literature. We selected and studied in detail 7628 original research articles from the Web of Science from 1974 to 2017".

In the paper titled "An Empirical Study on Visualizing the Intellectual Structure and Hotspots of Big Data Research from a Sustainable Perspective" [3], Hu et al. conducted a bibliometric study of big data literature from Web of Science (WoS) for the period 2002 to 2016, involving 4927 effective journal articles in 1729 journals contributed by 16,404 authors from 4137 institutions. The bibliometric outcomes show "the current annual publications distribution, journals distribution and co-citation network, institutions distribution and collaboration network, authors distribution, collaboration network and co-citation network, and research hotspots. The results can help researchers worldwide to understand the panorama of current big data research, to find the potential research gaps, and to focus on the future sustainable development directions".

López et al., in their paper titled "Semantic Modeling of Administrative Procedures from a Spanish Regional Public Administration" [4], explore how to improve public administration open data initiatives and help to develop their sustainability policies, such as improving decision-making procedures and administrative management sustainability. Using the methodology of a case study, the authors modelled public administrative processes and files in collaboration with a Regional Public Administration in Spain, the Principality of Asturias, which enabled access to its information systems.

Recognizing that little is known about the underlying relationship dynamics among the variables of transformational training programs, employee loyalty, and quality orientation of employees in the context of higher education institutions, Al Qudah et al. in their paper "Transformational Training Programs and Quality Orientation of Employees: Does Employees' Loyalty Matter?" [5] decide to explore the interplay of these variables. The outcomes of the research show "that both direct and indirect effects of transformational training programs on quality orientation of employees were significant. More specifically, the positive effects that transformational training programs have on quality orientation of employees are through employee loyalty. This finding significantly advances the existing body of knowledge and implies that transformational training programs enhance employees' loyalty which, in turn, escalates employees' orientations towards quality".

The paper titled "Ontology Design for Solving Computationally-Intensive Problems on Heterogeneous Architectures" [6] by Faheem et al. develops an ontology and shows "how we can use it to solve computationally-intensive problems from various domains. As a potential use for the idea, we present examples from the bioinformatics domain. Validation by using problems from the Elastic Optical Network domain has demonstrated the flexibility of the suggested ontology and its suitability for use with any other computationally-intensive problem domain".

Zhao et al., in their paper "Modeling and Quantifying User Acceptance of Personalized Business Modes Based on TAM, Trust and Attitude" [7], study the main factors influencing user acceptance of personalized business modes. The authors present a "research model that enhances the TAM (Technology Acceptance Model) model with trust and attitude to depict the influence from several variables to user acceptance of personalized business modes. Further, we use the structural equation method to conduct an empirical analysis on questionnaire data from the Internet. The results in terms of many kinds of data analysis show that trust and the TAM factors (perceived usefulness and perceived ease of use) have significant influence on user acceptance of personalized business modes".

In the paper titled "Existing Knowledge Assets and Disruptive Innovation: The Role of Knowledge Embeddedness and Specificity" [8], Lin et al. study the impact of "knowledge assets on disruptive innovation by analyzing the role of knowledge embeddedness and specificity. They conducted a hierarchical regression analysis by using survey data from 173 Chinese industrial firms to test the direct and indirect effects of knowledge embeddedness and specificity on disruptive innovation, which can be divided into outward-oriented and internal-oriented disruptive innovation. The results indicated that knowledge embeddedness not only played a positive role in knowledge specificity, but also had a positive effect on outward-oriented disruptive innovation".

In the paper "Business Intelligence Issues for Sustainability Projects" [9], Muntean argues that business intelligence is a term that encompasses strategies, technologies, and information systems used by companies and organizations "to extract from large and various data, according to the value chain, relevant knowledge to support a wide range of operational, tactical, and strategic business decisions. Sustainability, as an integrated part of the corporate business, implies the integration of the new approach at all levels: business model, performance management system, business intelligence project, and data model. Both business intelligence issues presented in this paper represent the contribution of the author in modeling data for supporting further BI approaches in corporate sustainability initiatives".

Liu et al., in the paper titled "Exploring the Technological Collaboration Characteristics of the Global Integrated Circuit Manufacturing Industry" [10], study the international technological collaboration characteristics of the integrated circuitry manufacturing industry based on patent analysis. The authors analyze "four aspects, which include collaboration patterns, collaboration networks, collaboration institutions, and collaboration impacts, by utilizing patent association analysis and social network analysis. The findings include the following: first, in regard to international technological collaboration, the USA has the highest level, while Germany has great potential for future development; second, Asia and Europe have already formed clusters, respectively, in the cooperative network; last, but not least, research institutions, colleges, and universities should also actively participate in international collaboration".

Hu et al., in their paper titled "A Hierarchical Feature Extraction Model for Multi-Label Mechanical Patent Classification" [11], developed "a hierarchical feature extraction model (HFEM) for multi-label mechanical patent classification, which is able to capture both local features of phrases as well as global and temporal semantics. First, a n-gram feature extractor based on convolutional neural networks (CNNs) is designed to extract salient local lexical-level features. Next, a long dependency feature extraction model based on the bidirectional long–short-term memory (BiLSTM) neural network model is proposed to capture sequential correlations from higher-level sequence representations. Then the HFEM algorithm and its hierarchical feature extraction architecture are detailed. We establish the training, validation and test datasets, containing 72,532, 18,133, and 2679 mechanical patent documents".

In the paper "Data Governance Taxonomy: Cloud versus Non-Cloud" [12], Al-Ruithe et al. admit that the only way to solve data problems is the implementation of effective data governance. The authors propose that "a taxonomy approach to define the different attributes of data governance is expected to make a valuable contribution to knowledge, helping researchers and decision makers to understand the most important factors that need to be considered when implementing a data governance strategy for cloud computing services. In addition to the proposed taxonomy, the paper clarifies the concepts of data governance in contracts with other governance domains".

Yu et al., in the paper "Knowledge Creation Process and Sustainable Competitive Advantage: the Role of Technological Innovation Capabilities" [13], study "the relationship between the knowledge creation process and technological innovation capabilities, and analyzes their effect on a firm's sustainable competitive advantage using a knowledge-based view theoretical framework. We conduct structural equation modeling analyses using survey data from 315 Chinese industrial firms to test the direct and indirect effects of the knowledge creation process on sustainable competitive advantage. Technological innovation capabilities—operationalized to reflect the dimensions of process innovation capability and product innovation capability—are used as the mediating variable for explaining the relationship between the knowledge creation process and sustainable competitive advantage. The results indicate that the knowledge creation process does not have a significant direct effect on sustainable competitive advantage".

Wu at al., in their paper titled "Top Management Teams' Characteristics and Strategic Decision-Making: A Mediation of Risk Perceptions and Mental Models" [14], propose that "strategic decision-making is a key factor of sustainability and development in enterprises. Moreover, the top management team (TMT) of an enterprise constitutes the base for decision-making. This study employed structural equation modeling to analyze questionnaires regarding TMTs' characteristics and strategic decision-making, and tested the mediating effects of risk perceptions and mental models and the moderating effects of psychological ownership. We investigated 289 valid questionnaires on TMTs completed by representatives from enterprises in China and found risk perceptions and mental models that serve as a mediating factor and are affected by the TMTs' characteristics and decision-making".

Roh et al. know the importance of text mining in patent analysis but also recognize its limitations. In their paper titled "Developing a Methodology of Structuring and Layering Technological Information in Patent Documents through Natural Language Processing" [15], the authors "structure meaningful keyword sets related to technological information from patent documents; then we layer the keywords, depending on the level of information. This research involves two steps. First, the characteristics of technological information are analyzed by reviewing the patent law and investigating the description of patent documents. Second, the technological information is structured by considering the information types, and the keywords in each type are layered through natural language processing. Consequently, the structured and layered keyword set does not omit useful keywords and the analyzer can easily understand the meaning of each keyword".

Can and Alatas, in the manuscript titled "Big Social Network Data and Sustainable Economic Development" [16], discuss that new information technologies contributed significantly to the rapid and effective growth of social networks. They propose that "the immediate or unpredictable effects of a wide array of economic activities on large masses and the reactions to them can be measured by using social media platforms and big data methods. Thus, it would be extremely beneficial to analyze the harmful environmental and social impacts that are caused by unsustainable business applications".

In the paper titled "Crowdsourcing Analysis of Twitter Data on Climate Change: Paid Workers vs. Volunteers" [17], Kirilenko et al. address the importance of web-based crowdsourcing for environmental data processing. The authors developed a study that compares "volunteer and paid processing of social media data originating from climate change discussions on Twitter. The same sample of Twitter messages discussing climate change was offered for processing to the volunteer workers through the Climate Tweet project, and to the paid workers through the Amazon MTurk platform. We found that paid crowdsourcing required the employment of a high redundancy data processing design to obtain quality that was comparable with volunteered processing. Among the methods applied to improve data processing accuracy, limiting the geographical locations of the paid workers appeared the most productive".

Zheng et al., in their paper titled "Impacts of Leadership on Project-Based Organizational Innovation Performance: The Mediator of Knowledge Sharing and Moderator of Social Capital" [18], explore the importance of innovation in the sustainable development of construction projects. In their paper, they study "various effects of different types of leadership on innovation performance in a construction project-based organization. Therefore, a theoretical model was constructed to explore the mediation mechanism and boundary condition of different types of leadership to improve innovation. The theoretical model was validated with empirical data covering project managers and engineers from the project-based organization in China via regression analysis and path analysis. The results show that transformational leadership and transactional leadership have some positively significant effects on knowledge sharing and innovation performance. Meanwhile, knowledge sharing partially mediates the relationship between transformational leadership and/or transactional leadership and innovation performance. Additionally, by considering different levels of social capital, transformational leadership is likely to have a strong positive impact on innovation performance through knowledge sharing".

García-Alcaraz et al., in the paper "Role of Human Knowledge and Communication on Operational Benefits Gained from Six Sigma" [19], focus on the implementation of a production philosophy called Six Sigma (SS). The paper proposes a "structural equation model integrating those aspects as latent variables and relating them with ten hypotheses. Data for hypothesis validation were gathered among 301 manufacturing companies, and assessed using partial least squares (PLS) to estimate direct, indirect, and total effects. As results, we found that access to reliable information, trusted analysis and knowledgeable management are crucial for SS implementation at the problem definition stage. Likewise, to execute and control SS projects, it is important to be trained in statistical techniques through clear didactic materials".

The paper "What Makes Firms Innovative? The Role of Social Capital in Corporate Innovation" [20] by Ahn and Kim developed a social capital explanation for the analysis of the relationship between human capital investment and the organizational innovation capability. The authors propose that "social capital plays a mediating role in the relationship between the level of individual knowledge of employees and organizations' innovation capabilities. The mediating mechanism is attributed to the role of social capital in knowledge exchange and combination that help enhance knowledge creation. They used a survey data of 319 manufacturing firms in Korea".

2. Conclusions

The excellent contributions from all over the world included in this Special Issue highlight diverse issues and topics that form the evolving field of knowledge management, big data, and innovation for the competitiveness of companies and economies. Thanks to this Special Issue, a substantial streamlining of the interrelation between big data and knowledge as triggers of innovation has been promoted. To capitalize on that work today, the imperative is to move to the next stage of the debate and explore more complex interrelations between these topics.

The overall contribution of this collection is the promotion of an integrative approach to the newly emerging, challenging research area of Big Data, Knowledge Management, and Innovation. What is evident from the relevant discussion is the promotion of four key aspects of novelty:

1. The fast developments in Artificial Intelligence and Machine Learning approaches.
2. The diffusion of numerous data and text mining techniques including sentiment analysis and social engineering.
3. The critical contribution of soft skills, and knowledge management process models to the efficiency of data-driven tasks and procedures.
4. The capacity of Big Data to support a multidisciplinary scientific intervention to real world social problems.

The new era of Big Data, Knowledge Management, and Innovation integration will be policy driven at a higher level of abstraction, where socially-inclusive economic growth and sustainability will be considered as top priorities. The multidisciplinary character of the phenomenon will be based on a critical and radical diffusion of Smart Machines and Artificial Intelligence [20–26] with a new Data–Knowledge–Wisdom ecosystem [See: Figure 1]; where knowledge artifacts and human and social entities will interact through new business models and applications powered by numerous new technologies such as augmented virtual reality, internet of things, sensors and 5G, cloud computing, anticipatory computing, and so on.

As a continuation of this Special Issue, we are planning a new Special Issue on Artificial Intelligence and Smart Machines for Sustainable Innovation Models. We would be happy to receiving the Sustainability community comments and feedback on the published research of this special issue.

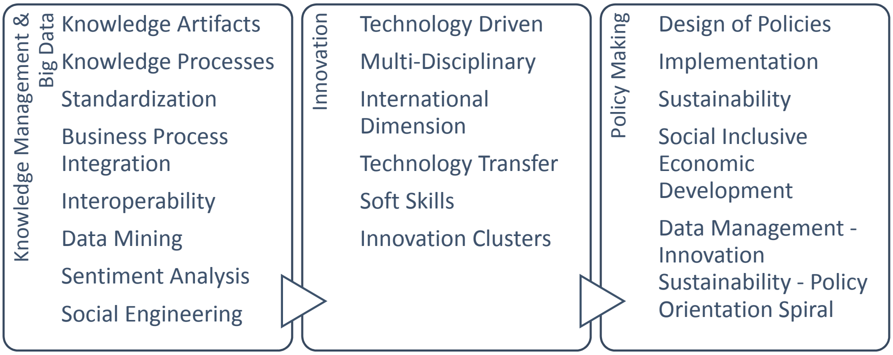


Figure 1. The Integrative KM, Innovation, Policy Making Research Landscape (source: The authors).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shan, W.; Zhang, C.; Wang, J. Internal Social Network, Absorptive Capacity and Innovation: Evidence from New Ventures in China. *Sustainability* **2018**, *10*, 1094. [CrossRef]
2. Wang, P.; Zhu, F.; Song, H.; Hou, J.; Zhang, J. Visualizing the Academic Discipline of Knowledge Management. *Sustainability* **2018**, *10*, 682. [CrossRef]
3. Hu, F.; Liu, W.; Tsai, S.; Gao, J.; Bin, N.; Chen, Q. An Empirical Study on Visualizing the Intellectual Structure and Hotspots of Big Data Research from a Sustainable Perspective. *Sustainability* **2018**, *10*, 667. [CrossRef]
4. López, F.J.H.; Gayo, J.E.L.; de Pablos, P.O. Semantic Modeling of Administrative Procedures from a Spanish Regional Public Administration. *Sustainability* **2018**, *10*, 633. [CrossRef]
5. Al Qudah, N.F.; Yang, Y.; Anjum, M.A. Transformational Training Programs and Quality Orientation of Employees: Does Employees' Loyalty Matter? *Sustainability* **2018**, *10*, 465. [CrossRef]
6. Faheem, H.M.; König-Ries, B.; Aslam, M.A.; Aljohani, N.R.; Katib, I. Ontology Design for Solving Computationally-Intensive Problems on Heterogeneous Architectures. *Sustainability* **2018**, *10*, 441. [CrossRef]
7. Zhao, J.; Fang, S.; Jin, P. Modeling and Quantifying User Acceptance of Personalized Business Modes Based on TAM, Trust and Attitude. *Sustainability* **2018**, *10*, 356. [CrossRef]
8. Lin, C.; Li, B.; Wu, Y.J. Existing Knowledge Assets and Disruptive Innovation: The Role of Knowledge Embeddedness and Specificity. *Sustainability* **2018**, *10*, 342. [CrossRef]
9. Muntean, M. Business Intelligence Issues for Sustainability Projects. *Sustainability* **2018**, *10*, 335. [CrossRef]
10. Liu, Y.; Yan, Z.; Cheng, Y.; Ye, X. Exploring the Technological Collaboration Characteristics of the Global Integrated Circuit Manufacturing Industry. *Sustainability* **2018**, *10*, 196. [CrossRef]
11. Hu, J.; Li, S.; Hu, J.; Yang, G. A Hierarchical Feature Extraction Model for Multi-Label Mechanical Patent Classification. *Sustainability* **2018**, *10*, 219. [CrossRef]
12. Al-Ruithe, M.; Benkhelifa, E.; Hameed, K. Data Governance Taxonomy: Cloud versus Non-Cloud. *Sustainability* **2018**, *10*, 95. [CrossRef]
13. Yu, C.; Zhang, Z.; Lin, C.; Wu, Y.J. Knowledge Creation Process and Sustainable Competitive Advantage: The Role of Technological Innovation Capabilities. *Sustainability* **2017**, *9*, 2280. [CrossRef]
14. Wu, T.; Wu, Y.J.; Tsai, H.; Li, Y. Top Management Teams' Characteristics and Strategic Decision-Making: A Mediation of Risk Perceptions and Mental Models. *Sustainability* **2017**, *9*, 2265. [CrossRef]
15. Roh, T.; Jeong, Y.; Yoon, B. Developing a Methodology of Structuring and Layering Technological Information in Patent Documents through Natural Language Processing. *Sustainability* **2017**, *9*, 2117. [CrossRef]
16. Can, U.; Alatas, B. Big Social Network Data and Sustainable Economic Development. *Sustainability* **2017**, *9*, 2027. [CrossRef]
17. Kirilenko, A.P.; Desell, T.; Kim, H.; Stepchenkova, S. Crowdsourcing Analysis of Twitter Data on Climate Change: Paid Workers vs. Volunteers. *Sustainability* **2017**, *9*, 2019. [CrossRef]

18. Zheng, J.; Wu, G.; Xie, H. Impacts of Leadership on Project-Based Organizational Innovation Performance: The Mediator of Knowledge Sharing and Moderator of Social Capital. *Sustainability* **2017**, *9*, 1893. [CrossRef]
19. García-Alcaraz, J.L.; Avelar-Sosa, L.; Latorre-Biel, J.I.; Jiménez-Macías, E.; Alor-Hernández, G. Role of Human Knowledge and Communication on Operational Benefits Gained from Six Sigma. *Sustainability* **2017**, *9*, 1721. [CrossRef]
20. Ahn, S.; Kim, S. What Makes Firms Innovative? The Role of Social Capital in Corporate Innovation. *Sustainability* **2017**, *9*, 1564. [CrossRef]
21. Lytras, M.D.; Raghavan, V.; Damiani, E. Big data and data analytics research: From metaphors to value space for collective wisdom in human decision making and smart machines. *Int. J. Semantic Web Inf. Syst.* **2017**, *13*, 1–10. [CrossRef]
22. Lytras, M.D.; Mathkour, H.I.; Abdalla, H.; Al-Halabi, W.; Yanez-Marquez, C.; Siqueira, S.W.M. Enabling technologies and business infrastructures for next generation social media: Big data, cloud computing, internet of things and virtual reality. *J. Univ. Comput. Sci.* **2015**, *21*, 1379–1384.
23. Lytras, M.D.; Mathkour, H.I.; Abdalla, H.; Al-Halabi, W.; Yanez-Marquez, C.; Siqueira, S.W.M. An emerging–Social and emerging computing enabled philosophical paradigm for collaborative learning systems: Toward high effective next generation learning systems for the knowledge society. *Comput. Hum. Behav.* **2015**, *51*, 557–561. [CrossRef]
24. Lytras, M.D.; Mathkour, H.; Torres-Ruiz, M. Innovative Mobile Information Systems: Insights from Gulf Cooperation Countries and All over the World. *Mob. Inf. Syst.* **2016**, *2016*, 2439389. [CrossRef]
25. Visvizi, A.; Lytras, M.D. Rescaling and refocusing smart cities research: From mega cities to smart villages. *J. Sci. Technol. Policy Mak.* **2018**. [CrossRef]
26. Lytras, M.D.; Visvizi, A. Who Uses Smart City Services and What to Make of It: Toward Interdisciplinary Smart Cities Research. *Sustainability* **2018**, *10*, 1998. [CrossRef]

sustainability

MDPI

Article

Social Networks Research for Sustainable Smart Education

Miltiadis D. Lytras [1,2,*], Anna Visvizi [1,3], Linda Daniela [4], Akila Sarirete [2] and Patricia Ordonez De Pablos [5]

[1] Deree College—The American College of Greece, Gravias 6, 153 42 Agia Paraskevi, Greece; avisvizi@acg.edu
[2] Effat College of Engineering, Effat University, P.O. Box 34689, Jeddah 21478, Saudi Arabia; asarirete@effatuniversity.edu.sa
[3] Effat College of Business, Effat University, P.O. Box 34689, Jeddah 21478, Saudi Arabia
[4] Faculty of Education, Psychology and Art, University of Latvia, Raiņa Bulvāris 19, Centra Rajons LV-1586 Rīga, Latvia; linda.daniela@lu.lv
[5] Faculty of Economics and Business, The University of Oviedo, Avda del Cristo, s/n. 33071 Oviedo-Asturias, Spain; patriop@uniovi.es
* Correspondence: mlytras@acg.edu; Tel.: +30-210-600-9800

Received: 11 July 2018; Accepted: 14 August 2018; Published: 21 August 2018

Abstract: Social networks research has grown exponentially over the past decade. Subsequent empirical and conceptual advances have been transposed in the field of education. As the debate on delivering better education for all gains momentum, the big question is how to integrate advances in social networks research, corresponding advances in information and communication technology (ICT) and effectively employ them in the domain of education. To address this question, this paper proposes a conceptual framework (maturity model) that integrates social network research, the debate on technology-enhanced learning (TEL) and the emerging concept of smart education.

Keywords: smart education; social networks; social media; conceptual maturity model; technology-enhanced learning process

1. Introduction

Social networks research has grown exponentially over the past decade. Subsequent empirical and conceptual advances have been transposed in the field of education. As the debate on delivering better education for all gains momentum, the big question is how to integrate advances in social networks research, corresponding advances in information and communication technology (ICT) and effectively employ them in the domain of education. To address this question, this paper proposes a conceptual framework (maturity model) that integrates social network research, the debate on technology-enhanced learning (TEL) and the emerging concept of smart education [1,2]. Social networks research is a rapidly emerging field of study, which in the context of education yields particular promise. Given the advances in ICT education providers worldwide, hence not only education institutions, have sought to employ social networks to boost the efficiency of teaching and learning. Several challenges and opportunities have been identified in that context, including:

Advanced learner profiling methods, developing active, self-directed, responsible learning context with technology [1,2], and integrating mobile applications and analysis tools are all examples of the studies concerning Smart Education Research [3,4]. Another important impact of social networks is setting very important milestones for Higher Education to advance social learning methodologies and practices [5] by the development of new strategies for student centric, and community centric learning. However, student centric and community centric learning is challenged by directions of Social Networking Research in Open and Distributed Learning [5].

Aspects of this reality are manifested in *Learning Analytics Research*, a branch that sets opportunities to reveal hidden pattern in large volume of data related to learners, academic institution, etc. Also, movement toward [6–10] more open learning using *MOOCs* (*Massive Open* Online Courses) is another area which offers full potential for researching student-centered learning analytics such as motivation effects and what to report on students learning. Another focus is *learning software provision*, which poses new significant challenges for policy makers, learning administrators, and faculty. Finally, the *fast movement* from centralized, controlled environments *towards collaborative distributed, integrated social learning systems* is an added direction to the list of focuses challenging the Social Network Research in Open and Distributed Learning.

In face of these challenges, it is important for universities and colleges worldwide to recognize whether they are ready to adopt flexible, decentralized and intelligent systems for learning and social networking or continue to perceive social networking as a standalone practice out of the typical, well set learning procedures. Within this context, the integration of *Social Networks Research* in *Smart Education Research* has to face, in our opinion, six critical challenges:

- The integration of advanced profiling techniques, learning objectives, and social networks;
- Effective use of learning analytics to boost the teaching and learning process;
- Advanced data mining techniques to support the teaching and learning process, on the one hand, and advanced management of teaching and learning, on the other hand;
- The use of data mining and data analytics to exploit synergies that the interaction of teachers and learners in the network environment create and develop strategies for collaborative active learning
- The use of data analytics and data mining to provide personalized learning assistance in context of global learning platforms.
- The use of data analytics and profiling methods examine and exploit the potential inherent in virtual and augmented reality as applied in teaching and learning;

These challenges directly pose a critical question: *Is it possible to realize the integration of all these changes in a sustainable plan for the evolution of Higher Education*? Do academic institutions have the innovators, the early adopters, or the policy makers capable of envisioning and preparing for such future?

The research is done in context of global challenges which rises several questions which authors of the research would like to make in front of future research directions. This paper addresses the following questions

(i) Do social media and social networking websites have a significant role in in educational institutions?
(ii) Is there a methodological framework capable of promoting the integration of the previous six challenges in current social networking capabilities of Technology Enhanced Learning and smart education platforms?
(iii) What are the policy making requirements for such strategic shift?

The reminder of this paper is structured as follows: Section 2 offers a thorough literature review of social networks research and links it to the debates on TEL and smart education. In the following section the research methodology employed in this study is outlined. Conclusions follow.

2. Literature Review on the Exploitation of Social Networking Technologies

Academics and researchers currently explore the capabilities of emergent technologies to support the learning process in an interactive learning environment. Table 1 summarizes recent relevant research on several dimensions of social networks and smart learning [11–28]:

The term "social networks" is broadly used in different contexts. Sometimes it is assumed that social networks are all online sites which can be accessed through the internet, sometimes it is suggested that there should be clear definitions how social networks are different form Learning Platforms, Learning Management Systems, Virtual Learning environment etc. Authors of this paper will not focus on defining these differences, but rather on benefits of social learning which is supported by using different kind of social media and ICT to scaffold Knowledge Building [29–37].

The discussion of social networking for learning purposes in multidimensional. Various comparative bibliometric studies about the connection of social networks, neuron networks and learning promote the debate for the added value of this integration. In the current thread of literature researchers integrate social networks research with the emerging learning analytics and big data domain. In the same direction there is a solid research area that is dealing with the examination of perceptions on learners and teachers about technologies including social networking technologies for learning. At a broader context other studies discuss the trends on digital campus and best practices on the integration of ICTs. In the current literature there is also a well-defined discussion on the connections of social networks research to recommendation systems in technology enhanced learning e.g., Development of a social recommender system based on Hadoop to reduce the gap between students and useful information for them. A very interesting finding in the literature is also the fact that students mainly use social networks for socialization reasons, not fully exploiting the potential of social networks as learning tool. From an applied point of view, the research community of social networking for learning is interested also in the specifications and the implementation of social learning systems. These systems also include Smart Learning solutions in diverse domains like Healthcare or services industries. Emerging technologies, including cloud systems as well as augmented and virtual reality enabled social networking services gain more interest.

Advanced learner profiling methods, developing active, self-directed, responsible learning context with technology [29–31], and integrating mobile applications and analysis tools are all examples of the studies concerning smart education research [32,33]. Another important impact of social networks is setting very important milestones for higher education to advance social learning methodologies and practices [34] by the development of new strategies for student centric, and community centric learning.

Aspects of this reality are manifested in *Learning Analytics Research*, a branch that sets opportunities to reveal hidden pattern in large volume of data related to learners, academic institution, etc. Also, movement toward [35–39] more open learning using *MOOCs* (*Massive Open Online Courses*) is another area which offers full potential for researching student-centered learning analytics such as motivation effects and what to report on students learning. Another focus is *learning software provision*, which poses new significant challenges for policy makers, learning administrators, and faculty. Finally, the *fast movement* from centralized, controlled environments *towards collaborative distributed, integrated social learning systems* is an added direction to the list of focuses challenging the Social Network Research in Open and Distributed Learning.

Table 1. Literature on social networks and smart learning.

Title	Authors	Main Research Issue	Main Contribution	Implications for a Holistic Model
Analysis of the scientific literature published on smart learning	Durán-Sánchez, A., Álvarez-García, J., Del Río-Rama, M.C., Sarango-Lalangui, P.O. (2018)	Analysis of state of the art of the field of Smart Learning	Comparative bibliometric study	Content Context Policy Making & Leadership/Innovation Integration/Sustainability
A bibliometric perspective of learning analytics research landscape	Waheed, H., Hassan, S.-U., Aljohani, N.R., and Wasif, M. (2018)	Analysis of learning analytics literature	Better understanding of current research on learning analytics and the importance of big data and data mining tools	Content Social Interaction Assessment Integration
Learning and teaching with social network sites: A decade of research in K-12 related education	Greenhow, C. and Askari, E. (2017)	Survey of educational research literature	Examination of perceptions on learners and teachers about technologies	Perceptions Integration/Sustainability
Smart learning in digital campus	Liu, D., Huang, R. and Wosinski, M. (2016)	Research on digital campus in China	Discussion of trends on digital campus and best practices	Content Context Collaboration & Social Interaction Development
Homogenizing social networking with smart education by means of machine learning and Hadoop: A case study	Jagtap, A., Bodkhe, B., Gaikwad, B., and Kalyana, S. (2016)	Educational activities, social networking environment and the interest of students for activities	Development of a social recommender system based on Hadoop to reduce the gap between students and useful information for them	Content Context Social Interaction Development Integration
Social media networks as a learning tool	Kolokytha, E., Loutrouki, S., Valsamidis, S. and Florou, G. (2015)	Examines if it is convenient for students the upload of e-learning content in social networks (like Facebook)	Students mainly use social networks for socialization reasons, not fully exploiting the potential of social networks as learning tool	Content Collaboration & Social Interaction & Leadership/Innovation
Analysis of collaborative learning in social network sites used in education	Al-Dhanhani, A., Mizouni, R., Otrok, H. and Al-Rubaie, A. (2015)	Comparative study between different social network sites and educational social networks sites	Development of an educational social network site based on the findings of the conducted study	Content Collaboration & Social Interaction Integration
Smart learning environments using social network, gamification and recommender system approaches in e-health contexts	Di Bitonto, P., Pesare, E., Rossano, V., and Roselli, T. (2015)	Creation of learning paths focused on the specific needs of individual, with the use of information technologies	Solutions of smart learning environment in the field of e-health	Gamification Context Smart Education
Social networks analysis and participation in learning environments to digital inclusion based on large-scale distance education	Da Silva, A.D.S., De Brito, S.R., Martins, D.L., (...), Costa, J.C.W.A. and Francês, C.R.L. (2014)	Evaluation and monitoring of programs designed for digital inclusion training	Identification challenges in these activities	Context Collaboration & Social Interaction Development & Assessment
The Social Network Learning Cloud: Architectural education for the 21st century	Schnabel, M. and Ham, J. (2014)	Social network learning cloud for architectural education and linking academic learning management systems and professional or private social networks	Ways of using social network cloud for other areas of the CV and future directions	Context Innovation Integration/Sustainability
Using smart mobile devices in social-network-based health education practice: A learning behavior analysis	Wu, T.-T. (2014)	Satisfaction of learning and learning behaviors	Empirical evidence show social networks can improve interactions between individuals in nursing education (students, educators)	Content Collaboration & Social Interaction Development & Assessment

Table 1. *Cont.*

Title	Authors	Main Research Issue	Main Contribution	Implications for a Holistic Model
The Use of Virtual Learning Environment (VLE) and Social Network Site (SNS) Hosted Forums in Higher Education: A Preliminary Examination	Hollyhead, A., Edwards, D.J. and Holt, G.D. (2012)	Use of asynchronous virtual learning environment forums and social network sites in higher education institutions	Lessons and also challenges for educators	Content Context Collaboration & Social Interaction Development & Assessment Innovation Integration/Sustainability
"How do social networks influence learning outcomes? A case study in an industrial setting"	Maglajlic, S. and Helic D. (2012)	Analysis of the impact of implicit social networks on the online learning outcome in an industrial context	Case study shows correlation between communication intensity and the outcome of the learning process	Development & Assessment Policy Making & Leadership/Innovation Integration/Sustainability
Reflective learning through social network sites in design education	Park, J.Y. and Kastanis, L. (2009)	Empirical study (survey) on reflective learning through social network sites in the context of two animation units	Findings show the importance of the learning circumstances of the students and the students' learning circumstances and the design of peer-to-peer interactions	Development & Assessment Policy Making & Leadership/Innovation Integration/Sustainability

Having a clear vision of the full potential of smart classroom environment is crucial for transitioning from conventional classroom to the smart classroom and thus driving the innovation in learning and higher education institutions. It is essential to integrate social networking in the smart education research domain and exploit the benefits of the use of social networks as learning tools in the context of smart classroom environment for smart universities. Emergent technologies offer valuable opportunities for using new learning methods with the focus on students. Higher education institution must be aware that pedagogy, and teaching and learning practices needs to adapt to these new tools too.

In our proposed model we highlight the importance of implementing a holistic approach to overcome the challenges of e- learning but also assist academics, deans and stakeholders involved in the design and reform of national educational systems as well as stakeholders working on national/EU/international online learning initiatives. Higher education institutions need to have a clear vision and leadership capabilities to accomplish this strategic transformation.

Participation and collaboration in social networks expands learning opportunities such as sharing, transfer and internalization of new knowledge by learners, which is essential in the online learning process. Higher education institutions (HEIs) must exploit the benefits of a holistic integration of initiatives towards the use of social networks as learning tools for students. And consider challenges in its implementation as well. Learning analytics, visual learning, cloud computing and emergent wearable technologies are key drivers for the successful achievement of the benefits of the online learning process [21,22,28,38,39]. At the same time researches show that neither academia [40], nor students are fully prepared to accept the challenges and use new possibilities meaningfully to construct new knowledge. There are concepts developed on blended learning where learning is supported by virtual and real environments are combined together. There are concepts developed on Knowledge Building [39] but they are missing the collective cognitive responsibility [41,42] as the learning environment is transforming and more responsibility is transferred to learners. This underlines the necessity to put more efforts in searching for new paradigms in learning and analysis of possibilities and challenges brought by social networking sites.

In order to create value through social networking sites in smart education and benefit students, academics and other stakeholders, policy makers and academic staff must understand the following key components and explore their interactions and implications: the content offering, context, collaboration, social interaction, development, assessment, policy making, levels of integration and finally leadership and innovation.

Social networks and technology together will provide different levels of interactivity during learning process, enriching teaching methods and developing students' skills as well as increasing the participation of students resulting in more active learning.

Feedback in a smart education environment is very important. Academic staff must receive feedback about students' performance in smart classroom which in our context is understood as learning space enhanced with diverse digital technologies and an environment of online learning. This feedback can be in form of data or information about the status of assigned tasks or results of an assessment.

Furthermore, another complexity in the smart education context must be emphasized: smart education requires innovative pedagogy methods and tools in order to maximize opportunities of active learning and exploit and enhance the creativity of students. At the moment there are some promising initiatives on pedagogical aspects for smart environment but this is still not enough [43,44].

Figure 1, summarizes the main components of value in social networking sites, helping policy makers at regional, national, and international level to understand the importance of components individually and as a whole. These components must be personalized for the specific context of smart learning environment enhancing the learning process in smart classrooms.

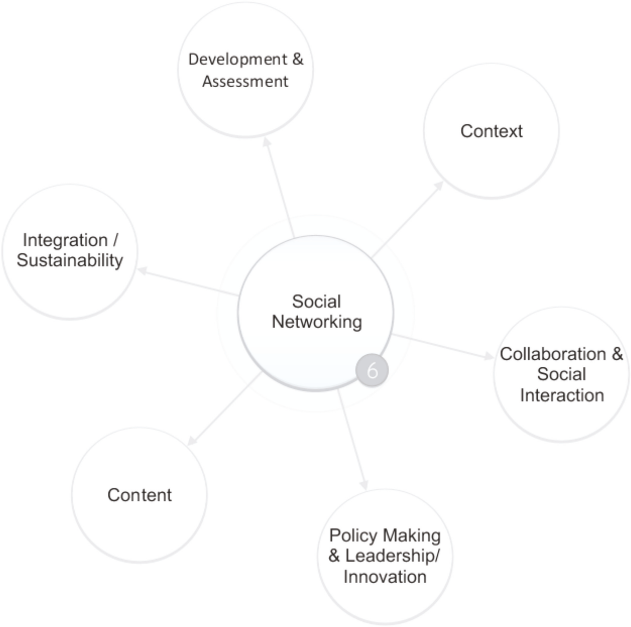


Figure 1. The value components of Social Networks in Education and Learning.

Here we present our innovative conceptual model based on our previous research experience and publications on smart education and technology enhanced learning for more than a decade as well as our daily teaching activities and utilization of new technologies and pedagogical tools to transform traditional classrooms into smart classrooms.

The research team is actively working on the final stages of design of a survey to test the conceptual model proposed here. The survey will be conducted in several countries in Europe and Gulf Region which will help us not only to test our theory and propositions but to gain deeper understanding of social networks and smart education's interrelations and also proving a comparative view across countries and regions.

Both the proposed theory and model as well as the geography regions we aim to cover in our survey will contribute to fill the gap in literature providing a holistic model to analyses and understand social networks in smart education environments.

3. Research Methodology

This research is part of an integrated research related to the International Technology Transfer and Best Practices in Higher Education. It serves as a follow-up, meta-research paper of a special issue recently published in IRRODL, International Review of Research in Open and Distributed Learning, on the theme of Social Networking for learning.

In Figure 2, we provide the overall research methodology adopted in this study. At stage 1, a combined and focused literature review focused on:

- The study of a rich literature on the use of emerging ICTs and their contribution to technology enhanced learning and smart education
- The understanding on how Social networking technologies are exploited in education
- The thorough analysis of policy making requirements and strategic propositions for Smart education and technology enhanced learning

At Stage 2, we drafted our key research problem which is how emerging technologies challenge the adoption of Social networks in Education? This approach integrates several of the new technologies like Learning Analytics, Virtual and Augmented Reality, Visual Learning and Cloud Learning Services. For this study the theoretical construct of Social Learning value components presented in the previous section was exploited further.

At Stage 3, a combination of qualitative and quantitative research was design. The main purpose was to run a quantitative questionnaire for the perception of uses of Technology enhanced learning services about the value of Social networks and emerging technologies as well as to adopt a desktop research on other studies. Soon we plan numerous qualitative interviews with experts of social networking services in education.

At Stage 4, our key theoretical contribution is presented a maturity model for value integration of Social networks in learning and education based on the contribution of emerging ICTs.

The main research questions that this article addresses include:

- What are the main aspects of adopting social networks in education?
- What is the strategic impact of social networks for teaching and learning at diverse levels and in diverse domains of education systems?
- Is there a maturity model that summarizes the value added of social networks in teaching and learning?
- What are the key challenges for the advancement of research geared toward integrating social networks in teaching and learning in higher education?

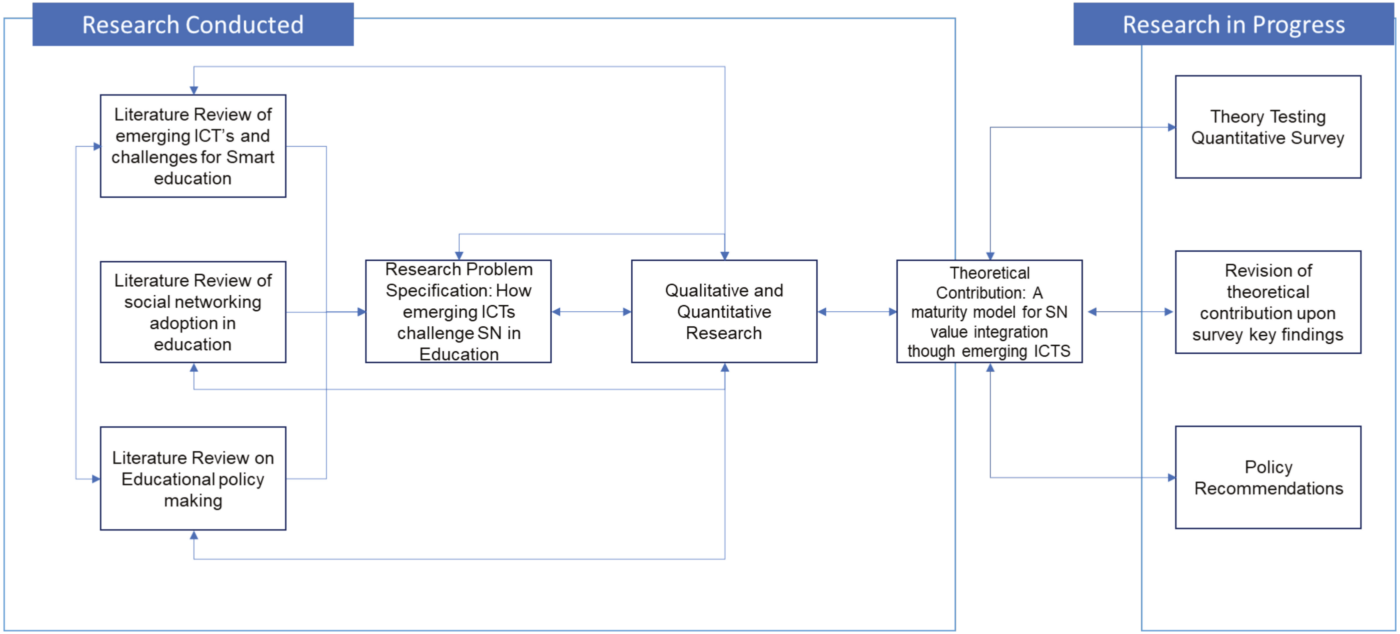

Research Conducted
Literature Review of emerging ICT's and challenges for Smart education
Literature Review of social networking adoption in education
Literature Review on Educational policy making
Research Problem Specification: How emerging ICTs challenge SN in Education
Qualitative and Quantitative Research
Theoretical Contribution: A maturity model for SN value integration though emerging ICTS
Research in Progress
Theory Testing Quantitative Survey
Revision of theoretical contribution upon survey key findings
Policy Recommendations

Figure 2. Our Research methodology for the integration of emerging technologies in Social Networks for Education.

4. Analysis and Main Findings–Social Learning Networking Strategic Shift

The development in Smart Education Research domain is a continuous evolving process towards sustainability in Education. In the analysis of it is critical to understand that the adoption of any emerging technology—along with the relevant experimentations, the lessons learnt and the analysis of contributions—do not have a significant impact unless they maintain a continuous value adding perspective for the future. In our analysis of the domain and in the rhetoric communicated in the special issue, Social Networking Research must be discussed in a context. This context should be associated with several maturity and growth stages, which reveal the hidden value of the application of social networks in Higher Education.

To find out current issues in HE in context of Technology Enhanced Learning the survey questionnaire was created using Google sheets to ensure that it was possible to get answers from different countries at the same time. Respondents were contacted through the researchers' personal contacts and asked for the questionnaires to be filled in by various faculty members, administrators, and students, thus gaining an opinion from all stakeholders involved in HE. The survey questionnaire consisted of 20 questions (in this paper there will be analyzed data which are important in context of role of Social Network in learning process); in the first part, respondents were asked to provide demographic information about themselves such as their current status in HE, country of origin, gender, and the field of science they represented. The following questions were administered regarding technologies used by respondents in their learning process, evaluating frequency of use on a Likert scale ranging from 1 (*never*) to 4 (*during each class*).

After these questions, respondents were asked about using different forms of online learning. They can elaborate on their experience using different kinds of learning such as Learning Management Systems, Social Media Applications and MOOCs.

Next part consisted of questions to evaluate possibilities to use different ICTs in learning process and possible reasons why ICT are not fully used to gain all the possible benefits of them.

All the quantitative data gathered was coded and inserted in SPSS program to make calculations to find out answers to research questions.

5. Results

140 respondents completed the questionnaires. 65 of them were women while 75 were men. Out of these respondents, 23 identified as students, 22 identified as researchers, 75 identified as professors etc., out of these four administration representatives identified themselves as students/researchers. Regarding the areas of science identified by respondents, the breakdown of the data was as follows: 67 identified as technology/IT/CS (Computer Sciences) experts; 40 identified as experts in the social sciences; 36 claimed that their educational studies were different but were currently engaged in TEL aspects of the learning process; 26 indicated that they were experts in learning theories; and 19 other areas were indicated with a small number of experts in these areas. In general, the survey was completed by respondents representing 38 countries, with the largest number of respondents from Latvia (24), followed by Pakistan with 15 respondents, then Greece and Poland, each with 11 respondents. Seventeen countries represented one respondent per country. Currently, there is not enough data from each country to perform data analysis by country. There were not also a sufficient number of respondents' views coming from different areas. Therefore the data was not analyzed in a comparative way, but in general way instead.

In this paper, authors analyze data where respondents had the opportunity to say whether or not they use some kind of Learning Management Systems, Social Media Applications and MOOCs and answers were coded by 0—if the system is not used by the respondent and by 1—if the system is used. The results are summarized in Table 2 and it can be concluded that most popular online learning is use of LMS (N 105) which provides the opportunity to organize teaching-learning process out of time and place, collect data on student activity, provide specific assignments by professors. The Social Media applications for learning were chosen by 57 respondents and MOOCs by 23 respondents of the survey.

These results show that structured learning platforms which are represented by LMS are preferred in HE.

Table 2. Descriptive Statistics on use of online learning possibilities.

	N	Sum
Learning management systems (LMS) (Blackboard, MOODLE etc.)	140	105
Social Media applications for learning	140	57
Massive Open Online Courses	140	23
Valid N (listwise)	140	

As the following step calculations were made on the answers about the possibilities in use of ICT and results are summarized in Table 3 where respondents could express their opinion by evaluating statements provided by researchers in Likert scale from 1–5 where 1 was disagree but 5–fully agree. The Mean and Standart Deviation was calculated to find out how the statements about the possibilities of ICT were evaluated. Results allowed to conclude that respondents are highly positive in their opinion that the use of ICT improves the effectiveness of teaching and learning (mean 4.46 and Std. dev. 0.714) and that the use of ICT in the teaching process promotes students' active engagement in the process (mean 4.35 and Std. dev. 0.758). The less positive opinion were expressed about the statement that the use of LMS fosters students' active engagement in the teaching and learning process (mean 3.76 and Std. dev. 1.001) and it enables the authors to conclude that different forms of online media should be used to foster active learning processes and that Social Media applications can be one of such possibilities. The last statement provided for evaluation was about not fully exploring the benefits of LMS use by professors and results show that the majority of respondents confirm that opinion (mean 4.09 and Std. dev. 0.928).

The next part of the questionnaire consisted of statements about the possible reasons why potentials of ICT are not fully used. The respondents were asked to evaluate given statements in Likert scale from 1–5 where 1—was for the opinion "strongly diagree" and 5—for the opinion that "it is the highest risk" and results are summarised in Table 4. It can be concluded that respondents believe that highest risks that professors are not aware of all the possibilities of ICT (mean 3.6) but these results are quite diverse because the Std. Deviation is 1.023. Also respondents as highly risky evaluate following aspects: There is not enough ICT available in educational environment (mean 3.24 and Std. dev. 1.085) and The ICT used in education are not interactive enough to ensure active learning processes (mean 3.19 and Std. dev. 1.175).

As the next steps authors grouped and coded 15 different fields indicated by respondents. In this stage there were 4 groups: students (N 29), academia (N 109), and administration (2). The group of students were made of respondents who do not have other connection with HE, if a respondent indicated that he or she is a student and researcher for HE, then they were coded for the group-academia where were included those who work in HE as lectures, professors, researchers and other academic personnel who are involved in academic responsibilities. The administration group consists of respondents who indicated themselves as administrators in HE but in group of others there are included those who have other roles but who cooperates with HE (IT consultant, business owner etc.). This grouping was necessary to find out is there differences among the groups on their opinion why there is not used the full potential of ICT. Results are summarized in Table 5. The mean calculation was chosen because the size of groups is not the same and Standard deviation is calculated to find out the diversity in respondents' opinions.

Table 3. Case Summaries of opinions on possibilities of use of ICT.

	The Use of ICT in Education Improves the Effectiveness of Teaching and Learning	The Use of ICT in the Teaching Process Promotes Students' Active Engagement in the Process	The Use of ICT in the Teaching Process Fosters Students' Creativity, Independent Thinking and Problem Solving Skills	The Use of ICT in Education Promotes Students Awareness and Willingness to Look for Additional Information in Other Sources	The Use of LMS Fosters Students' Active Engagement in the Teaching and Learning Process	Smart Use of ICT in the Teaching Process Might Foster the Development of Students Liberal Worldview, Open-Mindedness, Respect for Others	The Benefits of the Use of LMS Are Not Fully Explored by Professors
Mean	4.46	4.35	4.02	4.15	3.76	3.91	4.09
St. Deviation	0.714	0.758	0.925	0.856	1.001	0.948	0.928

Table 4. Case Summaries about the reasons of not using full potential of ICT.

	Students Get Bored Very Quickly	Students Lack the Necessary Skills to Use IT Enhanced Methods of Teaching	Professors Are Not Aware of All the Possibilities of ICT	There Is Not Enough ICT Available in Educational Environment	The ICT Used in Education Are Not Interactive Enough to Ensure Active Learning Processes
Mean	2.78	2.54	3.60	3.24	3.19
Std. Deviation	0.898	0.962	1.023	1.085	1.175

Table 5. Case Summaries about the reasons of not using full potential of ICT by groups of respondents.

Status		Students Get Bored Very Quickly	Students Lack the Necessary Skills to Use IT Enhanced Methods of Teaching	Professors Are Not Aware of All the Possibilities of ICT	There Is Not Enough ICT Available in Educational Environment	The ICT Used in Education Are Not Interactive Enough to Ensure Active Learning Processes
student	Mean	2.83	2.41	3.34	2.86	3.38
	Std. Deviation	0.966	1.018	1.173	1.187	1.293
academia	Mean	2.79	2.56	3.70	3.37	3.15
	Std. Deviation	0.882	0.946	0.979	1.010	1.141
administration	Mean	2.17	2.67	3.00	2.67	2.83
	Std. Deviation	0.753	1.033	0.894	1.506	1.329
other	Mean	3.50	3.00	4.00	4.00	4.00
	Std. Deviation	0.707	1.414	0.000	0.000	0.000
Total	Mean	2.78	2.54	3.60	3.24	3.19
	Std. Deviation	0.898	0.962	1.023	1.085	1.175

The analyses of results show that there is not big differences in opinions of students and academia for the statements. In students opinion the higher risk is that Professors are not aware of all the possibilities of ICT (mean 3.34) but results of Std. Deviation show that their opinion inside the group is quite diverse and it shows that they have different experience with professors. Representatives of academia assumes this risk as the most important (mean 3.7) and results of Std. Deviation shows that their opinion was more focused and can indicate that they feel that there can be done more to incorporate different ICT in learning process. Another quite interesting results are for statement "The ICT used in education are not interactive enough to ensure active learning processes" where results of students show that in their opinion it is the highest risk (mean 3.38) although the Std. Deviation shows the diverse opinion within the group, but for academia this risk is evaluated as third in line of importance (mean 3.15, Std. Deviation 1.141) and it confirms that the cycle of technology development influences the HE where processes of changes are slower than changes in possibilities provided by technologies and this uncertainty became more and more influential in teaching learning process and it also influences the use of social networking in learning process.

In Table 6, we introduce the Social Learning Networking Strategic Grid. In fact, a stage and growth model metaphor of strategic impact of Social Networking Research for Learning is introduced. The overall proposition is that nine key variables and dimensions of value delivery are integrated in Social Networking Research for Learning, namely: Content, Context, Collaboration, Social Interaction, Development, Assessment, Level of Integration, Policy Making, and Leadership/Innovation. These are the critical perceived value carriers and should be considered as critical success factors. Any initiative related to the adoption of Social Networks Research for learning should provide flexible methods, practices, and strategies for the realization of these factors. The current practice shows different approaches and extremely diversified value propositions. In a very abstract generalization for these eight value carriers, we define two perceptions about the strategic impact of their adoption. Their low and high strategic impact on learning quality. As we will present in abstract level, different strategic impact is linked to three Growth-Maturity stages:

- The Epos of Inquiry: Limited, not institutionally integrated social networking initiatives
- The Epos of Actualization: Integration of SN, in academic practice, towards active learning and engagement.
- The Epos of Value Delivery: Strategic use of SN, integrated with various other technological capabilities including Learning Analytics, Visual Learning, Cognitive Computing and Cloud.

Table 6. Social Learning Networking Strategic Grid.

Dimensions	**Strategic Impact of Social Networking for Learning**	
	LOW/Epos of Inquiry	***HIGH/Epos of Actualization***
Content	Packaging	Annotations; Dynamic Programs
Context	Static	Student-Centric
Collaboration	Social Networking	Social Enabled
Social Interaction	Instruction flow	Social Skills driven
Development	Knowledge Transfer	Problem Solving
Assessment	Content based	Critical Thinking
Policy making	Adoption	Evolution
Level of Integration	Course-based	Organization-wide
Leadership/innovation	No consideration	Entrepreneurship Driven

In the current era of evolution in Social Networking research for learning, there are some important facts. Most of the implementations in terms of content focus on packaging and the flexibility of delivering of micro contents. Most contexts for exploitation are static and predefined learning activities that provide a rather narrow environment for student engagement. Limited reflection on the results of collaboration can be understood. The focus of the collaboration is mostly facilitated by a given social networking strategy where static profiling of student characteristics provides the connectivity.

The emphasis on the social interaction of learners, professors and other stakeholders is focused to instruction flows. Additionally, the development strategy is mostly concerned with the knowledge transfer rather than problem solving capabilities. The assessment in most of the Social Networking for Learning is content based, and the strategizing of learning through adoption of social networks is in alignment with given narrow institutional policies. In most cases, there is also limited analysis of the linkage between social networking and innovation and leadership.

We call this maturity stage, '*Epos of Inquiry*'. In this stage academic institutions experiment at a limited base with Social Networking Tools. They do believe that there is a potential for this integration but they still have critical inquiries and questions to answer. The various initiatives are not integrated; there is not a concrete institutional strategy for the wide adoption of social networks in courses or programs.

The strategic impact of social networking for learning is realized in the next stage, which we call '*Epos of Actualization*'. In this maturity stage, social networking is exploited for the continuous creation of content annotations through collaborative filtering and profiling analysis permitting the dynamic construction of Dynamic Curricula across Programs. Social networking is no longer used as a typical facilitator of a technology driven-context, but the social characteristics of learners are exploited for dynamic provision of meaningful personalized context for learning. The modes of Social Interaction are also strategized toward the construction of Social Skills and not only as parts of a limited instruction flow and design. Developmental strategy in the Epos of Actualization is organized around Problem Solving advanced capabilities and social networks facilitate this. Assessment is promoting critical thinking and social networking tools exploited for delivering arguments, evidence and justifications. A collaborative, peer-based, systematic work is informing an Organization-wide level of integration permitting Evolving Policy Making. In this growth stage, Social Networks are integral parts of Institutional Strategies for Active Learning and Innovation Programs. The entire approach can be characterized as an out-of-the box Paradigm shift. (See Table 7). We need to highlight the importance of the alignment between this proposed conceptual framework for social networks and new technologies with the national educational strategy of a country and/or region.

The development of various novel technological capabilities in the last few years has a critical impact on the radical change of the previous two growth stages. The adoption of social networking research in Smart Education Research is entering a new phase of maturity and potential contribution in the higher education. This new stage, the Epos of Value Integration is powered by the introduction of Learning Analytics, Visual Learning, Cloud and Cognitive Computing solutions together with a new generation of wearable technologies and advanced Human Computer Interaction methods. In Table 7 below, we present the key characteristics of this new Maturity Level with a reference to the key enabling technologies. This is according to our perspective—the new challenging research context for Open and Distributed Learning.

Table 7. Social Learning Networking—Epos of Value Integration.

Dimensions	**Enabling Technologies and Strategic Impact of Social Networking for Learning**	
	Enabling Technologies	***TOO HIGH/Epos of Value Delivery***
Content	Analytics & Cloud Computing	Flexible, Different Media
Context	Mobile Learning Analytics & Visual Learning	Context-Aware
Collaboration	Cognitive Computing & analytics	Multimodal
Social Interaction	Analytics, VR and Cloud	Augmented, Enriched
Development	Cognitive Computing, Recommender Systems	Personalized
Assessment	Analytics & Cloud	Evolution, Personality
Policy making	Smart Cognition	Sustainability in Education
Level of Integration	Integral approach	Worldwide
Leadership/innovation	Integral Approach	New Radical Knowledge Creation

The diffusion and integration of the emerging technologies in the social networking research for Sustainable Higher Education will promote a number of radical changes. Pioneers in the strategic planning for the realization of this growth stage will be rewarded by the outcomes of the education systems in terms of both, critical thinking and creativity. In a way, the next generation social networking for higher education will:

- Promote dynamic learning contexts through advanced packaging, flexible designs, in different media with multiple annotation schemas that will focus on learning delivery and integration with other domains. The integration of learning domains with other significant human activities will provide a transparent, ubiquitous and pervasive infrastructure available anytime and from any place. Context-Awareness will exploit advanced learning analytics capabilities aiming to provide tailor made learning based on specific features of learner's profiles.
- Collaboration will be facilitated by multimodal social networking connections, which will be context aware and powered by sophisticated layers of analytics, most of them focusing in the value dimension of learning as perceived by learners and in the formulation of effective learning teams. This is one of the most promising area of this research which we will summarize in a next section of this vision article.
- The Social Interaction will be augmented and enriched with new learning experiences powered by wearable technologies and advanced Virtual Reality gadgets. It is our belief that in the next generation of Social Networking, the role of virtual and augmented reality will be crucial. Visual Learning will promote further video lectures, with the provision of advanced Learning Labs enabled by VR technologies. Consider for example the case of a Visual Learning Lab for Medical Training, where the social networking profiles of learners will be facilitating common virtual sessions and experiments.
- The Developmental dimension of Social Networking will be advanced and personalized. Sophisticated capabilities for portfolio management and repositories of active learning stories will maintain a systematic learning management system to analyze the learning requirements and personalized learning paths of learners.
- For the challenging dimension of Assessment, Higher Education institutions must adopt new ideas. Evolution and Personality will be the main factors for assessment methods. Given the fact that knowledge is available everywhere, and that social networking applications can promote effective update, retrieval, collaborative filtering, and rating of knowledge, the challenge then is on how to cultivate an assessment culture in which personality empowerment and evolution are promoted.
- Policy making related to the integration of Social Networks Research in Education will focus on Sustainability in Education. This is a critical step towards sustainable, strategic adoptions of information technologies in the educational context. The main dimensions of sustainability will be respect for the human entities, strategic and wise use of technical resources and embodiment of sustainable developmental ideas in designing programs and curricula.
- The integral exploitation of the emerging technologies will enable global initiatives putting together unexploited human and mental capacities for the fostering of innovation and entrepreneurship. This can be unpredictable in terms of impact. In future scenario, consider distributed academic programs where students will attend few courses from many institutions in the context of agreements and specializations.
- The maturity of this level will promote a strategized new Knowledge Creation campaign at worldwide level. The human capacities through social networking, mobile learning analytics, visual learning and cognitive capabilities of learning systems will bring together unexploited capacities of learning peers and institutions. Those who capable of envisioning this forthcoming reality earlier are going to benefit the most.

We understand that moving towards the Epos of Value Delivery is an uneasy case. Educational organizations suffer from several inefficiencies, most of them related to slow procedures, bureaucratic decision-making capabilities, slow adoption to the environments, and limited understanding mechanisms to the demands of the society and the industry. In this situation, though, we do believe that several pioneers and innovative institutions will lead the big change.

In the next section, we elaborate further in this vision. We provide the most promising areas for future research related to the four technologies already mentioned. It is in fact the next stage in our research methodology to test empirically these theoretical propositions aligned to the maturity model.

6. Future Research Directions

The next step of our research is to analyze how the key components of value in social networking can be mixed with the new value propositions of four technologies namely: Virtual and augmented reality, cloud services, learning analytics and visual learning. In Table 8, we summarize the main aspects of scenarios of services for social networking value adding services powered by emerging ICTs in connection to our theoretical proposition.

Table 8. Future ICT enabled research areas for social networking in education.

Social Networking Strategic Dimensions	Future ICT Enabled Research Areas for Social Networking in Education
Content	• Integration of value layers in content blocks • Packaging of Visual Learning sessions • Dynamic matching of learning paths to content • Distributed repositories of learned-generated content • Codification of reactions and learners' interventions • Feedback as learning content for future use. Mobile learning analytics.
Context	• Distributed Context for Open Learning with exploitation of wisdom gained from recording of learning stories • Active Learning over Augmented Reality Learning Networks • Social Networking for Community Building awareness • Social Responsibility as a context for Social Action • Decomposition of Academic Context for flexible learning
Collaboration	• Learning Analytics Strategies for enhanced Problem-Solving oriented professional learning • Multimodal Distributed Platforms for Exchanging Learning Experiences • Marketplaces of Collaborative Interventions • Agora of Collaborative Augmented Reality Learning Stories.
Social Interaction	• Visual Profiling • Social learning experiences • visual labs • Distributed, social learning networks • Open Learning Systems against poverty.
Development	• Massive Open Visual Learning Systems • Competencies models and assessment scenarios • Annotations of group skills • Organizational Development • Faculty Promotions • Global Faculty Research Networks • Social Responsibility Programs in Higher Education.
Assessment	• Distributed Assessments. Developmental plans of individual, groups and institutions • Institutional Assessments • Cloud Portfolios and Profiles • Backward Integration of Lessons Learnt.

Table 8. *Cont*.

Social Networking Strategic Dimensions	Future ICT Enabled Research Areas for Social Networking in Education
Policy Making	● Academia Industry Collaborations ● Experimentations at Postgraduate and Professional Education ● Quality of Education integration and Mobile Learning Analytics ● Higher Education Organizational Memories.
Level of Integration	● Global Distributed Learning Services ● Global Open Visual Labs.
Leadership/Innovation	● Social Networks of Innovators ● Global Training of Advanced Technologies and Competencies ● Smart Education Research alliances for New Knowledge Creation.

7. Discussion & Conclusions

Social networks will continue to play a significant role in Smart Education Research. The progress made the previous years has convinced several stakeholders for the critical need to support social learning interactions. Given the fact that most of the inefficiencies stated in education are related to limited collaboration between learners, narrow scenarios for active learning engagement, limited use of social media and rather limited exploitation of scientific knowledge available, thus we will shortly welcome a new Era of Open and Distributed Learning. More Open, Enriched, Global, Personalized, Social Engaged, with Social Responsibility and Sustainable. With only one prerequisite: That the coming change will not make afraid strict academic institutions and old fashion academic policy makers. Eventually the pressure set by learners eager to apply scientific knowledge into real problems for innovative solutions will cause a revolution to Education. We are looking forward to collaborating towards this new era of learning, knowledge and innovation.

The promotion of sustainability in Higher Education also requires a social inclusive participation in the new era of ICTs. Towards this direction critical policies are needed:

- Soft Skills training programs for faculty, administrators and students in the use of advanced SN services, as well as cloud computing, virtual reality, visual learning and learning analytics enabled services
- Feasibility and Sustainability studies for the contribution of Education and Technology enhanced learning to social inclusive economic growth
- Integration of Smart Learning to Smart Cities and regional development initiatives
- Policies to promote intercultural understanding and collaboration at educational level
- Policies to enhanced research collaboration and social impact
- Policies that develop and promote cross-cultural international networks of research and innovation excellence
- Development of transparent, open, distributed learning services with advanced accessibility and transparency
- Continuous improvement of learning infrastructures.

In our ongoing research, this is the ultimate objective: To draw the lines for a new era of sustainable adoption of emerging technologies in technology enhanced projects and initiatives. The maturity model introduced in this study is just an invitation to researchers, and scholars to understand the multidimensional character of Social Networking concept in Higher Education and its direct linkage to several emerging technologies. We are confident that social networks in higher education will be totally different from the current anticipation.

Author Contributions: All co-authors contributed equally to the design, implementation, conduct of research and analysis of data and the delivery of main findings. All co-authors, contributed equally in all the phases of this research.

Acknowledgments: The authors would like to thank Effat University in Jeddah, Saudi Arabia, for funding the research reported in this paper through the Research and Consultancy Institute.

Conflicts of Interest: The authors declare no conflict of interest. The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results".

Appendix A. Questionnaire

Demographics

1. Which of the following technologies/tools you think may be useful in the teaching and learning process and how frequently? (during each class, very frequently, frequently, rarely, never)

- internet, incl. YouTube and videos available on-line
- Social Networking applications
- Students' smart phones
- tablets
- personal computers
- educational games
- robotics
- virtual reality applications
- cloud applications
- Other

2. If you chose 'other', what would it be?
3. Can you outline an example of an innovative use of learning technologies in your class?
4. Do you find it difficult to use technology in the process of teaching in your class?
5. What hinders the use of technology-enhanced methods in your teaching?—OR—What stops you from enhancing the set of tools you already employ?
6. Which learning management systems (LMS) you use to support your teaching (or you use as a student)?
7. What is your perception of the use of information technologies in your teaching practice and strategy
8. Would you agree that the use of information technology in education improves the effectiveness of teaching and learning?
9. Would you agree that the use of information technology in the teaching process promotes students' active engagement in the process?
10. Would you agree that the use of information technology in the teaching process fosters students' creativity, independent thinking and problem solving skills?
11. Would you agree that the use of information technology in education promotes students awareness and willingness to look for additional information in other sources?
12. Would you agree that the use of LMS (Blackboard, MOODLE etc.) fosters students' active engagement in the teaching and learning process?
13. Would you agree that smart use of information technology in the teaching process might foster the development of students liberal worldview, open-mindedness, respect for others?
14. Would you agree that the benefits of the use of LMS (Blackboard, MOODLE etc.) are not fully explored by professors
15. In your teaching, have you ever used technology-enhanced approaches to boost students' awareness of their civic rights and responsibilities?
16. If you answered 'yes' above, can you tell you what did you do?

17. Please evaluate the following statements about the use of information technologies (ICT) in the teaching and learning process. The use of ICT in the teaching process.... (Strongly disagree, Mostly disagree, In some situations it can be so, I agree, Fully agree)

- helps students to better understand the topic and be prepared to use them in knowledge construction
- helps to ensure active learning processes for students
- provides additional opportunities to get access to knowledge for disadvantaged groups
- boosts the value of education
- depends on teaching strategies chosen by professor
- may contribute to the development of liberal, democratic worldviews and great civic engagement
- depends on the age and ability of students to use them
- depends on students' attitude to them
- depends on professors' attitude to them
- depends on the availability of infrastructure and the devices

18. Please evaluate the challenges/risks related to the use of ICT in the teaching and learning process (Scale: Strongly disagree, Mostly disagree, in some situation it can be a risk, In most situations it can be a risk, It is the highest risk)

- Students get bored very quickly
- Students lack the necessary skills to use IT enhanced methods of teaching
- Professors are not aware of all the possibilities of IIT
- There is not enough ICT available in educational environment
- The ICT used in education are not interactive enough to ensure active learning processes

19. Please evaluate the following statements regarding actions which might foster the use of ICT in the teaching and learning processes (Scale: It doesn't matter, it can be solved in some level, It is not a problem in our institution, It should be one of the first priorities, It is the highest priority)

- Professors should be trained to use ICT in teaching process
- There should be more cooperation among technology developers and educational institutions
- There should be more ICT available in educational environment
- The ICT used in education should been previously evaluated from the view of their sustainability
- The ICT used in education should be with high level of interactivity to ensure active learning processes

20. Would you prefer the 'old style' ICT-free teaching, i.e., no powerpoint, no youtube etc.

References

1. Lytras, M.D.; Mathkour, H.I.; Abdalla, H.; Al-Halabi, W.; Yanez-Marquez, C.; Siqueira, S.W.M. Enabling technologies and business infrastructures for next generation social media: Big data, cloud computing, internet of things and virtual reality. *JUCS J. Univ. Comput. Sci.* **2015**, *21*, 1379–1384.
2. Visvizi, A.; Lytras, M.D.; Daniela, L. (Re)Defining Smart Education: Towards Dynamic Education and Information Systems for Innovation Networks. In *Enhancing Knowledge Discovery and Innovation in the Digital Era*; IGI Global: Hershy, PA, USA, 2018. [CrossRef]
3. Lytras, M.D.; Mathkour, H.I.; Abdalla, H.; Al-Halabi, W.; Yanez-Marquez, C.; Siqueira, S.W.M. An emerging—Social and emerging computing enabled philosophical paradigm for collaborative learning systems: Toward high effective next generation learning systems for the knowledge society. *CiHB Comput. Human Behav.* **2015**, *51*, 557–561. [CrossRef]
4. Lytras, M.D.; Mathkour, H.; Torres-Ruiz, M. Innovative Mobile Information Systems: Insights from Gulf Cooperation Countries and All over the World. *MISY Mob. Inf. Syst.* **2016**, *2016*, 2439389. [CrossRef]

5. Siadaty, M.; Gasevic, D.; Hatala, M. Associations between technological scaffolding and micro-level processes of self-regulated learning: A workplace study. *Comput. Human Behav.* **2016**, *55*, 1007–1019. [CrossRef]
6. Siemens, G.; Gasevic, D. Guest Editorial—Learning and Knowledge Analytics. *Educ. Technol. Soc.* **2012**, *15*, 1–2.
7. Jdidou, Y.; Khaldi, M. Increasing the Profitability of Students in MOOCs using Recommendation Systems. *Int. J. Knowl. Soc. Res. IJKSR* **2016**, *4*, 75–85. [CrossRef]
8. Lytras, M.D.; Raghavan, V.; Damiani, E. Big data and data analytics research: From metaphors to value space for collective wisdom in human decision making and smart machines. *IJSWIS Int. J. Semant. Web Inf. Syst.* **2017**, *13*, 1–10. [CrossRef]
9. Lima, M.; Zorrilla, M. Social Networks and the Building of Learning Communities: An Experimental Study of a Social MOOC. *IRRODL Int. Rev. Res. Open Distrib. Learn.* **2017**, *18*, 40–64.
10. Laaser, W.L.; Concha, U.R. MOOCs, A Phenomenon with Many Faces: Success and Failures. *Int. J. Smart Educ. Urban Soc. IJSEUS* **2018**, *9*, 27–39. [CrossRef]
11. Al-Dhanhani, A.; Mizouni, R.; Otrok, H.; Al-Rubaie, A. Analysis of collaborative learning in social network sites used in education. *Soc. Netw. Anal. Min.* **2015**, *5*, 1–18. [CrossRef]
12. Da Silva, A.D.S.; De Brito, S.R.; Martins, D.L.; Costa, J.C.W.A.; Francês, C.R.L. Social networks analysis and participation in learning environments to digital inclusion based on large-scale distance education. *Int. J. Distance Educ. Technol.* **2014**, *12*, 1–25. [CrossRef]
13. Di Bitonto, P.; Pesare, E.; Rossano, V.; Roselli, T. Smart learning environments using social network, gamification and recommender system approaches in e-health contexts. *Smart Innov. Syst. Technol.* **2015**, *41*, 491–500. [CrossRef]
14. Durán-Sánchez, A.; Álvarez-García, J.; Del Río-Rama, M.C.; Sarango-Lalangui, P.O. Analysis of the scientific literature published on smart learning. *Espacios* **2018**, *39*, 18.
15. Greenhow, C.; Askari, E. Learning and teaching with social network sites: A decade of research in K-12 related education. *Educ. Inf. Technol.* **2017**, *22*, 623–645. [CrossRef]
16. Hollyhead, A.; Edwards, D.J.; Holt, G.D. The Use of Virtual Learning Environment (VLE) and Social Network Site (SNS) Hosted Forums in Higher Education: A Preliminary Examination. *Ind. High. Educ.* **2012**, *26*, 369–379. [CrossRef]
17. Jagtap, A.; Bodkhe, B.; Gaikwad, B.; Kalyana, S. Homogenizing social networking with smart education by means of machine learning and Hadoop: A case study. In Proceedings of the International Conference on Internet of Things and Applications (IOTA), Pune, India, 22–24 January 2016; pp. 85–90. [CrossRef]
18. Kolokytha, E.; Loutrouki, S.; Valsamidis, S.; Florou, G. Social media networks as a learning tool. *Procedia Econ. Financ.* **2015**, *19*, 287–295. [CrossRef]
19. Komninou, I. A Case Study of the Implementation of Social Models of Teaching in e-Learning: The Social Networks in Education. Online Course of the Inter-Orthodox Centre of the Church of Greece. *TechTrends* **2018**, *62*, 146–151. [CrossRef]
20. Liu, D.; Huang, R.; Wosinski, M. Smart learning in digital campus. *Lect. Notes Educ. Technol.* **2016**, 51–90. [CrossRef]
21. Lytras, M.D.; Mathkour, H.I.; Abdalla, H.; Yanez-Marquez, C.; Ordóñez de Pablos, P. The social media in academia and education research R-evolutions and a paradox: Advanced next generation social learning innovation. *J. Univ. Comput. Sci.* **2014**, *20*, 1987–1994.
22. Ordóñez de Pablos, P.; Lytras, M.D.; Xi Zhang, J.; Chui, K.T. *Opening Up Education for Inclusivity across Digital Economies and Societies*; IGI-Global: Hershey, PA, USA, 2019; forthcoming.
23. Park, J.Y.; Kastanis, L. Reflective learning through social network sites in design education. *Int. J. Learn.* **2009**, *16*, 11–22.
24. Schnabel, M.; Ham, J. The Social Network Learning Cloud: Architectural education for the 21st century. *Int. J. Arch. Comput.* **2014**, *12*, 225–241. [CrossRef]
25. Seid Maglajlic, S.; Helic, D. How do social networks influence learning outcomes? A case study in an industrial setting. *Interact. Technol. Smart Educ.* **2012**, *9*, 74–88. [CrossRef]
26. Waheed, H.; Hassan, S.-U.; Aljohani, N.R.; Wasif, M. A bibliometric perspective of learning analytics research landscape. *Behav. Inf. Technol.* **2018**. [CrossRef]
27. Wu, T.-T. Using smart mobile devices in social-network-based health education practice: A learning behavior analysis. *Nurse Educ. Today* **2014**, *34*, 958–963. [CrossRef] [PubMed]

28. Meishar-Tal, H.; Pieterse, E. Why Do Academics Use Academic Social Networking Sites? *IRRODL Int. Rev. Res. Open Distrib. Learn.* **2017**, *18*, 1–22. [CrossRef]
29. Wang, W.; Wu, J.; Yuan, C.H.; Xiong, H.; Liu, W.J. Use of Social Media in Uncovering Information Services for People with Disabilities in China. *IRRODL Int. Rev. Res. Open Distrib. Learn.* **2017**, *18*, 65–83. [CrossRef]
30. Durak, G. Using Social Learning Networks (SLNs) in Higher Education: Edmodo Through the Lenses of Academics. *IRRODL Int. Rev. Res. Open Distrib. Learn.* **2017**, *18*, 84–109. [CrossRef]
31. Macià, M.; García, J. Properties of Teacher Networks in Twitter: Are They Related to Community-Based Peer Production? *IRRODL Int. Rev. Res. Open Distrib. Learn.* **2017**, *18*, 110–140.
32. Raspopovic, M.; Cvetanovic, S.; Medan, I.; Ljubojevic, D. The Effects of Integrating Social Learning Environment with Online Learning. *IRRODL Int. Rev. Res. Open Distrib. Learn.* **2017**, *18*, 141–160. [CrossRef]
33. Shen, C.W.; Kuo, C.J.; Ly, P.T.M. Analysis of Social Media Influencers and Trends on Online and Mobile Learning. *IRRODL Int. Rev. Res. Open Distrib. Learn.* **2017**, *18*, 208–224. [CrossRef]
34. Mora, H.; Ferrández, A.; Gil, D.; Peral, J. A Computational Method for Enabling Teaching-Learning Process in Huge Online Courses and Communities. *IRRODL Int. Rev. Res. Open Distrib. Learn.* **2017**, *18*, 225–246. [CrossRef]
35. Ruiz-Calleja, A.; Asensio-Pérez, J.I.; Vega-Gorgojo, G.; Gómez-Sánchez, E.; Bote-Lorenzo, M.; Alario-Hoyos, C. Enriching the Web of Data with Educational Information Using We-Share. *IRRODL Int. Rev. Res. Open Distrib. Learn.* **2017**, *18*, 247–465. [CrossRef]
36. Gülbahar, Y.; Rapp, C.; Sitnikova, A. Enriching Higher Education with Social Media: Development and Evaluation of a Social Media Toolkit. *IRRODL Int. Rev. Res. Open Distrib. Learn.* **2017**, *18*, 23–39. [CrossRef]
37. Samra, H.E.; Li, A.S.; Soh, B.; AlZain, M.A. A Cloud-Based Architecture for Interactive E-Training. *Int. J. Knowl. Soc. Res. IJKSR* **2017**, *8*, 67–78. [CrossRef]
38. Tynan, B.; Ryan, Y.; Hinton, L.; Lamont Mills, A. Out of Hours: Final Report of the Project e-Teaching Leadership: Planning and Implementing a Benefits-Oriented Costs Model for Technology Enhanced Learning. 2012. Available online: http://www.olt.gov.au (accessed on 15 June 2018).
39. Cleveland, S.; Block, G. Toward Knowledge Technology Synchronicity Framework for Asynchronous Environment. *Int. J. Knowl. Soc. Res. IJKSR* **2017**, *8*, 23–33. [CrossRef]
40. Scardamalia, M.; Bereiter, C. A Brief History of Knowledge Building. *Can. J. Learn. Technol.* **2010**, *36*. Available online: https://www.cjlt.ca/index.php/cjlt/article/view/26367/19549 (accessed on 15 June 2018). [CrossRef]
41. Gutiérrez-Braojos, C.; Salmerón-Pérez, H. Exploring collective cognitive responsibility and its effects on students' impact in a knowledge building community. *Infanc. Aprendiz. J. Stud. Educ. Dev.* **2015**, *38*, 327–367. [CrossRef]
42. Gutiérrez-Braojos, C.; Montejo-Gámez, J.; Jiménez, A.E.; Martínez Martínez, A. Positive Interdependence in Blended Learning Environments: Is It Worth Collaborating? In *Learning Strategies and Constructionism in Modern Education Settings*; Daniela, L., Lytras, M., Eds.; IGI Global: Hershey, PA, USA, 2018; pp. 50–68. [CrossRef]
43. Zhu, Z.-T.; Yu, M.-H.; Riezebos, P. A research framework of smart education. *Smart Learn. Environ.* **2016**, *3*, 1–17. [CrossRef]
44. Spector, M. Conceptualizing the emerging field of smart learning environments. *Smart Learn. Environ.* **2014**, *1*, 2–10. [CrossRef]

sustainability

MDPI

Article

What Makes Firms Innovative? The Role of Social Capital in Corporate Innovation

Se-Yeon Ahn [1] and So-Hyung Kim [2,*]

[1] Management Research Center, Seoul National University, 1 Kwanak-ro, Kwanak-gu, Seoul 08826, Korea; ahn.seyeon@gmail.com

[2] Department of Business, Baekseok University, 76, Munam-ro, Dongam-gu, Chenan-si 31965, Korea

* Correspondence: shkim2@bu.ac.kr; Tel.: +82-10-7299-1736

Received: 7 August 2017; Accepted: 31 August 2017; Published: 3 September 2017

Abstract: This paper offers a social capital explanation for the purported relationship between human capital investment and an organization's innovation capability. We argue that social capital plays a mediating role in the relationship between the level of individual knowledge of employees and organizations' innovation capabilities. The mediating mechanism is attributed to the role of social capital in knowledge exchange and combination that help enhance knowledge creation. Using survey data of 319 manufacturing firms in Korea, we conducted structural equation modeling (SEM) analysis to verify the mediating role of social capital in firms' innovation performance. The results demonstrated that relational and cognitive dimensions of social capital are important mediators in realizing organizational innovation performance.

Keywords: innovation capability; social capital; knowledge creation; human capital investment

1. Introduction

What makes organizations innovative? The antecedents of organizational innovation are one of the most widely studied research topics among management scholars, as innovation is considered a core element for corporate growth and crucial for companies attaining sustainable competitive advantage [1–5]. Innovation capability is understood as being closely tied to a firm's ability to utilize its knowledge resources [1,2,6,7], and scholars have emphasized the role of individual manpower in the increase of innovation capability, in that the knowledge and skills brought into the firm by scientists and engineers can help firms innovate [7,8]. In this regard, it has been popular for firms to bring in innovative geniuses who possess the required technical skills and expertise to enhance their innovative capabilities.

However, nowadays many firms renowned for their excellence in innovation, such as Google, Facebook, and Amazon, among others, have begun to pay more attention to networks, instead of turning to a few innovative geniuses to enhance their innovation performance. Such knowledge networks are increasingly seen as central means to foster and enhance organizations' learning and knowledge sharing, and thus companies attempt to have channels where people can congregate. Google Cafés, which are designed to encourage interactions between employees within and across teams, and to spark conversation about work as well as play [9], are a good example. In line with this argument, previous research suggests that organizational innovation performance depends on an organizational culture or work practices that can foster innovation [10–14].

Various studies provide evidence that an individual's human capital contributes to the organization's innovative capability [15–20]. The concept of human capital refers to individuals' knowledge, skills and abilities embodied in people that compel them to act in new ways [21,22]. These knowledge, skills, and abilities represent capital because they enhance productivity [23]. Human capital theory posits that individuals with more or higher quality human capital will produce more desirable outcomes [20]. However, the question of how investment in human capital, such as training which intends to increase employees' level of knowledge, skills, abilities and values, contributes to the process of enacting innovation has eluded most scholars except for a few studies [24,25]. This research endeavors to fill that void. While it is intuitive that the knowledge and competence of employees contribute to the organization's innovative capability, it is less clear how having such efforts to enhance individual human capital might transform into organization-level effort toward innovation, thereby generating differences in innovative performance across firms.

This study thus starts from the premise that companies can enhance knowledge creation through social capital, which leads to superior innovation performance. Social capital is defined as "the sum of the actual and potential resources embedded within, available through, and derived from the network of relationships possessed by an individual or social unit" [26]. The central proposition in the social capital literature is that networks of relationships constitute resources that can be used for the good of the collective [27,28]. Accordingly, the social capital theory argues that the relationship resources existing among individuals or organizational units can be a source of knowledge creation and innovation at the organizational level [27–29]. Thus, in this paper we develop arguments that speak to the role of social capital as a catalyst for innovation and arguments as to how the different dimensions of social capital affect innovation.

In doing so, this study makes many contributions. First, this study investigates the underlying process of how investments in human capital lead to superior innovation performance through details that were noted as limitations in prior research [30,31]. Considering that maximizing the impact of human capital in organizations is one of the core inquiries in strategic human resource management [32,33] and that the underlying mechanisms that account for the effects of investments designed to build human capital on innovative performance are remaining as the "black box" [30,31,34], this study is expected to contribute greatly to the related research. Second, this study verifies a comprehensive mediating model with the three dimensions of social capital and provided empirical support for the theoretical proposition that social capital is a cornerstone of innovative capabilities [2,35], through a large-scale statistical examination, which to our knowledge is one of initial attempts. By synthesizing insights from studies on social capital, this study develops a research framework that conceptually delineates and can empirically test how different aspects of social capital and their interrelationships selectively influence organizational innovative capabilities that are critical in the development of sustainable competitive advantage.

The remainder of this article is organized as follows. The following section draws research hypotheses based on a literature review of existing studies regarding human capital, social capital and corporate innovation. The research setting and measurements of both the research data and variables are then presented. Finally, the study's research findings and implications are discussed.

2. Theory and Hypotheses

2.1. Knowledge Management, Human Capital, Social Capital and Innovation

The management of knowledge is frequently identified as an important antecedent to innovation. The concept of human capital and knowledge management is that people possess skills, experience and knowledge, and therefore contribute to organizational innovative performance [16,18,19,31,36].

Organizational innovation refers to the introduction of any new product, process, or system into an organization [37]. Many scholars posit that the exchange and combination of knowledge lead to the development of new products and services [1,36,38,39]. In other words, innovation is the outcome of exchanging and combining knowledge through interaction among various actors [40,41]. Naturally, it is important for companies to establish a climate of social exchange as it aids new knowledge creation to achieve innovation performance [42–44]. In this regard, researchers recently have noted the role of social climates or relationships that facilitate the development of employee capabilities to combine and exchange information to create new knowledge [13,40,45–47].

The social capital concept originated from community studies, and was extended to the organizational social capital field in the late 1990s to study social relations between organizations and the individuals within them [27,48,49]. Characterized by durable, interconnected relationships between humans, social capital is tightly bound with the strategy of the firm [27]. As such, the development of social capital within an organization is likely to be a source of sustainable competitive advantage [27,50–52]. According to [27], social interaction is an important feature of social capital and it strongly influences the extent to which interpersonal knowledge sharing occurs. Some studies indicated that improved mutual relationships among employees has a positive effect on attitudes toward knowledge sharing [53,54]. In this regard, there is a growing interest in the role of social capital in relation to the knowledge creation and transfer processes within organizational contexts [55–57]. Thus, this study identifies, from a perspective of knowledge creation, how social capital affects innovation performance through the process of exchanging and combining knowledge.

2.2. Investment in Human Capital and Social Capital

Investment in human capital refers to processes that relate to training, education, and other professional initiatives, to increase employees' levels of knowledge, skills, abilities, and values [58,59]. A growing body of studies recently attempted to understand the mechanisms through which the skill and expertise of employees related to training and education become the predictors that improve overall corporate innovative performance [25,60,61]. Employee training and education are especially important in that they can influence and modify employees' attitudes and capabilities [40,60,62]. The training literature suggests that when employees perceive a high degree of support to develop their skills and abilities, they feel more obligated to their organization and tend to increase reciprocal attitudes, help coworkers, and engage in more organizational relationships [63–65].

Social capital is centrally concerned with the significance of relationships as a resource for social action [21,49,66,67]. In integrating the existing studies, Nahapiet and Ghoshal [27] defined social capital as having three dimensions: structural, relational, and cognitive. First, the structural dimension is an overall pattern of connections between agents; it signifies who is reached and how they are reached. Representative structural dimensions include network ties, the network configuration, and appropriable organizations. Second, the relational dimension is formed from relationships, and acts as a lever. Representative relational dimensions are trust and trustworthiness, norms and sanctions, obligations and expectations, and identification. Third, the cognitive dimension provides shared representation, interpretations, and systems of meaning. Shared codes and languages as well as shared narratives fall under this category. Based on this classification and related previous studies, we set a community of practice, which represents the overall social interactions found in organizations [54,68], as the structural dimension variable. We also set trust as the relational dimension variable, and shared codes and languages as the cognitive dimension variable [24,48]. We comprehensively investigate in the following how investment in employee training and education influences the formation of structural, relational, and cognitive social capital, i.e., communities of practice, trust, and shared codes and language, which are expected to enhance corporate innovation performance.

Communities of Practice (CoPs). A CoP is defined as a group of people informally bound by their shared expertise and passion for a joint enterprise [69]. As participation in CoPs is generally based on voluntary contributions, members' motivation to participate in a CoP becomes a central determinant of the CoP's interactions [70,71]. According to [30], such human resource practices as training and education develop employees' motivation to participate in activities that enhance firm performance. Thus, we may conjecture that employees may become more willing to discuss problems in groups, and participate in such group activities as CoPs, if they are exposed to more training opportunities. We propose the following based on this line of reasoning:

Hypothesis 1a. *Investment in human capital will enhance an organization's structural social capital in the form of CoPs.*

Trust. Trust is an implicit set of beliefs that the other party will refrain from opportunistic behavior and will not take advantage of the situation [72,73]. Employees are more likely to trust one another if they have interacted [74]. Therefore, providing training and development opportunities for employees may foster increased trust among employees by providing chances for greater communication and interaction. Additionally, specific and close interpersonal contact and knowledge regarding the other party affect the formation of trust [75–77]. Thus, we posit the following:

Hypothesis 1b. *Investment in human capital will enhance an organization's relational social capital in the form of trust.*

Shared Codes and Languages. Employees acknowledge that the organization cares for them if the organization helps them to develop their personal and professional goals. With this support, an organization's members may become more loyal to the organization [64,78] and this interaction among the organization's members, through education and training, can trigger the sharing of vision and codes [79,80]. The members concur through social interaction, which, in turn, compels them to share their values, attitudes, and goals [81,82]. Therefore, we suggest the following hypothesis:

Hypothesis 1c. *Investment in human capital will enhance an organization's cognitive social capital in the form of shared codes and language.*

2.3. Social Capital and Innovation

According to [27], new knowledge is created within organizations through the process of exchange and combination among employees. This argument implies that the exchange and combination create new knowledge by connecting previously unconnected ideas and knowledge or recombining previously connected ideas and knowledge in novel ways [40,83,84]. According to previous research, characteristics of a corporate social environment facilitate such an exchange and combination [27,44,83], which is considered to be the outcome of utilization and creation of knowledge [85,86]. We explore below how structural, relational, and cognitive social capital facilitate corporate innovation activities by triggering knowledge exchange and combination.

Communities of Practice (CoPs). Establishing a knowledge management-friendly atmosphere, including the active promotion of CoPs, will increase peoples' awareness of the necessity to share knowledge in an organization [72,87]. CoPs by nature help people jointly develop new knowledge of an organization, and can thus be beneficial in starting new lines of business, quickly solving problems, transferring best practices, and developing professional skills [69,88]. We propose the following based on this line of reasoning:

Hypothesis 2a. *A CoP will enhance the organization's innovation performance through its effects on knowledge exchange and combination.*

Trust. A social climate of trust is widely observed as essential for increasing interaction and the likelihood of information exchange between individuals [27,89]. High levels of trust also increase employees' tendencies to seek and offer help, increasing the chances for exchange [90,91]. Therefore, trust should promote the exchange of valuable ideas between knowledge workers that will, in turn, lead to greater innovation. Therefore, we suggest the following hypothesis:

Hypothesis 2b. *Trust will enhance an organization's innovation performance through its effects on knowledge exchange and combination.*

Shared Codes and Language. While trust may increase the likelihood that exchange will occur, shared codes and language facilitate both access to information and the integration of exchanged knowledge [40]. A degree of shared knowledge or understanding is essential for individuals to comprehend and integrate new knowledge, which is acquired from exchanges with other employees [65,84,92]. A climate for shared codes and language provides a common base of understanding through which individuals with disparate experience, knowledge, and backgrounds can transfer and integrate new ideas [93,94]. Greater cognitive distance requires more effort to understand and absorb what others do, and to communicate [95]. Therefore, we propose the following.

Hypothesis 2c. *Shared codes and language will enhance an organization's innovation performance through its effects on knowledge exchange and combination.*

2.4. The Mediating Role of Social Capital

This study presumed that an indirect relationship exists between investment in human capital and its innovation performance; for example, a company's investment in human capital may not guarantee its innovation performance. Specifically, not all companies can improve their innovation performance simply by increasing their budget for employee training and education, or by hiring innovative geniuses. Rather, they can improve innovation performance only when their human capital investment affects the formation of employees' structural, relational, and cognitive social capital. The role of mediation of social capital can be seen from hypothesis 1 and 2. Namely, the following relationship will be established.

Hypothesis 3. *An organization's social capital will mediate the influence of its investment in human capital on its innovation performance.*

3. Method

3.1. Research Setting and Analysis

To verify our model, we used Human Capital Corporate Panel (HCCP) data provided by the Korea Research Institute for Vocational Education and Training (KRIVET). The HCCP data is based on a human resource management and development activities survey, which has been performed biannually since 2005, in partnership with the Ministry of Employment and Labor of Korea. We used the latest survey data on 319 manufacturing firms from 2013 (5th) to test our hypotheses. The sample for the survey is randomly drawn from the KIS corporate data of all private business organizations with 100 or more employees and with more than 300 million won of capital net worth in Korea. Also, the survey data are collected from multiple informants (i.e., HRM directors, employees, strategy directors, and department managers), thus helping avoid problems associated with common method bias. Most variables in the questionnaires are measured using a 5-point scale, ranging from 1 (strongly disagree) to 5 (strongly agree).

3.2. Focal Variables

Investment in human capital. To measure investment in human capital, which is the independent variable of this study, we used annual costs spent in group-based training and education. The survey asked each company to identify costs for the three types of employee training and education separately-i.e., costs for individual-based training and education (e-learning and online book-learning), costs for group-based training and education (internal and external), and financial support for external education. Among those, we utilized costs for group-based training and education, which are expected to have maximal effects on the formation of social capital at the organizational level. The usage of actual investment data in this study leads us to overcome the limitation of most previous studies, which used subjective measurements for human capital investment [96,97]. We calculated this variable by dividing the total amount of group-based training and education expenses by the number of total employees to obtain per capita spending on employee training [25].

Social Capital. To measure social capital, our mediating variable, we utilized the following survey data. Firstly, we assessed structural social capital—the level of CoP activities involvement of employees—with a questionnaire item, "In our company, employees are actively engaged in mutual learning among," adapted from [98]. Secondly, we measured relational social capital—the strength of trust among employees—with a questionnaire item, "In our company, employees have trusting relationships," following [99]. Lastly, we assessed cognitive social capital—the degree of shared code and language—with the following questionnaire item, "Our company shares company norms and information into details to the employees through management or the information system," based on study [27]. These three items were measured with a 5-point scale of (1) absolutely not; (2) preferably not; (3) neutral; (4) possibly; (5) definitely. As questionnaire items were limited and generally facet-based, we used a single measure approach instead of using multiple-item measures.

Innovation Performance. To measure innovation performance, we considered multiple dimensions of innovative performance such as new product development, product and service differentiation and new customer recruitment suggested by previous studies [100,101]. We used ratings for the following three questions: (1) "Our company had competitive advantage over other companies in new product development"; (2) "Our company had competitive advantage over other companies in introducing differentiation in the products and/or services"; and (3) "Our company had higher ratio of new customers compared to the industry average". The measurement was based on a 5-point scale (1 = much lower than the industry average, and 5 = much higher than the industry average) and was rated by employees. We aggregated individual responses to analyze the organization level of innovation performance.

3.3. Control Variables

To control for industry-level and firm-specific effects, which may influence corporate innovation performance, we included various control variables.

First, the nature of the industries that organizations compete in is known to influence their innovative capabilities [2,102,103]. Therefore, we controlled the effects of environmental change. The degree of environmental change was measured with a questionnaire item, "To what extent did the industry your company competes in experience demand changes in the last two years?" and a questionnaire item, "To what extent did the industry your company competes in experience new product development or introduction in the last two years?" The measurement was based on a 4-point scale (1 = not at all, and 4 = a great deal) and was rated by strategy directors. We aggregated ratings of the two questionnaire items.

Second, this study also controlled the overall level of human capital in a firm. Many previous studies posit that the overall skill, expertise, and knowledge levels of an organization's employees are crucial for organizational innovation performance [2,23,104,105]. We used survey ratings on the level of overall skills and expertise of employees in five functional areas- R&D, sales and service, engineering, management, and production. The measurement was based on a 5-point scale (1 = much lower than the industry average, and 5 = much higher than the industry average) and rated by employees at all levels. We used the average rating of the five questionnaire items.

Third, we controlled the effects of prior performance, considering that firms may increase innovation performance when they have greater slack resources [2,106,107]. The survey reported ratings on prior performance by strategy directors with a questionnaire item, "Our company lacked resource capacity due to economic downturn or management crisis." The measurement was based on a 5-point scale (1 = not at all, and 5 = a great deal) and was rated by strategy directors.

Finally, we controlled the size of an organization that may play a key role in innovative capability. Previous research assumes that the extensive resources of large organizations may be more likely to develop innovative capabilities [2,108]. We divided the size of firms based on the number of employees into below 300, 300–1000, and above 1000 and set dummy variables for each.

4. Results

Table 1 reports the means, standard deviations, and correlations of all variables. Generally, the results showed significant correlations between dependent and independent variables. The variance inflation factors (VIFs) are well below the cut-off point of 10, suggesting little multicolinearity in our data.

Table 1. Means, Standard Deviations, and Correlations.

	Mean	**SD**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1. Innovation Performance	10.24	1.36	1								
2. Human Capital Investment	0.11	0.15	0.19 **	1							
3. Structural Social Capital	2.63	0.38	0.23 **	0.19 **	1						
4. Relational Social Capital	3.53	0.35	0.34 **	0.25 **	0.48 **	1					
5. Cognitive Social Capital	3.27	0.47	0.30 **	0.18 **	0.41 **	0.66 **	1				
6. Environment Change	5.45	1.49	0.34 **	0.09	0.18 **	0.09 **	0.09	1			
7. Human Capital Level	2.99	0.48	0.32 **	0.10	0.17 **	0.22 **	0.15 **	0.10	1		
8. Prior Performance	2.50	1.07	−0.25 **	−0.10	−0.12 *	−0.14 *	−0.09	−0.34 **	−0.17 **	1	
9. Firm Size	750.9	1982.5	0.23 **	0.11	0.20 **	0.18 **	0.06	0.11	0.13 *	−0.08	1

* Correlation is significant at the 0.05 level (2-tailed); ** Correlation is significant at the 0.01 level (2-tailed).

We used structural equation modeling (SEM) to examine the relationships among investment in human capital, the mediating effects of social capital, and innovation performance. Figure 1 displays the full mediation model with standardized path coefficients. Results show that the model had a good fit to the data (χ = 37.465, df = 13, p < 0.001, RMSEA = 0.077, CFI = 0.946, NFI = 0.925, TLI = 0.815), supporting the hypothesized relationships of human capital investment with structural social capital (β = 0.48, p < 0.01), relational social capital (β = 0.58, p < 0.01) and cognitive social capital (β = 0.56, p < 0.01); relational social capital with innovation performance (β = 0.53, p < 0.05); and cognitive social capital with innovation performance (β = 0.37, p < 0.05). Therefore, hypotheses 1a, 1b, 1c, 2b and 2c were supported. However, the effect of structural social capital on innovative performance was non-significant, disconfirming Hypothesis 2a. Hence, hypothesis 3 was supported except for the variable of structural social capital.

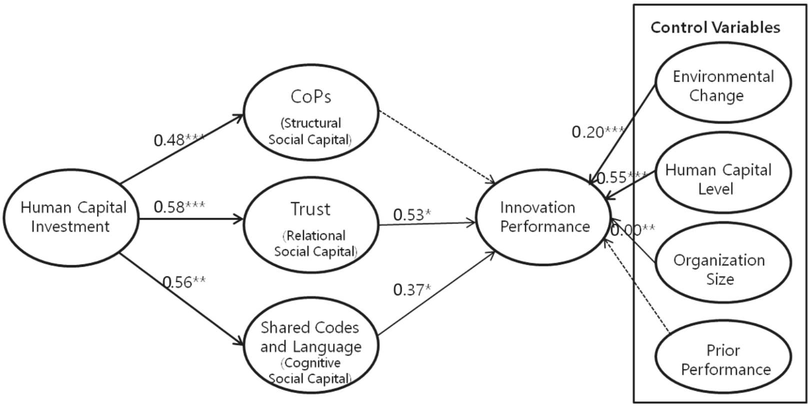


Figure 1. Structural Equation Model (Full Mediation). Note: Nonsignificant paths are depicted as dotted lines in the diagram. + p < 0.10; * p < 0.05; ** p < 0.01.

To further examine whether social capital fully or partially mediated the relationship between human capital investment and innovation performance, we tested an alternative model that included a direct path from the independent variable to the outcome variable. This partial mediation model also provided a good fit to the data (χ = 35.521, df = 12, p < 0.001, RMSEA = 0.079, CFI = 0.949, NFI = 0.929, TLI = 0.807), however, the fully mediated structural model demonstrated a better fit to the data than the partially mediated model (see Table 2).

Table 2. Fit indices for baseline, fully mediated, and partially mediated models.

	CMIN	df	CMIN/df	RMSEA	CFI	NFI	TLI	$\triangle\chi 2$
Partially mediated model	35.521	12	2.960/12	0.079	0.949	0.929	0.807	
Fully mediated model	37.465	13	2.882/13	0.077	0.946	0.925	0.815	1.94

In addition, as shown in Figure 1, several of our control variables exhibited significant effects on the innovation performance variable. Environmental change (β = 0.20, p < 0.01), human capital level (β = 0.55, p < 0.01), and firm size (β = 0.01, p < 0.01) were significantly related to innovation performance.

5. Discussion

The objective of this research was to verify the role of social capital in organization's innovation in relation to knowledge management. For this, we elaborated on how the different dimensions of social capital affect innovation and tested a detailed model of investment in human capital's effects on corporate innovation performance. The results demonstrated that social capital is an important mediator in realizing organizational innovation performance. Specifically, social capital played an essential role in connecting investment in human capital with an organization's innovative performance through its effects on knowledge exchange and combination.

The implications of this study can be summarized as follows. Theoretically, this study is meaningful in that it displayed and verified the mediating role of social capital in knowledge creation in all three dimensions—structural, relational, and cognitive—with large-scale statistical examination. The findings contribute to the social capital literature by enriching and providing strong empirical support for the theoretical proposition that social capital is a cornerstone of innovative capabilities [2,35]. Also, the findings provide support for the argument that social capital serves as a catalyst for organizational innovation by triggering knowledge exchange and combination among employees [27,50,51]. Also, this study verified the positive effect of group-based investment in human capital on organizational innovation performance. Considering that the theoretical framework for the relationship between investment in human capital and organizational innovation performance has been subject to considerable debate [25,109], this study is expected to make contributions to the strategic human resource management literature, laying a foundation on the efficacy of different types of investment in human capital on corporate innovation. In noting such findings, scholars may further elaborate upon the efficacy of group-based human capital investment in future studies. Lastly, the results indicate that to effectively leverage investments in human capital firms should invest in the organizational design that will enhance social capital among employees. That is, company manages may benefit from establishing a climate of social exchange that will encourage employees to build trust and share their expertise, to enhance organizational innovation performance that will be crucial to achieve the sustained competitive advantages.

Naturally, our study is subject to some limitations. The first concerns the cultural context of our data. As pointed out by some scholars previously, cultural differences may influence the formation of social capital [110,111]. Therefore, further study that utilizes cross-cultural data may be necessary to generalize the findings of this study. Secondly, the use of the archived data can be a disadvantage in that some data do not fully cover the domains of the constructs we attempted to measure. We used single-item measures instead of multiple-item scales to measure social capital. While some scholars argue for the practical virtues of single-item measures [112,113], future research may try multiple-item measures to better operationalize each dimension of social capital and to enhance reliability. In addition, we did not examine whether other organizational factors limited the positive impact of social capital on corporate innovation. Future research may look at potential moderators of the relationships between social capital and organizational innovation capability to determine what types of firms are most likely to benefit from organizational level arrangement for social exchange. Thirdly, it would be interesting to examine the effects of social capital on different types of corporate innovation capabilities [2,24] to verify the efficacy of each dimension of social capital on corporate innovation. Finally, we limit our study to cross-sectional designs. The implications may be further strengthened by the use of longitudinal data, as the impact of investment in human capital is not immediate but embedded and realized through time [114,115]. We hope that future research will challenge and verify these issues further.

Author Contributions: Se-Yeon Ahn and So-Hyung Kim contributed equally to this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Madhavan, R.; Grover, R. From Embedded Knowledge to Embodied Knowledge: New Product Development as Knowledge Management. *J. Mark.* **1998**, *62*, 1–12. [CrossRef]
2. Subramaniam, M.; Youndt, M.A. The Influence of Intellectual Capital on the Types of Innovation Capabilities. *Acad. Manag. J.* **2005**, *48*, 450–463. [CrossRef]
3. Damanpour, F. Organizational Innovation: A Meta-Analysis of Effects of Determinants and Moderators. *Acad. Manag. J.* **1991**, *34*, 555–590. [CrossRef]
4. Hurley, R.F.; Hult, G.T.M. Innovation, Market Orientation, and Organizational Learning: An Integration and Empirical Examination. *J. Mark.* **1998**, *62*, 42–54. [CrossRef]
5. Lew, Y.K.; Sinkovics, R.R. Crossing Borders and Industry Sectors: Behavioral Governance in Strategic and Product Innovation for Competitive Advantage. *Long Range Plan.* **2013**, *46*, 13–38. [CrossRef]
6. Mahr, D.; Lievens, A.; Blazevic, V. The value of customer cocreated knowledge during the innovation process. *J. Prod. Innov. Manag.* **2014**, *31*, 599–615. [CrossRef]
7. Romijn, H.; Albaladejo, M. Determinants of innovation capability in small electronics and software firms in southeast England. *Res. Policy* **2002**, *31*, 1053–1067. [CrossRef]
8. Ge, C.; Huang, K.W.; Png, I.P.L. Engineer/scientist careers; patents, online profiles, and misclassification bias. *Res. Policy* **2016**, *37*, 1053–1067. [CrossRef]
9. Lave, J.; Wenger, E. Situated learning: Legitimate peripheral participation. *Learn. Doing* **1991**, *95*, 138. [CrossRef]
10. Ahmed, P.K. European Journal of Innovation Culture and climate for innovation Culture and climate for. *Eur. J. Innov. Manag.* **1998**, *1*, 30–43. [CrossRef]
11. Martins, E.C.; Terblanche, F. Building organisational culture that stimulates creativity and innovation. *Eur. J. Innov. Manag.* **2003**, *6*, 64–74. [CrossRef]
12. McLean, L.D. Organizational Culture's Influence on Creativity and Innovation: A Review of the Literature and Implications for Human Resource Development. *Adv. Dev. Hum. Resour.* **2005**, *7*, 226–246. [CrossRef]
13. Donate, M.J.; Guadamillas, F. An empirical study on the relationships between knowledge management, knowledge-oriented human resource practices and innovation. *Knowl. Manag. Res. Pract.* **2013**, *13*, 134–148. [CrossRef]
14. Longo, M.C.; Narduzzo, A. Transactive knowledge from *communities of practice* to firms: An empirical investigation of innovative projects performance. *Eur. J. Innov. Manag.* **2017**, *20*, 291–311. [CrossRef]
15. Kazadi, K.; Lievens, A.; Mahr, D. Stakeholder co-creation during the innovation process: Identifying capabilities for knowledge creation among multiple stakeholders. *J. Bus. Res.* **2015**, *69*, 525–540. [CrossRef]
16. Bantel, K.A.; Jackson, S.E. Top management and innovations in banking: Does the composition of the top team make a difference? *Strateg. Manag. J.* **1989**, *10*, 107–124. [CrossRef]
17. Durham, K.; Kennedy, B. Intellectual capital: Realizing your company s true value by finding its hidden brainpower. *Res. Technol. Manag.* **1997**, *40*, 59.
18. Hitt, M.A.; Biermant, L.; Shimizu, K.; Kochhar, R. Direct and Moderating Effects of Human Capital on Strategy and Performance in Professional Service Firms: A Resource-Based Perspective. *Acad. Manag. J.* **2001**, *44*, 13–28. [CrossRef]

19. Skaggs, B.C.; Youndt, M. Strategic positioning, human capital, and performance in service organizations: A customer interaction approach. *Strateg. Manag. J.* **2004**, *25*, 85–99. [CrossRef]
20. Minh, N.V.; Badir, Y.F.; Quang, N.N.; Afsar, B. The impact of leaders' technical competence on employees' innovation and learning. *J. Eng. Technol. Manag-JET-M* **2017**, *44*, 44–57. [CrossRef]
21. Coleman, J.S. Social Capital in the Creation of Human Capital. *Am. J. Sociol.* **1988**, *94*, S95–S120. [CrossRef]
22. Coff, R.W. Human capital, shared expertise, and the likelihood of impasse in corporate acquisitions. *J. Manag.* **2002**, *28*, 107–128. [CrossRef]
23. Snell, S.; Dean, J.W. Integrated Manufacturing and Human Resource Management: A Human Capital Perspective. *Acad. Manag. J.* **1992**, *35*, 467–504. [CrossRef]
24. Chen, C.; Huang, J. Strategic human resource practices and innovation performance—The me,diating role of knowledge management capacity. *J. Bus. Res.* **2009**, *62*, 104–114. [CrossRef]
25. Sung, S.Y.; Choi, J.N. Do organizations spend wisely on employees? Effects of training and development investments on learning and innovation in organizations. *J. Organ. Behav.* **2014**, *35*, 393–412. [CrossRef] [PubMed]
26. Zheng, W. A social capital perspective of innovation from individuals to nations: Where is empirical literature directing us? *Int. J. Manag. Rev.* **2010**, *12*, 151–183. [CrossRef]
27. Nahapiet, J.; Ghoshal, S. Capital, Social Capital, Intellectual Capital, and Organizational Advantage. *Acad. Manag. Rev.* **1998**, *23*, 242–266. [CrossRef]
28. Dakhli, M.; De Clercq, D. Human capital, social capital, and innovation: A multi-country study. *Entrep. Reg. Dev.* **2004**, *16*, 107–128. [CrossRef]
29. Parker, A.; Halgin, D.S.; Borgatti, S.P. Dynamics of Social Capital: Effects of Performance Feedback on Network Change. *Organ. Stud.* **2015**, *37*, 375–397. [CrossRef]
30. Huselid, M.A.; Becker, B.E. Bridging micro and macro domains: Workforce differentiation and strategic human resource management. *J. Manag.* **2011**, *37*, 421–428. [CrossRef]
31. Jiang, K.; Lepak, D.P.; Hu, J.; Baer, J.C. How does human resource management influence organizational outcomes? A meta-analytic investigation of mediating mechanisms. *Acad. Manag. J.* **2012**, *55*, 1264–1294. [CrossRef]
32. Becker, B.; Huselid, M.A. Strategic Human Resources Management: Where Do We Go From Here? *J. Manag.* **2006**, *32*, 898–925. [CrossRef]
33. Crook, T.R.; Todd, S.Y.; Combs, J.G.; Woehr, D.J.; Ketchen, D.J., Jr. Does human capital matter? A meta-analysis of the relationship between human capital and firm performance. *J. Appl. Psychol.* **2011**, *96*, 443. [CrossRef] [PubMed]
34. Boxall, P.; Guthrie, J.P.; Paauwe, J. Editirial Introduction:progressing our understadong of the mediating variables linking HRM, employee well-being and organizational performance. *Hum. Resour. Manag. J.* **2016**, *26*, 103–111. [CrossRef]
35. Camps, S.; Marques, P. Exploring how social capital facilitates innovation: The role of innovation enablers. *Technol. Forecast. Soc. Chang.* **2014**, *88*, 325–348. [CrossRef]
36. Yang, Y.; Konrad, A.M. Diversity and organizational innovation: The role of employee involvement. *J. Organ. Behav.* **2011**, *32*, 1062–1083. [CrossRef]
37. Suranyi-Unger, T. *Innovation*; Encyclopedia of Economics, McGraw-Hill: New York, NY, 1994.
38. Nonaka, I.; Takeuchi, H. *The Knowledge-Creating: How Japanese Companies Create the Dynamics of Innovation*; Oxford University Press: Oxford, UK, 1995; Volume 3, pp. 25–27. [CrossRef]
39. Stewart, G. Supply-chain operations reference model (SCOR): The first cross-industry framework for integrated supply-chain management. *Logist. Inf. Manag.* **1997**, *10*, 62–67. [CrossRef]
40. Collins, C.J.; Smith, K.G. Knowledge Exchange and Combination: The Role of Human Resource Practices in the Performance of High-Technology Firms. *Acad. Manag. J.* **2006**, *49*, 544–560. [CrossRef]
41. Landry, R.; Amara, N.; Lamari, M. Does social capital determine innovation? To what extent. *Technol. Forecast. Soc. Chang.* **2002**, *69*, 681–701. [CrossRef]
42. Laursen, K.; Foss, N. New human resource management practices, complementarities and the impact on innovation performance. *Camb. J. Econ.* **2003**, *27*, 243–263. [CrossRef]
43. Scarbrough, H. Knowledge management, HRM and the innovation process. *Int. J. Manpow.* **2003**, *24*, 501–516. [CrossRef]

44. Hau, Y.S.; Kim, B.; Lee, H.; Kim, Y.G. The effects of individual motivations and social capital on employees' tacit and explicit knowledge sharing intentions. *Int. J. Inf. Manag.* **2013**, *33*, 356–366. [CrossRef]
45. Bowen, D.E.; Ostroff, C. Understanding HRM-firm performance linkages: The role of the "strength" of the HRM system. *Acad. Manag. Rev.* **2004**, *29*, 203–221. [CrossRef]
46. Van Reijsen, J.; Helms, R.; Batenburg, R.; Foorthuis, R. The impact of knowledge management and social capital on dynamic capability in organizations. *Knowl. Manag. Res. Pract.* **2015**, *13*, 401–417. [CrossRef]
47. Ahammad, M.F.; Tarba, S.Y.; Liu, Y.; Glaister, K.W. Knowledge transfer and cross-border acquisition performance: The impact of cultural distance and employee retention. *Int. Bus. Rev.* **2016**, *25*, 66–75. [CrossRef]
48. Tsai, W.; Ghoshal, S. Social capital and value creation: The role of intrafirm networks. *Acad. Manag. J.* **1998**, *41*, 464–476. [CrossRef]
49. Burt, R.S. *Structural Holes: The Social Structure of Competition*; Harvard University Press: Cambridge, MA, USA, 1992; pp. 38–40.
50. Adler, P.; Kwon, S. Social Capital: Prospects for a New Concept. *Acad. Manag. Rev.* **2002**, *27*, 17–40. [CrossRef]
51. Moran, P. Structural vs. relational embeddedness: Social capital and managerial performance. *Strateg. Manag. J.* **2005**, *26*, 1129–1151. [CrossRef]
52. Saeidi, S.P.; Sofian, S.; Saeidi, P.; Saeidi, S.P.; Saaeidi, S.A. How does corporate social responsibility contribute to firm financial performance? The mediating role of competitive advantage, reputation, and customer satisfaction. *J. Bus. Res.* **2015**, *68*, 341–350. [CrossRef]
53. Bock, G.-W.; Zmud, R.W.; Kim, Y.-G.; Lee, J.-N. Behavioral Intention Formation in Knowledge Sharing: Examining the Roles of Extrinsic Motivators, Social-Psychological Forces, and Organizational Climate. *MIS Q.* **2005**, *29*, 87–111. [CrossRef]
54. Chiu, C.-M.; Hsu, M.-H.; Wang, E.T.G. Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. *Decis. Support. Syst.* **2006**, *42*, 1872–1888. [CrossRef]
55. Zhang, M.; Lettice, F.; Zhao, X. The impact of social capital on mass customisation and product innovation capabilities. *Int. J. Prod. Res.* **2015**, *53*, 5251–5264. [CrossRef]
56. Bharati, P.; Zhang, W.; Chaudhury, A. Better knowledge with social media? Exploring the roles of social capital and organizational knowledge management. *J. Knowl. Manag.* **2015**, *19*, 456–475. [CrossRef]
57. Li, Y.; Chen, H.; Liu, Y.; Peng, M.W. Managerial ties, organizational learning, and opportunity capture: A social capital perspective. *Asia Pac. J. Manag.* **2014**, *31*, 271–291. [CrossRef]
58. Collins, C.J.; Clark, K.D. Strategic Human Resource Practices, Top Management Team Social Networks, and Firm Performance: The Role of Human Resource Practices in Creating Organizational Competitive Advantage. *Acad. Manag. J.* **2003**, *46*, 740–751. [CrossRef]
59. Wright, P.; Dunford, B.; Snell, S. Human resources and the resource based view of the firm. *J. Manag.* **2001**, *27*, 701–721. [CrossRef]
60. Marchington, M.; Grugulis, I. "Best practice" human resource management: Perfect opportunity or dangerous illusion? *Int. J. Hum. Resour. Manag.* **2000**, 37–41. [CrossRef]
61. Fulmer, I.S.; Gerhart, B.; Scott, K.S. Are the 100 best better? An empirical investigation of the relationship between being a "great place to work" and firm performance. *Pers. Psychol.* **2003**, *56*, 965–993. [CrossRef]
62. Lester, S. Professional standards, competence and capability. *High. Educ. Skill Work Learn. Inf.* **2014**, *4*, 31–43. [CrossRef]
63. Ahmad, K.Z.; Bakar, R.A. The association between training and organizational commitment among white-collar workers in Malaysia. *Int. J. Train. Dev.* **2003**, *7*, 166–185. [CrossRef]
64. Allen, D.G.; Shore, L.M.; Griffeth, R.W. The Role of Perceived Organizational Support and Supportive Human Resource Practices in the Turnover Process. *J. Manag.* **2003**, *29*, 99–118. [CrossRef]
65. Cohen, A. Commitment before and after: An evaluation and reconceptualization of organizational commitment. *Hum. Resour. Manag. Rev.* **2007**, *17*, 336–354. [CrossRef]
66. Baker, W.E. Market networks and corporate behavior. *Am. J. Sociol.* **1990**, *96*, 589–625. [CrossRef]
67. Stam, W.; Arzlanian, S.; Elfring, T. Social capital of entrepreneurs and small firm performance: A meta-analysis of contextual and methodological moderators. *J. Bus. Ventur.* **2014**, *29*, 152–173. [CrossRef]
68. Bolino, M.C.; Turnley, W.H. Bloodgood J.M. Citizenship behavior and the creation of social capital in organizations. *Acad. Manag. Rev.* **2002**, *27*, 505–522. [CrossRef]

69. Wenger, E.C.; Snyder, W.M. Communities of Practice: The Organizational Frontier. *Harv. Bus. Rev.* **2000**, *78*, 139–145. [CrossRef]
70. Teigland, R. *Knowledge Networking: Structure and Performance in Networks of Practice (Doctoral Dissertation)*; Institutional Business, Stockholm School of Economics: Stockholm, Sweden, 2003.
71. Pandey, S.C.; Dutta, A. *Communities of practice* and organizational learning: Case study of a global IT solutions company. *Strateg. HR Rev.* **2013**, *12*, 255–261. [CrossRef]
72. Ridings, C.M.; Gefen, D.; Arinze, B. Some antecedents and effects of trust in virtual communities. *J. Strateg. Inf. Syst.* **2002**, *11*, 271–295. [CrossRef]
73. Hosmer, L.T. Trust: The Connecting Link between Organizational Theory and Philosophical Ethics. *Acad. Mamag. Rev.* **1995**, *20*, 379–403. [CrossRef]
74. Whitener, E.M.; Brodt, S.E.; Korsgaard, M.A.; Werner, J.M. Managers as initiators of trust: An exchange relationship framework for understanding managerial trustworthy behavior. *Acad. Manag. Rev.* **1998**, *23*, 513–530. [CrossRef]
75. Bstieler, L. Trust formation in collaborative new product development. *J. Prod. Innov. Manag.* **2006**, *23*, 56–72. [CrossRef]
76. Bierly, P.E.; Damanpour, F.; Santoro, M.D. The application of external knowledge: Organizational conditions for exploration and exploitation. *J. Manag. Stud.* **2009**, *46*, 481–509. [CrossRef]
77. Kianto, A.; Ritala, P.; Spender, J-C.; Vanhala, M. The interaction of intellectual capital assets and knowledge management practices in organizational value creation. *J. Intellect. Cap.* **2014**, *15*, 362–375. [CrossRef]
78. Shore, L.M.; Wayne, S.J. Commitment and employee behavior: Comparison of affective commitment and continuance commitment with perceived organizational support. *J. Appl. Psychol.* **1993**, *78*, 774–780. [CrossRef] [PubMed]
79. Huh, M.G. Knowledge Search and Innovation. *J. Bus. Res.* **2011**, *40*, 1247–1271.
80. Reiche, B.S.; Harzing, A.; Pudelko, M. Why and how does shared language affect subsidiary knowledge inflows[quest] A social identity perspective. *J. Int. Bus. Stud.* **2015**, *46*, 528–551. [CrossRef]
81. Gulati, R.; Sytch, M. Does familiarity breed trust? Revisiting the antecedents of trust. *Manag. Decis. Econ.* **2008**, *29*, 165–190. [CrossRef]
82. Burgers, J.H.; Covin, J.G. The contingent effects of differentiation and integration on corporate entrepreneurship. *Strateg. Manag. J.* **2016**, *37*, 521–540. [CrossRef]
83. Kogut, B.; Zander, U. Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology. *Organ. Sci.* **1992**, *3*, 383–397. [CrossRef]
84. Roy, R.; Sarkar, M. Knowledge, firm boundaries, and innovation: Mitigating the incumbent's curse during radical technological change. *Strateg. Manag. J.* **2016**, *37*, 835–854. [CrossRef]
85. Cohen, W.M.; Levinthal, D.A. Absorptive Capacity: A New Perspective on Learning and Innovation. *Adm. Sci. Q.* **1990**, *35*, 128–152. [CrossRef]
86. Rao, H.; Drazin, R. Overcoming resource constraints on product innovation by recruiting talent from rivals: A study of the mutual fund industry, 1986–1994. *Acad. Manag. J.* **2002**, *45*, 491–507. [CrossRef]
87. Von Krogh, G. Care in Knowledge Creation. *Calif. Manag. Rev.* **1998**, *40*, 133–153. [CrossRef]
88. Ngah, R.; Tai, T.; Bontis, N. Knowledge Management Capabilities and Organizational Performance in Roads and Transport Authority of Dubai: The mediating role of Learning Organization. *Knowl. Process. Manag.* **2016**, *23*, 184–193. [CrossRef]
89. Mayer, R.C.; Davis, J.H.; Schoorman, F.D. an Integrative Model of Organizational Trust. *Acad. Manag. Rev.* **1995**, *20*, 709–734. [CrossRef]
90. Jones, G.R.; George, J.M. The experience and evolution of trust: Implications for cooperation and teamwork. *Acad. Manag. Rev.* **1998**, *23*, 531–546. [CrossRef]
91. McEvily, B.; Zaheer, A.; Kamal, D.K.F. Mutual and Exclusive: Dyadic Sources of Trust in Interorganizational Exchange. *Organ. Sci* **2017**, *28*, 74–92. [CrossRef]
92. Hansen, M.T. Knowledge Networks: Explaining Effective Knowledge Sharing in Multiunit Companies. *Organ. Sci.* **2002**, *13*, 232–248. [CrossRef]
93. Szulanski, G. Impediments to the transfer of best practice within the firm. *Strateg. Manag. J.* **1996**, *17*, 27–43. [CrossRef]
94. Malik, A. Knowledge integration mechanisms in high-technology business-to-business services vendors. *Knowl. Manag. Res. Pract.* **2016**, *14*, 537–564. [CrossRef]

95. Bogenrieder, I.; Nooteboom, B. Learning Groups: What Types are there? A Theoretical Analysis and an Empirical Study in a Consultancy Firm. *Organ. Stud.* **2004**, *25*, 287–313. [CrossRef]
96. Macky, K.; Boxall, P. The Relationship between "High.-Performance Work Practices" and Employee Attitudes: An Investigation of Additive and Interaction Effects. *Int. J. Hum. Resour. Manag.* **2007**, *18*, 537–567. [CrossRef]
97. Sels, L. More is not necessarily better: The relationship between the quantity and quality of training efforts. *Int. J. Hum. Resour. Manag.* **2002**, *13*, 1279–1298. [CrossRef]
98. McEvily, B.; Argote, L.; Reagans, R. Managing Knowledge in Organizations: An Integrative Framework and Review of Emerging Themes. *Manag. Sci.* **2003**, *49*, 571–582.
99. Mayer, R.C.; Davis, J.H. The Effect of the Performance Appraisal System on Trust for Management: A Field Quasi-Experiment. *J. Appl. Psychol.* **1999**, *84*, 123–136. [CrossRef]
100. Zhou, H.; Dekker, R.; Kleinknecht, A. Flexible labor and innovation performance: Evidence from longitudinal firm-level data. *Ind. Corp. Chang.* **2011**, *20*, 941–968. [CrossRef]
101. Shipton, H.; West, M.A.; Patterson, M.; Birdi, K.; Dawson, J. Organizational learning as a predictor of innovation. *Hum. Resour. Manag. J.* **2006**, *16*, 3–27. [CrossRef]
102. Benamati, J.; Lederer, A.L. Coping With Rapid Changes in It. *Commun. ACM* **2001**, *44*, 83–87. [CrossRef]
103. Langerak, F.; Hultink, E.J.; Robben, H.S.J. The mediating role of new product development in the link between market orientation and organizational performance. *J. Strateg. Mark.* **2007**, *15*, 281–305. [CrossRef]
104. Schultz, T.W. Investment in Human Capital. *Am. Econ. Rev.* **1961**, *51*, 1–17. [CrossRef]
105. Smith, K.; Collins, C.; Clark, K. Existing knowledge, knowledge creation capability, and the rate of new product introduction in high-technology firms. *Acad. Manag. J.* **2005**, *48*, 346–357. [CrossRef]
106. Autio, E.; Sapienza, H.J.; Almeida, J.G. Effects of age at entry, knowledge intensity, and imitability on international growth. *Acad. Manag. J.* **2000**, *43*, 909–924. [CrossRef]
107. Hill, C.W.L.; Rothaermel, F.T. The performance of incumbent firms in the face of radical technological innovation. *Acad. Manag. Rev.* **2003**, *28*, 257–274. [CrossRef]
108. Henderson, R.; Cockburn, I. Measuring Competence? Exploring Firm Effects in Pharmaceutical Research. *Strateg. Manag. J.* **1994**, *15*, 63–84. [CrossRef]
109. Nguyen, T.N.; Truong, Q.; Buyens, D. The relationship between training and firm performance: A literature Review. *Res. Pract. Hum. Resour. Manag.* **2010**, *18*, 36–45.
110. Allik, J.; Realo, A. Individualism-collectivism and social capital. *J. Cross Cult. Psychol.* **2004**, *35*, 29–49. [CrossRef]
111. Ji, Y.G.; Hwangbo, H.; Yi, J.S.; Rau, P.L.P.; Fang, X.; Ling, C. The influence of cultural differences on the use of social network services and the formation of social capital. *Int. J. Hum. Comput. Interact.* **2010**, *26*, 1100–1121. [CrossRef]
112. Postmes, T.; Haslam, S.A.; Jans, L. A single-item measure of social identification: Reliability, validity, and utility. *Br. J. Soc. Psychol.* **2013**, *52*, 597–617. [CrossRef] [PubMed]
113. Fisher, G.G.; Matthews, R.A.; Gibbons, A.M. Developing and Investigating the Use of Single-Item Measures in Organizational Research. *J. Occup. Health Psychol.* **2015**, *21*. [CrossRef] [PubMed]
114. Datta, D.K.; Guthrie, J.P.; Wright, P.M. Human resource management and labor productivity: Does industry matter? *Acad. Manag. J.* **2005**, *48*, 135–145. [CrossRef]
115. Tsui, A.S.; Wu, J.B. The new employment relationship versus the mutual investment approach: Implications for human resource management. *Hum. Resour. Manag.* **2005**, *44*, 115–121. [CrossRef]

sustainability

MDPI

Article

Existing Knowledge Assets and Disruptive Innovation: The Role of Knowledge Embeddedness and Specificity

Chunpei Lin [1], Baixun Li [2,*] and Yenchun Jim Wu [3,4]

[1] Business Management Research Center, School of Business Administration, Huaqiao University, Quanzhou 362021, China; alchemist@hqu.edu.cn

[2] School of Business Administration, Guangdong University of Finance and Economics, Guangzhou 510320, China

[3] Graduate Institute of Global Business and Strategy, National Taiwan Normal University, Taipei 10645, Taiwan; wuyenchun@gmail.com

[4] School of Business Administration, Huaqiao University, Quanzhou 362021, China

* Correspondence: libaixun@gdufe.edu.cn

Received: 31 December 2017; Accepted: 26 January 2018; Published: 29 January 2018

Abstract: Disruptive innovation has created a significant impact on management practices and academia. This study investigated the impact of existing knowledge assets on disruptive innovation by analyzing the role of knowledge embeddedness and specificity. We conducted a hierarchical regression analysis by using survey data from 173 Chinese industrial firms to test the direct and indirect effects of knowledge embeddedness and specificity on disruptive innovation, which can be divided into outward-oriented and internal-oriented disruptive innovation. The results indicated that knowledge embeddedness not only played a positive role in knowledge specificity, but also had a positive effect on outward-oriented disruptive innovation. Furthermore, knowledge specificity exhibited opposite functions in outward-oriented and internal-oriented disruptive innovation. In addition, knowledge specificity mediated the relationship between knowledge embeddedness and outward-oriented (internal-oriented) disruptive innovation.

Keywords: knowledge assets; knowledge embeddedness; knowledge specificity; disruptive innovation

1. Introduction

Today, more companies have become increasingly aware of social and environmental pressures. The technological revolution characterized by intelligence, greenness, and ubiquity is booming and profoundly affecting the environment in which we work and live [1]. For example, Green information technology can be deployed to support a variety of sustainability initiatives such as to measure carbon footprints, reduce waste in business processes, lower resource consumption, and reduce greenhouse gas emissions [2]. Many scholars and consultants have argued that this technological revolution offers terrific opportunities for progressive organizations [3–5], as innovation is one of the primary means by which companies can achieve sustainable growth [6]. In particular, disruptive innovation is an important type of enterprise innovation and an important strategic tool for enterprises to realize technological revolution and achieve sustainable growth [7–9].

By referring to Adner [7], Christensen [8], and Govindarajan et al. [9], disruptive innovation can be viewed as a new product or service that introduces a different set of performance attributes relative to what already exists, and this set of attributes is initially attractive to an emerging customer segment. Case studies from Rafii and Kampas [10], Husig and Hipp [11], and Keller and Hüsig [12] found that disruptive innovation can be seen either as a threat or opportunity. When a disruptive product is introduced, but does not erode the existing market space, it will be perceived as an opportunity to expand into emerging markets. Thus, the disruptive product is given priority for resource allocation. Conversely, when the introduction of a disruptive product erodes the existing market, it can be considered as a potential threat and may force the enterprise to adjust its existing operating mode [13]. However, due to the impact of organizational inertia, the enterprise may resist this self-adjustment. Based on the above analysis, we divided disruptive innovation into outward-oriented disruptive innovation and internal-oriented disruptive innovation. Outward-oriented disruptive innovation refers to the process of introducing disruptive products into external markets to open up a new market or erode a rival's market share. Internal-oriented disruptive innovation refers to the process of introducing disruptive products into the existing product market and even completely replacing the existing market. The significant opportunities and challenges that disruptive innovation brings to business have made it one of the most influential innovation management theories over the past decade. Some companies have succeeded in initiating disruptive innovation. For example, Apple successfully ported the "iPod + iTunes" model to the mobile phone market and launched the "iPhone + App Store" mode. This led other mobile phone manufacturers such as Nokia, Blackberry, and HTC to the plight of development. According to Apple's successful practices on disruptive innovation, one natural question has arisen: how do existing knowledge assets (e.g., "iPod + iTunes" model) affect disruptive innovation?

According to the resource-based view, the essence of business behavior is to find a competitive advantage that is largely determined by the resources they own and control [14–16]. In particular, existing knowledge assets such as valuable and rare resources can play a significant role in disruptive innovation. Some papers have shown that existing knowledge assets are not conducive to disruptive innovation. For example, Christensen and Raynor [17] and Christensen [18], found that the various existing knowledge embedded in individuals, products, practices, etc., prompts the firm to favor sustaining innovation in the allocation of resources rather than disruptive innovations. Assink [19] also verified similar findings that previous and successful designs and product concepts could adversely affect their disruptive innovation. However, other papers have reported opposite results. For example, Lindsay and Hopkins [20], through a case study of Kimberly-Clarks, found that low-cost intellectual assets such as patents, trademarks, and publications, were a "two-pronged" strategic intellectual asset that could be used both to help businesses defend against external destructive threats and to eliminate internal barriers to disruptive innovation. Wan et al. [21], using a set of case studies of Chinese firms, developed propositions about how novel research & development and production processes could foster disruptive innovation. To sum up, there is ambiguity over the role of existing knowledge assets in disruptive innovation. Recently, Fenech and Tellis [22] addressed the metrics, patterns, drivers, and predictive models of the dive and disruption of an existing product. Santoro et al. [23] investigated the relationship between a knowledge management system, open innovation, knowledge management capacity, and innovation capacity. Vecchiato [24] examined the relationship between managerial beliefs and the search processes for emerging markets, and found that in changing industries, the influence of prior history often increased the difficulty that decision-makers faced when seeking to respond to new events, and this difficulty then often resulted in organizational inertia and poor performance. These abovementioned studies tried to address the impact of existing knowledge assets on disruptive innovation, but the features of existing knowledge assets were rarely involved.

Some other scholars have introduced the feature of existing knowledge assets (such as observability, tacitness, and learnability), as antecedent or mediating variables into organizational behavior research to investigate their impact on knowledge transfer [25–28], organization structure [29], market performance [30], and innovation strategies [31]. In addition, other features of existing knowledge assets (e.g., knowledge embeddedness and knowledge specificity) have also attracted the attention of many researchers. In particular, knowledge embeddedness, as a recognized characteristic of knowledge, refers to the extent to which knowledge is embedded within an organization's individuals, dedicated assets, tools, organizational routines, and best practices as well as sub-networks [25,32].

Leszczyńska [33] argued that embedded knowledge and innovation influenced trajectory sequences in the long and discontinuous history of the cluster. Leszczyńska and Pruchnicki [34] investigated the impact of embedded knowledge on the efficiency of a localization choice made by a multinational corporation. Balland et al. [35] explained the formation of informal knowledge networks in clusters as an outcome of embeddedness, status, and proximity.

Meanwhile, asset specificity, as another important feature, refers to the notion that assets can only serve specific products and services, including site specificity, physical asset specificity, human asset specificity, dedicated asset specificity, and so on [36]. As an important type of intangible asset, knowledge also shows specificity. Moreover, knowledge specificity is not limited to specific products and services. There can also be specific businesses that consist of a series of products and services. Therefore, knowledge specificity can be referred to as the existing knowledge assets that can only serve the development of existing main business. Dibbern et al. [37] used data from 139 organizations on the sourcing of software development and maintenance services and found that production costs were generally lower in-house when knowledge specificity was high. Suh [38] investigated the role of excessive knowledge specificity in exhibiting the trusting ability of a firm. However, there have been few studies focusing on the role of knowledge embeddedness and knowledge specificity in affecting disruptive innovation.

To address these research gaps, based on the empirical analysis of data from 173 employees who engaged in product research and development, market monitoring, and product strategy formulation in Chinese manufacturing enterprises, we studied the impact of existing knowledge assets on disruptive innovation by taking into account the role of knowledge embeddedness and asset specificity. In summary, our paper makes two contributions. First, we divided disruptive innovations into outward-oriented disruptive innovation and internal-oriented disruptive innovation. This developed and supplemented Govindarajan and Kopalle's [9,39] scale for measuring disruptive innovation. Second, we added to the literature of disruptive innovation by investigating the impact of overall knowledge embeddedness and knowledge specificity on disruptive innovation, and provided a new perspective for the relationship between existing knowledge assets and disruptive innovation. The findings of this study revealed the role of knowledge embeddedness and knowledge specificity and provided practical information about understanding the impact of existing knowledge assets on disruptive innovation.

2. Theoretical Background and Research Hypotheses

According to Davenport and Prusak [40,41], knowledge is a flux mix of framed experiences, contextual information, values, and expert insight which provides a framework for evaluating and incorporating new experiences and information. Moreover, knowledge is dynamic since it is created in social interactions among individuals and organizations [42,43]. In particular, individual knowledge is the individual ability to draw distinctions within a collective domain of action under the appreciation of context or theory. Organizational knowledge is the capability developed by the members of an organization to draw distinctions in the process of carrying out their work [41]. According to the knowledge-based view, the overall intellectual capital can be defined as the sum of all the intangible and knowledge-related resources that an organization is able to use in its productive processes [44]. As an asset or capital, knowledge can play a significant role in disruptive innovation [45,46]. Therefore, we examined the impact of features of existing knowledge assets (e.g., knowledge embeddedness and knowledge specificity) on disruptive innovation.

2.1. Knowledge Embeddedness and Knowledge Specificity

According to Argote and Ingram [25] and Cummings and Teng [32], knowledge can be embedded in many different structural elements of an organization, such as in the people and their skills, the technical tools, and the routines and systems used by the organization, as well as in the networks formed between these elements. Moreover, Glisby and Holden [47] found that situational factors will constrain the process of knowledge creation and application. The higher the degree of knowledge embedded in culture, the more enterprises need to form a strong internal sharing network to communicate frequently and to ensure the effective flow of knowledge. In other words, knowledge specificity and application enhance the value of knowledge assets dedicated to a particular situation. However, if the external situation changes, such as the replacement of cultural context or external cooperation, the intrinsic value of the knowledge asset may be greatly reduced. Therefore, knowledge embeddedness will improve the situational applicability of knowledge assets and further strengthen knowledge specificity. On the other hand, Dayasindhu [48] and Jones et al. [49] pointed out that knowledge embeddedness would lead to the formation of various social mechanisms, such as restricted access (a limit on the number of members), collective sanctions (punishment meted out by constituents on erring partners), and reputation (the skills and reliability of the constituents), that coordinate and safeguard relations. The formation of these protection mechanisms restrains the conditions of occurrence and spatial extent of knowledge sharing, then enhances the extent to which these knowledge assets serve the firm's main businesses. As a result, knowledge embeddedness promotes the formation of protection mechanisms and further strengthens knowledge specificity. In summary, we proposed the following hypothesis:

Hypothesis 1. *Knowledge embeddedness has a positive effect on knowledge specificity.*

2.2. Knowledge Embeddedness and Disruptive Innovation

As mentioned above, disruptive innovation can be divided into outward-oriented disruptive innovation and internal-oriented disruptive innovation. We analyzed the role of knowledge embeddedness in outward-oriented disruptive innovation and internal-oriented disruptive innovation, respectively. As the extent of knowledge embeddedness increases, the knowledge can be mastered more deeply. This in-depth understanding enables companies to promptly transfer this knowledge into a usable form in response to changes in organizational needs and external market conditions to create more new products [50]. However, in the process of developing new products, knowledge embeddedness enables enterprises to search for solutions by using existing related knowledge, which presents obvious path-dependent characteristics. Such path-dependence will encourage enterprises to try hard to commercialize these high embedded knowledge assets in different industries, and then promote outward-oriented disruptive innovation. However, knowledge embeddedness also weakens the enterprise's self-willingness and motivation to replace the existing market, which is not conducive to the enterprise's internal-oriented disruptive innovation. Therefore, we proposed the following hypotheses:

Hypothesis 2a (H2a). *Knowledge embeddedness has a positive effect on outward-oriented disruptive innovation.*

Hypothesis 2b (H2b). *Knowledge embeddedness has a negative effect on internal-oriented disruptive innovation.*

2.3. Knowledge Specificity and Disruptive Innovation

Referring to Williamson's [51] definition of asset specificity, knowledge specificity is presented essentially as a lock-in effect. That is, once the knowledge assets dedicated to the development of the main businesses are identified, the relationship between the knowledge assets and main businesses will be continuously strengthened over time. Thus, the existing knowledge assets are locked [52]. If an enterprise tries to use these locked knowledge assets in other business areas, it will greatly discount its economic value. Therefore, the lock-in effect stipulates that companies must expand along the lines of their main businesses. In fact, this lock-in effect is a kind of path dependence and will have an impact on innovation activities. In other words, knowledge specificity can help enterprises continue to gain competitive advantages in the main business and guide enterprises in devoting more resources to carrying out innovation activities related to the main businesses. Moreover, these innovative activities are bound to promote the application of core technologies in different industries, thus promoting outward-oriented disruptive innovation [53–55]. However, as the extent of knowledge specificity is higher, it may lead to the rigidification of core competencies in the main business, and ignore the disruptive innovation opportunities that erode the market share of the main businesses [56]. Therefore, we proposed the following hypothesis:

Hypothesis 3a (H3a). *Knowledge specificity has a positive effect on outward-oriented disruptive innovation.*

Hypothesis 3b (H3b). *Knowledge specificity has a negative effect on internal-oriented disruptive innovation.*

2.4. The Mediating Role of Knowledge Specificity

Existing knowledge assets are the foundation of enterprise innovation, and enterprise innovation activity is an important manifestation of the application of existing knowledge assets. The features of the existing knowledge assets determine the direction and scope for enterprises to carry out innovative activities by using the existing knowledge assets. Knowledge embeddedness can detract from the flexibility of existing knowledge assets and let the application of existing knowledge assets focus on a particular area, thus limiting the direction and scope of an enterprise's innovation activities [31]. Moreover, highly embedded knowledge assets are often those that have been successful in helping enterprises achieve sustained growth and have path-dependent characteristics. Thus, it is inclined to outward-oriented disruptive innovation rather than internal-oriented disruptive innovation in the allocation of resources. This preference is largely due to knowledge specificity, which encourages companies to make innovative investments in areas they are better at or more familiar with to minimize risks and costs. On the other hand, Hsu and Wang [57] as well as Mura et al. [58] also found that knowledge assets embedded in tools, individuals, and relational networks bound the value realization of knowledge into specific situations. This lock-in effect of knowledge assets not only reduces the risk of knowledge spillover, but also strengthens the knowledge specificity. In summary, we proposed the following hypotheses:

Hypothesis 4a (H4a). *Knowledge specificity mediates the relationship between knowledge embeddedness and outward-oriented disruptive innovation.*

Hypothesis 4b (H4b). *Knowledge specificity mediates the relationship between knowledge embeddedness and internal-oriented disruptive innovation.*

Based on prior studies, this study proposed the conceptual model shown in Figure 1.

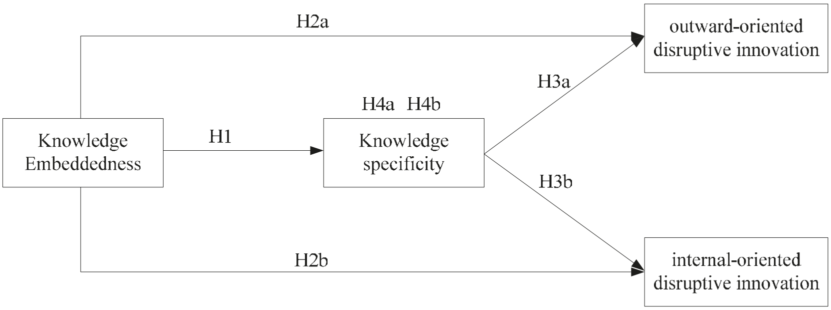

H2a
outward-oriented disruptive innovation
H3a
H4a H4b
Knowledge Embeddedness
H1
Knowledge specificity
H3b
H2b
internal-oriented disruptive innovation

Figure 1. Conceptual model.

3. Methods

3.1. Participants and Procedure

The participants were technical directors and senior presidents who engaged in product research and development, market monitoring, and product strategy formulation in Chinese manufacturing enterprises, and were familiar with product innovation activities. Two hundred and fifty questionnaires were sent, and the sample was reduced to 173 (69.2%) after deleting the forms returned by those who declined to participate or who failed to answer all of the items. Descriptive statistical analysis of valid samples found that the proportion of males (53.8%) was slightly higher than that of females (46.2%); the proportion of middle managers and top managers was 67.6% and 28.9%, respectively. As for the organizational characteristics of the surveyed enterprises, there were 126 (72.9%) enterprises that had been established for more than 10–15 years, 114 (65.9%) enterprises with 300 or more employees, and 142 (82.1%) private enterprises. In addition, these enterprises were mainly located in the specialized and general equipment manufacturing industries (12.7%), computer and telecommunications and other electronic equipment manufacturing industries (19.1%), automobile manufacturing industry (12.1%), chemical raw materials and chemical products manufacturing industry (9.2%), and the electrical machinery and equipment manufacturing industry (15.6%). Other manufacturing sectors amounted to less than 5% together.

In this study, we used three methods to collect samples. First, questionnaires were distributed on-site in local innovative training courses. These trained personnel mainly came from manufacturing enterprises in Quanzhou, Xiamen, Fuzhou, and other places in Fujian Province. Second, we commissioned a third-party platform (e.g., WenJuan Xing, the largest questionnaire distribution platform in China) to search the right sample companies and contact product development staff to complete the questionnaire. Third, we directly handed out questionnaires to related staff in qualified companies during our visits and project surveys.

3.2. Variables and Measures

3.2.1. Knowledge Embeddedness

To measure knowledge embeddedness, we modified the scales presented by Cummings [59] to fit the Chinese scenes, and designed four sample items including "It is difficult for a competitor to obtain the know-how of the company through field observation", "It is difficult for a competitor to obtain the know-how by studying production equipment", "It is difficult for a competitor to obtain the know-how by testing and using the product", and "It is difficult for a competitor to know how it works only by the company's activities, tasks, and procedures". The Cronbach's alpha for the scale was 0.806.

3.2.2. Knowledge Specificity

Knowledge specificity intends to measure the extent to which the existing knowledge assets specifically service main businesses. By referring to Cable and DeRue [60], we developed the scale and five sample items including "The main businesses provide ample opportunities for the use of existing knowledge assets", "The existing knowledge assets have made significant contributions to the development of main businesses", "The existing knowledge assets have been widely used in the main businesses", "The existing knowledge assets provide value to the enterprise through the main businesses", and "The existing knowledge assets increase with the development of the main businesses". The Cronbach's alpha for the scale was 0.772.

3.2.3. Disruptive Innovation

The scales of outward-oriented and internal-oriented disruptive innovation were referred from Christensen [8], Markides [61], Govindarajan and Kopalle [39], and Schmidt [62]. In particular, outward-oriented disruptive innovation included two sub-constructs of disruptive innovations: targeting new markets and targeting a competitor. The former had four items, including "Disruptive products target potential customers", "Disruptive products aim to predict future market needs", "Disruptive products aim to open up a new market", and "We often develop disruptive products for new markets". The Cronbach's alpha for the scale was 0.856. The latter had three items, including "Disruptive products aim to substitute the competing products", "Disruptive products aim to reduce competitor's market share", and "Disruptive products aim to pose a market threat for competitors". The Cronbach's alpha for the scale was 0.711. As for internal-oriented disruptive innovation, there were also three measurement items including "Disruptive products decrease the market share of the existing products", "Disruptive products substitute existing products", and "Disruptive products decrease the sales of the existing products". The Cronbach's alpha for the scale was 0.772.

In addition, Chandy and Tellis [63] as well as Govindarajan and Kopalle [39] found that the strategic business unit (SBU) and the characteristics of the firm such as firm size and strategic autonomy also had an impact on the firm's innovation behaviors. In order to better observe the relationship between the main variables, we controlled the impact of establishment time, staff size, property rights, R&D investment, and strategic autonomy. To avoid involving trade secrets, we measured the establishment time and staff size of the company by using the Likert four-point scale and measured the strategic autonomy by using the Likert seven-point scale, and investigated the potential impact of property rights by setting two dummy variables of state-owned enterprises and private enterprises. The measurement items of all of the constructs are presented in Appendix A, Table A1.

4. Results and Discussion

4.1. Reliability and Validity Analysis

The EFA (Exploratory Factor Analysis) method was used to test the structural validity of the scales. The results are shown in Table 1. In particular, the KMO (Kaiser-Meyer-Olkin) value of 19 sample items was 0.752, and the Chi-square value of Barlett's spherical test was 1378.183 (the degree of freedom was 171). Therefore, there were common factors in the correlation matrix and the factor analysis was suitable. Five factors were obtained from the factor analysis, which included disruptive innovation targeting new markets, disruptive innovation targeting competitors, internal-oriented disruptive innovation, knowledge embeddedness, and knowledge specificity. Meanwhile, the extent of common method variance (CMV) was examined using Harman's one-factor test. The results showed that five factors were extracted by principal component analysis, and explained the total variance of 66.037%, while all factor loads after rotation were over 0.59. In particular, one factor (i.e., disruptive innovation targeting new markets) explained a total variance of 23.53%, which was less than half of the total variance. As a result, a single factor did not explain the vast majority of the amount of variation, and the common method variance was largely controlled.

Table 1. Exploratory factor analysis and Cronbach's alpha.

Measurement Item	Component					Explained Variance (%)	Cronbach's Alpha
	1	2	3	4	5		
ODI 1	0.777	0.232	0.141	0.102	0.106	15.333	0.856
ODI 2	0.790	0.178	−0.046	0.089	0.198		
ODI 3	0.873	0.100	0.071	−0.058	0.148		
ODI 4	0.730	−0.061	0.227	0.097	0.353		
KS 1	0.124	0.672	0.120	−0.194	0.051	14.162	0.772
KS 2	−0.009	0.598	0.051	−0.342	0.306		
KS 3	0.137	0.767	−0.020	−0.166	−0.022		
KS 4	−0.009	0.766	0.113	0.082	0.040		
KS 5	0.160	0.653	0.054	−0.089	0.059		

Table 1. *Cont.*

Measurement Item	Component					Explained Variance (%)	Cronbach's Alpha
	1	2	3	4	5		
KE 1	0.078	0.182	0.717	0.181	−0.020	13.823	0.806
KE 2	0.203	0.039	0.737	−0.236	−0.015		
KE 3	−0.012	−0.002	0.869	−0.006	0.044		
KE4	0.051	0.081	0.820	−0.021	−0.003		
IDI 1	0.203	−0.168	−0.158	0.546	0.385	11.737	0.772
IDI 2	0.096	−0.171	0.034	0.882	−0.008		
IDI 3	−0.003	−0.167	0.007	0.885	0.041		
ODI 5	0.267	0.135	−0.012	0.039	0.704	10.928	0.711
ODI 6	0.116	0.185	−0.015	−0.046	0.857		
ODI 7	0.384	−0.053	0.063	0.222	0.636		
Cumulative Explained variance (%)						66.037	
Cronbach's alpha of the total scale							0.782

Note: Extraction Method: Principal Component Analysis; Rotation Method: Varimax with Kaiser Normalization. N = 173. KE = knowledge embeddedness, KS = knowledge specificity, ODI = outward-oriented disruptive innovation, IDI = internal-oriented disruptive innovation.

4.2. Confirmatory Factor Analyses

To enhance the stability of the model fitting results by averaging the highest and the lowest factor loading, we grouped and packaged the measurement items of disruptive innovation targeting new markets and disruptive innovation targeting competitors, and took the average of the group scores as the index value of outward-oriented disruptive innovation. As a result, we obtained four measurement items. On this basis, we checked the discriminant validity by confirmatory factor analyses. The results are shown in Table 2. It can be seen from Table 2 that the χ^2/df of the four-factor model was 1.86, the goodness of fit index (GFI) and non-normed fir index (NNFI) were close to 0.9, the comparative fit index (CFI) and incremental fit index (IFI) were greater than 0.9, the adjusted goodness of fit index (AGFI) was greater than 0.5, and the root mean square error of approximation (RMSEA) was less than 0.1. The overall goodness of fit was good. Moreover, when compared with the other three models, the four-factor model was the best fitting model. This indicated that the four factors involved in this study had good discriminant validity and represented four different concepts.

Table 2. Results of confirmatory factor analyses.

Model	Factor	χ^2	df	χ^2/df	GFI	CFI	NNFI	IFI	AGFI	RMSEA
Model 1	4 Factor: KS; KE; ODI; IDI	182.56	98	1.86	0.89	0.91	0.89	0.91	0.85	0.07
Model 2	3 Factor: KS + KE; ODI; IDI	583.13	103	5.66	0.68	0.48	0.40	0.49	0.58	0.17.
Model 3	2 Factor: KS + KE; ODI+IDI	379.43	101	3.76	0.79	0.70	0.64	0.71	0.71	0.13
Model 4	1 Factor: KS + KE + ODI + IDI	735.41	104	7.07	0.61	0.32	0.22	0.33	.049	.019

Note: N = 173. KE = knowledge embeddedness, KS = knowledge specificity, ODI = outward-oriented disruptive innovation, IDI = internal-oriented disruptive innovation, GFI = goodness of fit index, CFI = comparative fit index, NNFI = non-normed fit index, IFI = incremental fit index, AGFI = adjusted goodness of fit index, RMSEA = root mean square error of approximation.

4.3. Correlation Analysis

The mean, standard deviation, and correlation coefficients of the variables in this study are shown in Table 3. As seen from Table 3, knowledge embeddedness had a significant positive correlation with knowledge specificity ($r = 0.19$, $p < 0.05$). As for the correlation between knowledge embeddedness and disruptive innovation, knowledge embeddedness had a significant positive correlation with outward-oriented disruptive innovation ($r = 0.18$, $p < 0.05$), and had a negative correlation with internal-oriented disruptive innovation, but did not reach the significant level. In addition, knowledge specificity presented positive correlations with outward-oriented disruptive innovation (r = 0.27, p < 0.01), but revealed negative correlations with internal-oriented disruptive innovation (r = −0.33, p < 0.01). Moreover, we found that the correlation coefficient (r = 0.22, p < 0.01) between outward-oriented disruptive innovation and internal-oriented disruptive innovation was relatively low. Therefore, these two types of disruptive innovation are independent. These results laid the foundation for the following regression analysis.

Table 3. Descriptive statistics and correlation analysis.

Variables	M	SD	1	2	3	4	5	6	7	8
1. Age of Enterprise	3.00	0.869	1							
2. Size of Enterprise	2.89	0.796	0.46 **	1						
3. Ownership of Enterprise	0.18	0.385	0.23 **	0.16 *	1					
4. Strategic Autonomy	5.14	0.817	0.13	0.15	−0.04	1				
5. Knowledge Embeddedness	4.76	0.947	0.07	0.19 *	−0.01	0.17 *	1			
6. Knowledge Specificity	5.76	0.617	0.11	0.07	−0.06	0.32 **	0.19 *	1		
7. Outward-Oriented Disruptive Innovation	5.12	0.913	0.03	0.04	−0.03	0.23 **	0.18 *	0.27 **	1	
8. Internal-Oriented Disruptive Innovation	4.12	1.165	−0.12	−0.23 **	−0.12	−0.14	−0.06	−0.33 **	0.22 **	1

Note: N = 173. * $p < 0.05$, ** $p < 0.01$.

4.4. Hierarchical Regression Analysis and Discussion

In this section, we conducted hierarchical regression analysis to test our hypotheses. As shown in Table 4, all models (i.e., from Model 1 to 10) were used to test the abovementioned hypotheses. In particular, Models 1 (M1), 3 (M3), and 7 (M7) were denoted the benchmark models with controlled variables and investigated the impact of the enterprise's age, size, property rights, and strategic autonomy. Model 2 (M2) was used to verify the impact of knowledge embeddedness on knowledge specificity. Models 4 (M4) and 5 (M5) were used to verify the impact of knowledge embeddedness and knowledge specificity on outward-oriented disruptive innovation. Models 8 (M8) and 9 (M9) were used to verify the impact of knowledge embeddedness and knowledge specificity on internal-oriented disruptive innovation. Model 6 (M6) was used to verify the mediation effect of knowledge specificity on the relationship between knowledge embeddedness and outward-oriented disruptive innovation. Model 10 (M10) was used to test the mediation effect of knowledge specificity on the relationship between knowledge embeddedness and internal-oriented disruptive innovation. We found that across all models, the variance inflation factor was less than 10, which indicated that there was no multi-collinearity in the model and our results were reliable.

Table 4. Results of hierarchical regression analysis.

Variables	KS		ODI				IDI			
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
AE	0.09	0.10	0.01	0.01	−0.02	−0.01	0.02	0.02	0.04	0.05
SE	−0.01	−0.03	0.01	−0.02	0.01	−0.02	−0.20*	−0.20 *	−0.21 *	−0.21 ***
OE	−0.07	−0.07	−0.02	−0.02	−0.01	−0.01	−0.10	−0.10	−0.12	−0.12
SA	0.31 ***	0.29 ***	0.22 **	0.20*	0.16 *	0.14 $^{+}$	−0.11	−0.11	−0.01	−0.02
KE		0.14 $^{+}$		0.15 *		0.12		−0.10		0.04
KS					0.22 **	0.20 *			−0.32 ***	−0.33 ***
R2	0.11	0.13	0.05	0.07	0.10	0.11			0.17	0.17
F	5.36 ***	5.1 ***	2.25 $^{+}$	2.61 *	3.40 **	3.30 **	3.31 *	2.63 *	6.50 ***	5.51 ***
VIF	1.039 ≦ VIF ≦ 1.332									

Note: N = 173. AE = age of enterprise, SE = size of enterprise, OE = ownership of enterprise, SA = strategic autonomy, KE = knowledge embeddedness, KS = knowledge specificity, ODI = outward-oriented disruptive innovation, IDI = internal-oriented disruptive innovation. *** Significant at $p < 0.001$, ** significant at $p < 0.01$, * significant at $p < 0.05$, $^{+}$ significant at $p < 0.1$.

4.4.1. The Role of Knowledge Embeddedness in Knowledge Specificity

Comparing Model 1 with Model 2, knowledge embeddedness had a significant positive influence on knowledge specificity (β = 0.14, $p < 0.1$). Therefore, Hypothesis 1 was supported. This result indicates that if the knowledge assets were embedded in many different structural elements of an organization, it was more likely for the knowledge assets to serve the firm's main business. The findings complemented the results of Cummings and Teng [33], Un and Asakawa [64], and Lin et al. [31], where on the one hand, knowledge embeddedness makes knowledge asset transfer more difficult. On the other hand, knowledge embeddedness will attract companies to pay attention to strengthening their knowledge specificity.

4.4.2. The Role of Knowledge Embeddedness in Disruptive Innovation

Comparing Model 3 with Model 4 (alternatively Model 7 with Model 8), we can see that knowledge embeddedness had a significant positive impact on the outward-oriented disruptive innovation (β = 0.15, $p < 0.05$). In contrast, the knowledge embeddedness had a negative impact on the internal-oriented disruptive innovation. However, it did not reach a significant level. Therefore, Hypothesis 2a was supported and Hypothesis 2b was not supported. This result demonstrated that the enterprises would pay more attention to more highly embedded knowledge, and more likely carry out disruptive innovation induced by highly embedded knowledge in various industries. Moreover, our result supports Lindsay and Hopkins's [20] finding that enterprises could replicate existing success patterns to other industry areas by using highly embedded knowledge. However, the negative impact of knowledge embeddedness on internal-oriented disruptive innovation was not supported. This may be due to the fact that highly embedded knowledge can be related to the main businesses or not. For example, some novel ideas embedded in individuals may not be related to the current main businesses and can be ignored at present, even if it will play a significant role in future internal-oriented disruptive innovation [20]. Therefore, we should analyze the importance and contribution of existing knowledge assets to the development of main businesses when we investigate the impact of knowledge embeddedness on disruptive innovation.

4.4.3. The Role of Knowledge Specificity in Disruptive Innovation

Comparing Model 3 with Model 5 (alternatively Model 7 with Model 9), it was found that knowledge specificity had a significant positive effect on the outward-oriented disruptive innovation (β = 0.22, $p < 0.01$), and had a significant negative effect on the internal-oriented disruptive innovation (β = −0.32, $p < 0.001$). Therefore, both Hypotheses 3a and 3b were supported. This indicates that as the extent of knowledge specificity becomes higher, the enterprises will devote more resources to carrying out disruptive innovation activities associated with the main business and ignore those disruptive innovations that may conflict with the main businesses. Our results support the research of Christensen and Raynor [17] and Assink [19], which indicated that the impact of knowledge specificity on disruptive innovation is related to the nature of the markets. In other words, as the knowledge assets are dedicated to the enterprise's main businesses, it can actively encourage enterprises to develop disruptive products that target competitors' markets or new markets. However, the knowledge assets focusing on the main businesses are often closely related to existing products, and restrain internal-oriented disruptive innovation. Therefore, this result also provides an explanation for the relationship between knowledge embeddedness and internal-oriented disruptive innovation.

4.4.4. The Mediating Role of Knowledge Specificity

Referring to the procedure proposed by Baron and Kenny [65], we tested the mediating effect of knowledge specificity. Based on the above analysis, we found that the prerequisites of verifying the mediate variables could be satisfied. Then, we verified the mediating effect of knowledge specificity on the relationship between knowledge embeddedness and outward-oriented disruptive innovation. The results are shown in Table 4. Comparing Model 6 with Model 4, the coefficient of knowledge embeddedness decreased significantly (β = 0.15→0.12) and did not reached a significant level. Therefore, the mediating effect could be verified and Hypothesis 4a was supported. However, as knowledge embeddedness had no significant effect on the internal-oriented disruptive innovation, then the mediating effect of knowledge specificity on the relationship between knowledge embeddedness and internal-oriented disruptive innovation did not exist. Therefore, Hypothesis 4b could not be supported. This result once again shows that knowledge specificity can play a significant role in disruptive innovation. It can act as a bridge to transfer the positive effect induced by knowledge embeddedness on outward-oriented disruptive innovation.

5. Conclusions

In this paper, we investigated the impact of existing knowledge assets on disruptive innovation by analyzing the role of knowledge embeddedness and specificity. Based on an empirical analysis of the data from 173 employees who engaged in product research and development, market monitoring, and product strategy formulation in Chinese manufacturing enterprises, we found that first, knowledge embeddedness not only played a positive role in knowledge specificity, but also had a positive effect on outward-oriented disruptive innovation. Second, knowledge specificity exhibited opposite functions on outward-oriented and internal-oriented disruptive innovation. In particular, knowledge specificity showed remarkable and direct positive effects on outward-oriented disruptive innovation, but notably had negative effects on internal-oriented disruptive innovation. This revealed that an enterprise with higher knowledge specificity could allocate more resources to the main businesses. These results provide a possible explanation for understanding the different viewpoints of Christensen [8], Lindsay and Hopkins [20], and Assink [19]. In addition, we also verified the mediating effect of knowledge specificity on the relationship between knowledge embeddedness and outward-oriented disruptive innovation. To sum up, with consideration of the features of knowledge assets, we expanded the research coverage on disruptive innovation and supplemented the findings of Christensen [8], Assink [19], and Wagner [66] regarding the relationship between knowledge assets and disruptive innovation. It can be concluded that knowledge embeddedness and knowledge specificity are two critical features of knowledge assets influencing disruptive innovation.

The research conclusion presents significant inspiration for knowledge asset management and innovation management. First, existing knowledge assets are the basis of innovation development, but there is not an apparently positive relationship between them. It is crucial to analyze the features of knowledge assets when we consider disruptive innovation. Second, with regard to knowledge asset management, enterprises should establish knowledge asset evaluation systems to dynamically track and evaluate knowledge embeddedness and knowledge specificity to then provide guidance for developing related disruptive innovation. Third, for innovation management, by making full use of the specialized knowledge assets of the main businesses, enterprises should promote the disruptive innovation of the existing knowledge assets in different industries.

However, this study had some limitations. First, this study adopted a convenient sample research approach with strong geographical features and industry limitations. In the future, we will expand the geographical scope of the samples. Second, embeddedness and specialization are only two of the features of knowledge assets. There also exist other important features which have had a significant impact on knowledge transfer and innovation. Subsequent studies can further explore the impact of other features on disruptive innovations. Third, the external environment is an important factor that affects enterprises in carrying out innovative activities [67]. Fourth, we should more comprehensively and holistically investigate the determinants of innovation and input more results to a mathematical model in future research. Therefore, further research should probe into the multi-layered factors and mechanisms of innovation management.

Acknowledgments: This study was supported financially by the Natural Science Foundation of China (71302163, 71401042, 71401151), Fund for Social Science Projects of Fujian Province in China (FJ2017B012), and the Humanities and Social Science Foundation Supported by the Ministry of Education in China (16YJC630160). Thanks are extended to Associate Professor Na Wang for language revision and suggestion on the paper's introduction.

Author Contributions: Chunpei Lin, Baixun Li, and Yenchun Jim Wu conceived and designed the experiments; Chunpei Lin performed the experiments; Chunpei Lin and Baixun Li analyzed the data; Yenchun Jim Wu contributed reagents/materials/analysis tools; Chunpei Lin and Baixun Li wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Measures for key constructs.

Construct		Measurement Items
Outward-oriented disruptive innovation	Targeting new markets	Disruptive products target potential customers. Disruptive products aim to predict future market needs. Disruptive products open up a new market. We often develop disruptive products for new markets.
	Targeting competitors' markets	Disruptive products aim to substitute the competing products. Disruptive products aim to reduce competitor's market share. Disruptive products aim to pose a market threat for competitors.
Internal-oriented disruptive innovation		Disruptive products decrease the market share of the existing products. Disruptive products substitute existing products. Disruptive products decrease the sales of the existing products.
Knowledge specificity		The main businesses provide ample opportunities for the use of existing knowledge assets. The existing knowledge assets have made significant contributions to the development of main businesses. The existing knowledge assets have been widely used in the main businesses. The existing knowledge assets provide value to the enterprise through the main businesses. The existing knowledge assets increase with the development of the main businesses.
Knowledge embeddedness		It is difficult for a competitor to obtain the know-how of the company through field observation. It is difficult for a competitor to obtain the know-how by studying production equipment. It is difficult for a competitor to obtain the know-how by testing and using the product. It is difficult for a competitor to know how it works only by the company's activities, tasks, and procedures.

References

1. Carrasco, J.L.; Careaga, M.; Badilla-Quintana, M.G. The New Pyramid of Needs for the Digital Citizen: A Transition towards Smart Human Cities. *Sustainability* **2017**, *12*, 2258. [CrossRef]
2. Mohan, K.; Ramesh, B.; Cao, L.; Sarkar, B. Managing Disruptive and Sustaining Innovations in Green IT. *IT Prof.* **2012**, *6*, 22–29. [CrossRef]
3. Watson, R.T.; Boudreau, M.-C.; Li, S.; Levis, J. Telematics at UPS: En Route to Energy Informatics. *MIS Q. Exec.* **2010**, *1*, 1–11.
4. Wu, Y.C.J.; Pan, C.I.; Yuan, C.H. Attitudes towards the use of information and communication technology in management education. *Behav. Inf. Technol.* **2017**, *3*, 243–254. [CrossRef]
5. Abel, M.-H. Knowledge map-based web platform to facilitate organizational learning return of experiences. *Comput. Hum. Behav.* **2015**, *1*, 960–966. [CrossRef]
6. Hall, J.; Vredenburg, H. The challenges of innovating for sustainable development. *MIT Sloan Manag. Rev.* **2003**, *1*, 61–68.
7. Adner, R. When are technologies disruptive? A demand-based view of the emergence of competition. *Strateg. Manag. J.* **2002**, *8*, 667–688. [CrossRef]

8. Christensen, C.M. *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*; Harvard Business School Press: Boston, MA, USA, 1997.
9. Govindarajan, V.; Kopalle, P.K.; Danneels, E. The Effects of Mainstream and Emerging Customer Orientations on Radical and Disruptive Innovations. *J. Prod. Innov. Manag.* **2011**, *28*, 28,121–132. [CrossRef]
10. Rafii, F.; Kampas, P.J. How to identify your enemies before they destroy you? *Harv. Bus. Rev.* **2002**, *80*, 115–123. [PubMed]
11. Husig, S.; Hipp, C.; Dowling, M. Analyzing the disruptive potential: The case of wireless local area network and mobile communications network companies. *R D Manag.* **2005**, *35*, 17–35. [CrossRef]
12. Keller, A.; Hüsig, S. Ex-ante identification of disruptive innovations in the software industry applied to web applications: The case of Microsoft's vs. Google's office applications. *Technol. Forecast. Soc.* **2009**, *76*, 1044–1054. [CrossRef]
13. Gilbert, C.; Bower, J.L. Disruptive change. When trying harder is part of the problem. *Harv. Bus. Rev.* **2002**, *80*, 94–101. [PubMed]
14. Wernerfelt, B.A. A resource-based view of the firm. *Strateg. Manag. J.* **1984**, *5*, 171–180. [CrossRef]
15. Barney, J. Firm Resources and Sustained Competitive Advantage. *J. Manag.* **1991**, *17*, 90–120. [CrossRef]
16. Grant, R.M. Prospering in Dynamically-Competitive Environments: Organizational Capability as Knowledge Integration. *Organ. Sci.* **1996**, *7*, 375–387. [CrossRef]
17. Christensen, C.M.; Raynor, M.E. *The Innovator's Solution: Creating and Sustaining Successful Growth*; Harvard Business School Press: Boston, MA, USA, 2003.
18. Christensen, C.M. The Ongoing Process of Building a Theory of Disruption. *J. Prod. Innov. Manag.* **2006**, *23*, 39–55. [CrossRef]
19. Assink, M. Inhibitors of disruptive innovation capability: A conceptual model. *Eur. J. Innov. Manag.* **2006**, *9*, 215–233. [CrossRef]
20. Lindsay, J.; Hopkins, M. From experience: Disruptive Innovation and the Need for Disruptive Intellectual Asset Strategy. *J. Prod. Innov. Manag.* **2010**, *27*, 283–290. [CrossRef]
21. Wan, F.; Williamson, P.J.; Yin, E. Antecedents and implications of disruptive innovation: Evidence from China. *Technovation* **2015**, *39*, 94–104. [CrossRef]
22. Fenech, J.P.; Tellis, G.J. The Dive and Disruption of Successful Current Products: Measures, Global Patterns, and Predictive Model. *J. Prod. Innov. Manag.* **2016**, *33*, 53–68. [CrossRef]
23. Santoro, G.; Vrontis, D.; Thrassou, A.; Dezi, L. The Internet of Things: Building a knowledge management system for open innovation and knowledge management capacity. *Technol. Forecast.* **2017**, in press. [CrossRef]
24. Vecchiato, R. Disruptive innovation, managerial cognition, and technology competition outcomes. *Technol. Forecast.* **2016**, *116*, 116–128. [CrossRef]
25. Argote, L.; Ingram, P. Knowledge Transfer: A Basis for Competitive Advantage in Firms. *Organ. Behav. Hum. Decis. Process.* **2000**, *82*, 150–169. [CrossRef]
26. Mcevily, B.; Argote, L.; Reagans, R. Managing Knowledge in Organizations: An Integrative Framework and Review of Emerging Themes. *Manag. Sci.* **2003**, *49*, 571–582.
27. Mciver, D.; Lepisto, D.A. Effects of knowledge management on unit performance: Examining the moderating role of tacitness and learnability. *J. Knowl. Manag.* **2017**, *21*, 796–816. [CrossRef]
28. Mostafa, R.; Klepper, S. Industrial Development through Tacit Knowledge Seeding: Evidence from the Bangladesh Garment Industry. *Manag. Sci.* **2017**, in press. [CrossRef]
29. Birkinshaw, J.; Nobel, R.; Ridderstråle, J. Knowledge as a contingency variable: Do the characteristics of knowledge predict organization structure? *Organ. Sci.* **2002**, *13*, 274–289. [CrossRef]
30. Chong, W.K.; Bian, D.; Zhang, N. E-marketing services and e-marketing performance: The roles of innovation, knowledge complexity and environmental turbulence in influencing the relationship. *J. Mark. Manag.* **2016**, *32*, 149–178. [CrossRef]
31. Lin, H.E.; McDonough, E.F.; Yang, J.; Wang, C.Y. Aligning Knowledge Assets for Exploitation, Exploration, and Ambidexterity: A Study of Companies in High-Tech Parks in China. *J. Prod. Innov. Manag.* **2017**, *34*, 122–140. [CrossRef]
32. Cummings, J.L.; Teng, B.S. Transferring R&D knowledge: The key factors affecting knowledge transfer success. *J. Eng. Technol. Manag.* **2003**, *20*, 39–68.
33. Leszczyńska, D. Historical trajectory and knowledge embeddedness: A case study in the French perfume cluster. *Manag. Organ. Hist.* **2013**, *8*, 290–305.

34. Leszczyńska, D.; Pruchnicki, E. The evolution of knowledge transfer and the location of a multinational corporation: Theory and mathematical model. *Multinatl. Bus. Rev.* **2015**, *23*, 111–129. [CrossRef]
35. Balland, P.A.; Belsomartínez, J.A.; Morrison, A. The Dynamics of Technical and Business Knowledge Networks in Industrial Clusters: Embeddedness, Status, or Proximity? *Econ. Geogr.* **2016**, *92*, 35–60. [CrossRef]
36. Joskow, P.L. Asset Specificity and the Structure of Vertical Relationships: Empirical Evidence. *J. Law Econ. Organ.* **1988**, *4*, 95–117.
37. Dibbern, J.; Chin, W.W.; Kude, T. The Sourcing of Software Services: Knowledge Specificity and the Role of Trust. *Data Base Adv. Inf. Syst.* **2016**, *47*, 36–57. [CrossRef]
38. Suh, T. Exhibited trust and excessive knowledge specificity: A competitive altruism hypothesis. *Ind. Mark. Manag.* **2016**, *62*, 51–60. [CrossRef]
39. Govindarajan, V.; Kopalle, P.K. The usefulness of measuring disruptiveness of innovations ex post in making ex ante predictions. *J. Prod. Innov. Manag.* **2006**, *23*, 12–18. [CrossRef]
40. Davenport, T.H.; Prusak, L. *Working Knowledge*; Harvard University Press: Cambridge, MA, USA, 1998.
41. Tsoukas, H.; Vladimirou, E. What is organizational knowledge? *J. Manag. Stud.* **2001**, *7*, 973–993. [CrossRef]
42. Nonaka, L.; Takeuchi, H.; Umemoto, K. A theory of organizational knowledge creation. *Int. J. Technol. Manag.* **1996**, *11*, 833–845.
43. Hassan, S.-U.; Haddawy, P. Measuring International Knowledge Flows and Scholarly Impact of Scientific Research. *Scientometrics* **2013**, *1*, 163–179. [CrossRef]
44. Kianto, A.; Ritala, P.; Spender, J.C.; Vanhala, M. The interaction of intellectual capital assets and knowledge management practices in organizational value creation. *J. Intellect. Cap.* **2014**, *3*, 362–375. [CrossRef]
45. Scaringella, L. Knowledge, knowledge dynamics, and innovation: Exploration of the internationalization of a multinational corporation. *Eur. J. Innov. Manag.* **2016**, *3*, 337–361. [CrossRef]
46. Kang, M.; Lee, M.J. Absorptive capacity, knowledge sharing, and innovative behaviour of R&D employees. *Technol. Anal. Strateg.* **2017**, *29*, 219–232.
47. Glisby, M.; Holden, N. Contextual constraints in knowledge management theory: The cultural embeddedness of Nonaka's knowledge-creating company. *Knowl. Process Manag.* **2003**, *10*, 29–36. [CrossRef]
48. Dayasindhu, N. Embeddedness, knowledge transfer, industry clusters and global competitiveness: A case study of the Indian software industry. *Technovation* **2002**, *22*, 551–560. [CrossRef]
49. Jones, C.; Hesterly, W.S.; Borgatti, S.P. A General Theory of Network Governance: Exchange Conditions and Social Mechanisms. *Acad. Manag. Rev.* **1997**, *22*, 911–945.
50. Nonaka, I. Toward Middle-Up-Down Management: Accelerating Information Creation. *MIT Sloan Manag. Rev.* **1988**, *29*, 9–18.
51. Wlliamson, O.E. *The Economic Institutions of Capitalism*; The Free Press: New York, NY, USA, 1985.
52. Fitzroy, F.R.; Mueller, D.C. Cooperation and Conflict in Contractual Organization. *Q. Rev. Econ. Bus.* **1984**, *24*, 23–49.
53. Allarakhia, M.; Walsh, S. Managing knowledge assets under conditions of radical change: The case of the pharmaceutical industry. *Technovation* **2011**, *31*, 105–117. [CrossRef]
54. Di Guardo, M.C.; Harrigan, K.R. Shaping the path to inventive activity: The role of past experience in R&D alliances. *J. Technol. Transf.* **2016**, *41*, 1–20.
55. Zahra, S.; George, G. Absorptive capacity: A review, reconceptualization, and extension. *Acad. Manag. Rev.* **2002**, *27*, 185–203.
56. Wu, F.S.; Haak, R. Innovation mechanisms and knowledge communities for corporate central R&D. *Creat. Innov. Manag.* **2013**, *22*, 37–52.
57. Hsu, L.C.; Wang, C.H. Clarifying the effect of intellectual capital on performance: The mediating role of dynamic capability. *Br. J. Manag.* **2012**, *23*, 179–205. [CrossRef]
58. Mura, M.; Radaelli, G.; Spiller, N.; Lettieri, E.; Longo, M. The effect of social capital on exploration and exploitation. *J. Intellect. Cap.* **2014**, *15*, 430–450. [CrossRef]
59. Cummings, J.L. *Knowledge Transfer across R&D Units: An Empirical Investigation of the Factors Affecting Successful Knowledge Transfer across Intra- and Inter-Organizational Units*; UMI: Honk Kong, China, 2002; pp. 62–65.
60. Cable, D.M.; DeRue, D.S. The convergent and discriminant validity of subjective fit perceptions. *J. Appl. Psychol.* **2002**, *87*, 875–884. [CrossRef] [PubMed]

61. Markides, C. Disruptive innovation: In need of better theory. *J. Prod. Innov. Manag.* **2006**, *23*, 19–25. [CrossRef]
62. Schmidt, G.M.; Druehl, C.T. When is a disruptive innovation disruptive? *J. Prod. Innov. Manag.* **2008**, *25*, 347–369. [CrossRef]
63. Chandy, R.K.; Tellis, G.J. Organizing for radical product innovation: The overlooked role of willingness to cannibalize. *J. Mark. Res.* **1998**, *35*, 474–487. [CrossRef]
64. Un, C.A.; Asakawa, K. Types of R&D collaborations and process innovation: The benefit of collaborating upstream in the knowledge chain. *J. Prod. Innov. Manag.* **2015**, *32*, 138–153.
65. Baron, R.M.; Kenny, D.A. The moderator mediator variable distinction in social psychological research. *J. Personal. Soc. Psychol.* **1986**, *51*, 1173–1182. [CrossRef]
66. Wang, Y.; Vanhaverbeke, W.; Roijakkers, N. Exploring the impact of open innovation on national systems of innovation—A theoretical analysis. *Technol. Forecast. Soc.* **2012**, *79*, 419–428. [CrossRef]
67. Bouali, S.; Buscarino, A.; Fortuna, L.; Frasca, M.; Gambuzza, L.V. Emulating complex business cycles by using an electronic analogue. *Nonlinear Anal. Real World Appl.* **2012**, *13*, 2459–2465. [CrossRef]

sustainability

MDPI

Article

Role of Human Knowledge and Communication on Operational Benefits Gained from Six Sigma

Jorge L. García-Alcaraz [1,*], Liliana Avelar-Sosa [1], Juan I. Latorre-Biel [2], Emilio Jiménez-Macías [3] and Giner Alor-Hernández [4]

[1] Department of Industrial Engineering and Manufacturing, Universidad Autónoma de Ciudad Juárez, Av. Del Charro 450 Norte, Col. Partido Romero, P.C. 32310 Ciudad Juárez, Mexico; liliana.avelar@uacj.mx

[2] Department of Mechanical Engineering, Energy and Materials, Universidad de Navarra, Avda. de Tarazona s/n, P.C. 31500 Tudela, Spain; juanignacio.latorre@unavarra.es

[3] Department of Electrical Engineering, Universidad de La Rioja, C/San José de Calasanz 31, P.C. 26004 Logroño, Spain; emilio.jimenez@unirioja.es

[4] Division of Research and Postgraduate Studies, Instituto Tecnológico de Orizaba, Oriente 9, Emiliano Zapata Sur, P.C. 94320 Orizaba, Mexico; galor@itorizaba.edu.mx

* Correspondence: jorge.garcia@uacj.mx; Tel.: +52-656-688-4843

Received: 9 August 2017; Accepted: 21 September 2017; Published: 26 September 2017

Abstract: Six Sigma (SS) is a production philosophy focused on human experiences and knowledge, aimed to minimize defects of products and services. The appropriate implementation of SS requires an education process, reliable data analysis, efficient didactic material, statistical techniques and human knowledge to improve communication and operational benefits. In this article, we present a structural equation model integrating those aspects as latent variables and relating them with ten hypotheses. Data for hypothesis validation were gathered among 301 manufacturing companies, and assessed using partial least squares (PLS) to estimate direct, indirect, and total effects. As results, we found that access to reliable information, trusted analysis and knowledgeable management are crucial for SS implementation at the problem definition stage. Likewise, to execute and control SS projects, it is important to be trained in statistical techniques through clear didactic materials.

Keywords: training; education; communication; six sigma; structural equation model

1. Introduction

To stay afloat in this globalized context, companies must adopt and implement tools, techniques, and methods that have helped other companies to succeed. One of the most popular of these methodologies is Six Sigma (SS), which is part of Total Quality Management (TQM). SS has its origins in the 1980s, in the Motorola industry and, due to its success, years later, other companies such as Allied Signal, Bombardier, Siebe, Sony, Polaroid, Toshiba and Texas Instruments implemented it. However, its recognition as an efficient methodology for the control of variability was obtained in General Electric [1].

The application of SS evolved rapidly to other sectors, and, already in the year 2000, its applications were associated with plant distribution [2] and education [3], and it was formally established as a methodology for control of quality [4,5]. According to their use, SS can be defined as a methodology for the continuous improvement of customer satisfaction thanks to a reduction of defects in products/services to increase economic benefits [6].

Although SS can be defined in many ways, from a statistical viewpoint, it is a controlled production process methodology with a near-perfect rate of 3.3 defects per million opportunities. This is perhaps the most important definition of SS as a methodology, since it implies that processes must be appropriately standardized [7]. However, as a procedure, SS is also treated as an integrated methodology consisting of two sub-methodologies: defining, measuring, analyzing, improving, and controlling (DMAIC), used when products or processes are in existence; and defining, measuring, analyzing, designing and verifying (DMADV), used when products or processes are not in existence and the company needs to develop them [8]. In this research, SS is considered as a methodology integrating other techniques, but this research is only focused on DMAIC, due to the geographical context.

Products/services defects are areas of opportunity that can be addressed with SS. Examples of companies having successfully implemented this methodology in the 1990s include Motorola, General Electrics, and Allied Signals [9]. Investments in education and training in these companies showed remarkable cost–benefit relationships. Unfortunately, since not every organization implementing SS obtains satisfying results, it is important to identify the critical success factors (CSF) of this methodology to prevent companies from throwing SS projects away.

It is common for managers to think that SS is not a suitable methodology to be implemented because its benefits are obtained only in specific contexts. Fortunately, current studies have demonstrated that SS is not only appropriate for manufacturing systems, but also for the services sector [10].

Research has provided an overview of the critical success factors of SS for countries such as India [11], Brazil [12,13], Malaysia [14], Italy [15], and Mexico [7]. Some factors seem to be consistent across countries, whereas others greatly differ, especially in industrial contexts. For instance, authors such as Chakraborty and Chuan Tan [16] reported the 19 critical success factors of SS for the services industry in Singapore, while Jayaraman et al. [17] found 25 of these factors for Malaysian industries. As regards Australian industries, a study conducted by Kumar et al. [18] discussed 14 critical success factors for SS implementation. Finally, a study led by Lande et al. [19] managed to compare the key success factors from different countries thanks to a collection of 63 SS-related scientific articles. Thus, there are many reports from industries that have successfully implemented SS, and for a full report, please consult Kwak and Anbari [20].

Organizational learning, a culture of innovation and change, leadership, consistency, *Communication*, integration, understanding of the SS principles, and managerial commitment are some of the most reported critical success factors of SS [19]. However, being SS a methodology, we believe that the first implementation stages have an impact on subsequent ones. Therefore, it is important to analyze, from a quantitative perspective, how and to what extent SS implementation stages are interrelated.

The literature also reports education and training as key factors to explain the success of SS. In this sense, some companies measure the cost–benefit relationship of investments in education and training as an efficiency indicator [21]. In any case, an educated and trained workforce, which is also expert in the necessary tools for SS deployment, helps companies improve *Communication* and solve problems more efficiently. Consequently, the learning process and feedback are facilitated thanks to shared experiences [22]. Obviously, companies invest in a complete education and training process to obtain the best *Operational benefits*, which would reflect on product quality and process cycle times. However, we should bear in mind that, to obtain such benefits and a strong competitive advantage, teamwork is also crucial [23].

Although many studies have identified the critical success factors of SS across contexts and industrial sectors, few of them have managed to find how these factors are interrelated and to what extent, especially in Mexican maquiladora sector. Moreover, there seems to be no consensus regarding the level of importance of these critical success factors in the SS implementation process in the manufacturing industry. To address these gaps, our research quantifies the impact of education and training (access to information, *Didactic material*, and understanding of statistical tools) on *Communication* and operational performance using a structural equation model with ten hypotheses. The model is tested statistically with information from Mexican maquiladora and focused on DMAIC only, because there is a product or process in existence and the problem to solve is the quality in production lines.

The main motivation for this research is that, in Ciudad Juárez (Mexico), there are 326 maquiladora companies, which are subsidiaries that belong to foreign companies established in other countries, but that perform activities of assembly of products in other countries. Those maquiladoras are characterized by having an extensive import of raw materials, and export of finished products, since the production and assembly process is carried out in Mexico, and SS is considered an efficient methodology to ensure quality. Nevertheless, there is an special emphasis on the role of access to information, training in *Statistical techniques* and teaching material, as well as the administration and *Communication* of knowledge to achieve *Operational benefits*.

The remainder of this paper is structured as follows: Section 2 proposes and justifies the research hypotheses; Section 3 describes the methods used to test such hypotheses; Section 4 presents our findings; and Section 5 discusses the research conclusions and industrial implications of results.

2. Background and Hypotheses Formulation

The objective of this study is to quantify the relationships among education, training, *Communication*, and *Operational benefits* under a SS implementation scenario. In this section, we describe these four variables and explain how they can be measured.

2.1. Education in Six Sigma

Education is a pillar of SS implementation. In this research, education has been divided into three latent variables: access to *Information*, use of *Statistical techniques* and use of *Didactic material*.

2.1.1. Access to Information

It is important to exploit all the resources to gather information for the SS project at a planning stage before its execution. In this sense, it is common to consult data on similar SS initiatives previously taken on with success [24]. Black Belts (BBs) and Green Belts (GBs) should assist the information retrieval process, since they know best the company's achievements. However, if it is difficult to find what information is needed for a specific project; project leaders and members must make use of all the information and communication technologies put at their disposal [25].

Access to *Information* is essential when implementing a SS project. If gathered data are not reliable, this might compromise the project's ability to solve the problem that it aims at tackling [26]. However, we should bear in mind that BBs and GBs have the responsibility to control the kind of information that can be accessed and how it may be disclosed [27].

To measure *Information* as a latent variable, the following items, previously studied in other research works, were taken into account:

- Easy access to information [6,14]
- Relevant information found in the company's databases [15,19,28]
- Protection of information obtained from SS projects [15,29]
- Access to other company departments when information is not available in one of them [6,15]
- Rules for information protection and confidentiality [15,30]

2.1.2. Statistical Techniques for Six Sigma

Once the company has defined the problem to be solved, project leaders must find in which *Statistical techniques* project members must be trained [31]. For the basic techniques, it is important to be familiar with key concepts such as measures of central tendency and measures of dispersion [32]. By internalizing these notions, operators can more easily understand what is meant by simple regression, multiple regression, experiments design, and variance analysis, among a few concepts.

To analyze this variable, we took into account the following items:

- Use of root cause analysis tools [14]
- Use of the DMAIC technique [14,19]
- Problem identification through statistical tools [6,30]
- Use of specialized software [19,21]
- Use of graphs and statistics [30]

The use of one or another statistical technique always depends on the type of problem to be solved. For this reason, we propose our first working hypothesis as follows:

Hypothesis 1 (H_1). *Analyzing Information for SS projects has a positive direct impact on the types of Statistical techniques to be taught.*

2.1.3. Didactic Material

Education and training programs are designed according to the very specific needs of each project. This means that *Didactic material* (DMs) greatly vary across organizations [33], yet all of them must aim at helping workers better understand the SS methodology. Likewise, all education and training resources must address both theoretical and practical aspects to allow employees to present some of the company's success stories. Moreover, DMs should not be used for a single project; instead, information contained in a given material should be available for future projects [31]. This education and training process plays a key role in company performance [34].

To measure this latent variable, we considered the following items:

- DMs contribute to a better understanding of how SS works and is used [6]
- BBs appropriately explain the objective of the DM [14,21]
- DMs help execute an SS project [6,19]
- DMs are useful in other SS projects [30,35]
- DMs and BBs instructions are understandable [6,36]

Since the focus of DMs varies depending on the type of problem to be solved and the kind and amount of information that is available, we propose the second working hypothesis as follows:

Hypothesis 2 (H_2). *Information available for SS projects has a positive direct impact on the types of Didactic material to be used during the education and training process*.

The content of DMs must deal with the *Statistical techniques* to be taught. Similarly, facilitators (BBs and GBs) have to master all the contents. For these reasons, we propose our third research hypothesis below:

Hypothesis 3 (H_3). *The Statistical techniques of SS to be taught have a positive direct impact on the Didactic material.*

2.2. Communication

Communication—both horizontal and vertical—is another critical success factor of SS. Because this methodology does not work in isolation, an appropriate flow of *Information* inside the organization is of vital importance [37]. As a means to reach or improve *Communication*, BBs and GBs can organize meetings with their group members on a regular basis to discuss the progress of projects that they have taken on and provide/receive feedback. Such reunions would allow SS teams to make the necessary changes to the projects and to appropriately allocate resources to reach the goals planned [38].

To measure the level of *Communication* that companies have during the SS implementation process, we took into account the following variables:

- BBs, GBs, and project leaders organize meetings [28,29]
- The BB and GB provide support in measuring variables and obtaining information [15,28]
- Group members inform of their progress to their peers [28,35]
- Group members talk about their problems with an SS project [6,19]
- Work teams share their experience among them [21,30]

The *Communication* flow within an organization depends on many factors. Perhaps one of the most important is the extent to which employees receive assistance when analyzing and interpreting information and when identifying measurement and control variables [39]. Considering this fact, we propose our fourth research hypothesis as follows:

Hypothesis 4 (H_4). *The amount of Information available during SS implementation has a positive direct impact on Communication inside work teams.*

The types of *Statistical techniques* used for SS implementation are another factor contributing to an appropriate *Communication* flow. Techniques that are easy to understand will be quickly communicated horizontally among group members; however, more complex procedures for statistical analysis may require greater vertical *Communication* between team members and BBs or GBs [40]. Similarly, *Communication* can be compromised when team members are novice users of some specialized software. In such cases BBs and GBs, which know best how to use such software, have to assist their fellow team members during the information analysis process. This would ensure that figures and mathematical models contain accurate information and present it appropriately [41].

Considering the impact of *Statistical techniques* on the *Communication* flow in organizations implementing SS, we propose our fifth working hypothesis as follows:

Hypothesis 5 (H_5). *Statistical techniques used in SS implementation have a positive direct effect on the quality of Communication*.

Didactic material is a third factor contributing to effective *Communication*. Questions may arise among team members regarding how such materials must be interpreted and analyzed; therefore, it is important to plan regular meetings to provide training and feedback as well as allow team members to share their experiences using such materials [42]. Similarly, to avoid misunderstandings, managers must make sure that all team members are familiar with basic SS concepts [43]. Finally, research has demonstrated that clear *Didactic material* increases self-confidence when initiating new projects and contributes to effective *Communication* [44].

Thus, considering the impact of training and education resources on *Communication*, we propose the sixth working hypothesis below:

Hypothesis 6 (H_6). The *Didactic material* used during the SS implementation process has a positive direct impact on *Communication.*

2.3. Operational Benefits of Six Sigma

One of the main contributions of this research is that we relate the critical success factors of SS to their corresponding benefits. *Operational benefits* are perhaps the most commonly reported of all benefits of SS, although there is also a great amount of literature regarding *Economic Benefits* [20]. Among the main *Operational benefits* of SS, we can find:

- Quality or service perceived by customers [45]
- Cycle time reduction [46]
- Increased employee performance [23,45]
- World-class standards [20,47]
- Waste reduction [23,47]
- Increased teamwork [45,48]
- Multifunctional employees [46,47]

These *Operational benefits* are one of the main reasons for SS implementation. However, companies must know which the most efficient ways to obtain them are. Which activities guarantee them? Which critical success factors favor them? A great number of factors may allow companies to boost their operational performance, yet *Information* is perhaps the most crucial of these factors. If *Information* is not reliable, a project may be poorly designed or the possible alternatives to a problem can be incorrectly analyzed [49]. Likewise, problems may be incorrectly defined if information is not accurate, thereby leading to a loss of money and time.

Considering the impact of having reliable and enough *Information* on *Operational benefits*, our seventh working hypothesis reads as follows:

Hypothesis 7 (H_7). *Information used during SS implementation has a positive direct impact on Operational benefits.*

Statistical techniques are another source of operational performance at all SS implementation stages: problem definition, project execution, and project control [21]. In addition, besides supporting the problem identification process, *Statistical techniques* support other tools, such as the DMAIC approach [50]. This approach is an equivalent of the scientific method applied to the industrial context. The DMAIC approach is commonly adopted by industries worldwide, since its implementation, combined with SS, has reported remarkable success stories.

Therefore, since we believe that *Statistical techniques* have a positive impact on *Operational benefits* under a SS implementation scenario, we propose our eighth research hypothesis as follows:

Hypothesis 8 (H_8). *Statistical techniques used during SS implementation have a positive direct effect on Operational benefits.*

The *Operational benefits* of SS also depend on the quality of the *Didactic material* employed to teach and train workers in the SS methodology. As previously mentioned, if this material lacks clarity, team members may analyze and/or interpret results incorrectly [51]. The responsibility of BBs and GBs is therefore to ensure that every team member is provided with clear and correct instructions during the training sessions [34,43]. This would contribute to minimizing waste, which is one of the objectives of SS from a statistical point of view [37]. This discussion regarding the role of *Didactic material* in *Operational benefits* under a SS implementation scenario allows us to propose our ninth research hypothesis below:

Hypothesis 9 (H_9). *The Didactic material used to teach and train in SS has a positive direct impact on Operational benefits*.

Communication is another critical success factor of SS with a strong impact on *Operational benefits.* If BBs and GBs do not organize meetings with their fellow team members, they may be missing important opportunities to supervise projects, provide feedback, and take timely decisions [52]. Similarly, training and education sessions are key moments to communicate the different notions of quality, and they would indirectly minimize conflicts and crises in the company, since products and services will be standardized as a result of *Communication* [53]. Finally, *Communication* is the means to reach collaboration [38]. Companies that do not communicate their success may cause their employees to work on their own, which contradicts the objective of SS [54]. Following this discussion, we propose our tenth and last working hypothesis as follows:

Hypothesis 10 (H_{10}). *Communication in a SS implementation scenario has a positive direct impact on Operational benefits*.

Figure 1 depicts the ten research hypotheses proposed and previously discussed into a structural equation model, integrated according to dependence established among latent variables.

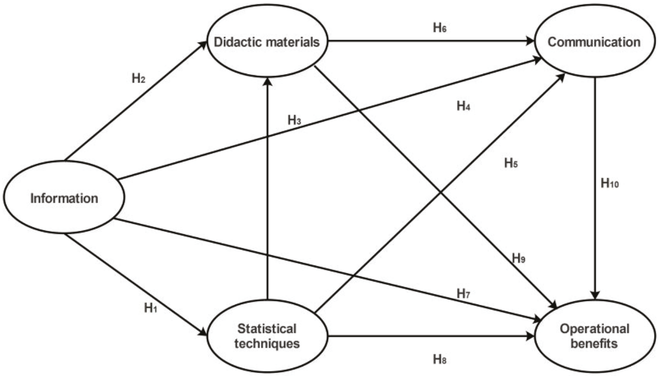


Figure 1. Proposal model.

3. Methodology

To reach our objective and validate the ten hypotheses depicted in Figure 1, we adopted the following methodology.

3.1. Defining Variables and Creating the Survey

In this research, we discuss three basic concepts related to SS: *Education and Training* (*Information, Didactic material*, and *Statistical techniques*), *Communication*, and *Operational benefits.* To find the observed variables of each of these concepts, we conducted a literature review on databases such as Sciencedirect, Ingenta, and Ebscohost, among others.

Once we identified the observed variables, we constructed a survey. The literature review thus represented the rational validation of this survey [55,56], which was composed of three sections. The first section was aimed at collecting sociodemographic data of the participants, and the second section assessed the critical success factors of SS. As regards the third section, it analyzed the different types of SS benefits; however, for the purpose of this study, only took into account items collecting data on *Operational benefits* were taken into account.

The first version of the survey was submitted to a content validation among a group of subject-matter experts. The purpose of the experts validation was to make sure that our instrument had been appropriately adapted to the context of the research [57,58], since data had been collected from previous studies conducted worldwide, not just in Mexico.

Finally, the survey had to be answered using a five-point Likert scale, where a one value indicated that an SS activity was not important or an SS benefit was never obtained. On the other hand, the highest value indicated that an SS activity was highly important or an SS benefit was always obtained. Values of two are used for activities frequently executed or benefits gained, while three for regular, and four for usually.

3.2. Survey Administration

The model and its hypotheses were validated with data collected from the Mexican maquiladora industry during May to July 2016. More specifically, we administered the survey to manufacturing company managers, GBs, BBs, Champions, and group leaders having at least two years of experience in SS or who had participated in SS initiatives at least three times. A list of maquiladoras, manager names and contact information was provided by AMAC (Maquiladoras Association AC) and then, survey administration meetings were scheduled in advance to managers having experience in SS implementation, and the questionnaire was to be answered individually. After a first interview with managers, the snowball method was used to identify other possible responders (managers, green belts, black belts and champions).

However, for participants who cancelled the meeting three times, we stopped insisting due to time restrictions.

3.3. Data Capture and Validation

We constructed an electronic database with gathered data using SPSS software. Then, we performed a screening process to identify missing values and outliers, which were replaced by the median value of items. However, cases or surveys showing more than 10% of missing values were removed.

Latent variables were validated through the following indices:

- Cronbach's alpha and composite reliability index: Used to measure internal validity. Acceptable values must be higher than 0.7 [57,59].
- R-Squared (R^2) and Adjusted R-Squared: Used to measure the predictive validity of dependent latent variables from a parametric perspective. Acceptable values must be higher than 0.2 [60].
- Q-Squared (Q^2): Used to measure predictive validity from a non-parametric perspective. Acceptable values of Q^2 must be similar to their corresponding R^2 and adjusted R^2 values [60].
- Average Variance Extracted (AVE): Used to assess convergent validity, setting 0.5 as the threshold [61].
- Variance Inflation Factors (VIFs): Used as a measure of collinearity. Acceptable values must be below 3.3 [62].

Note that sometimes it is possible to increase the Cronbach's alpha value in a latent variable after removing items that seem to compromise its validity. For this reason, we ran several iterations to validate each latent variable.

3.4. Descriptive Analysis of the Sample

At this stage, we created contingency tables to analyze trends in the sample characteristics. As previously mentioned, we analyzed the genre of participants, number of SS projects that they had taken on, company size, and industrial subsector, among a few.

3.5. Descriptive Analysis of Items

Since we worked with ordinal data, we used the median as a measure of central tendency and the interquartile range (IQR) as a measure of data dispersion. Both measures helped us identify which SS activities and benefits are the most important to manufacturing companies, from the sample's viewpoint. High median values indicated that an SS activity was important to the sample or an SS benefit was always obtained, whereas low median values indicated that an SS activity was not important to the sample or an SS benefit was not obtained. As regards the IQR, high values revealed low consensus among respondents regarding the median value of an item. Low IQR values revealed high consensus among respondents with respect to the median value of an item.

3.6. Hypotheses Validation

To validate the research hypotheses, we created a structural equation model (SEM) using WarpPLS v.5. This software relies on partial least squares (PLS) and is regularly recommended for studies working with Likert scales, small samples, and non-normal data [39]. Likewise, WarpPLS v.5 has been a useful tool for validating theories in SS research. For instance, it was reportedly employed to know the impact of knowledge created in SS on organizational performance [45].

Before interpreting the model, we analyzed its efficiency by computing the model fit and quality indices in Table 1 [63]:

Table 1. Indexes for model validation.

Index	Acceptable If	Description
Average Path Coefficient (APC)	p-value < 0.05	Hypotheses significance
Average R-Squared (ARS)	p-value < 0.05	Predictive model validity
Average Adjusted R-Squared (AARS)	p-value < 0.05	Predictive model validity
Average block Variance Inflation Factor (AVIF)	<3.3	Collinearity among latent variables
Average Full collinearity VIF (AFVIF)	<3.3	Collinearity among latent variables
Tenenhaus Goodness of Fit (GoF)	>0.36	Data fit to model
Simpson's Paradox Ratio (SPR)	>0.7	Direction in relationship
R-Squared Contribution Ratio (RSCR)	>0.7	Direction in relationship
Statistical Suppression Ratio (SSR)	>0.7	Direction in relationship
Nonlinear Bivariate Causality Direction Ratio (NLBCDR)	>0.7	Direction in relationship

Once the model proved to be statistically stable, we proceeded to interpret it. For this interpretation, we analyzed three types of effects in every relationship:

Direct effects: They validate hypotheses presented in Figure 1. Every direct effect corresponds to a hypothesized relationship between latent variables.
Indirect effects: These occur between two latent variables through a mediating variable.
Indirect effects are always interpreted using two or more model paths.
Total effects: They are the sum of direct and indirect effects in a relationship.

All effects were associated with a beta (β) value—expressed in standard deviations—and a p-value for the statistical significance of effects at a 95% confidence level, thus setting 0.05 as the cutoff and testing the null hypothesis: $\beta = 0$, against the alternative hypothesis: $\beta \neq 0$. Finally, every effect also included an effect size (ES) to represent the amount of R^2 or explained variance contained in dependent latent variables [64].

4. Results and Discussion

After three months of administering the survey, from May to July 2016, we collected 323 surveys or cases, but only 301 of them were analyzed, since the remaining ones presented more than 10% missing values and were excluded. Results from the data analysis are discussed in the following subsections.

4.1. Descriptive Analysis of the Sample

Table 2 presents the sample's characteristics regarding surveyed industries and job positions. As it can be observed, 289 participants reported information on these two aspects, and the automotive industry was the most surveyed. Similarly, men represented the majority of the sample, whereas only 83 women participated in the study. As regards job positions, 88 Champions, 69 Master Black Belts (MBs), 64 Black Belts (BBs), and 80 Green Belts (GBs) formed the sample.

Table 2. Industrial sector and years of experience.

Industrial Sector	Years or Experience on SS					Total
	2–3	2–4	4–5	5–10	>10	
Automotive	49	48	41	27	16	181
Electric	13	4	6	5	5	33
Machining	8	7	10	3	2	30
Electronic	9	3	7	1	2	22
Medical	1	6	4	2	3	16
Aeronautic	2	0	2	1	2	7
Total	82	68	70	39	30	289

4.2. Descriptive Analysis of Items

Table 3 presents the descriptive analysis of the latent variables and their corresponding items, also known as observed variables. Data are organized in descending order, according to the median value of items. In this sense, it should be noted that none of the items from the critical success factors of SS showed a median value above four, although they are all higher than three. On the other hand, two *Operational benefits* had a median value greater than four, meaning that they are usually obtained in Mexican manufacturing industries.

Table 3. Measures of central tendency and dispersion.

Latent Variable	Observed Variables (Items)	Percentile			IR
		25	50	75	
Information	Rules for information protection and confidentiality	2.893	3.811	4.658	1.766
	Protection of information obtained from SS projects	2.737	3.659	4.521	1.784
	Access to other company departments when information is not available in one of them.	2.684	3.659	4.545	1.861
	Relevant information found in the company's data bases	2.538	3.506	4.426	1.889
	Easy access to information	2.353	3.311	4.270	1.917
Statistical techniques	Use of graphs and statistics	3.022	3.956	4.750	1.728
	Use of root cause analysis tools	2.838	3.776	4.624	1.786
	Problem identification through statistical tools	2.774	3.693	4.556	1.782
	Use of specialized software	2.602	3.654	4.570	1.968
	Use of the DMAIC technique	2.670	3.651	4.543	1.874
Didactic material	DMs and BBs instructions are understandable	2.937	3.717	4.528	1.591
	DMs contribute to a better understanding of how SS works and is used	2.676	3.667	4.526	1.850
	DMs help execute an SS project	2.718	3.659	4.540	1.821
	DMs are useful in other SS projects	2.754	3.642	4.497	1.743
	BBs appropriately explain the objective of the DM	2.669	3.629	4.512	1.843
Communication	The BB and GB provide support in measuring variables and obtaining information	2.626	3.639	4.533	1.907
	Work teams share their experience among them	2.629	3.635	4.540	1.911
	Group members talk about their problems with an SS project	2.662	3.592	4.459	1.797
	BBs, GBs, and project leaders organize meetings and reunions	2.564	3.572	4.481	1.917
	Group members inform of their progress to their peers	2.544	3.552	4.487	1.943
Operational benefits	Quality or service perceived by customers	3.139	4.112	4.846	1.708
	Cycle time reduction	3.140	4.010	4.763	1.622
	Waste reduction	3.085	3.988	4.734	1.648
	World-class standards	3.069	3.964	4.742	1.673
	Increased employee performance	3.044	3.918	4.721	1.677
	Increased teamwork	3.009	3.888	4.704	1.695
	Multifunctional employees	2.797	3.836	4.716	1.918

In addition, from this descriptive analysis it is important to highlight the following results:

- *Information* should be highly accessible, yet in this research it holds the last position. On the other hand, it seems that such *Information* is highly protected, since this item holds the first place.
- Graphs and statistics seem to be common tools when deploying SS projects in Mexican manufacturing companies, since this item was ranked first within latent variable *Statistical techniques*. However, the use of the DMAIC approach seems to be far less important, since it was ranked last. Unfortunately, when companies do not promote the use of this approach, operators may feel afraid of joining work and improvement teams.
- As regards *Didactic materials*, the item referring to clear instructions from instructors showed the highest median. This implies that leaders responsible for training and education programs are competent professionals.
- Support and assistance from BBs and GBs is the most important item from latent variable *Communication.* On the other hand, progress monitoring and supervising held the last place in this analysis. Unfortunately, if SS projects are not regularly monitored and supervised, it may be difficult to detect deviations on time.
- As regards *Operational benefits*, results indicate that Mexican manufacturing companies remarkably improved their product quality as a result of SS implementation. Such results reflect the objective of SS, which is to minimize defects in products and services. However, the analysis also indicates that a multidisciplinary workforce is a much less common benefit of SS.

4.3. Validation of Latent Variables

Table 4 introduces results obtained from the data validation process. Based on such results, the following conclusions were reached:

- All latent variables had enough predictive validity from both parametric and non-parametric perspectives, since all R^2, Adjusted R^2, and Q^2 values were higher than 0.2. Moreover, all Q^2 values were similar to their corresponding R^2 values.
- All latent variables had enough internal validity, since the Cronbach's alpha and the composite reliability index were higher than 0.7 in all cases.
- All latent variables reported enough convergent validity, since all AVE values were higher than 0.5.
- All latent variables were free from collinearity problems, since the VIFs values were lower than 3.3.

Table 4. Validation for latent variables.

Index	Latent Variable				
	Didactic Material	*Statistical Techniques*	*Communication*	*Operational Benefits*	*Information*
R-squared	*0.652*	*0.553*	*0.636*	*0.603*	
Adjusted R-squared	0.65	0.552	0.632	0.597	
Composite reliability	0.913	0.915	0.947	0.925	0.905
Cronbach´s alpha	0.872	0.883	0.93	0.906	0.869
Average variance extracted	0.723	0.683	0.781	0.639	0.657
Full collinearity variance inflation factor	3.261	3.109	2.973	2.491	2.868
Q-squared	0.653	0.552	0.635	0.606	

4.4. SEM Evaluation

The model depicted in Figure 1 was tested using the model fit and quality indices described in the methodology. The list below shows the obtained values for these indices. In addition, the tested version of the model is presented in Figure 2:

- Average path coefficient (APC) = 0.333, $p < 0.001$
- Average R-squared (ARS) = 0.611, $p < 0.001$
- Average adjusted R-squared (AARS) = 0.608, $p < 0.001$
- Average block VIF (AVIF) = 2.810, ideally $\leq$3.3
- Average full collinearity VIF (AFVIF) = 2.940, ideally $\leq$3.3
- Tenenhaus GoF (GoF) = 0.652, large $\geq$0.36
- Sympson's paradox ratio (SPR) = 1.000, ideally = 1
- R-squared contribution ratio (RSCR) = 1.000, ideally = 1
- Statistical suppression ratio (SSR) = 1.000, acceptable if $\geq$0.7
- Nonlinear bivariate causality direction ratio (NLBCDR) = 1.000, acceptable if $\geq$0.7

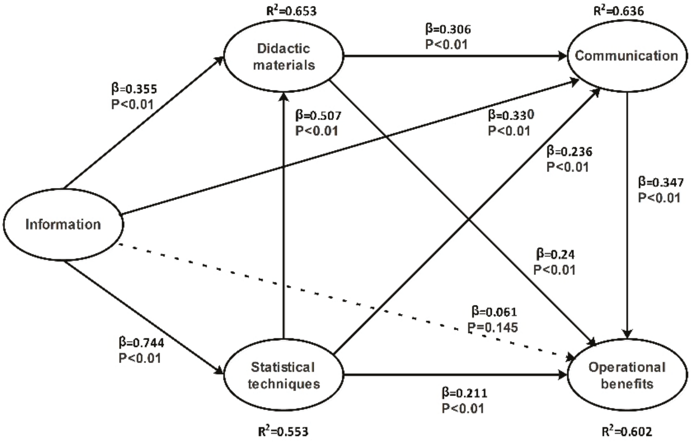


Figure 2. Evaluated model.

These results suggested the existence of no collinearity problems in the proposed model; thus, we could successfully analyze and interpret its values. In this sense, in Figure 2, we associated every relationship with a beta (β) value and a *p*-value. The former represents a dependency measure whereas the latter is the significance value for the hypothesis testing. In addition, each dependent latent variable included an R^2 to indicate its amount of explained variance. After analyzing such values, we found nine statistically significant relationships, depicted as solid lines, and one statistically not significant relationship, illustrated as a dotted line.

4.4.1. Direct Effects: Hypotheses Evaluation

Table 5 presents the results from the hypotheses evaluation. For a hypothesis to be statistically significant at a 95% confidence level, its corresponding *p*-value had to be lower than 0.05. For every hypothesis, the table specifies which dependent (DLV) and independent (ILV) latent variables were involved, the beta and its *p*-value, and its acceptance into or rejection from the model.

Table 5. Conclusions about hypotheses.

Hypothesis	ILV	DLV	β	*p*-Value	Conclusion
H_1	*Information*	*Statistical techniques*	0.744	<0.01	Accepted
H_2	*Information*	*Didactic material*	0.355	<0.01	Accepted
H_3	*Statistical techniques*	*Didactic material*	0.507	<0.01	Accepted
H_4	*Information*	*Communication*	0.330	<0.01	Accepted
H_5	*Statistical techniques*	*Communication*	0.236	<0.01	Accepted
H_6	*Didactic material*	*Communication*	0.306	<0.01	Accepted
H_7	*Information*	*Operational benefits*	0.061	0.145	Rejected
H_8	*Statistical techniques*	*Operational benefits*	0.211	<0.01	Accepted
H_9	*Didactic material*	*Operational benefits*	0.240	<0.01	Accepted
H_{10}	*Communication*	*Operational benefits*	0.347	<0.01	Accepted

4.4.2. Effect Sizes

Effect sizes indicate the contribution of an independent latent variable to the R-Squared coefficient of its corresponding dependent latent variable. Table 6 shows the effects size for every relationship. It should be noted that latent variable *Statistical techniques* was 55.3%, explained by a single independent latent variable: *Information,* whereas in the remaining relationships, the dependent latent variables were explained by two or more independent latent variables. For instance, we found that *Didactic material* was 65.3%, explained by *Statistical techniques* and *Information,* the former being responsible for 39.2% of the variability, and the latter explaining 26.1%. The remaining relationships were similarly interpreted.

In addition, considering the ES values, we reached the following conclusions:

- Latent variable *Statistical techniques* is the most important when explaining the variability of *Didactic material*, being ES = 0.392.
- Latent variable *Information* is key to explaining the variability of *Communication* (ES = 0.242), although *Didactic material* seems to have a similar effect on it (ES = 0.224).
- Latent variable *Communication* is a crucial element for obtaining *Operational benefits*, since ES = 0.249.

Table 6. R-squared contribution to dependent latent variable.

To	From				R^2
	Didactic Material	***Statistical Techniques***	***Communication***	***Information***	
Didactic material		0.392		0.261	0.653
Statistical techniques				0.553	0.553
Communication	0.224	0.17		0.242	0.636
Operational benefits	0.168	0.146	0.249	0.039	0.602

4.4.3. Sum of Indirect Effects

Indirect effects between two latent variables occur through a mediating variable. Table 7 presents the sum of indirect effects between latent variables, the *p*-value for the statistical hypothesis testing, and the effect size.

After analyzing the research hypotheses (see Section 4.4.1), we found that latent variable *Information* did not have a statistically significant effect on *Operational benefits*; however, its indirect effect is statistically significant, and it explained up to 35.1% of the variability of *Operational benefits.* This is the largest indirect effect reported in the model, with β = 0.585. From a similar perspective, we found that *Information* had a statistically significant indirect effect on *Communication* thanks to latent variables *Statistical techniques* and *Didactic material.* In this case, although the direct effect also reported a statistically significant value (β = 0.330), the indirect effect was much larger (β = 0.399).

In conclusion, latent variable *Information* had statistically significant indirect effects on all the remaining latent variables, and such effects were the largest ones. For this reason, it was placed in the top left-hand corner of the model.

Table 7. Sum of indirect effects.

To	From		
	Didactic Material	*Statistical Techniques*	*Information*
Didactic material			0.377 ($p < 0.001$), ES = 0.277
Communication		0.155 ($p < 0.001$), ES = 0.112	0.399 ($p < 0.001$), ES = 0.293
Operational benefits	0.106 (p = 004), ES = 0.074	0.257 ($p < 0.001$), ES = 0.178	0.585 ($p < 0.001$), ES = 0.351

4.4.4. Sum of Total Effects

The total effects of a relationship are the sum of its direct and indirect effects. As Table 8 demonstrates, the ten total effects that we found were statistically significant at a 95% confidence level, since all *p*-values were below 0.05. In addition, the analysis revealed that latent variable *Information* affected all the other latent variables and also caused the largest total effects. In this sense, its total effects on *Statistical techniques* are worth being highlighted. *Information* directly influenced *Statistical techniques* in 0.744, but it was also indirectly responsible for its variability in 53.8% because the effect size is 0.538. Such high value in that relationship indicates that *Information* is a pillar of SS, but also it is required a good *Statistical technique* for analysis.

Table 8. Sum of total effects.

To	From			
	Didactic Material	*Statistical Techniques*	*Communication*	*Information*
Didactic material		0.507 ($p < 0.001$) ES = 0.392		0.732 ($p < 0.001$) ES = 0.538
Statistical techniques				0.744 ($p < 0.001$) ES = 0.553
Communication	0.306 ($p < 0.001$) ES = 0.224	0.391 ($p < 0.001$) ES = 0.282		0.729 ($p < 0.001$) ES = 0.535
Operational benefits	0.346 ($p < 0.001$) ES = 0.243	0.468 ($p < 0.001$) ES = 0.324	0.347 ($p < 0.001$) ES = 0.249	0.646 ($p < 0.001$) ES = 0.421

Currently, it is not enough to have access to information to guarantee the SS success, a deep analysis is required and modern techniques must be used, such as big data. In addition, the relationship between *Information* and *Communication* is very high with a beta value of 0.729, indicating that the knowledge must be transmitted and saved an important resource.

5. Conclusions and Industrial Implications

In this study, we assessed 20 activities related to educational processes (*Information, Statistical techniques*, and *Didactic material)* and *Communication* as critical success factors of SS. We associated these activities with seven *Operational benefits* of SS. In the multivariate analysis performed on 301 surveys, all SS activities showed a median value higher than 3 but lower than 4, implying that they are regularly performed in the Mexican manufacturing sector. On the other hand, two of the seven *Operational benefits* reported a median value higher than 4, thereby implying that they are always obtained.

Although Mexican manufacturing companies seem to rely on effective rules to guarantee confidentiality of *Information*, SS team members report that access to such data as resource material to plan and start SS projects is not easily granted. Organizations should further analyze this issue, since it may affect employee engagement in SS initiatives. In other words, it is good to motivate employees to improve the production process, but it is equally important to grant them access to the necessary data and information, especially during the first implementation stages of an SS project. If companies do not work on this, SS projects are likely to be incorrectly planned, because of a lack of information related to the production process status. However, this can be a hard activity, because that production process can generate much information for different departments and data analysis requires big data techniques.

As regards *Statistical techniques*, graphs and figures seem to be the main statistical tool to support SS, whereas the DMAIC approach proved to have a less significant place among Mexican manufacturing companies. Such results suggest that organizations approach SS as a statistical technique rather than as a problem-solving methodology. It is important for companies to find a balance between these two conceptions [20], otherwise SS may become a significant obstacle for those who are not experts in statistics [65]. In this sense, other studies, having detected an imbalance between the different ways SS can be approached, also report that, in such cases, SS projects are more often abandoned [66,67] and that is why this is an opportunity for techniques such as big data or novel techniques to help analyze information obtained from production process, because it allows finding trends, as it is applied in education [68], and SS is a philosophy based in education and training.

In addition, our study reports that GBs and BBs usually provide clear instructions on how to work with *Didactic material*. However, we also found that this material is not always useful for future projects, and team members must thus be trained every time they initiate a new project with SS, which increases final costs. Concerning *Communication*, GBs and BBs in Mexican manufacturing companies seem to provide appropriate and sufficient assistance to team members; nevertheless, when it comes to supervising and monitoring SS projects, their performance appears to be less regular. Sadly, employees may lose their motivation when they perceive a lack of consistency when it comes to monitoring projects [69].

Finally, product quality and process cycle time reduction seem to be the main *Operational benefits* of SS, whereas teamwork and multifunctional skills are less common. In this case, it is important to add human attributes to SS, since human resources know the administrative procedures, the production processes, and the company's opportunities for improvement [21,70].

After analyzing the relationships between the latent variables, the following conclusions regarding direct effects were reached:

- *Information* available to solve problems defines which *Statistical techniques* employees will be trained and how *Didactic material* must be designed, so they can be clear to all team members and reusable in future projects. However, big data can be implemented as a *Statistical technique* for yellow belts, green belts and champions, because it can help to find hidden patterns into *Information*.
- Having available *Information* does not automatically guarantee *Economic Benefits*, because direct effect is statistically not significant. First, managers must be focused on teaching and training employees in the use of the necessary *Statistical techniques* through clear and meaningful *Didactic material*. The fact that *Information* does not have a direct impact on *Economic benefits* implies that it is not appropriately analyzed or it cannot directly become a benefit. *Communication* and education are therefore required, and managers should be part of appropriate *Communication* channels and training sessions. However, managers must also encourage that workers integrate in SS projects and share the knowledge gained among them as a way to disseminate their experiences solving problems.
- *Information* has a direct effect on *Statistical techniques*, which denotes the importance of this variable at the first implementation stages of an SS project, where employees identify the problem and define it. In addition, *Information* has the largest direct effects on all subsequent latent variables, meaning that, if *Information* is not reliable or easily accessible, companies must have problems implementing SS and probably, they abandon this philosophy, resulting in a lack of quality in their production process.
- Another important relationship involves *Statistical techniques* and *Didactic material*. This relationship again reflects the importance of education and training for an appropriate SS implementation. In fact, the three largest effects, considering the β values, involve the three latent variables that make up the education process.

- Managers and SS members must strive to implement an appropriate education and training scheme in which *Information* is reliable but also easily accessible. Likewise, education and training must focus on the use of basic *Statistical techniques* through clear *Didactic material*, since these three variables affect *Communication*, both vertical and horizontal. In other words, without a suitable education and training process, *Information* cannot properly flow; consequently, all involved variables may diminish the indirect effects that *Information* has on *Economic benefits* through *Communication* as the mediating variable.

6. Limitations and Future Research

The set of hypotheses that have been proposed in the model that associates access and quality of information with *Didactic material* and *Statistical techniques*, as well as with *Communication* processes and *Operational benefits*, have been tested with information from Mexican maquiladora industry, thus, in another sector and country, it is possible to obtain different results.

In future work, we will get information from industries in other regions of the country to perform comparative analyses to identify the factors that are best associated with SS implementation process. Similarly, given that maquiladoras are a globalized phenomenon, comparisons can be made between countries in South America.

Acknowledgments: Jorge L. García-Alcaraz gratefully acknowledges financial support from the Autonomous University of Ciudad Juarez (IITCA93ET1DEc-19) and Program for Professional Teacher Development (PRODEP).

Author Contributions: In this research, Jorge L. García-Alcaraz and Liliana Avelar-Sosa conceived and designed the structural equation model and collected information for its validation; Juan I. La Torre-Biel and Emilio Jimenez-Macías reviewed the manuscript; and Giner Alor-Hernandez helped in analysis and conclusions.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Furterer, S.L. Lean Six Sigma roadmap. In *Lean Six Sigma Case Studies in the Healthcare Enterprise*; Springer: London, UK, 2014; pp. 11–62.
2. Han, C.; Lee, Y.-H. Toward intelligent integrated plant operation system for six sigma. *IFAC Proc. Vol.* **2001**, *34*, 15–28. [CrossRef]
3. Joseph Gordon, M., Jr. Chapter 5—Six sigma education and using the existing quality methods and procedures. In *Six Sigma Quality for Business and Manufacture*; Elsevier Science B.V.: Amsterdam, The Netherlands, 2002; pp. 171–245.
4. Joseph Gordon, M., Jr. Chapter 8—Six sigma keys to success are control, capability and repeatability. In *Six Sigma Quality for Business and Manufacture*; Elsevier Science B.V.: Amsterdam, The Netherlands, 2002; pp. 333–392.
5. Sharon, H.N.; Miller, M.A.; Stimart, R.P. Demonstrating the value of the user interface design process using six sigma methodology. *Comput. Stand. Interfaces* **1999**, *21*, 167. [CrossRef]
6. De Jesus, A.R.; Antony, J.; Lepikson, H.A.; Peixoto, A.L.A. Six sigma critical success factors in Brazilian industry. *Int. J. Qual. Reliab. Manag.* **2016**, *33*, 702–723. [CrossRef]
7. Tlapa, D.; Limon, J.; García-Alcaraz, J.L.; Baez, Y.; Sánchez, C. Six sigma enablers in Mexican manufacturing companies: A proposed model. *Ind. Manag. Data Syst.* **2016**, *116*, 926–959. [CrossRef]
8. Shaaban, S.; Darwish, A.S. Production systems: Successful applications and new challenges part one—Lean, six sigma, inventory, JIT and TOC. *Prod. Plan. Control* **2016**, *27*, 539–540. [CrossRef]
9. Firka, D. Six sigma: An evolutionary analysis through case studies. *TQM J.* **2010**, *22*, 423–434. [CrossRef]
10. Tsironis, L.K.; Psychogios, A.G. Road towards Lean Six Sigma in service industry: A multi-factor integrated framework. *Bus. Process Manag. J.* **2016**, *22*, 812–834. [CrossRef]
11. Prabhushankar, G.V.; Devadasan, S.R.; Shalij, P.R. Six sigma in Indian automotive components sector: A survey. *ICFAI J. Oper. Manag.* **2008**, *7*, 19–37.

12. De Carvalho, M.M.; Ho, L.L.; Pinto, S.H.B. The six sigma program: An empirical study of Brazilian companies. *J. Manuf. Technol. Manag.* **2014**, *25*, 602–630. [CrossRef]
13. De Jesus, A.R.; Antony, J.; Lepikson, H.A.; Teixeira Cavalcante, C.A.M. Key observations from a survey about six sigma implementation in Brazil. *Int. J. Product. Perform. Manag.* **2015**, *64*, 94–111. [CrossRef]
14. Habidin, N.F.; Yusof, S.M. Critical success factors of Lean Six Sigma for the Malaysian automotive industry. *Int. J. Lean Six Sigma* **2013**, *4*, 60–82. [CrossRef]
15. Brun, A. Critical success factors of six sigma implementations in Italian companies. *Int. J. Prod. Econ.* **2011**, *131*, 158–164. [CrossRef]
16. Chakraborty, A.; Chuan Tan, K. Case study analysis of six sigma implementation in service organisations. *Bus. Process Manag. J.* **2012**, *18*, 992–1019. [CrossRef]
17. Jayaraman, K.; Leam Kee, T.; Lin Soh, K. The perceptions and perspectives of Lean Six Sigma (LSS) practitioners: An empirical study in Malaysia. *TQM J.* **2012**, *24*, 433–446. [CrossRef]
18. Kumar, M.; Khurshid, K.K.; Waddell, D. Status of Quality Management practices in manufacturing SMEs: A comparative study between Australia and the UK. *Int. J. Prod. Res.* **2014**, *52*, 6482–6495. [CrossRef]
19. Lande, M.; Shrivastava, R.L.; Seth, D. Critical success factors for Lean Six Sigma in SMEs (small and medium enterprises). *TQM J.* **2016**, *28*, 613–635. [CrossRef]
20. Kwak, Y.H.; Anbari, F.T. Benefits, obstacles, and future of six sigma approach. *Technovation* **2006**, *26*, 708–715. [CrossRef]
21. Marzagão, D.S.L.; Carvalho, M.M. Critical success factors for six sigma projects. *Int. J. Proj. Manag.* **2016**, *34*, 1505–1518. [CrossRef]
22. Ismyrlis, V.; Moschidis, O. Six sigma's critical success factors and toolbox. *Int. J. Lean Six Sigma* **2013**, *4*, 108–117. [CrossRef]
23. Jacobs, B.W.; Swink, M.; Linderman, K. Performance effects of early and late six sigma adoptions. *J. Oper. Manag.* **2015**, *36*, 244–257. [CrossRef]
24. Zhang, W.; Xu, X. Six sigma and information systems project management: A revised theoretical model. *Proj. Manag. J.* **2008**, *39*, 59–74. [CrossRef]
25. Sabry, A. Factors critical to the success of six-sigma quality program and their influence on performance indicators in some of Lebanese hospitals. *Arab Econ. Bus. J.* **2014**, *9*, 93–114. [CrossRef]
26. De Mast, J.; Lokkerbol, J. An analysis of the Six Sigma DMAIC method from the perspective of problem solving. *Int. J. Prod. Econ.* **2012**, *139*, 604–614. [CrossRef]
27. Wojtaszak, M.; BiaŁY, W. Problem solving techniques as a part of implementation of six sigma methodology in tire production. Case study. *Manag. Syst. Prod. Eng.* **2015**, *19*, 133–137.
28. Arumugam, V.; Antony, J.; Kumar, M. Linking learning and knowledge creation to project success in six sigma projects: An empirical investigation. *Int. J. Prod. Econ.* **2013**, *141*, 388–402. [CrossRef]
29. Wang, F.-K.; Yeh, C.-T.; Chu, T.-P. Using the design for six sigma approach with TRIZ for new product development. *Comput. Ind. Eng.* **2016**, *98*, 522–530. [CrossRef]
30. Atmaca, E.; Girenes, S.S. Lean Six Sigma methodology and application. *Qual. Quant.* **2013**, *47*, 2107–2127. [CrossRef]
31. Fouweather, T.; Coleman, S.; Thomas, A. Six sigma training programmes to help SMEs improve. In *Intelligent Production Machines and Systems*; Elsevier Science Ltd.: Oxford, UK, 2006; pp. 39–44.
32. Tang, L.C.; Goh, T.N.; Lam, S.W. Fortifying six sigma with OR/MS tools. In *Six Sigma*; John Wiley & Sons, Ltd.: Chichester, UK, 2006; pp. 49–69.
33. Kavčič, K.; Gošnik, D. Lean Six Sigma education in manufacturing companies: The case of transitioning markets. *Kybernetes* **2016**, *45*, 1421–1436. [CrossRef]
34. McCrie, R. 4—training and development for high performance. In *Security Operations Management*, 3rd ed.; Butterworth-Heinemann: Boston, MA, USA, 2016; pp. 113–143.
35. Swink, M.; Jacobs, B.W. Six sigma adoption: Operating performance impacts and contextual drivers of success. *J. Oper. Manag.* **2012**, *30*, 437–453. [CrossRef]
36. Ho, Y.-C.; Chang, O.-C.; Wang, W.-B. An empirical study of key success factors for Six Sigma Green Belt projects at an Asian MRO company. *J. Air Trans. Manag.* **2008**, *14*, 263–269. [CrossRef]
37. Cherrafi, A.; Elfezazi, S.; Chiarini, A.; Mokhlis, A.; Benhida, K. The integration of lean manufacturing, six sigma and sustainability: A literature review and future research directions for developing a specific model. *J. Clean. Prod.* **2016**, *139*, 828–846. [CrossRef]

38. Kuvvetli, Ü.; Firuzan, A.R.; Alpaykut, S.; Gerger, A. Determining six sigma success factors in turkey by using structural equation modeling. *J. Appl. Stat.* **2016**, *43*, 738–753. [CrossRef]
39. Kock, N.; Verville, J.; Danesh-Pajou, A.; DeLuca, D. Communication flow orientation in business process modeling and its effect on redesign success: Results from a field study. *Decis. Support Syst.* **2009**, *46*, 562–575. [CrossRef]
40. Saidi-Mehrabad, M.; Paydar, M.M.; Aalaei, A. Production planning and worker training in dynamic manufacturing systems. *J. Manuf. Syst.* **2013**, *32*, 308–314. [CrossRef]
41. Ordaz, N.; Romero, D.; Gorecky, D.; Siller, H.R. Serious games and virtual simulator for automotive manufacturing education & training. *Procedia Comput. Sci.* **2015**, *75*, 267–274.
42. Chauhan, R.; Ghosh, P.; Rai, A.; Shukla, D. The impact of support at the workplace on transfer of training: A study of an Indian manufacturing unit. *Int. J. Train. Dev.* **2016**, *20*, 200–213. [CrossRef]
43. Onimole, S.O.; Zekeri, A. The impact of systematic training approach on the operational performance of manufacturing and engineering industries in southwest zone of Nigeria. *IFE PsychologIA* **2012**, *20*, 119–127.
44. Ikegami, K.; Tahara, H.; Yamada, T.; Mafune, K.; Hiro, H.; Nagata, S. Effects of a mental health training program for manufacturing company managers. *J. UOEH* **2010**, *32*, 141–153. [CrossRef] [PubMed]
45. Boon Sin, A.; Zailani, S.; Iranmanesh, M.; Ramayah, T. Structural equation modelling on knowledge creation in Six Sigma DMAIC project and its impact on organizational performance. *Int. J. Prod. Econ.* **2015**, *168*, 105–117. [CrossRef]
46. Shafer, S.M.; Moeller, S.B. The effects of six sigma on corporate performance: An empirical investigation. *J. Oper. Manag.* **2012**, *30*, 521–532. [CrossRef]
47. Ertürk, M.; Tuerdi, M.; Wujiabudula, A. The effects of six sigma approach on business performance: A study of white goods (home appliances) sector in turkey. *Procedia Soc. Behav. Sci.* **2016**, *229*, 444–452. [CrossRef]
48. De Freitas, J.G.; Costa, H.G.; Ferraz, F.T. Impacts of Lean Six Sigma over organizational sustainability: A survey study. *J. Clean. Prod.* **2017**, *156*, 262–275. [CrossRef]
49. Vinod, M.; Devadasan, S.R.; Sunil, D.T.; Thilak, V.M.M. Six sigma through Poka-Yoke: A navigation through literature arena. *Int. J. Adv. Manuf. Technol.* **2015**, *81*, 315–327. [CrossRef]
50. Moosa, K.; Sajid, A. Critical analysis of six sigma implementation. *Total Qual. Manag. Bus. Excell.* **2010**, *21*, 745–759. [CrossRef]
51. Boothby, D.; Dufour, A.; Tang, J. Technology adoption, training and productivity performance. *Res. Policy* **2010**, *39*, 650–661. [CrossRef]
52. Luxton, S.; Reid, M.; Mavondo, F. Integrated marketing communication capability and brand performance. *J. Advert.* **2015**, *44*, 37–46. [CrossRef]
53. Zaumane, I. The internal communication crisis and its impact on an organization's performance. *J. Bus. Manag.* **2016**, 24–33.
54. Deering, S.; Johnston, L.C.; Colacchio, K. Multidisciplinary teamwork and communication training. *Semin. Perinatol.* **2011**, *35*, 89–96. [CrossRef] [PubMed]
55. Yaşlıoğlu, M.M.; Şap, Ö.; Toplu, D. An investigation of the characteristics of learning organizations in Turkish companies: Scale validation. *Procedia Soc. Behav. Sci.* **2014**, *150*, 726–734. [CrossRef]
56. Garcia-Alcaraz, J.L.; Maldonado-Macias, A.A.; Hernandez-Arellano, J.L.; Blanco-Fernandez, J.; Jimenez-Macias, E.; Saenz-Diez Muro, J.C. The impact of human resources on the agility, flexibility and performance of wine supply chains. *Agric. Econ.* **2017**, *63*, 175–184.
57. Nunnally, J. *Psychometric methods*; McGraw-Hill: New York, NY, USA, 1978.
58. Garcia-Alcaraz, J.L.; Maldonado-Macias, A.A.; Alor-Hernandez, G.; Sanchez-Ramirez, C. The impact of information and communication technologies (ICT) on agility, operating, and economical performance of supply chain. *Adv. Prod. Eng. Manag.* **2017**, *12*, 29–40. [CrossRef]
59. García-Alcaraz, J.L.; Adarme-Jaimes, W.; Blanco-Fernández, J. Impact of human resources on wine supply chain flexibility, quality, and economic performance. *Ing. Investig.* **2016**, *36*, 74–81. [CrossRef]
60. Hair, J.F.; Black, W.C.; Babin, B.J.; Anderson, R.E. *Multivariate Data Analysis*; Prentice Hall: Upper Saddle River, NJ, USA, 2009.
61. Nitzl, C. The use of partial least squares structural equation modelling (PLS-SEM) in management accounting research: Directions for future theory development. *J. Acc. Lit.* **2016**, *37*, 19–35. [CrossRef]
62. Kock, N. Using warppls in e-collaboration studies: Mediating effects, control and second order variables, and algorithm choices. *Int. J. e-Collab.* **2011**, *7*, 1–13. [CrossRef]

63. Kock, N. *Single Missing Data Imputation in PLS-SEM*; ScriptWarp Systems: Laredo, TX, USA, 2014.
64. Hayes, A.F.; Preacher, K.J. Quantifying and testing indirect effects in simple mediation models when the constituent paths are nonlinear. *Multivar. Behav. Res.* **2010**, *45*, 627–660. [CrossRef] [PubMed]
65. Antony, J.; Krishan, N.; Cullen, D.; Kumar, M. Lean six sigma for higher education institutions (HEIS): Challenges, barriers, success factors, tools/techniques. *Int. J. Product. Perform. Manag.* **2012**, *61*, 940–948. [CrossRef]
66. Lodgaard, E.; Ingvaldsen, J.A.; Gamme, I.; Aschehoug, S. Barriers to lean implementation: Perceptions of top managers, middle managers and workers. *Procedia CIRP* **2016**, *57*, 595–600. [CrossRef]
67. Yamada, T.T.; Poltronieri, C.F.; Gambi, L.D.N.; Gerolamo, M.C. Why does the implementation of quality management practices fail? A qualitative study of barriers in Brazilian companies. *Procedia Soc. Behav. Sci.* **2013**, *81*, 366–370. [CrossRef]
68. Laux, C.; Li, N.; Seliger, C.; Springer, J. Impacting big data analytics in higher education through six sigma techniques. *Int. J. Product. Perform. Manag.* **2017**, *66*, 662–679. [CrossRef]
69. Coronado, R.B.; Antony, F. Critical success factors for the successful implementation of six sigma projects in organisations. *TQM Mag.* **2002**, *14*, 92–99. [CrossRef]
70. Harrison, A. Chapter 13—An application of six sigma in human resources. In *World Class Applications of Six Sigma*; Butterworth-Heinemann: Oxford, UK, 2006; pp. 224–238.

sustainability

MDPI

Article

Impacts of Leadership on Project-Based Organizational Innovation Performance: The Mediator of Knowledge Sharing and Moderator of Social Capital

Junwei Zheng [1], **Guangdong Wu** [2] **and Hongtao Xie** [1,*]

[1] Faculty of Civil Engineering and Mechanics, Kunming University of Science and Technology, Kunming 650500, China; zjw1989@kmust.edu.cn
[2] Department of Construction Management, Jiangxi University of Finance and Economics, Nanchang 330013, China; gd198410@163.com
* Correspondence: xhtkmust@kmust.edu.cn; Tel.: +86-137-5914-3607

Received: 21 September 2017; Accepted: 17 October 2017; Published: 20 October 2017

Abstract: With the increasing importance of leadership in project-based organizations, innovation is essential for the sustainable development of construction projects. Since few studies have explored the relationship between leadership and innovation in construction projects, this study fills this research gap and makes a significant theoretical contribution to the existing body of literature. Based on a knowledge-rated and resource-based view, this study aims to investigate various effects of different types of leadership on innovation performance in a construction project-based organization. Therefore, a theoretical model was constructed to explore the mediation mechanism and boundary condition of different types of leadership to improve innovation. The theoretical model was validated with empirical data covering project managers and engineers from the project-based organization in China via regression analysis and path analysis. The results show that transformational leadership and transactional leadership have some positively significant effects on knowledge sharing and innovation performance. Meanwhile, knowledge sharing partially mediates the relationship between transformational leadership and/or transactional leadership and innovation performance. Additionally, by considering different levels of social capital, transformational leadership is likely to have a strong positive impact on innovation performance through knowledge sharing. Our findings ensure a better understanding of the role of leadership, knowledge management, and social capital in the innovation process of construction projects. Therefore, project managers should promote a higher stimulation of a leadership behavior, encouraging knowledge management, and establishing the social capital, thus improving the innovation performance in the project-based organizations in construction projects.

Keywords: leadership; innovation performance; knowledge sharing; social capital; project-based organization

1. Introduction

With the rapid development of the construction technology, the construction industry has become more knowledge-intensive, thus it became imperative to carry on innovation by the sustainability trends [1]. Innovation in construction is beneficial for the sustainability (i.e., competitiveness improvement) of the construction industry and firms, and the sustainability (i.e., quality and technical level) of the construction projects. The innovation level in the construction industry attracts both criticism [2] as well as praise [3]. On the one hand, the initial criticism was that construction lags behind the innovativeness of the manufacturing and service sectors [4]. On the other hand, Pries and Janszen [5] stated that construction is inherently innovative. The form of project-based organization was adopted by increasing construction companies to improve project efficiency and performance [6], which was seen as a form of the organization where different parties participated. Faced with a different complexity of the various construction projects and considering sustainability demand, the construction project-based organization would start addressing the challenge of sharing knowledge and coordinating the relationship to carry out innovation [7]. The project-based character of construction activities that are practiced due to a strong price orientation, are by several scholars seen as for a lack of knowledge transfer and innovation [8]. Furthermore, various factors are hampering innovation in construction, such as the conservation of established practices, fear of future collaboration, perceived high financial investment needed in innovation, and limited time-span and resources [9]. As a result, how to drive innovation in a construction project-based organization has become an emerging topic and pressing issue.

Consist with OECD definition [10], innovation is defined as a new or significantly improved product (good or service), process (production or delivery method), marketing method or managerial method [11]. Park et al. [12] defined innovation as the generation, development, and implementation of ideas that are new to an organization and have practical or commercial benefits. Thus, the term innovation is closely related to organization. Meanwhile, it is well documented that innovation is essential to various aspects of the organizational performance [9]. Previous studies investigated many factors that predicted organizational innovation, including organizational strategy [13], organizational learning [14], communication and engagement of stakeholders [15] and positive expectations of innovation from the team [16]. First, leadership is essential for enacting and implementing an innovative organizational strategy and structure, promoting organizational learning, and motivating team visions towards innovation. Therefore, leadership has been regarded as one of the most important factors in determining the degree to which employees strive for innovation. Although previous studies identify leadership as a critical factor in the development of innovation in construction [17,18], detailed or in-depth studies on the effectiveness of different types of leadership on innovation performance in construction projects are still in their infancy. Besides, a large part of the literature focused on innovation at the firm level [19], thus a limited number of studies mentioned innovation from an inter-organizational perspective in construction underlying the cooperation of members engaged in construction projects [20].

Second, knowledge sharing has acted as an important enabler for innovation [1,21]. Innovation is associated with the change of information [22]. Knowledge sharing enables a transfer of experience which avoids a repetition of mistakes in a construction project-based organization [23]. According to the knowledge-based view, the project-based organization in construction acts as the knowledge platform for the members and the basis for integrating multidisciplinary expertise [7,24]. Knowledge exchange provides the information channels to facilitate communication between partners in organization [25]. Although there is an increasing attention of knowledge sharing in the organization, the focus on knowledge sharing of construction project-based organization still remains rather poor [26].

Third, the construction project-based organization, acts as the temporary organization in a construction project, in which the members and resources are aggregated. The way individual linked, the tangible and intangible resources exchanged between members, are referred to as project social capital [27]. Although previous studies found that social capital benefits organizational outcomes, whether and how the social capital of construction projects influences the innovation performance in the project-based organization setting is still a topic to be considered. Besides, access to knowledge is important for the organizational performance, and the project social capital provides an arena for members to use knowledge as resources. Thus, it is essential to explore the inner relationship between knowledge sharing and social capital in the construction project-based organization.

This study considers a project-based organization from an inter-organization perspective to emphasize the importance of coordination, knowledge exchange and mutual relationship among different parties in the construction projects. Hence, this study mainly focuses on the inter-organization level of construction projects, aiming to fill this research gap and by investigating the mechanism between different types of leadership and construction project-based organizational innovation performance. To obtain an intensive understanding of the impact of different types of leadership on innovation performance in construction projects, this study reveals the internal mechanism used for fostering knowledge exchange and identifying the association with social capital. The reliability test, validity test, regression analysis and moderated path analysis are incorporated in this study to analyze the relationships among different leadership styles, knowledge sharing, social capital, and project-based organizational innovation performance in construction projects. The main theoretical implications of our findings are the following: (i) Transformational leadership and transactional leadership act as the precursor of knowledge sharing and innovation performance, while knowledge sharing partially transmits the effect of leadership on innovation performance. (ii) Project social capital acts as the moderation mechanism, in which the effect of transformational leadership on innovation performance via knowledge sharing are amplified. To sum up, this study not only addresses both key leadership (i.e., transformational leadership and transactional leadership) in construction project-based organization, but also investigates the mediation role of knowledge sharing and the boundary condition of social capital to capture the mechanism used to improve innovation performance in a construction project-based organization. This study establishes a concept model to introduce the leadership styles into a construction project-based organization, and by employing the contingent model it tests how transformational leadership, transactional leadership and knowledge sharing interact with social capital to influence innovation performance. This study provides a more integrative view of how leaders may stimulate project-based organizational innovation in construction projects by facilitating knowledge sharing and improving relationships with the team members and stakeholders.

The rest of the paper is structured as follows. Section 2 discusses the literature review and various hypothesis formulations. Section 3 deals with sample selection, scales design, reliability test, factor analysis, and methodology. Section 4 deals with measurement models assessment and hypothesis testing. Section 5 discusses the results. Finally, conclusions, implications, limitations and future directions are drawn in the last section.

2. Literature Review and Research Hypotheses

Previous studies on organizational leadership highlight that a critical mission of leaders is to facilitate followers to engage in activities that strive for the achievement of group and/or organization goals [28], although not all leaders are equally inclined to motivate the followers to be innovative. In this study, different leadership styles, i.e., transformational leadership and transactional leadership, are proposed as key variables to encourage and motivate followers to engage in innovative actions in construction projects. The central factors of the mechanism transferring or regulating the effects of two leadership types are knowledge sharing and social capital. The hypothesized relationships are depicted in Figure 1, which describes the theoretical framework used in this study.

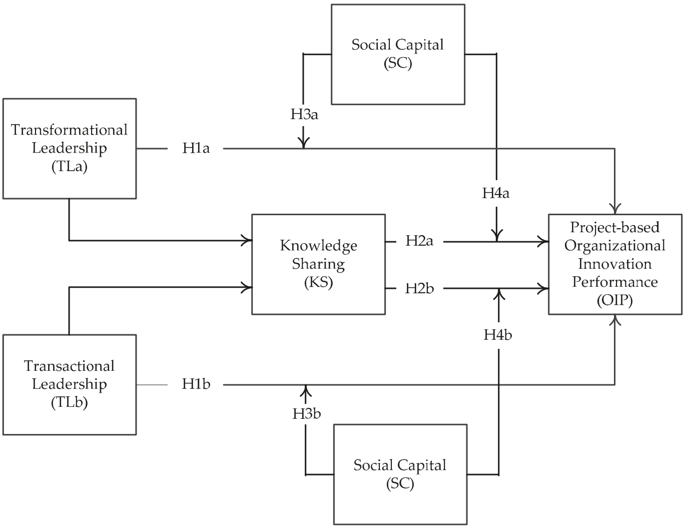

Social Capital (SC)
Transformational Leadership (TLa)
H3a
H1a
H4a
Knowledge Sharing (KS)
H2a
H2b
Project-based Organizational Innovation Performance (OIP)
H4b
Transactional Leadership (TLb)
H1b
H3b
Social Capital (SC)

Figure 1. The theoretical model.

2.1. Main Effects: Leadership Styles and Project-Based Organizational Innovation Performance

The concepts of transformational and transactional leadership have long been adopted in previous studies [29,30]. Transformational and transactional leadership are also the two main leadership styles on which we have focused in this study. On the one hand, transformational leadership refers to leaders focusing on meeting the higher-order intrinsic needs of their followers, resulting in followers identifying themselves with the needs of their leader [31]. It has four dimensions: charisma (or idealized influence), inspirational stimulation, intellectual stimulation and individualized consideration [30]. On the other hand, transactional leadership refers to leaders focusing on satisfying the extrinsic needs of their subordinates, such a focus results in the subordinates performing the tasks that their leader requires [31]. It involves contingent rewards and management by exception [30].

The effects that different types of leadership might have on innovation vary in the literature. The transformational leader motivates the employees to strive for the collective goals, and stimulating followers' focus and understanding towards the organizational vision [32]. For example, Bass and Riggio [33] suggested that transformational leadership enhances the creative effort in an organization and contributes to the innovative goal. Transformational leadership is a strong supporter of the unconventional things that foster innovation and improve performance [34]. According to Bass and Riggio [33], a leader having a contingent reward behavior would obtain employees' prior agreement on the job to be done in exchange of rewards for delivering the performance within a time frame. Moreover, a leader promoting an active management by exception supervises employees, identifies errors or mistakes, and then takes corrective actions. Additionally, the project' success partially depends on the manager's leadership style [35], thus the leadership has a great influence on the performance of the construction work [36]. Empirical studies exhibited the effects of transformational or transactional leadership on the innovation and performance of organizations [37,38]. The empirical evidence in the relevant literature is still ambiguous (e.g., [9,39]), and the conclusion always emphasize the positive influence of transformational leadership when compared with transactional leadership [40]. Furthermore, studies focusing on the relationships among transformational leadership, transactional leadership and innovation performance at the inter-organization level are still rare, referring to the construction project-based organization in which different parties of construction projects are engaged. Thus, it is commonly believed that transformational leadership and transactional leadership will have a significant and positive impact on organizational performance, mainly on the innovation performance in the project-based organization.

Hypothesis 1a. *Transformational leadership is positively related to innovation performance of the construction project-based organization.*

Hypothesis 1b. *Transactional leadership is positively related to innovation performance of the construction project-based organization.*

2.2. Mediation Effects: Knowledge Sharing

Innovation is one of the most knowledge-intensive activities [41]. Knowledge-intensive activities are simultaneously essential for innovation in the organization operation process [42], and enable interaction and promote the connection among actors in the organization [43]. Knowledge management is a pre-requisite for creating, sharing, and storing creative ideas. Effective leadership plays a significant role in promoting a supportive climate for exposing knowledge into organization innovation [34]. Knowledge management refers to all managerial activities which helps individuals in the organization to create new knowledge and share this knowledge with others in order to improve the performance of the organization [27]. Both knowledge sharing and knowledge application have been known to facilitate the creation of new ideas and processes so that it can improve the performance of the organizations.

Furthermore, Birasnav [44] proposed that knowledge management plays a mediation role in the relationship between transformational leadership and organizational performance, when controlling the impact of transactional leadership. Singh [45] investigated the role of leadership in the knowledge management process. Han [46] also exhibited the effect of transformational leadership and knowledge sharing. In addition, Bryant [47] proposed that leaders contribute to improving performance through exploiting knowledge in the organizations. Moreover, the previous studies focus on the associations between leadership and knowledge sharing at the firm level [48], and an important strand of literature concluded that knowledge management or knowledge sharing contribute significantly to innovation efforts and help ameliorate organizational performance at the firm level [49]. However, this type of studies in the construction projects is still rare. Hence, transformation leaders nurture the intrinsic needs to share knowledge while transaction leaders involve in providing rewards to share knowledge. It is also expected that such leaders enhance innovation performance of project-based organization through the process of knowledge sharing in the construction projects. Based on the above discussion, the hypotheses developed are the following.

Hypothesis 2a. *Knowledge sharing would mediate the relationship between transformational leadership and project-based organizational innovation performance.*

Hypothesis 2b. *Knowledge sharing would mediate the relationship between transactional leadership and project-based organizational innovation performance.*

2.3. Moderating Effects: Social Capital

According to the social capital theory, social capital is both a tangible and intangible resource to organizations to be used appropriately. Social capital is associated with the extent to which people share information, and are concerned with the resources embedded in the relationship network [50]. Further, it is possible to obtain the necessary resources for new technology adoption and technology improvement [51]. Capital acts like a precursor to organizational innovation and performance [52]. Previous studies focusing on the impact of transformational leadership have predominantly studied human capital rather than social capital [53,54]. Social capital theory suggests that social relationships among organizational members and those with outside actors, confer vital resources such as information or advices among others, all representing important preconditions for information sharing, knowledge creation, and innovation [55]. The leadership style is a combination of characteristics, skills and behavior that the managers uses to interact with employees [56]. Leaders manage and influence a vital part of the resources through social capital, though a few studies addressed the effects of transformational leadership and transactional leadership on social capital [57]. It is essential to study how different leadership styles leverage social capital to foster innovation. Moreover, social capital research has been somewhat neglected in the construction projects. Thus, the research hypotheses used in our research are the following.

Hypothesis 3a. *Social capital moderates the positive relationship between transformational leadership and project-based organizational innovation performance, so that this relationship is stronger in the presence of higher social capital.*

Hypothesis 3b. *Social capital moderates the positive relationship between transactional leadership and project-based organizational innovation performance so that this relationship is stronger in the presence of higher social capital.*

Social capital can facilitate access to information and vital sources in order to promote performance [58]. For example, Golmoradi et al. [59] stressed that an organization with powerful social capital could have immediate access to a wide range of information in order to create an innovative performance. Social capital is useful in the process of sharing knowledge through some components such as trust and cooperation, thus it will improve the innovative performance of the organization [60]. Additionally, social capital could also impact organizations' efficiency in different ways, by using knowledge sharing and innovation [61]. Alvani et al. [62] noted that social capital is considered as a value, shared by the people who are involved in social networks due to common cultural norms, effective interactions, mutual trust and personal relationships.

As articulated in the previous section, it is logical to further predict that the heightened innovation performance considers the role of social capital, resulting from the positive path between leadership styles and knowledge sharing, and the positive interactions between knowledge sharing and social capital. Thus, there could be registered some moderated mediation effects [63]. Based on this argument, we expect social capital to moderate the indirect effect of leadership on project-based organizational innovation performance via knowledge sharing.

Hypothesis 4a. *The indirect transformational leadership on project-based organizational innovation performance via knowledge sharing is moderated by social capital, so that the indirect effect is more positive when social capital is high than when it is low.*

Hypothesis 4b. *The indirect transactional leadership on project-based organizational innovation performance via knowledge sharing is moderated by social capital, so that the indirect effect is more positive when social capital is high than when it is low.*

3. Methods

3.1. Sample

The sample for the current study was composed of project managers or technicians joining the construction projects in China. Project managers and technicians can be grouped into the same project team or different project teams. Our data collection investigated the team members from various construction projects. A total of 340 questionnaires are distributed to potential respondents. The questionnaires include the basic information of project and measured items, such as types of construction projects, transformational leadership, transactional leadership, social capital, knowledge sharing, and project-based organizational innovation performance. All 288 managers or technicians completed this final survey. We eliminated 20 participants for improperly completing the questionnaires, e.g., the missing data and the regular answers. Continuing, 28.73% (77) of the final sample worked on the building construction projects, 31.72% (85) were engaged in the highway construction projects, and 21.64% (58) joined in the railway construction projects. The population of the project-based organization was usually aged 20–50 (29.48%). The investment of the construction projects was usually between RMB 100 million yuan and RMB 1 billion yuan (42.91%). The respondents' profile is shown in Table 1.

Table 1. Descriptive statistics of sample frame.

The Basic Information of Construction Projects	Number	Percentage
The project type joined in		
Building construction project	77	28.73
Municipal infrastructure	18	6.72
Highway construction project	85	31.72
Railway construction project	58	21.64
Hydraulic construction project	24	8.96
others	6	2.24
The population of project-based organization		
200+	49	18.28
100–200	52	19.40
50–100	63	23.51
20–50	79	29.48
0–20	25	9.33
The investment (RMB, yuan) of the construction project engaged		
1 billion+	87	32.46
100 million–1 billion	115	42.91
50–100 million	28	10.45
10–50 million	29	10.82
0–10 million	9	3.36
Total	268	100

3.2. Measures

We used five-point Likert scales ranging from 1 = "strongly disagree" to 5 = "strongly agree" to measure the observed variables. All items in the survey were presented in Chinese because our respondents were Chinese. Since the original scales were mostly developed in English, all of the items underwent a back-translation process [64].

3.2.1. Transformational Leadership

The transformational leadership was rated by the members of the construction project-based organizations with a five-item scale developed by Bass et al. [65] and McColl-Kennedy et al. [66], combining with the Chinese context. The Cronbach's alpha was 0.913. A sample item was "The leader could increase my level of enthusiasm". EFA led to a one-factor solution (eigenvalue = 3.722, each factor loading >0.800, explaining 74.444% of the total variance). The CFA result (χ^2/df = 1.793, RMSEA = 0.054, CFI = 0.995, TLI = 0.991) shows that the transformation leadership measure (5 items) fits well in our data.

3.2.2. Transactional Leadership

Transactional leadership was a four-item scale developed by Podsakoff et al. [67]. The Cronbach's alpha was 0.903. A sample item was "The leader always gives me positive feedback when I perform well". The EFA yielded a single-factor solution with each factor loading >0.800, explaining the 77.524%

of the total variance. The CFA result (χ^2/df = 3.691, RMSEA = 0.100, CFI = 0.992, TLI = 0.977) shows that the transactional leadership measure fits well in our data.

3.2.3. Social Capital

Social capital was assessed within a project of four-item scale developed by Chen et al. [68] and Tsai et al. [61]. The scale focuses on the relationship regarding the capital dimension. A representative item was "We support each other when facing change". The Cronbach's alpha was 0.865. EFA led to a one-factor solution (eigenvalue = 2.853, each factor loading >0.700, explaining 71.320% of the total variance). CFA results show that the social capital measure almost fits our data (χ^2/df = 4.536, RMSEA = 0.115, CFI = 0.986, TLI = 0.959).

3.2.4. Knowledge Sharing

Knowledge sharing was assessed using the four-item measure validated by Collins et al. [69]. This scale is focused on the knowledge sharing behavior dimension. Respondents were asked to indicate at work their agreement with statements such as "In the project, we are willing to exchange and combine ideas to find solutions to problems". The Cronbach's alpha for knowledge sharing was 0.827. The EFA yielded a single-factor solution with each factor loading >0.600, explaining 66.099% of the total variance. CFA results indicate that the knowledge sharing measure almost fits our data (χ^2/df = 4.773, RMSEA = 0.119, CFI = 0.982, TLI = 0.946).

3.2.5. Project-Based Organizational Innovation Performance

A five-item measure was used to assess project-based organizational innovation performance, being developed by Gu et al. [70]. One sample item is "The innovativeness has been improved in the implement process of the project". The Cronbach's alpha for the scale was 0.920. EFA led to a one-factor solution (eigenvalue = 3.795, each factor loading >0.800, explaining 75.902% of the total variance). The CFA result (χ^2/df = 3.008, RMSEA = 0.087, CFI = 0.989, TLI = 0.979) indicates that the measure fits well in our data. These results provided construct validity evidence of the project-based organizational innovation performance measure in a Chinese context.

The results of confirmatory factor analysis (CFA) were displayed in Table 2.

Table 2. Results of confirmatory factor analysis.

Construct	CFA Loading
Transformational leadership (α = 0.913; χ^2/df = 1.793, RMSEA = 0.054, CFI = 0.995, TLI = 0.991)	
TLa1 The leader of project-based organization has a clear understanding of where we are going.	0.798
TLa2 The leader of project-based organization transmits a sense of mission to us.	0.818
TLa3 The leader of project-based organization could increase my level of enthusiasm.	0.855
TLa4 The leader of project-based organization shows respect for the personal feelings of others.	0.817
TLa5 The leader of project-based organization encourages us to be team players.	0.837
Transactional leadership (α = 0.903; χ^2/df = 3.691, RMSEA = 0.100, CFI = 0.992, TLI = 0.977)	
TLb1 The leader of project-based organization always gives me positive feedback when I perform well.	0.807
TLb2 The leader of project-based organization gives me special recognition when my work is good.	0.919
TLb3 The leader of project-based organization commends me when I do a better than average job.	0.868
TLb4 The leader of project-based organization frequently acknowledges my good performance.	0.757
Social capital (α = 0.865; χ^2/df = 4.536, RMSEA = 0.115, CFI = 0.986, TLI = 0.959)	
SC1 We build relationship with the organizational individuals in order to exchange idea and information.	0.817
SC2 We support each other when facing change.	0.886
SC3 We seek the support from leaders in resources.	0.739
SC4 We are open to try out new ways of doing things.	0.707
Knowledge sharing (α = 0.827; χ^2/df = 4.773, RMSEA = 0.119, CFI = 0.982, TLI = 0.946)	
KS1 The members of project-based organization are capable of sharing their expertise to bring new initiatives to fruition.	0.541
KS2 We feel that we have learned from each other by sharing information or ideas.	0.754
KS3 We are willing to share information or ideas with the other member of project-based organization.	0.852
KS4 In the project, we are willing to exchange and combine ideas to find solutions to problems.	0.805

Table 2. *Cont.*

Construct	CFA Loading
Project-based organizational innovation performance (α = 0.920; χ^2/df = 3.008, RMSEA = 0.087, CFI = 0.989, TLI = 0.979)	
OIP1 The innovativeness has been improved in the implement process of the project.	0.833
OIP2 The decision making process has been optimized.	0.789
OIP3 The quality of construction project has been improved.	0.883
OIP4 The cost of construction project has been decreased.	0.864
OIP5 The profit of construction project has been increased.	0.812

3.2.6. Control Variables

The industrial characteristics of construction projects, the number of project departments, and the investment of construction projects were controlled in this study. The surveyed respondents engaged in five types of projects (building construction project, municipal infrastructure, highway construction project, railway construction project, and hydraulic construction project), numbered 1–5. We also divided the population of the project department into five levels: 1 (less than 20), 2 (20–50), 3 (50–100), 4 (100–200) and 5 (more than 200). We performed the same action for the investment (RMB, yuan) of the joined projects: 1 (less than ¥10 million), 2 (from ¥10 million to ¥50 million), 3 (from ¥50 million to ¥100 million), 4 (from ¥100 million to ¥1 billion) and 5 (more than ¥1 billion). When testing our hypotheses, we controlled these variables.

3.3. Analytic Strategies

First, the reliability test, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) for each measure are conducted to assess whether their factor structures conform to the anticipated results. We applied the reliability test and exploratory factor analysis (EFA) with the software SPSS 22.0 (Armonk, NY, USA), and discussed these results for each above-mentioned measure.

Second, we conducted a series of CFAs to confirm the distinctiveness of the multi-item variables in the study, including transformational leadership, transactional leadership, social capital, knowledge sharing, and project-based organizational innovation performance. The confirmatory factor analysis (CFA) was assessed using a χ^2 test and multiple practical fit indices (TLI, CFI, RMSEA and SRMR) that allow an evaluation of different aspects of the model fit (absolute fit, comparative fit, and parsimony-adjusted fit) [71]. The CFA was performed using the software MPLUS 7.0 (Los Angeles, CA, USA) to test the measurement model including all variables of our study. The CFA results revealed that the hypothesized five-factor model fits our data in the best way, compared to other alternative models. These results provide evidence for the discriminant validity of our measures.

Third, we applied a linear regression analysis using the software Stata 13.0 (College Station, Texas, USA) to test our hypotheses, including the main effect, as well as the mediation and moderation effects. We took knowledge sharing and project-based organizational innovation performance as dependent variables. Then the main effect models were tested with two leadership styles as independent variables. The mediating model took knowledge sharing as the mediator. The moderating model took social capital as the moderator. All regression models' results were assessed in terms of the significance of the coefficients and R-square. Besides, the indirect effects of moderated mediation were tested according to the approach proposed by Edwards and Lambert [63].

The analysis undertaken after data collection, as illustrated in Figure 2, was used to assess the scales that satisfy the requirements of reliability, validity, and test the hypotheses.

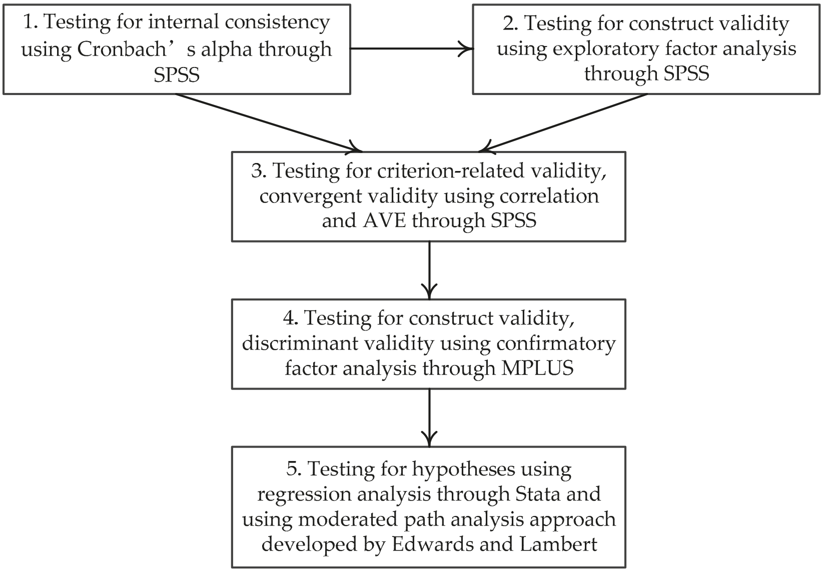

1. Testing for internal consistency using Cronbach’s alpha through SPSS
2. Testing for construct validity using exploratory factor analysis through SPSS
3. Testing for criterion-related validity, convergent validity using correlation and AVE through SPSS
4. Testing for construct validity, discriminant validity using confirmatory factor analysis through MPLUS
5. Testing for hypotheses using regression analysis through Stata and using moderated path analysis approach developed by Edwards and Lambert

Figure 2. The analysis process.

4. Results

Table 3 shows the descriptive statistics and correlation matrix. Compared to the square root of AVE and correlations, the results provided a convergent validity of the study variables.

Table 3. Descriptive statistics and correlations.

Variables	Mean	SD	nature	number	invest	TLa	TLb	SC	KS	OIP
1. nature	2.820	1.403	(–)							
2. number	2.920	1.262	−0.425 **	(–)						
3. invest	2.100	1.077	−0.351 **	0.504 **	(–)					
4. TLa	2.637	0.799	0.068	−0.025	0.026	(0.825)				
5. TLb	2.676	0.911	0.093	0.080	0.127 *	0.437 **	(0.840)			
6. SC	2.592	0.833	−0.014	0.131 *	0.164 **	0.365 **	0.337 **	(0.791)		
7. KS	2.504	0.711	−0.116	0.158 **	0.132 *	0.413 **	0.424 **	0.409 **	(0.748)	
8. OIP	2.696	0.818	0.035	0.055	0.116	0.568 **	0.556 **	0.456 **	0.544 **	(0.837)

Note: Sample size = 268; the square root of average variance extraction are in parentheses on the diagonal. SD, standard deviation; Nature, the industry nature of projects; Number, the number of employees in the project department; Invest, the investment of projects; TLa, transformational leadership; TLb, transactional leadership; SC, social capital; KS, knowledge sharing; OIP, project-based organizational innovation performance.* $p < 0.05$; ** $p < 0.01$. Two-tailed tests.

4.1. Assessment of the Measurement Model

As shown in Table 4, the CFA indicated that the expected model with five factors—transformational leadership, transactional leadership, social capital, knowledge sharing, and project-based organizational innovation performance (Model 1)—demonstrated an excellent fit. All other alternative models (Models 2–4) demonstrated a rather poor fit compared to the hypothesized five-factor model. This is shown by the increased values of χ^2, decreased values of TLI and CFI, and increased values of RMSEA and SRMR.

Table 4. A conducted to examine factor structure of the scales used in the study.

Model	Description	χ^2	*df*	RMSEA [90% CI]	TLI	CFI	SRMR
Model 1	5 factors: TLa, TLb, SC, KS, OIP	640.729 ***	199	0.091 [0.083; 0.099]	0.898	0.912	0.041
Model 2	3 factors: leader-rated variables (combined: TLa and TLb), social SC with KS, OIP	1117.998 ***	206	0.129 [0.121; 0.136]	0.796	0.818	0.068
Model 3	2 factors: leader-rated variables (combined: TLa and TLb), and OIP combined SC with KS	1335.677 ***	208	0.142 [0.135; 0.150]	0.751	0.775	0.076
Model 4	1 factor: all items loading on the same factor	1352.570 ***	209	0.143 [0.136; 0.150]	0.748	0.772	0.077

Note: Sample size = 268; TLa, transformational leadership; TLb, transactional leadership; SC, social capital. KS, knowledge sharing; OIP, project-based organizational innovation performance. Model fit was assessed using the recommended cut-offs: 0.90 for TLI [72] and CFI [73]; 0.05 as an indicator of good fit and 0.10 as the upper limit of acceptable fit for RMSEA [74], and 0.08 as an indicator of good fit for SRMR [73]. *** $p < 0.001$.

4.2. Tests of the Research Hypotheses

Table 5 provides the results of the tests performed on the research hypotheses. We tested the hypotheses in three steps: (1) $Mediator = X_1 + X_2$ (see Model 1 with knowledge sharing as an outcome in Table 5), *Mediator* means the mediating variable; (2) $Y = X_1 + X_2 + Mediator$ (the main effect model and the indirect effect model, respectively; see Model 2 and Model 3 with project-based organizational innovation process as an outcome in Table 5); and (3) $Y = X_1 + X_2 + Mediator + Moderator + X_1 \times Moderator + X_2 \times Moderator + Mediator \times Moderator$ (the interactive model, see Model 6 in Table 5), *Moderator* means the moderating variable. The model observed in the first step provided evidence regarding the main effects of two leadership styles on project-based organizational innovation performance (Hypothesis 1). The indirect effect model observed at Step 2 provided evidence regarding the mediating effect of knowledge sharing on the link between leadership and project-based organizational innovation performance (Hypothesis 2). Finally, models observed for all three steps provided evidence of the interactive effect of leadership and knowledge sharing on project-based organizational innovation performance via social capital (Hypothesis 3).

4.2.1. Main Effect of Transformational Leadership and Transactional Leadership

As shown by the results registered for Model 2 in Table 5, the main impact of transformational and transactional leadership on project-based organizational innovation performance was positive and statistically significant ($\beta_{TLa} = 0.379$, $p < 0.001$; $\beta_{TLb} = 0.621$, $p < 0.001$). This yields support for Hypotheses 1a and 1b.

4.2.2. Indirect Effects of Transformational Leadership and Transactional Leadership on Project-Based Organizational Innovation Performance via Knowledge Sharing

To test Hypothesis 2 by predicting the mediation effect of knowledge sharing between two leadership styles and project-based organizational innovation performance, we have estimated this indirect effect as a product of three paths [75]: the direct impact of the independent variable, the direct path from the independent variable to the mediator, and path regarding the impact of the mediator. The estimation of knowledge sharing as the mediator was positive and statistically significant ($\beta_{KS} = 0.103$, $p < 0.01$), as shown by the results of Model 3 in Table 5.

Table 5. Regression analysis results: Main, mediation and moderation effects.

Variables	Knowledge Sharing (Mediator)	Project-Based Organizational Innovation Performance (Dependent Variable)								
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**	**Model 9**	**Model 10**
Intercept	1.234 ***	0.286 *	0.140	0.263	0.363 *	0.234	−0.271	0.280	0.817 **	0.345
Control variables										
Nature	−0.157 **	−0.048	−0.032	0.063	−0.032	−0.034	0.020	−0.049	0.018	−0.031
Number	−0.070	0.015	0.022	−0.005	0.054	0.021	0.008	0.045	0.002	0.054
Invest	0.028	−0.018	−0.021	−0.076 +	−0.013	−0.020	−0.036	−0.005	−0.032	−0.013
Independent variables										
TLa	0.207 ***	0.379 ***	0.357 ***	0.588 ***		0.627 ***	0.687 ***		0.492 ***	
TLb	0.506 ***	0.621 ***	0.569 ***		0.750 ***	0.442 ***		0.875 ***		0.735 ***
Mediator										
KS			0.103 **	0.339 ***	0.179 ***	−0.137			−0.161 +	0.176 +
Moderator										
SC						0.035	0.681 ***	0.252 *	0.116	0.028
Interactions										
TLa × SC						−0.472 **	−0.361 **			
TLb × SC						0.101		−0.238		
LS × SC						0.451 *			0.518 **	−0.008
R^2	0.448	0.822	0.828	0.681	0.755	0.836	0.750	0.743	0.759	0.755

Note: Sample size = 268; Standardized beta coefficients and unstandardized intercept value are reported. Nature, the industry nature of projects; Number, the number of employees in the project department; Invest, the investment of projects; TLa, transformational leadership; TLb, transactional leadership; KS, knowledge sharing; SC, social capital. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.1$.

A partial mediation effect was also registered. More specifically, the impact of the transformational and transactional leadership as independent variables on knowledge sharing as the mediator was positively significant ($\beta_{\mathrm{TLa}} = 0.207$, $p < 0.001$; $\beta_{\mathrm{TLb}} = 0.506$, $p < 0.001$), reflected by the results of Model 1. In Model 2, there was observed a significant relationship between the independent variables and dependent variable ($\beta_{\mathrm{TLa}} = 0.379$, $p < 0.001$; $\beta_{\mathrm{TLb}} = 0.621$, $p < 0.001$). After entering the mediator knowledge sharing into the regression equation, in which innovation performance was regressed on two leadership styles, the beta values of transformational leadership and transactional leadership significantly decreased ($\beta_{\mathrm{TLa}} = 0.357$, $p < 0.001$; $\beta_{\mathrm{TLb}} = 0.569$, $p < 0.001$), as seen in Model 3. The results show that knowledge sharing plays a partial mediator role in the impact of transformational leadership and transactional leadership on project-based organizational innovation performance. This set of analyses supports Hypotheses 2a and 2b.

Meanwhile, as shown by the results obtained for Models 4 and 5 in Table 5, there were registered significant relationships between transformational leadership or transactional leadership and project-based organizational innovation performance via knowledge sharing ($\beta_{\mathrm{TLa}} = 0.588$, $p < 0.001$ and $\beta_{\mathrm{KS}} = 0.339$, $p < 0.001$; $\beta_{\mathrm{TLb}} = 0.750$, $p < 0.001$ and $\beta_{\mathrm{KS}} = 0.179$, $p < 0.001$). This provides additional support for the mediating Hypotheses 2a and 2b.

4.2.3. Moderating the Direct Effects of Social Capital

On the one hand, Hypothesis 3a stated that social capital amplifies the relationship between transformational leadership and project-based organizational innovation performance. The results obtained for Model 6 in Table 5 show that the interaction between transformational leadership and social capital was significant, though negatively related to the project-based organizational innovation performance ($\beta_{\mathrm{TLa}\times\mathrm{SC}} = -0.472$, $p < 0.01$). In addition, as shown in Model 7, after controlling the effects of transactional leadership, the interaction of transformational leadership and social capital was also statistically significant but negative ($\beta_{\mathrm{TLa}\times\mathrm{SC}} = -0.361$, $p < 0.01$). Therefore, Hypothesis 3a was not supported.

On the other hand, Hypothesis 3b proposed that social capital augments the relationship between transactional leadership and project-based organizational innovation performance. The results of Model 6 in Table 5 show that the interaction of social capital with transactional leadership was not statistically significant ($\beta_{\mathrm{TLb}\times\mathrm{SC}} = 0.101$, $p > 0.05$). In addition, as shown in Model 8, after controlling the effects of transformational leadership, the interaction between transactional leadership and social capital was not statistically significant ($\beta_{\mathrm{TLb}\times\mathrm{SC}}= -0.238$, $p > 0.05$). This set of analyses does not yield support for Hypothesis 3b.

Further, to demonstrate the interaction, we computed a series of estimates of social capital at low (mean – 1 SD) and high (mean + 1 SD) values of the moderator [76,77], and we conducted the interaction plots. Figure 3 provides two interaction plots with simple slopes for social capital at one standard deviation below the mean and social capital at one standard deviation above the mean. (i) The results of this additional analysis suggest that transformational leadership and social capital interact to negatively impact the project-based organizational innovation performance, which contradicts Hypothesis 3a. (ii) The results indicate that transactional leadership and social capital interact to not affect the project-based organizational innovation performance, which does not support Hypothesis 3b.

Besides, we estimated the indirect effect through the moderated path analysis approach using a Bootstrap technique [63], as can be seen in Table 6. The results reported in Table 6, reveal that the estimation of the direct impact of transformational leadership on project-based organizational innovation performance at low social capital was positive and statistically significant ($\beta = 0.237$, 95% CI [0.091; 0.355], not containing zero). However, the estimation at high social capital was positive but non-significant ($\beta = 0.014$, 95% CI [-0.239; 0.283], containing zero). This set of analyses is consistent with the plot in Figure 3a, suggesting that the moderating effect of social capital becomes noticeable at lower levels. However, the direct moderation effects were negative and significant because the paths from transformational leadership to project-based organizational innovation performance differed significantly across different levels of social capital (Δβ = −0.223, 95% CI [−0.333; −0.050], not containing zero). This result coincides with the regression analysis of Model 7 in Table 5, contradicting Hypothesis 3a.

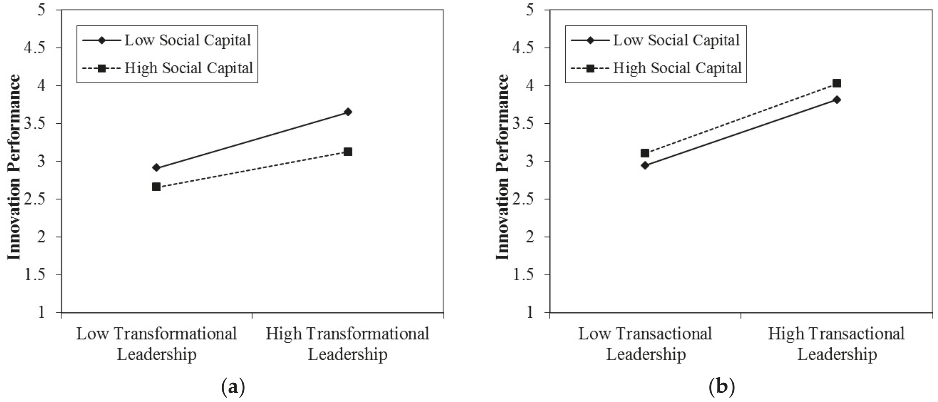


Figure 3. The interactive effects of: (**a**) social capital and transformational leadership on project-based organizational innovation performance; and (**b**) social capital and transactional leadership on project-based organizational innovation performance.

Table 6. Results of the moderated path analysis.

Moderator Variable: Social Capital	**Transformational Leadership (X_1)→Knowledge Sharing (M)→Project-Based Organizational Innovation Performance (Y)**		
	Second Stage [95% CI]	**Direct Effect [95% CI]**	**Indirect Effect [95% CI]**
Low Social Capital (−1 SD)	0.352 [0.208; 0.502]	0.237 [0.091; 0.355]	0.179 [0.116; 0.264]
High Social Capital (+1 SD)	0.603 [0.300; 0.854]	0.014 [−0.239; 0.283]	0.308 [0.201; 0.441]
Differences	0.252 [0.102; 0.383]	−0.223 [−0.333; −0.050]	0.128 [0.075; 0.183]
Moderator Variable: Social Capital	**Transactional Leadership (X_2)→Knowledge Sharing (M)→Project-Based Organizational Innovation Performance (Y)**		
	Second Stage [95% CI]	**Direct Effect [95% CI]**	**Indirect Effect [95% CI]**
Low Social Capital (−1 SD)	0.282 [0.112; 0.447]	0.526 [0.341; 0.695]	0.175 [0.077; 0.294]
High Social Capital (+1 SD)	0.379 [0.043; 0.688]	0.374 [0.012; 0.667]	0.235 [0.391; 0.806]
Differences	0.097 [−0.100; 0.250]	−0.152 [−0.317; 0.033]	0.060 [−0.056; 0.163]

Note: Low moderator variable refers to one standard deviation below the mean of the moderator, while high moderator variable refers to one standard deviation above the mean of the moderator; 95% CI means 95% confidence interval; it does not contain zero, it means it is significant and vice versa; bootstrap equals 1000; and these were performed using bias corrected percentile method (BC).

In terms of Hypothesis 3b, as shown in Table 6, the direct effect of transactional leadership on project-based organizational innovation performance via social capital was significant (β = 0.526, 95% CI [0.341; 0.695], not containing zero) when social capital was low and statistically significant (β = 0.374, 95% CI [0.012; 0.667], not containing zero) when social capital was high. However, the difference in the direct effect of the social capital was not significant (Δβ = −0.152, 95% CI [−0.317; 0.033], containing zero), providing additional support to the results registered for Model 8 in Table 5. Thus, Hypothesis 3b was not supported. We plotted the moderated direct effects in Figure 3b.

4.2.4. Moderated Indirect Effects by Social Capital

Hypothesis 4 predicted that social capital moderated the indirect effect of transformational leadership/transactional leadership on project-based organizational innovation performance via knowledge sharing. As expected, the results of Model 6 in Table 5 indicated that social capital had a positive moderating impact ($\beta_{KS \times SC} = 0.451$, $p < 0.05$).

More specifically, the moderated path analysis results are registered in Table 6. The second-stage moderation effect, which reveals that the path from transformational leadership to project-based organizational innovation performance differs significantly (Δβ = 0.252, 95% CI [0.102; 0.383], not containing zero). However, we do not register a statistically significant result in terms of the path from transactional leadership to project-based organizational innovation performance across different types of social capital (Δβ = 0.097, 95% CI [−0.100; 0.250], containing zero).

Meanwhile, the indirect effect of transformational leadership on project-based organizational innovation performance via knowledge sharing was statistically significant (β = 0.179, 95% CI [0.116; 0.264], not containing zero) when social capital was low and significant (β = 0.308, 95% CI [0.201; 0.441], not containing zero) when social capital was high. Overall, the difference in the indirect effect of social capital was statistically significant (Δβ = 0.128, 95% CI [0.075; 0.183], not containing zero). Thus, Hypothesis 4a was supported. The results in Table 6 also indicate that the indirect effect of transactional leadership on project-based organizational innovation performance via knowledge sharing was significant (β = 0.235, 95% CI [0.391; 0.806], not containing zero) under high social capital, whereas it was statistically significant (β = 0.175, 95% CI [0.077; 0.294], not containing zero) under low social capital. Overall, the difference in the indirect effects by social capital was not statistically significant (Δβ = 0.060, 95% CI [−0.056; 0.163], containing zero). Hence, Hypothesis 4b was not supported.

Overall, our results provided evidence for the second-stage moderation and moderated the indirect effects by social capital from transformational leadership to project-based organizational innovation performance via knowledge sharing.

To summarize, the results of the hypotheses testing are shown in Table 7.

Table 7. Results of the hypothesized model.

Hypotheses	Path Analysis Results	Moderated Path Analysis Results	Test Results
H1 The main effect of two leadership styles on project-based organizational innovation performance			
H1a TLa→OIP	0.207 ***		supported
H1b TLb→OIP	0.506 ***		supported
H2 The mediating effect of knowledge sharing between two leadership styles on project-based organizational innovation performance			
H2a TLa→KS→OIP	0.339 ***		supported
H2b TLb→KS→OIP	0.179 ***		supported
H3 The moderated effect of social capital on the path from leadership to project-based organizational innovation			
H3a Tla × SC→OIP	−0.361 **	−0.223 [−0.333; −0.050]	unsupported
H3b TLb × SC→OIP	−0.238	−0.152 [−0.317; 0.033]	unsupported
H4 The moderated mediation effect of social capital on the path from leadership to project-based organizational innovation performance via knowledge sharing			
H4a TLa→KS × SC→OIP	0.518 **	0.128 [0.075; 0.183]	supported
H4b TLb→KS × SC→OIP	−0.008	0.060 [−0.056; 0.163]	unsupported

Note: TLa, transformational leadership; TLb, transactional leadership; KS, knowledge sharing; SC, social capital; OIP, project-based organizational innovation performance; *** $p < 0.001$, ** $p < 0.01$; and 95% CI means 95% confidence interval.

5. Discussion

The main objective of this study was to explore how different leadership styles exchange knowledge, coordinate resources and improve innovation in a temporary project-based organization.

Based on the social capital and knowledge exchange perspective, we proposed knowledge sharing as the mediating mechanism and social capital as the boundary condition for the impact that transformational leadership and transactional leadership on project-based organizational innovation performance.

5.1. The Differentiation of the Two Leadership Styles

For construction projects, a leadership style that encourages participation and ideas from followers is expected to determine a higher efficiency in facilitating the project's success [78]. Because knowledge sharing is an enabler of innovation, the leadership style that promotes an open communication of innovative ideas and approaches is essential for organizational innovation [20]. Transformational leaders that nurture the intrinsic development of employees by considering their participation and ideas are essential for an organization [9]. With transformational leadership, innovation is inspired and stimulated among followers.

Compared to transformational leadership, transactional leadership also motivates followers with extrinsic rewards. The construction industry is composed of project-based organizations in which project members who come from different units cooperate as a team in their actions. Moreover, the construction projects are completed within a limited timeframe and resources, thus leaders may find it difficult to apply transformational leadership to nurture the intrinsic long-term needs of employees. Instead, for leaders to execute a less participating leadership behavior may be more effective. This study also found that the transactional leadership style adopting contingent rewards are positively associated with the innovation performance in the construction field, this result being also coherent with previous findings [79]. Meanwhile, the impact of transactional leadership is more than that of transformational leadership style, indicating that, under the context of the limited resources and time, the reward measures could motivate the followers to strive for the short-term goals. Considering innovation as a multistage process, transformational leaders may intervene at a later date in the project's management.

5.2. The Mediating Role of Knowledge Sharing

Implementing the knowledge sharing activity is essential for project organizations to facilitate innovation. In this direction, transformational or transactional project managers could develop and combine the knowledge, skills, and capabilities of project members by transferring the knowledge to others and applying that knowledge for completing projects and creating new ideas for organizational or technology innovation. Scholars have provided support for this theory that leadership is positively related to the knowledge management process [44,80]. Besides, results advocated that leaders improve the innovation performance via knowledge sharing because the R square of Model 3 is higher than that of Model 1 in Table 5. Thus, the way in which knowledge sharing occurs help project managers to improve the organizational performance. The internal interaction of knowledge and the external acquisition of knowledge also benefits for stimulating innovation to achieve the sustainability of construction projects and construction innovation [81]. Transformational and transactional leadership in the project organization are the sources of improvement in innovation performance either directly or indirectly through knowledge sharing. These findings suggest that leadership is an important predecessor of knowledge sharing. Thus, transformational or transactional project managers can encourage knowledge sharing in the management process.

5.3. The Moderating Effect of Social Capital

First, this study extends the leadership examinations regarding capital. Although transformational leadership is known to positively impact innovation by transmitting through social capital [57], few researchers have tested different leadership styles and their impact within the different levels of social capital. In the project-based organization composed of many stakeholders, our findings support social capital as a significant predecessor in the improvement of innovation performance in the construction projects. Nevertheless, the effects of leadership combined with social capital are not as expected. Social capital did not help project managers or leaders to improve innovation performance. More specifically, the transformational leaders' expert a negative impact on innovation performance across different levels of social capital, while the transactional leaders did not vary significantly across different levels of social capital. Besides, our findings also indicate a positive relationship between projects' social capital and innovation performance. Thus, the empirical results show the negative effect or non-significant effect of leadership on innovation performance, demonstrating that there may be other components capable of explaining this mechanism.

Second, the current study extends the leadership and innovation literature by demonstrating a boundary condition regarding the impact of leadership on innovation performance, namely social capital. Our findings indicate that when there is a high social capital relationship in the project organization, leaders are more likely to improve innovation performance through knowledge sharing. This finding suggests that the role of social capital in construction projects is far more complicated than previously considered [27]. Thus, social capital may also serve as a contextual factor facilitating knowledge sharing. In the temporary project-based organization, people work together on complex innovative tasks for a well-defined limited period. It is likely that projects will become highly embedded in a set of project-specific relationships. This determines transformational leaders to promote information generation and exchange within the project-based organizations, where employees are likely to engage in social activities [82]. In addition, there is no significant impact of social capital on the path from transactional leadership to innovation performance via knowledge sharing. Project leaders may provide rewards to employees to share knowledge and enhance performance by implementing some part of the social capital. However, such a condition may be minimal in comparison to the contribution of transformational leaders on innovation performance via knowledge sharing in the project-based context. Thus, by considering the condition of social capital, the project managers promoting a transformational leader behavior have a high potential to contribute to knowledge sharing and innovation performance compared to the effects determined by transactional leaders.

6. Conclusions and Implications

6.1. Conclusions

Extending towards research leadership and innovation, the present study highlights the role of knowledge sharing and social capital in facilitating innovation performance. Moreover, the construction industry has long since been recognized as an industry with low innovation. To improve sustainability and promote innovation in the construction projects, leadership, knowledge management process and social capital should be enhanced. First, the results of this study reveal that transformation leadership and transaction leadership are positively associated with innovation performance. The results show the need for academics and practitioners in the construction industry to play a greater importance on leadership. Second, to further enhance innovation in the project-based organizations, knowledge sharing should be promoted during the different levels of social capital, as knowledge sharing was found to have a significant impact on innovation. Additionally, the study empirically shows that knowledge sharing partially mediates the relationship between transformational leadership or transactional leadership and innovation performance. Third, the present study highlights that it is transformational project managers, who make a significant contribution to innovation performance through increasing social capital level in the project-based organizations beyond the contributions of transactional leadership on exchanging and sharing knowledge process.

6.2. Practical Implications

The results of this study generate several valuable insights with interesting theoretical implications. Compared to the literature, this study developed a conceptual model by integrating a series of concepts, namely transformational leadership, transactional leadership, knowledge sharing, social capital and project-based organizational innovation performance. Although a number of investigations indicate the positive effects of different leadership styles on organizational performance, our study built the conceptual model specifying the mediating role of knowledge sharing and the moderating role of social capital. In addition, this study empirically tested the conceptual model and proved five hypotheses. More precisely, transformational leadership and transactional leadership have a direct impact on project-based organizational innovation performance, in the same time observing the indirect impacts on performance through implementing knowledge sharing. Thus, this study provides strong evidence in showing the interrelationships between two leadership styles and knowledge sharing and between two leadership styles and project-based organizational innovation performance. In addition, this study tested the moderated mediation effect of social capital, examining the effects of transformational leadership on innovation performance via knowledge sharing across different levels of social capital. As such, this study emphasized the importance of developing two leadership styles to improve project-based organizational innovation performance through implementing knowledge sharing and building social capital relationships.

Our findings also have important managerial implications. It may be helpful for project leaders or individuals who manage teams in project-based organization setting, providing important insights on the management of inter- or intra-organizational knowledge exchange and network establishment. Our results suggest that through an appropriate leadership style, project-based organization can increase their coordination and knowledge sharing with the management of social capital, thus producing and improving the high levels of innovation performance. The uncovered moderating effect of social capital on the indirect relationship between transformational leadership and project-based organizational innovation performance, leads us to the conclusion that construction project-based organizations could strengthen the relationships among project managers and other members of the organization by launching different interventions. The intervention could be implemented in such a way that it amplifies the number of connections, providing mutual support between managers and other organization members.

6.3. Limitations and Future Research

This study has several limitations. First, in this field of study, the surveys used to measure transformational leadership, transactional leadership, knowledge sharing, social capital, and innovation performance were collected concomitantly, although the data were collected from employees engaged in different construction projects, and the same potential source bias was controlled so it did not represent a severe problem. However, in this case, it was extremely difficult to prove causality between different concepts. Therefore, future research should focus on collecting longitudinal data to test the relationships among leadership styles, knowledge management, social capital and innovation performance.

Secondly, the study was performed in China, a market dominated by emerging projects where leadership is essential for accessing external and internal resources. However, the concepts of transformational leadership and transactional leadership come from Western countries, which may not be appropriate in the Chinese construction projects context. Thus, there could be a differential impact between the two leadership styles, while the transformational leadership exerts a weak positive impact compared with transactional leadership in construction project-based setting. Future research could extend the sample to other industries in China or replicate the study to other countries in order to validate the results.

Acknowledgments: This study is supported by the National Natural Science Foundation of China (71701083, 71561009 and 71761021), China Postdoctoral Science Foundation (2016M590605 and 2017T100477), Postdoctoral Science Foundation of Jiangxi Province (2016KY27), Social Science Planning Foundation of Jiangxi Province (16GL32), Natural Science Foundation of Jiangxi Province (S2017QNJJB0493), and Talent Training Foundation of Yunnan Province (KKSY201606041).

Author Contributions: Junwei Zheng designed this study and completed the paper in English; Guangdong Wu participated in drafting the paper and revised it critically for important intellectual content; and Hongtao Xie gave many good research advices and revised the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wen, Q.; Qiang, M. Coordination and knowledge sharing in construction project-based organization: A longitudinal structural equation model analysis. *Autom. Constr.* **2016**, *72*, 309–320. [CrossRef]
2. Blayse, A.M.; Manley, K. Key influences on construction innovation. *Constr. Innov.* **2004**, *4*, 143–154. [CrossRef]
3. Seaden, G.; Manseau, A. Public policy and construction innovation. *Build. Res. Inf.* **2001**, *29*, 182–196. [CrossRef]
4. Reichstein, T.; Salter, A.J.; Gann, D.M. Last among equals: A comparison of innovation in construction, services and manufacturing in the UK. *Constr. Manag. Econ.* **2005**, *23*, 631–644. [CrossRef]
5. Pries, F.; Janszen, F. Innovation in the construction industry: The dominant role of the environment. *Constr. Manag. Econ.* **1995**, *13*, 43–51. [CrossRef]
6. Buvik, M.P.; Rolfsen, M. Prior ties and trust development in project teams—A case study from the construction industry. *Int. J. Proj. Manag.* **2015**, *33*, 1484–1494. [CrossRef]
7. Zhang, P.; Ng, F.F. Explaining knowledge-sharing intention in construction teams in Hong Kong. *J. Constr. Eng. Manag.* **2013**, *139*, 280–293. [CrossRef]
8. Havenvid, M.I. Competition versus interaction as a way to promote innovation in the construction industry. *IMP J.* **2015**, *9*, 46–63. [CrossRef]
9. Chan, I.Y.S.; Liu, A.M.M.; Fellows, R. Role of leadership in fostering an innovation climate in constrction firms. *J. Manag. Eng.* **2014**, *30*, 06014003. [CrossRef]
10. The Organization for Economic Co-operation and Development (OECD). *Oslo manual: Guidelines for Collecting and Interpreting Technological Innovation Data*; OECD: Paris, France, 2005.
11. Manley, K.; McFallan, S.; Kajewski, S. Relationship between construction firm strategies and innovation outcomes. *J. Constr. Eng. Manag.* **2009**, *135*, 764–771. [CrossRef]
12. Park, M.; Nepal, M.P.; Dulaimi, M.F. Dynamic modeling for construction innovation. *J. Manag. Eng.* **2004**, *20*, 170–177. [CrossRef]
13. Naranjo-Gil, D. The influence of environmental and organizational factors on innovation adoptions: Consequences for performance in public sector organizations. *Technovation* **2009**, *29*, 810–818. [CrossRef]
14. Bates, R.; Khasawneh, S. Organizational learning culture, learning transfer climate and perceived innovation in Jordanian organizations. *Int. J. Train. Dev.* **2005**, *9*, 96–109. [CrossRef]
15. Widén, K.; Olander, S.; Atkin, B. Links between Successful Innovation Diffusion and Stakeholder Engagement. *J. Manag. Eng.* **2014**, *30*, 04014018. [CrossRef]
16. Carlfjord, S.; Lindberg, M.; Bendtsen, P.; Nilsen, P.; Andersion, A. Key factors influencing adoption of an innovation in primary health care: A qualitative study based on implementation theory. *BMC Fam. Pract.* **2010**, *11*, 60. [CrossRef] [PubMed]
17. Amabile, T.M.; Schatzel, E.A.; Moneta, G.B.; Karmer, S.J. Leader behaviors and the work environment for creativity: Perceived leader support. *Leadersh. Q.* **2004**, *15*, 5–32. [CrossRef]
18. Peterson, H.C. Tranformational suply chanis and the 'wicked problem' of sustainability: Aligning knowledge, innovation, entrepreneurship, and leadership. *J. Chain Netw. Sci.* **2009**, *9*, 71–82. [CrossRef]
19. Ozorhon, B. Analysis of Construction Innovation Process at Project Level. *J. Manag. Eng.* **2013**, *29*, 455–463. [CrossRef]
20. Ozorhon, B.; Abbott, C.; Aouad, G. Integration and leadership as enablers of innovation in construction: Case study. *J. Manag. Eng.* **2014**, *30*, 256–263. [CrossRef]
21. Wu, G. Knowledge collaborative incentive based on inter-organizational cooperative innovation of project-based supply chain. *J. Ind. Eng. Manag.* **2013**, *6*, 1065–1081. [CrossRef]
22. Brook, J.W.; Pagnanelli, F. Integrating sustainability into innovation project portfolio management—A strategic perspective. *J. Eng. Technol. Manag.* **2014**, *34*, 46–62. [CrossRef]

23. Reich, B.H.; Gemino, A.; Sauer, C. Knowledge management and project-based knowledge in it projects: A model and preliminary empirical results. *Int. J. Proj. Manag.* **2012**, *30*, 663–674. [CrossRef]
24. Javerick-Will, A. Motivating knowledge sharing in engineering and construction organizations: Power of social motiations. *J. Manag. Eng.* **2011**, *28*, 193–202. [CrossRef]
25. Benavides-Espinosa, M.D.M.; Ribeiro-Soriano, D. Cooperative learning in creating and managing joint ventures. *J. Bus. Res.* **2014**, *67*, 648–655. [CrossRef]
26. Zhang, L.; He, J.; Zhou, S. Sharing tacit knowledge for integrated project team flexibility: Case study of integrated project delivery. *J. Constr. Eng. Manag.* **2012**, *139*, 795–804. [CrossRef]
27. Di Vincenzo, F.; Mascia, D. Social capital in project-based organizations: Its role, structure, and impact on project performance. *Int. J. Proj. Manag.* **2012**, *30*, 5–14. [CrossRef]
28. Ilies, R.; Judge, T.; Wagner, D. Making sense of motivaitonal leadership: The trail from transformational leaders to motivated followers. *J. Leadersh. Organ. Stud.* **2006**, *13*, 1–22. [CrossRef]
29. Avolio, B.J.; Bass, B.M. *Multifactor Leadership Questionnaire, Manual and Sampler Set*, 3rd ed.; Mind Garden Inc.: Menlo Park, CA, USA, 2004.
30. Bass, B.M. *Leadership and Performance beyond Expectations*; Free Press: New York, NY, USA, 1985.
31. Bygballe, L.E.; Ingemansson, M. The logic of innovation in construction. *Ind. Mark. Manag.* **2014**, *43*, 512–524. [CrossRef]
32. Jiang, W.; Zhao, X.; Ni, J. The impact of transformational leadership on employee sustainable performance: The mediating role of organizational citizenship behavior. *Sustainability* **2017**, *9*, 1567. [CrossRef]
33. Bass, B.M.; Riggio, R.E. *Transformational Leadership*; Lawrence Erlbaum Associates: London, UK, 2006.
34. Uddin, M.A.; Fan, L.; Das, A.K. A study of the impact of transformational leadership, organizational learning, and knowledge management on organizational innovation. *Manag. Dyn.* **2017**, *16*, 42–54.
35. Tabassi, A.A.; Roufechasei, K.M.; Ramli, M.; Bakar, A.H.A.; Ismail, R.; Pakir, A.H.K. Leadership competences of sustainable construction project managers. *J. Clean. Prod.* **2016**, *124*, 339–349. [CrossRef]
36. Odusami, K.T.; Iyagba, R.R.O.; Omirin, M.M. The relationship between project leadership, team composition and construction project performance in Nigeria. *Int. J. Proj. Manag.* **2003**, *21*, 519–527. [CrossRef]
37. Chang, Y.-Y. Multilevel transformational leadership and management innovation: Intermediate linkage evidence. *Leadersh. Organ. Dev. J.* **2016**, *37*, 265–288. [CrossRef]
38. Tajasom, A.; Hung, D.K.M.; Nikbin, D.; Hyun, S.S. The role of transformational leadership leadership in innovation performance of Malaysian SMEs. *Asian J. Technol. Innov.* **2015**, *23*, 172–188. [CrossRef]
39. Shin, S.J.; Zhou, J. Transformational leadership, conservaiton, and creativity: Evidence from Korea. *Acad. Manag. J.* **2003**, *46*, 703–714.
40. Liu, A.M.M.; Chan, I.Y.S. Understanding the interplay of organizational climate and leadership in construction innovation. *J. Manag. Eng.* **2017**, *33*, 04017021. [CrossRef]
41. Overall, J. A conceptual framework of innovation and performance: The importance of leadership, relationship quality, and knowledge management. *Acad. Entrep. J.* **2015**, *21*, 41–54.
42. Mas-Tur, A.; Ribeiro-Soriano, D. The level of innovation among young innovative companies: The impacts of knowledge-intensive services use, firm characteristics and the entrepreneur attributes. *Serv. Bus.* **2014**, *8*, 51–63. [CrossRef]
43. Mas-Verdu, F.; Ribeiro-Soriano, D.; Roig Dobon, S. Regional development and innovation: The role of services. *Serv. Ind. J.* **2010**, *30*, 633–641. [CrossRef]
44. Birasnav, M. Knowledge management and organizational performance in the service industry: The role of transformational leadership beyond the effects of transactional leadership. *J. Bus. Res.* **2014**, *67*, 1622–1629. [CrossRef]
45. Singh, S.K. Role of leadership in knowledge management: A study. *J. Knowl. Manag.* **2008**, *12*, 3–15. [CrossRef]
46. Han, S.H.; Seo, G.; Yoon, S.W.; Yoon, D.-Y. Transformational leadership and knowledge sharing. *J. Work. Learn.* **2016**, *28*, 130–149. [CrossRef]
47. Bryant, S.E. The role of transformational and transactional leadership in creating, sharing and exploiting organizational knowledge. *Leadersh. Organ. Stud.* **2003**, *9*, 32–44. [CrossRef]
48. Suparak, S.; Avery, G. Sustainable leadership practices driving financial performance: Empirical evidence from Thai SMEs. *Sustainability* **2016**, *8*, 327.

49. Noruzy, A.; Dalfard, V.M.; Azhdari, B.; Nazari-Shirkouhi, S.; Rezazadeh, A. Relations between transformational leadership, organizational learning, knowledge management, organizational innovaion, and organizational performance: An empirical investigation of manufacturing firms. *Int. J. Adv. Manuf. Technol.* **2013**, *64*, 1073–1085. [CrossRef]
50. Wang, Y.; Ho, C. No money? No problem! The value of sustainability: Social capital drives the relationship among customer identification and citizenship behavior in sharing economy. *Sustainability* **2017**, *9*, 1400. [CrossRef]
51. Solé Parellada, F.; Ribeiro-Soriano, D.; Huarng, K.-H. An overview of the service industries' future (priorities: Linking past and future). *Serv. Ind. J.* **2011**, *31*, 1–6. [CrossRef]
52. Sánchez, A.A.; Marín, G.S.; Morales, A.M. The mediating effect of strategic human resource practices on knowledge management and firm performance. *Rev. Eur. Dir. Econ. Empres.* **2015**, *24*, 138–148. [CrossRef]
53. Birasnav, M.; Rangnekar, S.; Dalpati, A. Transformational leadership and human capital benefits: The role of knowledge management. *Leadersh. Organ. Dev. J.* **2011**, *32*, 106–126. [CrossRef]
54. Peachey, J.W.; Burton, L.J.; Wells, J.E. Examining the influence of transformational leadership, organizational commitment, job embeddedness, and job search behaviors on turnover intentions in intercollegiate athletics. *Leadersh. Organ. Dev. J.* **2014**, *35*, 740–755. [CrossRef]
55. Zhen, Z.; Peterson, S.J. Advice networks in teams: The role of transformational leadership and memebers' core self-evaluations. *J. Appl. Psychol.* **2011**, *96*, 1004–1017.
56. Lussier, R. *Human Relations in Organizations: Applications and Skill Building*, 7th ed.; McGraw-Hill Education: New York, NY, USA, 2006.
57. Chen, L.; Zheng, W.; Yang, B.; Bai, S. Transformational leadership, social capital and organizational innovation. *Leadersh. Organ. Dev. J.* **2016**, *37*, 843–859. [CrossRef]
58. Johnson, S.G.; Schnatterly, K.; Hill, A.D. Board composition beyond independence: Social capital, human capital and demographics. *J. Manag.* **2013**, *39*, 232–262. [CrossRef]
59. Golmoradi, R.; Ardabili, F.S. The effects of social capital and leadership styles on organizational learning. *Procedia—Soc. Behav. Sci.* **2016**, *230*, 372–378. [CrossRef]
60. Turkina, E.; Thai, M.T.T. Social capital, networks, trust and immigrant entrepreneurship: A cross-country analysis. *J. Enterp. Communities People Places Glob. Econ.* **2013**, *7*, 108–124. [CrossRef]
61. Tsai, W.; Ghoshal, S. Social capital and value creation: The role of intrafirm networks. *Acad. Manag. J.* **1998**, *41*, 464–476. [CrossRef]
62. Alvani, M.; Nategh, T.; Farahi, M. The role of social capital in developing organizaitonal knowledge management. *J. Iran's Manag. Sci.* **2007**, *2*, 35–70.
63. Edwards, J.R.; Lambert, L.S. Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychol. Methods* **2007**, *12*, 1–22. [CrossRef] [PubMed]
64. Brislin, R.W. The wording and translation of research instrument. In *Field Methods in Cross-Cultural Research*; Lonner, W.J., Berry, J.W., Eds.; Sage: Beverly Hills, CA, USA, 1986; pp. 137–164.
65. Bass, B.M.; Avolio, B.J. *Multifactor Leadership Questionnaire*; Consulting Psychologists Press: Palo Alto, CA, USA, 1996.
66. McColl-Kennedy, J.R.; Anderson, R.D. Impact of leadership style and emotions on subordiante performance. *Leadersh. Q.* **2002**, *13*, 545–559. [CrossRef]
67. Podsakoff, P.M.; MacKenzie, S.B.; Moorman, R.H.; Fetter, R. Transformational leader behaviors and their effects on followers' trust in leader, satisfaction, and organizational citizenship behaviors. *Leadersh. Q.* **1990**, *1*, 107–142. [CrossRef]
68. Ming-Huei, C.; Yuan-Chieh, C.; Shih-Chang, H. Social capital and creativity in R&D project teams. *R D Manag.* **2007**, *38*, 21–34.
69. Collins, C.J.; Smith, K.G. Knowledge exchange and combination: The role of human resource practices in the performance high-technology firms. *Acad. Manag. J.* **2006**, *49*, 544–560. [CrossRef]
70. Qinxuan, G.; Yishi, F.; Aimin, H. knowledge sharing and organizational performance: The role of knowledge-driven human resource management practice. *Nankai Bus. Rev.* **2009**, *12*, 59–66.
71. Brown, T.A. *Confirmatory Factor Analysis for Applied Research*; Guilford Press: New York, NY, USA, 2006.
72. Bentler, P.M.; Bonett, D.G. Significance tests and goodness of fit in the analysis of covariance structures. *Psychol. Bull.* **1980**, *88*, 588–606. [CrossRef]

73. Hu, L.; Bentler, P.M. Cuttoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Model. A Multidiscip. J.* **1999**, *6*, 1–55. [CrossRef]
74. Browne, M.W.; Cudeck, R. Alternative ways of assessing model fit. *Sociol. Methods Res.* **1992**, *21*, 230–258. [CrossRef]
75. Baron, R.M.; Kenny, D.A. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* **1986**, *51*, 1173–1182. [CrossRef] [PubMed]
76. Hayes, A.F. An index and test of linear moderated mediation. *Multivar. Behav. Res.* **2015**, *50*, 1–22. [CrossRef] [PubMed]
77. Preacher, K.J.; Rucker, D.D.; Hayes, A.F. Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivar. Behav. Res.* **2007**, *42*, 185–227. [CrossRef] [PubMed]
78. Naoum, S. *People and Organizational Management in Construction*; Thomas Telford: London, UK, 2001.
79. Chan, A.T.S.; Chan, E.H.W. Impact of perceived leadership style on work outcomes: Case of building professionals. *J. Constr. Eng. Manag.* **2005**, *131*, 413–422. [CrossRef]
80. Zhu, W.; Chew, I.K.H.; Spangler, W.D. CEO transformational leadership and organizational outcomes: The mediating role of human-capital-enhancing human resource management. *Leadersh. Q.* **2005**, *16*, 39–52. [CrossRef]
81. Lin, H.; Hu, T. Knowledge interaction and spatial dynamics in industrial districts. *Sustainability* **2017**, *9*, 1421. [CrossRef]
82. Bresnen, M.; Goussevskaia, A.; Swan, J. Embedding new management knowledge in project-based organizations. *Organ. Stud.* **2004**, *25*, 1535–1555. [CrossRef]

sustainability

Article

Top Management Teams' Characteristics and Strategic Decision-Making: A Mediation of Risk Perceptions and Mental Models

Tungju Wu [1,2], Yenchun Jim Wu [3,4,*], Hsientang Tsai [5] and Yibin Li [1,2]

[1] College of Business Administration, Huaqiao University, Quanzhou 362021, China; tjwu@hqu.edu.cn (T.W.); liyibin@hqu.edu.cn (Y.L.)
[2] East Business Management Research Center, Huaqiao University, Quanzhou 362021, China
[3] National Taiwan Normal University, Taipei 10645, Taiwan
[4] National Taipei University of Education, Taipei 10671, Taiwan
[5] Department of Business Management, National Sun Yat-sen University, Kaohsiung 80424, Taiwan; htt@mail.nsysu.edu.tw
* Correspondence: wuyenchun@gmail.com; Tel.: +886-2-7734-3996

Received: 2 November 2017; Accepted: 5 December 2017; Published: 7 December 2017

Abstract: Strategic decision-making is a key factor of sustainability and development in enterprises. Moreover, the top management team (TMT) of an enterprise constitutes the base for decision-making. This study employed structural equation modeling to analyze questionnaires regarding TMTs' characteristics and strategic decision-making, and tested the mediating effects of risk perceptions and mental models and the moderating effects of psychological ownership. We investigated 289 valid questionnaires on TMTs completed by representatives from enterprises in China and found risk perceptions and mental models that serve as a mediating factor and are affected by the TMTs' characteristics and decision-making. We also found that psychological ownership exerts moderating effects between TMTs' characteristics and decision-making. This paper concludes with a discussion of theoretical and managerial implications for enterprise owners.

Keywords: top management team; strategic decision-making; risk perception

1. Introduction

Enterprises are expanding and shifting toward group management amidst perpetual economic growth and industrial development. In addition, enterprise management has gradually shifted from employing leader-oriented strategies to employing co-management strategies in the form of top management teams (TMTs). A TMT comprises key executives within an organization who are responsible for the planning and execution of organizational strategies. Previous studies have noted that strategic decision-making is affected by the behavioral factors of TMTs [1,2]. Hambrick and Mason [3] introduced the upper echelon theory, stating that TMT characteristics (e.g., age, international experience, educational background) are closely associated with vender strategy selection and are reflected in the major decisions of many organizations.

A number of previous studies on the effects of TMT factors on decision-making in enterprises have asserted that international experience among top managers [4], tenure [4,5], scale of the organization [4], and number of reappointed executives in the TMT [6] all influence the members of a TMT when making strategic decisions. The present study focused on examining the effects of TMTs and job-related characteristics. In contrast to previous studies that have discussed TMTs and job-related characteristics in relation to enterprises in emerging markets, the present study adopted tenure heterogeneity [4] to reflect TMT members' levels of understanding of their organizations [7], and employed international experience, educational background heterogeneity, and functional background heterogeneity to represent TMT members' training, professional knowledge, and skill levels.

Regardless of style, team composition, and atmosphere, leadership is a crucial variable influencing organizational performance and decision-making. Many previous studies have observed that leadership style is the foremost environmental factor that directly affects performance, morale, and sense of satisfaction among subordinates [3]. The personality traits of top executives are the most prominent factors that influence the formation of leadership style and organizational culture. In the management pyramid, management is categorized into three levels, namely first-, middle-, and top-level management. Top managers are typically involved in the direction, strategy, leadership, effectiveness, and philosophy of a company. The duties of these managers are to characterize the vision of the organization, establish common organizational values and concepts, capitalize on opportunities, promote innovation, evaluate risk, and lead the organization in fulfilling its organizational objectives. Therefore, top managers are the most crucial assets of organizations, and their leadership styles are key to the successful or failed formation of organizational culture and the implementation of systems, which consequently affects business performance. Organizational culture is a type of consensus among employees in an organization that prompts them to think and act in unison. According to previous studies, organizational culture can be defined as the beliefs, expectations, and values formed through cumulative communication, exchange, and transfer among organizational employees over time. The leadership styles of managers not only generate unique organizational cultures but also influence the work attitudes of subordinates.

In strategic decision-making, management team members' strategic decisions influence not only the future implementation of strategies but also the future survival and development of the enterprise. Therefore, a tacit understanding between team members in terms of their perceptions, methods of risk assessment and control, and feelings is a crucial factor that influences decision-making. Although many studies have investigated strategic decision-making in enterprises, most have been based on decision content [8], core resources [8,9], and network relationships [10]. The decision-making behaviors of TMTs are influenced by objective facts and conditions such as investment environment factors, and also personal factors among team members such as the personality traits, thinking logic, and cognitive understanding of managers. Several studies have analyzed the selection of business strategies based on the characteristics of management teams [11,12] or risk perception [13]. However, few studies have simultaneously explored the causal relationships between TMT members' experience characteristics, mental models, risk perception, and strategic decision-making. Managers' backgrounds and experiences influence the strategic decisions they make [12]. In addition, managers apply mental models to comprehend various situations and predict strategic decisions [14]. Therefore, clarifying managers' experience characteristics and the mental models they use to interpret objective environmental situations and subsequently determining correlations between these characteristics and models could help to elucidate the effects of management team members' experiences on their decision-making processes and cognitive understanding. Second, decision makers' behaviors are a type of risk. Decision makers draw on their personal experiences when perceiving diverse and extensive sets of information and, consequently, form various interpretations of information and scenarios or adopt various mental models. The human limitation of being able to selectively perceive or remember only certain information causes decision makers to perceive risk differently from one another. Therefore, clarifying the correlation between managers' experience characteristics and their risk perceptions could help to elucidate the effects of experiences on team decision-making processes for risk assessment. Third, previous studies have found that experience characteristics, mental models, and risk perceptions directly influence strategic decision-making. However, the findings of these studies have failed to reveal whether the effects of experience characteristics on strategic decision-making are mediated by mental models or risk perceptions. The present study attempted to clarify the causal relationships between these factors (e.g., experience characteristics, mental models, and risk perceptions) to determine the strategic decision-making processes of enterprise TMTs. Finally, an increasing number of enterprises are shifting toward directly appointing managers from outside the enterprise to serve in TMTs rather than promoting managers from within the organization. Therefore, determining managers' sense of belonging within TMTs is imperative. In this study, psychological ownership was selected as the measure of the sense of belonging to determine whether psychological ownership influences strategic decision-making. The causal relationships between psychological ownership and strategic decision-making were examined to elucidate the role of psychological ownership in the strategic decision-making processes of TMTs.

2. Literature Review

2.1. Managers' Decision-Making Behaviors

A review of previous studies on managers' decision-making behaviors showed that the majority of these studies have proposed various influence factors based on perspectives such as transaction cost efficiency, resource-based value, network relationships, and core logic. These factors are mostly objective facts andconditions and variables of operating environments and enterprise resources. These studies have largely focused on the effects of these variables on managers' decision-making behaviors. However, analyses have revealed that the effects of these factors on various decision-making behaviors are largely associated with managers' characteristics and perceptions [11,13]. Conventional economic analysis assumes that managers make decisions to maximize economic benefits. However, several scholars have noted that in addition to economic analysis, other decision-making models exist [15], and different decision makers differ in their risk perceptions and decision-making styles [16,17]. When reviewing a large amount of diversified information, decision makers draw on selective cognition or memory based on their personal qualities and experiences. Their interpretations of different types of information are influenced by their understanding of and feelings regarding the information. Mental models control the coding and storage of information in managers' memories, thereby enabling them to quickly select and process useful information and make suitable decisions in specific situations [18]. When enterprises engage in strategic decision-making, their decisions are generally influenced by a number of uncertainty factors. These factors, coupled with time constraints and inadequate information, force decision makers to make subjective judgments and decisions based on their experiences. Therefore, different decision makers often make different investment decisions. Understanding the effects of managers' decision-making behaviors on their strategic decisions helps managers to avoid potential cognitive biases and behavioral traps in the decision-making process. In reality, the decision-making processes and outcomes of enterprises are influenced by not only objective transaction facts and environments but also the mentalities and personality traits of management team members. Consequently, the subjective traits and objective conditions of managers may affect or interfere with their decisions [1]. Strategic decisions are often affected by managers' personal experiences or other members of the management team, and this can lead to judgment errors. In addition, misjudgment may result from inadequate or erroneous data collected by the management team. Therefore, analyzing the differences among the background characteristics and subjective experiences of team members can elucidate the effects of human factors on strategic decision-making in relation to transaction costs.

The resource-based view emphasizes efficiency. The abilities, knowledge, and resources of individuals and enterprises are acquired and accumulated over time. Learning and development pathways differ among individuals and enterprises, as may the rate in which knowledge and abilities are accumulated [19]. Enterprises comprise management teams and employees. Management teams are responsible for making enterprise decisions. The cognitive foundations, values, and observations of team members and the interactive processes of these factors affect organizational competitiveness and behaviors. Strategy formulation is a key process in the development of an enterprise. Strategy formulation processes involve the knowledge and cognitive processes of team members and the sharing of mental models among team members. Management teams apply different models and select different competitive strategies for different environments and markets [20]. The resource-based view can be applied to explain the differences between the core resources that facilitate business performance. Managers' decision-making behaviors can explain the differences in core resources selected by team members based on their thought processes, cognitive models, and decision-making behaviors when enterprises have similar explicit behaviors. The network view focuses on relationships and ties. Networks are continuous and repetitive exchange relationships within organizations, between organizations, and between organizations and external environments embedded within organizations [21]. Enterprises can establish relationships and acquire valuable resources by being active in their networks, thereby establishing prominence within their networks [21,22]. Investment activities are vital channels for enterprises to establish network relationships with the external environment. Such relationships are maintained through interaction and exchange between enterprises. Subsequently, management team members form cognitive pools by sharing their trial and error experiences to achieve the sharing of mental models. Socialization enables management team members to share their business know-how, industrial norms and technologies, and differences between consumption characteristics. Therefore, the network view explains that enterprises acquire valuable market resources through network relationships. By contrast, managers' decision-making behaviors explain how management team members perceive markets, respond to demand, and make relevant strategic decisions in relation to the network relationships of enterprises.

2.2. TMT Experience Characteristics and Strategic Decision-Making Behaviors

A TMT is a group of key managers responsible for strategy formulation, planning, and execution. Generally, the upper echelons theory can be used to explain how TMTs make strategic decisions and the relationship between those decisions and business performance. Validating TMT characteristics can facilitate the prediction of organizational strategic decisions and business performance. Experience characteristics refer to members' cognitive foundations and values at a psychological level, as well as their age, professional history and experiences, professional background, organizational tenure, and international experiences [23,24]. Murray [25] asserted that any manager of the rank of deputy general manager or higher could be a TMT member [7]. Team members spend less time communicating and coordinating with other team members when the team has high age homogeneity. Conflict is also less likely to occur in such teams, and thus such teams are more likely to adopt high-risk strategies. Tenure experience refers to the amount of time a manager has held a single management position within a company [26]. Managers with less tenure experience require more time to learn enterprise practices. Therefore, teams with such managers tend to make conservative decisions. Management team members' professional histories and experiences refer to their previous professional domains. Managers typically draw on their experiences in these domains when making decisions. Management teams with a background in productivity primarily emphasize manufacturing, research and design, accounting, data processing, and information. Therefore, managers in such teams attach great value to organizational control and business efficiency and adopt strategies that provide them with a greater degree of control [7,27]. International experience refers to business management experience in a variety of countries and contexts. General international experience refers to management team members' ability to oversee the business activities of various overseas markets [28]. When members are familiar with the cultures and policies of other countries and maintain a broad worldview, they are more accepting of risk and tend to formulate more radical strategies [28]. Previous studies have reported that TMTs are more likely to prefer radial and high-risk strategies when members are similar in age, have more experience, come from productivity-related professional backgrounds, and possess substantial international experience. Therefore, the following hypothesis was formulated:

Hypothesis 1. *TMT members' preferences for making risky decisions are positively correlated to the level of similarity among their experience characteristics.*

2.3. Mental Models and Strategic Decision-Making Behaviors

Mental models are intrinsic cognitive processes applied by individuals when performing an action in a specific environment. From the perspective of cognitive psychology, mental models are defined as an individual's internal representation of external reality [18]. Mental models are often used to predict specific behaviors under different conditions. Scholars often use mental models to interpret the causes of specific phenomena rather than religiously following the formal rules of logical reasoning [20]. A major reason that people rely on mental models stems from the limitations of individuals' underlying work-related memories. Applying formal rules of logical reasoning requires individuals to consider many possibilities. Using mental models to explain an individual's behavior reduces the level of conjecture regarding the individual's cognitive behaviors. However, this inference approach may not yield entirely correct or comprehensive findings [29]. In other words, people may develop different perceptions of the same matter because of differences in their mental models. Greca and Moreira [18] described mental models as models that people develop to explain real-world situations or events. Individuals apply mental models to ingeniously understand and explain phenomena. These models also serve as a reference for predicting individuals' behaviors.

In an organization, top managers' mental or cognitive understanding affects their understanding and selection of information, thereby influencing their strategic decisions. TMT members generally make decisions based on their existing cognitive foundations. Therefore, TMTs in different organizations usually create different organizational cultures. Previous studies have verified that managers' mental models influence their decision-making behaviors and organizational output [30]. TMTs make strategic decisions to achieve the development objectives of the enterprise. Thus, managers' mental models are vital factors that influence strategic decision-making [30]. Furthermore, a study on the effects of managers' mental models on strategic decision-making found that when TMTs value market competition, they are likely to adopt radical and high-risk strategies to quickly defend or retaliate against market competitors and ensure the survival of the company in a fiercely competitive market [31]. When TMTs value clients, they tend to formulate conservative strategies based on client demands to conform to consumer rights. When TMTs value their own opinions, they tend to formulate high-risk strategies that afford them more control rather than those that focus on competitors' behaviors or clients' opinions. Enterprise strategies change with the external environment; for example, collecting information on competitors and appropriately increasing client value are crucial elements for enterprises that require competitive strategies. Therefore, team members who pursue development prefer strategies that provide them with more control. Thus, the following hypothesis was formulated:

Hypothesis 2. *TMT members' preferences for high-risk strategic decision-making are positively correlated with their preferences for development strategies.*

2.4. Risk Perceptions and Strategic Decision-Making

Risk perception concepts suggest that the actions of decision makers may produce unexpected results. Therefore, decision makers' behaviors are a type of risk. Previous studies have noted that managers tend to define risk as the sum of possible losses, including financial, labor, and other tangible and intangible losses. Managers are often confident that they can hone their management skills and talent to control the occurrence of risk while continually revising their methods of evaluating target risk until the evaluation conforms to their original expected results. This concept has crucial implications for strategic decision-making [32]. Global trends are key factors that influence TMTs' international decisions. Therefore, TMT members' risk perceptions concerning external environments significantly influence their enterprises' strategic decision-making. Moreover, business experience, cultural differences, industrial structures, and systematic risk all influence the risk perceptions of management team members [33].

Business experience refers to experiences in business activity accumulated over time by TMTs. When external environments are controllable and market and competitor information are accessible, managers are better able to predict the potential outcomes of their decisions. Under such circumstances, managers are more likely to adopt radical high-risk strategies. Furthermore, in a highly competitive business environment, TMTs must be attentive to the number of competitors and the strategies adopted by competitors. Moreover, local politics and institutional environments or risks stemming from political unrest [34] are also factors requiring consideration when making strategic decisions. According to the resource-based view, whether the local government supports enterprise development is a major factor of consideration for TMTs during the decision-making process. If a government promotes policies that benefit business operations, such as tax exemptions on raw material imports, the provision of components and equipment, low wages, outstanding talent, and government subsidies, this reduces risk as perceived by TMTs and prompts TMTs to adopt radical and high-risk strategies. These uncertainties affect the future development of an enterprise, and the risk perceptions of TMTs affect strategic decision-making. Based on the preceding discussion, the following hypothesis was formulated:

Hypothesis 3. *TMT members' preferences for high-risk strategic decision-making are negatively correlated with their risk perceptions.*

2.5. Mediating Effect of Mental Models and Risk Perceptions

The compositional characteristics of TMTs significantly influence enterprise strategies and business performance [35]. Mental models are extrinsic behavioral manifestations of decision makers' cognitive understanding as influenced by their experiences [31]. Team members share their experiences with one another to determine common experiences, which are subsequently applied in mental models for market development and new strategy formulation [36]. According to social identity theory, individuals interact and communicate better and have less conflict with others of a similar age. Therefore, a team composed of members of similar ages is more likely to agree on a common decision [37]. Team members with more tenure experience are more familiar with the internal management of the enterprise and have more freedom to make decisions and more power to influence strategic orientation. Therefore, such team members are more capable of making decisions that are beneficial for the organization [1] and long-term coworkers are more likely to concede to strategic selections made by such team members. TMT members with business experience in overseas markets enhance the ability of the enterprise to handle overseas affairs and improve financial performance through international development [38]. Amidst economic globalization, development strategies are fundamental for ensuring the survival and competitiveness of enterprises. Members of a management team who are similar in age, have more tenure experience, and possess extensive international experience can share and accumulate similar experiences consistently, rapidly acquire the latest market information and competition conditions, make effective judgments, and quickly and effectively communicate and coordinate with one another. Most crucially, proposed development strategies are more likely to yield consensus in such a team. Therefore, the following hypothesis was formulated:

Hypothesis 4. *TMT members' preferences for development strategies are positively correlated with similarities in their experience characteristics.*

Decision makers may overestimate or underestimate risk or misjudge uncertainty, thereby causing them to make poor decisions. Subsequently, the characteristics and experiences of TMTs affect the risk perceptions of the team members [27]. Perceived differences may vary depending on the managers' backgrounds and experiences, eventually influencing their decision-making behaviors. When team members are similar in age, the similarities in their thought processes, backgrounds, and experiences enable them to more easily communicate, coordinate, and resolve team problems and conflicts. Improved communication quality reduces managers' risk perceptions, thereby increasing their willingness to take risks. Teams with more tenure experience have more time to establish a consensus because team members can draw on team experiences when making decisions, thereby reducing their perceived risk [2,24]. Management teams with a prominent background in productivity are concerned with manufacturing, research and development, innovation and improvement, and long-term investment, all of which facilitate innovation. Such teams are more accepting of potential losses, and thus have lower risk perceptions and higher willingness to take risks [27]. In teams with extensive international experience, members generally specialize in specific regional markets or overall international trends and are more confident operating in overseas markets. Such teams can minimize the perceived cost of uncertainty, and thus members often have lower risk perceptions. Considering this context, the following hypothesis was formulated:

Hypothesis 5. *When formulating enterprise development strategies, TMT members' risk perceptions are negatively correlated with the similarities in their experience characteristics.*

2.6. Moderating Effect of Psychological Ownership

Psychological ownership primarily refers to individuals' intrinsic (emotional and cognitive) sense of possession (i.e., mine-ness or our-ness) [39]. The definition of psychological ownership is slightly different to that of ownership in terms of the concept and motivations of possession. The sense of possession is omnipresent in psychological ownership, and the target of such possession may be tangible or intangible. This possession can be characterized as ownership that is seemingly legitimate but not governed by any existing laws [40]. When an individual's sense of possession is combined with his or her psychological feelings, the members of the organization may instinctively sense this ownership and attempt to apply these feelings to gain satisfaction, particularly when the target is part of the owner's psychological identification. This is psychological ownership [41]. When an individual's sense of possession is combined with his or her psychological feelings, members of the organization may instinctively sense this ownership and attempt to apply these feelings to gain satisfaction [42]. Psychological ownership is a type of attitude. Attitudes stem from individuals' personal values. For employees to apply their values to their team or organization, some employees require motivational stimuli, whereas others require the ascension of perception. In other words, employees are likely to exhibit psychological ownership when the following three basic requirements are met: (1) Workplace identification: a sense of home or belonging in the workplace; (2) efficiency and effectiveness: the right to control is satisfied; and (3) self-identity: expression of unique self-value and self-recognition [41,43].

Psychological ownership is essentially different from possession in terms of concept and motivation. The feeling of possession is omnipresent and can be associated with tangible or intangible "targets". Such ownership may or may not have a legitimate basis. In psychology, "target" refers to any tangible or intangible object of attachment. Attachment can stem from an individual or group. The object of attachment can be as small as a seat or as large as an organization or enterprise [42]. Previous studies have reported that psychological ownership influences individuals' behaviors and performance within an organization and their relationships with their work and their organization [40]. Therefore, ownership is mutually linked to individuals' behaviors and can be used to effectively predict individuals' actions. In other words, the feeling of self-regulation is part of an individual's psychological composition, or in other words, the result of self-efficacy [44]. A sense of control and action are the elements that enable tasks to be successfully completed. Therefore, when TMT members acknowledge their organization and its importance to them, they perceive organizational objectives as personal objectives [45]. Consequently, these objectives enhance members' self-efficacy and sense of responsibility and influence their subsequent behaviors, performance [42], and decisions. Employees assume responsibility for their duties and strive to produce the best possible results. Thus, employees with a high sense of psychological ownership have a greater ability to acknowledge the objectives of their organization and formulate appropriate organizational development strategies. Considering this context, the following hypothesis was formulated:

Hypothesis 6. *TMT members' preferences for high-risk strategic decision-making are negatively correlated with the similarities in their experience characteristics and psychological ownership perceptions.*

3. Method

3.1. Research Variables

In this study, the effects of similarities in TMT members' experience characteristics, their strategic mental preferences, and their risk perceptions when making strategic decisions were analyzed. In addition, psychological ownership was adopted as a mediator to elucidate the effects of psychological ownership on the relationship between team members' experience characteristics and strategic decision-making. A questionnaire survey was conducted and scored using a 7-point Likert scale. The experience characteristics proposed by Terpstra and Yu [46] were adopted in this study. These were age, work background, international experience, and length of tenure. The effects of these characteristics on strategic decision-making were analyzed. TMT members were defined as deputy general managers, department directors, and other high-ranking managers. Similarities in experience characteristics were examined based on four items, namely age homogeneity, length of tenure, productivity-related work background, and extent of international experience. The average Cronbach's α coefficient of these items was 0.89. The concepts proposed by Maignan and Lukas [47] were adopted to measure mental models. The strategic mental preferences of team members were observed, including competitor-based and client-based strategic objectives, the effects of these strategies on company decisions, and the environmental pressures stemming from these strategies. The mental model concepts proposed by Maignan and Lukas [47] were used to define team members' preferences for development strategies, or more specifically, preferences toward competitor-based or client-based strategies. The following four items were developed: actively guarding against competitors, enhancing client value, information acquired from strategies significantly influences company decisions (non-self-centered), and high environmental pressures stemming from strategies (market-driven). The average Cronbach's α coefficient of these items was 0.91. The risk perception concepts proposed by Brouthers [48] were adopted in this study to observe team members' risk perceptions regarding overseas investment and the effects of these perceptions on strategic decision-making. Five items concerning the risk perceptions of host countries were formulated, namely investment and business experience, cultural differences, industrial concentration, system stability of the host country, and ownership risk. The average Cronbach's α coefficient of these items was 0.92. Psychological ownership is a psychological state where individuals develop a sense of possession toward specific targets based on their experiences. Targets can be tangible objects or intangible ideas. The definition of psychological ownership proposed by Pierce et al. [41] was adopted in the present study. Psychological ownership was measured using three items. The average Cronbach's α coefficient of these items was 0.93. The concepts of strategic decision-making proposed by Garnsey et al. [49] were adopted in the present study. Garnsey et al. explained that managers evaluate the strengths and weaknesses of an enterprise and the opportunities and threats of the environment before making a suitable strategic decision. Strategic decision-making was measured using three items. The average Cronbach's α coefficient of these items was 0.89.

3.2. Scope and Subjects

The TMTs of China's top 1000 companies listed on the China Stock Exchange were selected as the research subjects. Companies' public statements and annual financial reports were analyzed to form a list of companies with TMTs. Questionnaires were submitted to these companies. A total of 698 companies satisfied the inclusion criteria and one questionnaire was submitted to each company. Thus, 698 questionnaires were administered. The questionnaires were administered via surface mail or email. Slight bias may exist in the research data because the identities of the respondents could not be validated. A total of 371 questionnaires were returned. After excluding invalid and incomplete questionnaires, 289 were retained. Among the respondents, 79% were top managers and were men. The average age of the respondents was 53 years, and most worked in the electronics industry (41%). The average length of employment was 17 years.

3.3. Methodologies and Tools

Structural equation modeling was conducted to process the data collected from the questionnaires. LISREL 8.54 software was adopted as the analysis tool for confirmatory factor analysis (CFA) and theoretical causal model analysis to validate the correlation and path coefficients between TMT members' experience characteristics, mental models, risk perceptions, psychological ownership, and strategic decision-making.

4. Results

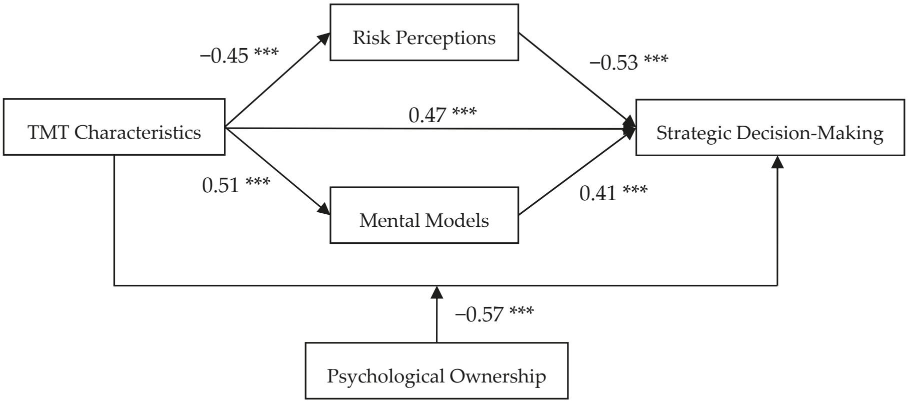
CFA was performed to evaluate the overall measurement model. The results indicated excellent goodness of fit (χ^2 = 99.86, df = 94, p = 0.07, χ^2/df = 1.06, goodness of fit index (GFI) = 0.93, adjusted goodness-of-fit index (AGFI) = 0.86, comparative fit index (CFI) = 0.97, non-normed fit index (NNFI) = 0.98, root mean square error of approximation (RMSEA) = 0.026, standardized root mean square residual (SRMR) = 0.031). The factor loadings (λ) of the various measurement indices ranged between 0.73 and 0.92. All indices achieved a factor loading of 0.7 or higher and a p value of 0.001 or lower. The squared multiple correlation (SMC) was adopted as the reliability coefficient. All indices achieved an SMC value of 0.5 or higher, suggesting that the measurement indices adopted in this study had excellent reliability, effectively reflected the latent variables, and achieved excellent convergence validity. In addition, the construct reliability and average variance extracted (AVE) coefficients of the various latent variables were within an acceptable range. The construct reliability results showed that all the latent variables achieved a value of 0.7 or higher, suggesting excellent internal consistency. The AVE results revealed that all potential variables achieved a value of 0.5 or higher, suggesting that the measurement variables could appropriately explain the latent variables. Therefore, the latent variables achieved excellent construct reliability and validity. To achieve acceptable discriminate validity, the results for the between-group relationships of the latent variables in the measurement model were required to be lower than those of their in-group relationships. A relationship matrix was developed to test the between-group relationships of the latent variables. The results showed that the square root of the AVE was higher than in the between-group coefficients of the latent variables, suggesting excellent discriminate validity (Table 1).

Table 1. Correlation coefficient matrix of the latent variables.

Variable	Experience Characteristics	Mental Models	Risk Perceptions	Psychological Ownership	Strategic Decision-Making
Experience characteristics	**0.85**				
Mental models	0.47 **	**0.82**			
Risk perceptions	−0.41 **	−0.45 **	**0.84**		
Psychological ownership	0.39 **	0.47 **	−0.44 **	**0.87**	
Strategic decision-making	0.43 **	0.41 **	−0.46 **	−0.43 **	**0.89**

Note: n = 289; ** $p < 0.01$; The blackbody is the square root of the interpreted variance; the data below the diagonal is the correlation coefficient between the variables.

When the final model and its goodness of fit had been validated, the relationships between the latent variables were examined. Figure 1 shows that significant causal relationships existed among the latent variables. The analysis results indicated that the similarities in the team members' experience characteristics positively influenced their strategic decision-making. The path coefficient was 0.47, indicating statistical significance (t = 4.357, $p < 0.001$). Thus, H1 was supported. Moreover, team members' strategic mental preferences positively influenced their strategic decision-making. The path coefficient was 0.41, indicating statistical significance (t = 4.064, $p < 0.001$). Thus, H2 was supported. TMT members' risk perceptions negatively influenced their strategic decision-making. The path coefficient was −0.53, indicating statistical significance ($t = -5.968$, $p < 0.001$). Thus, H3 was supported. The path coefficient for the similarities of TMT members' experience characteristics and their strategic mental preferences was 0.51, indicating statistical significance ($t = 5.847$, $p < 0.001$). Thus, H4 was supported. The similarities of TMT members' experience characteristics were negatively influenced by their risk perceptions. The path coefficient was −0.45, indicating statistical significance ($t = -4.378$, $p < 0.001$). Thus, H5 was supported. Finally, the path coefficient concerning the moderating effect of psychological ownership on the relationship between similarities in TMT members' experience characteristics and their strategic decision-making was −0.57, indicating statistical significance ($t = 6.018$, $p < 0.001$). Thus, H6 was supported (Figure 1).



*** $p < 0.001$

Figure 1. Model path diagram and standardized parameter estimations. TMT = top management team.

The multi-mediator model proposed by Brown [50] was used to evaluate the mediating effect of strategic preferences and risk perceptions. The model effects were classified into direct effects, total effects, total indirect effects, and individual indirect effects. The path coefficient for experience characteristics, strategic mental preferences, and strategic decision-making achieved statistical significance. The individual indirect effect of strategic mental preferences was 0.21 (0.41 × 0.51), which was lower than the direct effect of 0.47, suggesting that the absolute mediating effect of strategic mental preferences was extremely limited. Furthermore, the path coefficient for experience characteristics, risk perceptions, and strategic decision-making achieved statistical significance. The individual indirect effect of risk perceptions was 0.24 (−0.53 × (−0.45)), which was lower than the direct effect of 0.47, indicating the presence of an absolute mediating effect of risk perceptions. A summary of the overall model revealed that the indirect effect of risk perceptions on strategic decision-making (0.24) was greater than that of strategic mental preferences on strategic decision-making (0.21). In other words, risk perception is the key mediator affecting the relationship between the similarities of TMT members' experience characteristics and their strategic decision-making.

5. Discussion

Previous studies on strategic decision-making have largely focused on transaction costs and the styles of decision makers. Although these studies have explained the objective environmental factors that influence enterprises and the core resources that influence strategic decisions, they have failed to discuss the influence of TMT members, who are the principal decision makers in a company, on strategic decision-making. The present study provided an alternative theoretical perspective by focusing on TMT members rather than non-enterprise individuals. This study analyzed several decision-making behavior variables from the perspective of decision-making (e.g., similarities in team members' experience characteristics, mental models, and risk perceptions) to examine strategic decision-making in companies. The findings confirmed that TMT members' experience characteristics, strategic mental preferences, and risk perceptions, as well as their enterprise ownership perceptions, influence their strategic decision-making processes. In addition, the findings revalidated the claims of previous studies that management team members' characteristics [38], mental models [20], and risk perceptions [33] are crucial factors that influence team members' strategic decisions. Therefore, team members draw on previously accumulated experiences to make strategic decisions. These decisions are the product of a strategic consensus fueled by team members' experiences and cognitive processes. They are the outcomes that pose the lowest risk as evaluated by team members. This study focused on the factors that affect decision-making processes rather than those that affect decision-making performance, particularly the effects of team members' cumulative experiences on their decisions concerning the arrangement of transactional activities or processes of value creation. For example, psychological ownership concepts were examined to determine whether TMT members' psychological ownership perceptions influenced their decision-making processes. The findings showed that when team members had high ownership perceptions of their enterprise (i.e., they perceived the enterprise as theirs), they were more careful when making strategic decisions and less likely to formulate radical high-risk strategies. Therefore, enterprises should strive to create an environment that generates a sense of belonging and prompts team members to identify with the organization and develop willingness to contribute to the organization unconditionally. Team members who establish a symbiotic bond with their enterprises are more careful when making strategic decisions and less likely to act impulsively. When making strategic decisions, team members' cumulative experiences and mutual understanding and awareness facilitate rapid judgment. However, these factors may also cause heuristic biases. A team with members who have more tenure experience, international experience, and similarities in terms of their professional backgrounds is more likely to achieve consensus. However, these factors could lead to member herding or confirmation bias, which refers to members selecting or collecting beneficial information and neglecting unfavorable or contradictory information to support themselves or reinforce team consensus, thereby causing determination biases. Nevertheless, when team members' experience characteristics are extremely different from one another, the team can acquire a broader range of information and achieve greater innovation. However, such differences can create more conflict. In teams with such differences, effective screening of critical information is crucial to eliminate bias. Understanding the composition and characteristics of team members' experiences can effectively prevent misjudgment and facilitate suitable decision-making. These concepts help enterprises identify the reasons for ineffective decisions and minimize erroneous decisions.

Finally, the findings revealed that in the decision-making processes of TMTs, members' experience characteristics, strategic mental preferences, and risk perceptions influence strategic decision-making. Moreover, the influence of experience characteristics on decision-making is mediated by risk perceptions. In other words, TMT members are generally more tolerant to external risk and more likely to formulate radical and high-risk strategies when their experience characteristics are similar. External environments are influenced by policies, supply and demand, and global trends. TMT composition characteristics should be associated with these factors. The risk awareness of members with an extensive worldview or more tenure experience is influenced by their cognitive experiences. Therefore, such members are more likely to formulate conservative strategies.

6. Conclusions

This study comprises three major practical implications. First, it serves as a basis for explaining the effects of similarities in management team members' experience characteristics, their strategic mental preferences, and their risk perceptions when making strategic decisions. In other words, similarities in management team members' experience characteristics, their strategic mental preferences, and their risk perceptions are all major factors that influence decision-making. Therefore, enterprises should consider not only objective factors such as market transactions and enterprise resources but also factors concerning team members.

Second, risk perception is a mediator in the relationship between similarities in management teams' experience characteristics and their decision-making behaviors. Risk perceptions are team members' subjective opinions regarding objective facts, and these opinions are affected by members' background characteristics and experiences. Therefore, different members adopt different cognitive understanding models and evaluation methods, both of which are reflected in their decision-making behaviors. Third, psychological ownership was confirmed to influence the similarities in team members' experience characteristics and their decision-making behaviors. A team generally comprises several members, each of whom has his or her own set of experience characteristics. Therefore, each member has a different sense of belonging in the enterprise. Crucial company decisions are typically formulated by the TMT. Members who are unable to identify with their organization are more likely to form erroneous strategies, consequently hindering company performance or causing problems. Therefore, enterprises should not only value the experience characteristics of team members but also strive to create a comfortable, respectful, and trusting work environment to enhance members' ownership perceptions, thereby improving the quality of their strategic decision-making.

The respondents of this study were members of TMTs in China's top 1000 enterprises. Management teams generally comprise two or more members, and different members may have different subjective opinions and, therefore, provide different responses to questionnaire items. Future studies could adopt management teams as the unit for analysis to compare the characteristics of members of the same team and elucidate how TMT members with different backgrounds form and share mental models within the team, as well as how conflicts stemming from different risk perceptions are resolved and how they affect final decisions. Moreover, the questionnaire recovery rate in this study was relatively low because of limited funding and time. Nevertheless, the outcomes of the statistical analysis performed in this study were acceptable, and the reliability and validity results were influential. Future studies could endeavor to increase sample recovery to enhance the reference value of research outcomes.

Companies may have standard procedures for formulating decisions. However, decisions in Taiwanese enterprises or TMTs are ultimately made by people. Therefore, to some extent, decisions are affected by the personality traits and experiences of the decision maker. These traits and experiences influence decision makers' thought processes and incite them to form different perceptions of specific events or objects. Consequently, these perceptions influence team decisions. Therefore, this study focused on the subjective experiences and cognitive factors of TMT members and decision makers. However, the factors that influence strategic decision-making include objective conditions and subjective behaviors. Future studies could combine all these factors when examining strategic decision-making to validate the relationships between the two factor types and elucidate the effects of these relationships on decision-making.

Acknowledgments: The authors are grateful to the valuable comments made by the reviewers. This research was supported by the National Natural Science Foundation of China (71702059, 71571074), Social Science and Humanity on Young Fund of the Ministry of Education (16YJC630136), and Education and Research Projects of Young and Middle-aged Teachers in Fujian Province (JAS160039), Ministry of Science & Technology, Taiwan (104-2511-S-003-031-MY3 & 106-2511-S-003-029-MY3).

Author Contributions: Tungju Wu and Yenchun Jim Wu conceived, designed, and wrote this paper. Yibin Li searched the research data, refined the collected data, and manipulated the data collected in software tools. Hsientang Tsai realized the analysis of the combinations and interpreted the data. All authors contributed to the closing of the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Barker, V.L., III; Mueller, G.C. CEO characteristics and firm R&D spending. *Manag. Sci.* **2002**, *48*, 782–801. [CrossRef]
2. Colbert, A.E.; Barrick, M.R.; Bradley, B.H. Personality and leadership composition in top management teams: Implications for organizational effectiveness. *Pers. Psychol.* **2014**, *67*, 351–387. [CrossRef]
3. Hambrick, D.C.; Mason, P.A. Upper echelons: The organization as a reflection of its top managers. *Acad. Manag. Rev.* **1984**, *9*, 193–206. [CrossRef]
4. Wowak, A.J.; Gomez-Mejia, L.R.; Steinbach, A.L. Inducements and motives at the top: A holistic perspective on the drivers of executive behavior. *Acad. Manag. Ann.* **2017**, *11*, 669–702. [CrossRef]
5. Jaw, Y.L.; Lin, W.T. Corporate elite characteristics and firm's internationalization: CEO-level and TMT-level roles. *Int. J. Hum. Resour. Manag.* **2009**, *20*, 220–233. [CrossRef]
6. Liu, Y.; Valenti, M.A.; Yu, H.Y. Presuccession performance, CEO succession, top management team, and change in a firm's internationalization: The moderating effect of CEO/chairperson dissimilarity. *Can. J. Adm. Sci.* **2012**, *29*, 67–78. [CrossRef]
7. Hambrick, D.C.; Cho, T.S.; Chen, M.J. The influence of top management team heterogeneity on firms' competitive moves. *Adm. Sci. Q.* **1996**, 659–684. [CrossRef]
8. Williams, D.W.; Grégoire, D.A. Seeking commonalities or avoiding differences? Re-conceptualizing distance and its effects on internationalization decisions. *J. Int. Bus. Stud.* **2015**, *46*, 253–284. [CrossRef]
9. Sui, S.; Baum, M. Internationalization strategy, firm resources and the survival of SMEs in the export market. *J. Int. Bus. Stud.* **2014**, *45*, 821–841. [CrossRef]
10. Peng, M.W.; Wang, D.Y.; Jiang, Y. An institution-based view of international business strategy: A focus on emerging economies. *J. Int. Bus. Stud.* **2008**, *39*, 920–936. [CrossRef]
11. Abatecola, G. Untangling self-reinforcing processes in managerial decision making. Co-evolving heuristics? *Manag. Decis.* **2014**, *52*, 934–949. [CrossRef]
12. Kouamé, S.; Oliver, D.; Poisson-de-Haro, S. Can emotional differences be a strength? Affective diversity and managerial decision performance. *Manag. Decis.* **2015**, *53*, 1662–1676. [CrossRef]
13. Laufs, K.; Bembom, M.; Schwens, C. CEO characteristics and SME foreign market entry mode choice: The moderating effect of firm's geographic experience and host-country political risk. *Int. Mark. Rev.* **2016**, *33*, 246–275. [CrossRef]
14. Jarratt, D.; Fayed, R. The impact of market and organisational challenges on marketing strategy decision-making: A qualitative investigation of the business-to-business sector. *J. Bus. Res.* **2001**, *51*, 61–72. [CrossRef]
15. Kumar, V.; Subramanian, V. A contingency framework for the mode of entry decision. *J. World Bus.* **1997**, *32*, 53–72. [CrossRef]
16. Baron, R.A.; Franklin, R.J.; Hmieleski, K.M. Why entrepreneurs often experience low, not high, levels of stress: The joint effects of selection and psychological capital. *J. Manag.* **2016**, *42*, 742–768. [CrossRef]
17. Wennberg, K.; Delmar, F.; McKelvie, A. Variable risk preferences in new firm growth and survival. *J. Bus. Ventur.* **2016**, *31*, 408–427. [CrossRef]
18. Greca, I.M.; Moreira, M.A. Mental models, conceptual models, and modelling. *Int. J. Sci. Educ.* **2000**, *22*, 1–11. [CrossRef]
19. Cohen, W.M.; Levinthal, D.A. Adsorptive capacity: A new perspective on learning. *Adm. Sci. Q.* **1990**, *35*, 128–152. [CrossRef]
20. Karakaya, F.; Yannopoulos, P. Defensive strategy framework in global markets: A mental models approach. *Eur. J. Mark.* **2010**, *44*, 1077–1100. [CrossRef]
21. Park, S.H.; Luo, Y. Guanxi and organizational dynamics: Organizational networking in Chinese firms. *Strateg. Manag. J.* **2001**, *22*, 455–477. [CrossRef]
22. Wu, T.J.; Tsai, H.T.; Yeh, S.P. The role of manager's locus of control between perceived guanxi and leadership behavior in family business. *Rev. Int. Sociol.* **2014**, *72*, 87–104. [CrossRef]

23. Naranjo-Gil, D. The role of management control systems and top teams in implementing environmental sustainability policies. *Sustainability* **2016**, *8*, 359. [CrossRef]
24. Hambrick, D.C.; Humphrey, S.E.; Gupta, A. Structural interdependence within top management teams: A key moderator of upper echelons predictions. *Strateg. Manag. J.* **2015**, *36*, 449–461. [CrossRef]
25. Murray, A.I. Top management group heterogeneity and firm performance. *Strateg. Manag. J.* **1989**, *10*, 125–141. [CrossRef]
26. Tihanyi, L.; Ellstrand, A.E.; Daily, C.M.; Dalton, D.R. Composition of the top management team and firm international diversification. *J. Manag.* **2000**, *26*, 1157–1177. [CrossRef]
27. Herrmann, P.; Datta, D.K. CEO experiences: Effects on the choice of FDI entry mode. *J. Manag. Stud.* **2006**, *43*, 755–778. [CrossRef]
28. Jahanshahi, A.A.; Brem, A. Sustainability in SMEs: Top Management Teams Behavioral Integration as Source of Innovativeness. *Sustainability* **2017**, *9*, 1899. [CrossRef]
29. Mercier, H.; Sperber, D. Why do humans reason? Arguments for an argumentative theory. *Behav. Brain Sci.* **2011**, *34*, 57–74. [CrossRef] [PubMed]
30. Gary, M.S.; Wood, R.E.; Pillinger, T. Enhancing mental models, analogical transfer, and performance in strategic decision making. *Strateg. Manag. J.* **2012**, *33*, 1229–1246. [CrossRef]
31. Santos, M.V.; Garcia, M.T. Organizational change: The role of managers' mental models. *J. Chang. Manag.* **2006**, *6*, 305–320. [CrossRef]
32. Duanmu, J.L. The effect of corruption distance and market orientation on the ownership choice of MNEs: Evidence from China. *J. Int. Manag.* **2011**, *17*, 162–174. [CrossRef]
33. Cesinger, B.; Hughes, M.; Mensching, H.; Ricarda, B.; Viktor, F.; Sascha, K. A socioemotional wealth perspective on how collaboration intensity, trust, and international market knowledge affect family firms' multinationality. *J. World Bus.* **2016**, *51*, 586–599. [CrossRef]
34. Brouthers, K.D.; Brouthers, L.E. Why service and manufacturing entry mode choices differ: The influence of transaction cost factors, risk and trust. *J. Manag. Stud.* **2003**, *40*, 1179–1204. [CrossRef]
35. Myung, J.K.; Choi, Y.H.; Kim, J.D. Effects of CEOs' negative traits on corporate social responsibility. *Sustainability* **2017**, *9*, 543. [CrossRef]
36. Murray, J.Y.; Gao, G.Y.; Kotabe, M. Market orientation and performance of export ventures: The process through marketing capabilities and competitive advantages. *J. Acad. Mark. Sci.* **2011**, *39*, 252–269. [CrossRef]
37. Burt, R.S. The network structure of social capital. *Res. Organ. Behav.* **2000**, *22*, 345–423. [CrossRef]
38. Nielsen, S. Top management team internationalization and firm performance. *Manag. Int. Rev.* **2010**, *50*, 185–206. [CrossRef]
39. Pierce, J.L.; Rubenfeld, S.A.; Morgan, S. Employee ownership: A conceptual model of process and effects. *Acad. Manag. Rev.* **1991**, *16*, 121–144. [CrossRef]
40. Van Dyne, L.; Pierce, J.L. Psychological ownership and feelings of possession: Three field studies predicting employee attitudes and organizational citizenship behavior. *J. Organ. Behav.* **2004**, *25*, 439–459. [CrossRef]
41. Pierce, J.L.; Kostova, T.; Dirks, K.T. Toward a theory of psychological ownership in organizations. *Acad. Manag. Rev.* **2001**, *26*, 298. [CrossRef]
42. Avey, J.B.; Avolio, B.J.; Crossley, C.D.; Luthans, F. Psychological ownership: Theoretical extensions, measurement and relation to work outcomes. *J. Organ. Behav.* **2009**, *30*, 173–191. [CrossRef]
43. Peng, H.; Pierce, J. Job-and organization-based psychological ownership: Relationship and outcomes. *J. Manag. Psychol.* **2015**, *30*, 151–168. [CrossRef]
44. Downes, P.E.; Kristof-Brown, A.L.; Judge, T.A.; Darnold, T.C. Motivational mechanisms of self-concordance theory: Goal-specific efficacy and person–organization fit. *J. Bus. Psychol.* **2017**, *32*, 197–215. [CrossRef]
45. Galvin, B.M.; Lange, D.; Ashforth, B.E. Narcissistic organizational identification: Seeing oneself as central to the organization's identity. *Acad. Manag. Rev.* **2015**, *40*, 163–181. [CrossRef]
46. Terpstra, V.; Yu, C.M. Determinants of foreign investment of U.S. Advertising Agencies. *J. Int. Bus. Stud.* **1988**, *19*, 33–46. [CrossRef]
47. Maignan, I.; Lukas, B.A. Entry mode decisions: The role of managers' mental models. *J. Glob. Mark.* **1997**, *10*, 7–22. [CrossRef]

48. Brouthers, K.D. The influence of international risk on entry mode strategy in the computer software industry. *Manag. Int. Rev.* **1995**, *35*, 7–28.
49. Garnsey, E.; Stam, E.; Heffernan, P. New firm growth: Exploring processes and paths. *Ind. Innov.* **2006**, *13*, 1–20. [CrossRef]
50. Brown, R.L. Assessing specific mediational effects in complex theoretical models. *Struct. Equ. Model. A Multidiscip. J.* **1997**, *4*, 142–156. [CrossRef]

sustainability

MDPI

Article

Knowledge Creation Process and Sustainable Competitive Advantage: the Role of Technological Innovation Capabilities

Chuanpeng Yu [1], Zhengang Zhang [2], Chunpei Lin [3] and Yenchun Jim Wu [4,5,*]

[1] School of Economics and Commerce, South China University of Technology, Guangzhou 510006, China; yucp2015@scut.edu.cn

[2] School of Business Administration, South China University of Technology, Guangzhou 510641, China; adgzhang@scut.edu.cn

[3] Business Management Research Center, School of Business Administration, Huaqiao University, Quanzhou 362021, China; alchemist@hqu.edu.cn

[4] Graduate Institute of Global Business and Strategy, National Taiwan Normal University, Taipei 10645, Taiwan

[5] College of Innovation and Entrepreneurship, National Taipei University of Education, Taipei 10671, Taiwan

[*] Correspondence: wuyenchun@gmail.com; Tel.: +886-2-7734-3996

Received: 25 November 2017; Accepted: 7 December 2017; Published: 11 December 2017

Abstract: This study examines the relationship between the knowledge creation process and technological innovation capabilities, and analyzes their effect on a firm's sustainable competitive advantage using a knowledge-based view theoretical framework. We conduct structural equation modeling analyses using survey data from 315 Chinese industrial firms to test the direct and indirect effects of the knowledge creation process on sustainable competitive advantage. Technological innovation capabilities—operationalized to reflect the dimensions of process innovation capability and product innovation capability—are used as the mediating variable for explaining the relationship between the knowledge creation process and sustainable competitive advantage. The results indicate that the knowledge creation process does not have a significant direct effect on sustainable competitive advantage. Rather, the knowledge creation process can only influence the sustainable competitive advantage through the mediating effect of technological innovation capabilities completely. Consequently, the knowledge creation process favors the development of technological innovation capabilities for processes and products, because processes and products can lead to a sustainable competitive advantage.

Keywords: knowledge creation process; process innovation capability; product innovation capability; sustainable competitive advantage

1. Introduction

The question of what constitutes a sustainable competitive advantage (SCA) is a primary topic in current business strategy management research, and the knowledge-based view (KBV) of the source of SCA has received substantial attention. Indeed, as global competition, environmental turbulence, and the knowledge economy continue to grow, knowledge is deemed a strategic asset that enables businesses to develop a competitive edge [1]. To achieve and maintain competitiveness and sustainable growth, companies must constantly absorb existing knowledge, create new knowledge, and pursue practical wisdom [2]. Moreover, the rapid development of information technologies, such as big data and cloud computing, has promoted an exponential accumulation of social knowledge and knowledge stock; thus, being able to tap into a wealth of knowledge and information and create knowledge resources that can foster a company's development has become central to business competitiveness [3].

Numerous arguments have been proposed as to the role of a company's knowledge creation in its competitive advantage. Penrose first formulated the KBV of business competitive advantage, suggesting that the interaction between productive services and knowledge creation underlies a company's development [4]. Prahalad and Hamel attributed the source of a company's SCA to its core competencies, and defined core competencies as "the company's collective knowledge about how to coordinate diverse production skills and technologies" [5]. Leonard-Barton indicated that the knowledge a company accumulates throughout its development, particularly the tacit knowledge that is difficult for its rivals to replicate, is essential to its SCA [6]. Grant maintained that knowledge creation and utilization are essential to businesses; compared with their markets, businesses can create and expand knowledge more effectively to facilitate the application and protection of their intellectual property; business competitive advantage is based on the capability to establish a system for generating and protecting knowledge resources; and such a capability facilitates the integration of knowledge and other resources to yield economic rents that are larger than average profits [7].

Although knowledge creation has been widely recognized as the key to SCA, few studies have discussed, at the micro level, the mechanism by which a company develops this advantage when it creates knowledge [8]. This literature gap has two possible explanations. First, whether knowledge creation always leads to a SCA lacks robust evidence. Given the overriding focus of a business on profit generation, the outcomes of knowledge creation as a business behavior should be examined through input–output comparisons. Second, a SCA represents the continued competitiveness of a product or service in the market [9]. From a resource-advantage perspective, knowledge is characteristically tacit, non-dynamic, and is difficult to transfer or disseminate [7]. Therefore, businesses should create knowledge to produce innovations, such as technologies and products, thereby translating their knowledge resources into SCAs [10]. In summary, research on the mechanism in the relationship between knowledge creation and SCA formation requires greater rigor and greater depth. On the basis of recent studies, the present study used technological innovation theories to investigate how the knowledge creation process (KCP) affects SCA through process and product innovation capabilities. Employing a mediation approach offers the benefits of a more comprehensive understanding of the mechanism underlying the KCP—SCA relationship. Specifically, we firstly investigate how KCP promotes firms' technological innovation capability (technological IC) and SCA directly. Secondly, we explore the relationship between firm's technological IC and SCA. Then, we further examine whether technological IC mediates the relationship between KCP and SCA. The remainder of the paper is structured as follows: Section 2 considers the theoretical background and sets out the study hypotheses; Section 3 details the research methods; Section 4 presents the analysis and results of the empirical study; and Section 5 presents the discussion and conclusions.

2. Theoretical Background and Research Hypotheses

Theories on knowledge creation are largely applied to explicate the formation processes and outcomes of various types of knowledge within an organization [11]. Theoretically, models can be constructed to explore knowledge creation at an individual, group, or organizational level. Learning model types employed at the individual level are predominantly single or double-loop [12], which illustrate the criteria required to facilitate an individual's knowledge development. Generally, research into knowledge creation at the group level focuses on knowledge development and conversion, and argues for a close relationship between individual and organizational learning [13]. Studies on knowledge creation at the organizational level have suggested that we have at least two perspectives to understand the process of knowledge creation in organizations: the internal view and the ecosystem view [14]. Research studies that focused on the internal view emphasized that new knowledge begins with intuitive metaphors that link contradictory concepts [15], and may be created when prior knowledge is shared and transferred among members of an organization [16]. Thus, the internal view highlighted the crucial role of the creative and absorptive capacity of individuals within the organization [17]. Researchers that focused on the ecosystem view argued that a singular organization is embedded in an ecosystem [14], and much knowledge creation happens between organizations or industries, or by networks of organizations including suppliers, users, competitors, universities, public research centers, etc. [18,19]. Firms have increasingly co-created knowledge with external stakeholders during the innovation process to expand their knowledge base [20], and interorganizational learning is the key to the accomplishment of knowledge co-creation process [8]. Our goals in this article are to clarify the mechanisms through which firms can establish SCA through KCP, and we adopt the opinion that a firm's SCA is mainly determined by the endogenous force from resources and capabilities [9]. Thus, our work will focus only on the internal dimension of knowledge creation. Regarding research studies from the internal perspective, the SECI model, which defines KCP as a spiral process of socialization, externalization, combination, and internalization, developed by Nonaka and Takeuchi is the most widely used model for analyzing knowledge creation [21]. Focusing on the conversion between tacit and explicit knowledge, this model divides knowledge creation into the processes of socialization, externalization, combination, and internalization, suggesting that businesses should promote conversion between their tacit and explicit knowledge to foster their innovation and development [16,21].

The SECI model defines the ranges of tacit and explicit knowledge to identify the mechanisms underlying knowledge creation. Explicit knowledge refers to knowledge that can be articulated using formal and systematic language and shared in the form of statistics, scientific formulae, specifications, and manuals [22,23]. It can be readily processed, communicated, and stored. Tacit knowledge is highly personal and difficult to codify and communicate; it is acquired through experience. Moreover, such knowledge is disseminated through activities under certain conditions and locations; acquired through observation, imitation, and practice; communicated through apprentice training and face-to-face interaction; and transferred through the movement of individuals between organizations [23]. Subjective perception, intuition, and instinct are encompassed by tacit knowledge [24,25].

The SECI model of knowledge creation is illustrated as follows. Socialization refers to the conversion between and the dissemination of tacit and explicit knowledge within an organization (i.e., the process of sharing experiences to create tacit knowledge, such as shared mental models and technical skills) [21]. Externalization is the conversion of tacit knowledge to explicit knowledge or the articulation of tacit knowledge into explicit concepts; it is the essence of knowledge creation that involves using metaphors, analogy, concepts, hypotheses, or models to explicate tacit knowledge [24]. Combination is the conversion from explicit to tacit knowledge or the organization of concepts into a knowledge system [24]. This process occurs primarily through formal education and training. Internationalization is the conversion from explicit to tacit knowledge or the embodiment of explicit knowledge into tacit knowledge [25]. The current study adopted the SECI model to depict knowledge creation within organizations and explore the role of the model in SCA.

2.1. KCP and Technological IC

Among numerous classifications of innovations, one of the most commonly accepted is that of the OECD (Organization for Economic Co-operation and Development) in the Oslo Manual [26], which distinguishes four types of innovation: process, product, marketing, and organizational [27]. Technological innovation involves process and product innovations, whereas non-technological innovation involves marketing and organizational innovations [26]. This paper focuses on the two types of innovation that the OECD considered technological. The first, technological innovation capability for product (Product IC), refers to a firm's capability to create, design, and develop new products to satisfy customer needs [27]. Depending on its novelty, a product innovation can be either incremental or radical. Incremental innovation refers to a product with slightly altered technology, functionality, and appearance, whereas radical innovation refers to a product characterized by thorough, innovative, and distinct technical alterations [28]. The second, technological innovation capability for process (Process IC), is the ability to improve product or work processes through technical advances [29].

The KBV states that a business is a "unique sum of heterogeneous knowledge; its primary function is to create, integrate, and utilize knowledge; and communication and interaction outside and inside the business is characteristically 'a process of knowledge flow'" [7]. For businesses, knowledge is fundamental for undertaking technological innovation, expanding the scope of knowledge integration, and improving the ability to create knowledge; this is because knowledge contributes significantly to the improvement and technical level of products [30]. The existence of a business depends on whether it can efficiently create, apply, and commercialize knowledge related to technologies and markets [31]. Knowledge creation can be seen as explicating the knowledge of individuals in an organization into group knowledge. Therefore, knowledge creation at the individual level underpins that at the organizational level; exchanging and sharing knowledge within an organization promotes the explication, transmission, and integration of tacit knowledge within the organization. These processes internalize newly created knowledge in individual employees, thereby completing the cycle of knowledge creation [11,21]. From the perspective of knowledge stock, a business capable of creating knowledge can constantly generate the knowledge resources required to upgrade its processes and products [20,32]. This is particularly true in increasingly competitive industries, where businesses that exchange and integrate knowledge more frequently and are more capable of knowledge creation are more efficient at research and development, as well as innovating their products more effectively [33]. From the perspective of knowledge flow, knowledge creation by an organization is based on the organization's pursuit of creativity and innovation, which encourages its employees to create knowledge, promotes knowledge exchange between employees and across teams, and fosters new ideas and solutions to reduce redundant knowledge and increase non-redundant and heterogeneous knowledge resources that are conducive to product and process innovation [34,35]. Thus, we believe that a firm's KCP is positively related to its technological innovation capability. On this basis, we proposed the following two hypotheses:

Hypotheses 1a (H1a). *A positive relationship exists between a firm's KCP and its development of technological IC for process.*

Hypotheses 1b (H1b). *The effect of a firm's KCP on technological IC for product is mediated partially by its generation of technological IC for process.*

2.2. KCP, Technological IC, and SCA

Numerous studies have suggested that knowledge creation is essential to businesses gaining a SCA [24,36]. As globalization continues to intensify, numerous businesses have become knowledge-intensive, and now compete with "brains" not "brawn"; against this backdrop, knowledge is deemed the most crucial factor in distinguishing a company from its competitors [37]. Therefore, businesses with greater means to acquire knowledge and a greater ability to integrate and create knowledge are more efficient at identifying and responding to rapid changes in the market and resolving the limitations of their knowledge resources to outclass rivals [36]. On this basis, knowledge creation plays a crucial role in enhancing a SCA [38]. Knowledge creation can be seen as a process through which knowledge is constantly transferred and integrated among businesses, functional departments, and individuals, or a process involving repeated conversions of tacit and explicit knowledge [21,25]. This process leads to new knowledge resources, which include new approaches to solving problems and boosting performance, new work methods, new products, new concepts, and new lines of thinking [39]. These knowledge resources enable a company to improve its efficiency, reduce its costs, or refine its products. Through this, the company improves its ability to create value for its customers, serve customer needs, and increase employee and customer satisfaction [38], thereby developing a competitive edge [40]. Thus, we believe that a firm's KCP is positively related to its SCA. On this basis, we proposed the following hypothesis:

Hypotheses 2 (H2). *A positive relationship exists between a firm's KCP and its SCA.*

A SCA is the ultimate embodiment of a business's capabilities, resources, and activities, and it is a crucial criterion for whether the business allocates its resources appropriately and what outcomes it achieves accordingly. Researchers have argued that a business that is unable to continually innovate cannot operate in an increasingly competitive market, and will consequently lose its competitive advantage [41]. Additionally, with the growing penetration of big data and the Internet, knowledge is rapidly and frequently being replaced and updated; changes and developments occur every day, and business operators should constantly formulate or refine strategies to develop products or services and acquire and maintain a competitive edge [38]. From the perspective of business operation, innovating technologically—particularly through process innovation to introduce and implement "lean production" and "total quality management"—can improve production and operating efficiencies and substantially reduce costs [42]. From the perspective of target markets, technological innovation allows a business to satisfy not only customer needs for its existing products and services, but also new customer needs. Moreover, businesses that are highly capable of technological innovation are more likely to push beyond the boundaries of their capabilities and markets to identify new markets and capture the opportunities that they afford [40]. From the perspective of business models, such businesses tend to enter unfamiliar domains to create and commercialize products; the services that they postulate are groundbreaking, and the businesses pilot and promote them in existing markets [43]. Reducing costs, exploring opportunities in new markets, and providing new products or services are all outcomes of SCAs. Thus, we believe that a firm's technological innovation capability is positively related to SCA. On this basis, we proposed the following two hypotheses:

Hypotheses 3a (H3a). *A positive relationship exists between technological IC for process and SCA.*

Hypotheses 3b (H3b). *A positive relationship exists between technological IC for product and SCA.*

Process innovation was demonstrated to exert a significant direct influence on product innovation [27]. Improving process innovation capability, particularly through the innovation and optimization of the product development process, enables a company to expedite its product research and development, reduce its research and development (R&D) costs, and enhance its capability to innovate products. For example, the efficiency and quality of product innovation can be further improved by implementing product development in an integrated rather than stage-gate manner. Thus, we believe that process innovation capability is positively related to product innovation capability. On this basis, we proposed the following hypothesis:

Hypotheses 3c (H3c). *The effect of technological IC for process on SCA is mediated partially by the development of technological IC for product.*

The conceptual model of this research is shown in Figure 1.

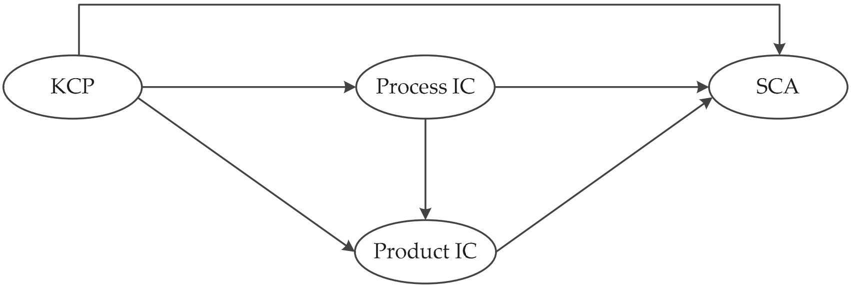


Figure 1. Conceptual model.

3. Research Methods

3.1. Data Collection and the Sample

To verify the hypotheses, we applied a questionnaire survey to collect data. Since the technological innovations are actually deployed in manufacturing industries, manufacturing firms in China were chosen as the research setting. China is a vast country that encompasses a wide range of regions. Different regions have different cultures, government policies, and locational conditions. In order to reduce the influences of these situational factors on the research results, we strategically selected a typical manufacturing region representing the new economic development stage in China, which is the Pearl River Delta (in southern China).

The survey was carried out between 1 July and 30 November 2016. The research participants were employees in firms in the industries of communication and computer-related equipment, electrical machinery and equipment, machinery and engineering, instruments and related products, metal products, and so on. The data collection procedure included three phases. Firstly, we developed the original English questionnaire based on previous studies, and translated it into Chinese using collaborative and iterative translation. Three management scholars with rich research experience in the knowledge and innovation management research field translated the questionnaire into Chinese. Then, we implemented two preliminary assessments to refine the item wording of the Chinese questionnaire. Three manufacturing managers and three professors reviewed the pre-questionnaire, and resolved any unfamiliar or unclear wording to improve clarity and identify. Next, we conducted a pre-test in six manufacturing firms. Based on the feedback, we detected any possible misunderstandings caused by the translation and further modified the questionnaire to make sure that the questionnaire was understandable and relevant to practices in China.

Secondly, we selected 1000 firms randomly from a list of manufacturing firms provided by the science and technology service departments of local government as our sampling frame. Following the suggestion of Frohlich [44], selected firms were contacted in advance to identify the key respondents. In order to ensure the reliability of the data regarding firms' knowledge creation processes, technological innovation capabilities, and sustainable competitive advantage, one respondent who was familiar with these activities (e.g., the top management team member, the manager of a manufacturing or R&D department, or a leader of process and product research projects) was chosen as the key respondent.

Thirdly, email or online surveys were sent out to the respondents with a cover letter that briefly introduced the objective, outlined the study, and ensured confidentiality. To encourage participation, we also offered a summary report of the study's conclusions to each respondent. In total, 343 questionnaires were collected. After deleting the responses with missing data, we received 312 valid questionnaires, yielding an effective response rate of 31.2%. The detailed characteristics of sampled firms are shown in Table 1, indicating a wide variety of sizes and industries.

We conducted several multivariate analysis of variance (MANOVA) tests to investigate the potential non-response bias [45]. An analysis of differences between early and late responses for all of the variables indicated no statistical differences, suggesting that non-response bias was not a major concern in our study.

Common method variance (CMV) was a concern in this study, as each questionnaire was finished by a single respondent [46]. We tried to reduce the potential influence of CMV by carefully selecting scale items and separating them within the fairly lengthy questionnaire. Then, two diagnostic tests were conducted to further evaluate the possibility of CMV. First, Harman's single-factor test was conducted [46]. The results revealed that no single factor emerged, and the first factor only accounted for 24.83% of the total 76.6% explained variance, indicating that CMV was not a serious concern. Secondly, we conducted confirmatory factor analysis (CFA) on Harman's single-factor model [47]. The model's fit indices of $\chi^2/\text{df} = 8.483$, CFI = 0.456, GFI = 0.412, TLI = 0.425, IFI = 0.458 and RMSEA = 0.155 were unacceptable, as they were considerably worse than those of the measurement model. This suggested that the single-factor model was not acceptable, further indicating that CMV was not a serious issue.

Table 1. Profile of sampled firms.

Characteristics of Firms	Frequency	Percentage (%)
Industry		
Communication and computer-related equipment	62	19.87
Electrical machinery and equipment	56	17.95
Machinery and engineering	50	16.03
Instruments and related products	45	14.42
Metal products	48	15.38
Others	51	16.35
Firm age		
1–5 years	66	21.15
6–10 years	69	22.12
11–15 years	115	36.86
>15 years	62	19.87
Number of employees		
Large size (>1000)	101	32.37
Medium size (300–1000)	145	46.48
Small size (<300)	66	21.15
Annual sales (million RMB)		
Large size (>400)	116	37.18
Medium size (20–400)	142	45.51
Small size (<20)	54	17.31
Ownership		
State-owned	97	31.09
Private-owned	180	57.69
Foreign-owned	35	11.22

Note: n = 312; we defined large, medium, and small sizes according to the standards issued by China's Ministry of Industry and Information Technology.

3.2. Variables and Measures

The questionnaire in this study consisted of four construct measurements: knowledge creation process, process innovation capabilities, product innovation capabilities, and sustainable competitive advantage. All of the measures were adapted from existing scales found in previous studies. The survey was a "tick the box" survey, and all of the items were measured using a seven-point Likert scale that ranged from "strongly disagree" (1) to "strongly agree" (7). The measurement items of all of the constructs are presented in Appendix A Table A1.

3.2.1. Knowledge Creation Process

The knowledge creation process has been conceptualized as a multidimensional construct in prior studies [21,25,48]. Consistent with these previous research studies, we measured the knowledge creation process in four dimensions: socialization, externalization, combination, and internalization. These four dimensions have four, five, four, and three items respectively, and were measured using scales adapted from previous work [48]. The socialization process converts individuals' tacit knowledge into new tacit knowledge through shared experiences and joint activities [21] such as cooperative projects across directorates, employee rotation across areas, etc. The externalization process articulates the tacit knowledge into comprehensible forms that are more understandable to others [25] through adopting various tools such a problem-solving system, collaboration learning tools, etc. The combination process converts explicit knowledge collected from inside or outside the organization into more complex and systematic explicit knowledge [24] through using web pages, databases, etc. The internalization process transfers explicit knowledge into tacit knowledge [48] through on-the-job training, learning by doing, and learning by observation.

3.2.2. Technological Innovation Capabilities

According to previous literature [26,49], "technological innovation capabilities" have also been conceptualized as a multidimensional construct, which includes two dimensions: process innovation capability and product innovation capability. Process innovation capability is defined as a firm's ability to develop new or significantly changed productive and technological processes. The measurement scale used in prior work [27] consisted of 11 reflective items, which assessed the extent to which process innovation capability constitutes a particular strength for the firm in comparison with its main competitors. Product innovation capability is defined as a firm's ability to develop new or significantly improved products [26,49,50]. According to prior work [27], the measurement scale was made up by five reflective items, which assessed the extent to which product innovation capability constitutes a particular strength for the firm in comparison with its main competitors.

3.2.3. Sustainable Competitive Advantage

Wiggins and Ruefli [51] defined sustainable competitive advantage as a firm's capability to achieve a series of temporary advantages over time. The measurement of sustainable competitive advantage contained six items [52], which assessed respondents' perceptions of the extent to which their firm performed competitively in various fields (i.e., R&D, managerial capability, profitability, etc.) in comparison with its main competitors.

3.2.4. Control Variables

Previous studies have suggested that a firm's innovation and competitive advantage may be influenced by firm age, firm size, annual sales, ownership, and environmental uncertainty [27,53]. Accordingly, we included these control variables in the study. Firm age was measured using a four-point Likert scale according to the time since the firm was established. Firm size was measured using a three-point Likert scale according to the firm's number of employees. Annual sales were measured using a three-point Likert scale according to the firm's total revenue last year. Ownership was operationalized as three dummy variables, with state-owned as the baseline. Environmental uncertainty was measured by a three-item scale used in prior work [54].

3.3. Reliability and Validity

The reliability of the multi-item scale for each dimension was assessed using Cronbach's α coefficient and composite reliability (CR). Table 2 showed each construct's Cronbach's α and CR values. The Cronbach's α values of all of the constructs ranged from 0.833 to 0.957, exceeding the recommended minimum standard of 0.70 [55]. All of the CR values were larger than 0.85, which is greater than the minimum acceptable value of 0.7. So, the reliability of the measurement in this study is acceptable.

Rigorous processes were employed to evaluate the validity of this study. First, in order to ensure the content validity, we carefully extracted scales from existing constructs based on an extensive search of the literature in our study. Moreover, we conducted several iterative reviews of the questionnaire by executives and academics to clarify the item wording. Content validity was such established.

Secondly, confirmatory factor analysis (CFA) was executed to assess the convergent validity [56]. The model fit indices were as follows: $\chi^2 = 730.689$, degree of freedom (df) = 646, $p < 0.05$; $\chi^2/\text{df} = 1.131$; comparative fit index (CFI) = 0.990; Tucker–Lewis index (TLI) = 0.989; incremental fit index (IFI) = 0.990; root mean square error of approximation (RMSEA) = 0.021. All of the indices were above the minimum acceptable values, and all of the factor loadings were higher than 0.70 and significant at the $p < 0.001$ level (see Table 2), indicating strong convergent validity [57]. In addition, the average variance extracted (AVE) values of all of the constructs exceeded the minimum threshold of 0.50 (see Table 2), as advocated by Fornell and Larcker [55], which supports the convergent validity of the measures.

Table 2. Standardized item loadings, Cronbach's α, composite reliability (CR) and average variance extracted (AVE) values.

Constructs	Items	λ	Cronbach's α	CR	AVE
Socialization (SOC)	SOC1	0.821	0.900	0.902	0.697
	SOC2	0.837 ***			
	SOC3	0.870 ***			
	SOC4	0.809 ***			
Externalization (EXT)	EXT1	0.837	0.912	0.912	0.676
	EXT2	0.855 ***			
	EXT3	0.782 ***			
	EXT4	0.818 ***			
	EXT5	0.818 ***			
Combination (COM)	COM1	0.770	0.855	0.856	0.598
	COM2	0.715 ***			
	COM3	0.812 ***			
	COM4	0.793 ***			
Internalization (INT)	INT1	0.851	0.833	0.835	0.629
	INT2	0.706 ***			
	INT3	0.815 ***			
Process innovation capability (Process IC)	Process IC1	0.873	0.957	0.958	0.673
	Process IC2	0.846 ***			
	Process IC3	0.769 ***			
	Process IC4	0.846 ***			
	Process IC5	0.790 ***			
	Process IC6	0.811 ***			
	Process IC7	0.801 ***			
	Process IC8	0.809 ***			
	Process IC9	0.847 ***			
	Process IC10	0.813 ***			
	Process IC11	0.811 ***			
Product innovation capability (Product IC)	Product IC1	0.844	0.930	0.931	0.729
	Product IC2	0.868 ***			
	Product IC3	0.864 ***			
	Product IC4	0.835 ***			
	Product IC5	0.857 ***			
Sustainable competitive advantage (SCA)	SCA1	0.837	0.934	0.934	0.703
	SCA2	0.801 ***			
	SCA3	0.823 ***			
	SCA4	0.835 ***			
	SCA5	0.856 ***			
	SCA6	0.877 ***			

Note: *** $p < 0.001$.

In order to assess the discriminant validity, we calculated each construct's square root of AVE and compared them with correlations between pairs of constructs [55]. As shown in Table 3, the results indicated that the square root of each construct's AVE value was higher than its correlation with any other construct. Therefore, the discriminant validity was established in this study. Based on the above results, the reliability and validity of the measurements in this study are acceptable.

Table 3. Discriminant validity test.

Variables	Mean	SD	SOC	EXT	COM	INT	PCIC	PDIC	SCA
SOC	4.768	1.286	**0.835**						
EXT	4.747	1.257	0.193	**0.822**					
COM	4.660	1.163	0.118	0.116	**0.773**				
INT	4.736	1.350	0.239	0.104	0.122	**0.793**			
PCIC	4.618	1.147	0.094	0.046	0.226	0.228	**0.820**		
PDIC	4.391	1.380	0.095	0.005	0.176	0.135	0.404	**0.854**	
SCA	4.246	1.132	0.012	−0.028	0.170	0.025	0.376	0.420	**0.838**

Note: PCIC: process IC, PDIC: product IC, SD: standard deviation, bold numbers on the diagonal line represent the square root of AVE.

4. Results

We employed structural equation modeling (SEM), using AMOS 18.0, to test the hypotheses. A set of fit indices was used to examine the structural model, which supported a good model fit (χ^2 = 491.970, df = 404, $p < 0.01$; χ^2/df = 1.218; CFI = 0.986; GFI (goodness-of-fit index) = 0.909; TLI = 0.984; IFI = 0.986; RMSEA = 0.026). Table 4 presents the results of the SEM mediation analysis. We also demonstrate the results of the full model in Figure 2. Of the six hypotheses, only five hypotheses have been supported by the empirical data. Specifically, as can be observed in Table 4 and Figure 1, the KCP has a significant effect on process IC (β = 0.333, $p < 0.001$), and H1a is strongly supported. Also, process IC has a significant effect on product IC (β = 0.359, $p < 0.001$). However, the direct effect of KCP on product IC is not significant (β = 0.125, $p > 0.05$). These results confirm H1b, which posits an indirect effect of KCP on product IC through process IC.

With regard to SCA, the KCP does not have a significant effect on SCA ($\beta = -0.118$, $p > 0.05$), indicating that H2 is not supported. Thus, we find that KCP cannot influence SCA directly. Process IC has a significant effect on SCA (β = 0.257, $p < 0.001$), which gives support to H3a. Product IC also has a significant effect on SCA (β = 0.355, $p < 0.001$), supporting H3b. Considering the significant effect of process IC on product IC, therefore, as stated in H3c, the effect of process IC on SCA is mediated partially by the development of product IC. This means that process IC positively influences SCA directly, as well as positively influences it indirectly through Product IC. Based on the above results, we can conclude that KCPs can only foster SCA indirectly through technological innovation capabilities.

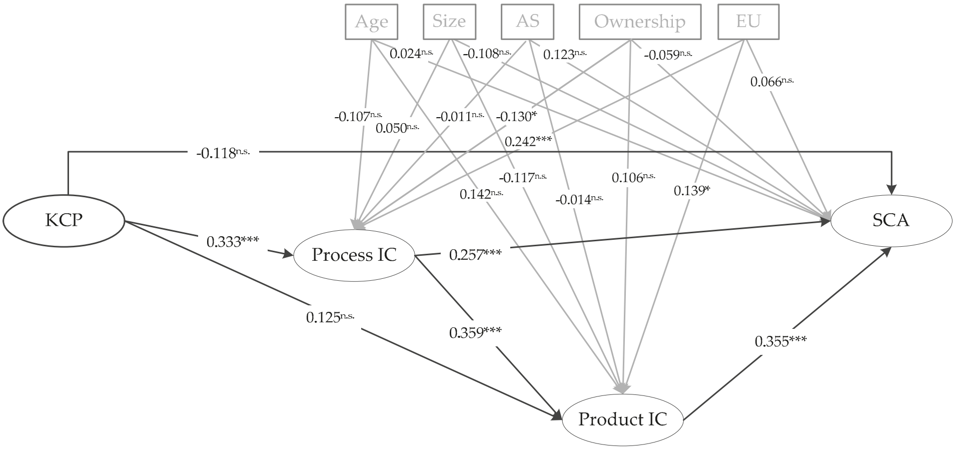


Figure 2. The results of the full model. Note: * $p < 0.05$, *** $p < 0.001$, $^{\text{n.s.}}$ non-significant, AS: annual sale, EU: environment uncertainty, KCP: knowledge creation process.

Table 4. Structural equation model results.

Structural Path	Proposed Effect	Path Coefficient	Results
Direct effects			
KCP→Process IC	+	0.333 ***	H1a supported
KCP→Product IC		0.125 n.s.	
Process IC→Product IC		0.359 ***	
KCP→SCA	+	−0.118 n.s.	H2 not supported
Process IC→SCA	+	0.257 ***	H3a supported
Product IC→SCA	+	0.355 ***	H3b supported
Indirect effects			
KCP→Process IC→Product IC	+	0.120	H1b supported
Process IC→Product IC→SCA	+	0.127	H3c supported

Table 4. *Cont.*

Structural Path	Proposed Effect	Path Coefficient	Results
Non-hypothesized (control variables)			
Age→Process IC		$-0.107^{\text{n.s.}}$	
Size→Process IC		$0.050^{\text{n.s.}}$	
Annual sale→Process IC		$-0.011^{\text{n.s.}}$	
Ownership→Process IC		$-0.130^{\text{n.s.}}$	
Uncertainty→Process IC		0.242 ***	
Age→Product IC		$0.142^{\text{n.s.}}$	
Size→Product IC		$-0.117^{\text{n.s.}}$	
Annual sale→Product IC		$-0.014^{\text{n.s.}}$	
Ownership→Product IC		$0.106^{\text{n.s.}}$	
Uncertainty→Product IC		0.139 *	
Age→SCA		$0.024^{\text{n.s.}}$	
Size→SCA		$-0.108^{\text{n.s.}}$	
Annual sale→SCA		$0.123^{\text{n.s.}}$	
Ownership→SCA		$-0.059^{\text{n.s.}}$	
Environment uncertainty→SCA		$0.066^{\text{n.s.}}$	

Note: * $p < 0.05$, *** $p < 0.001$, $^{\text{n.s.}}$ non-significant, + the path coefficient is positive.

5. Conclusions

5.1. Discussion

This study combined the KBV with a technological innovation perspective and experimentally confirmed that technological innovation capability mediated the influence of the KCP on SCA. On the basis of the OECD definition of technological innovation, technological innovation was divided into the dimensions of process and product innovation. This study discovered that technological innovation capability mediated the role of the KCP in the development of SCA. These findings provide valuable insights for manufacturing firms to foster their SCA through implementing KCPs and strengthening their technical innovation abilities.

Considering the direct effect of the KCP on SCA, our findings indicated that KCPs do not have a direct positively significant effect on SCA. This conclusion was inconsistent with our hypotheses, probably because, from the KBV, knowledge creation is the core task of a business and the knowledge created by members of a business through social interaction (particularly tacit knowledge, which is difficult to articulate, imitate, and disseminate) is a crucial source of SCA [7,21,58]. However, knowledge creation does not necessarily lead to stronger SCA. One immediate outcome of exploiting that advantage is that a business can respond to market changes effectively and develop products or services continually to satisfy customer needs, dominate market shares, and achieve a greater operational performance than its competitors [9,59]. This allows a business to acquire business value. For businesses, knowledge creation lays the groundwork for new technologies, new products, or new services, which help the firms achieve a positional advantage in obtaining SCA [60]. However, this does not mean that all of the companies can achieve a SCA successfully through creating new knowledge. Many businesses that create knowledge do not achieve successful commercialization; therefore, this study empirically verified that KCPs do not directly strengthen SCA. Two examples are cited to illustrate this point. Kodak, despite creating knowledge about digital photography and inventing digital photography techniques, did not benefit from its brainchild. Although Apple created knowledge about personal computers and inaugurated the world's first, it was IBM that first achieved commercial success in this domain.

The analysis of the relationships among the KCP, technological innovation capability, and SCA yielded the following findings. The KCP favors the development of process innovation capability, but doesn't directly positively affect product innovation capability. This result indicates that the relationship between KCP and product innovation capability is completely mediated by process innovation capability. In addition, the positive impact of KCP on SCA is not significant, but both process and product IC had a significant direct effect on SCA, as did process innovation capability on product innovation capability. These findings highlight that simply implementing a KCP is not sufficient to favor SCA; it would be necessary to put it through process and product innovation capability. Accordingly, the KCP encourages a business to improve its technological innovation capability, thereby gaining a SCA. This conclusion, which is derived from an empirical verification of the mechanism through which the KCP influences the SCA, contrasts with quantitative investigations on what underlies the relationship between the KCP and SCA [4,5,7].

Numerous studies have suggested that as global competition continues to intensify, the more knowledge businesses create, the faster they can identify trends in technological and market development and act accordingly to secure advantageous positions in the market [36,61]. Additionally, businesses that are more capable of knowledge creation typically wield stronger SCAs; therefore, academic research has often found knowledge creation and SCA to be positively related. However, their statistically proven correlation does not accurately explain how knowledge creation, which leads to new knowledge resources (such as product concepts, manners of work, and lines of thinking), translates into the SCA. This study argued that only when process and product IC are improved do these knowledge resources influence SCAs. This conclusion contributes theoretically to the understanding of how businesses acquire SCAs through knowledge creation.

5.2. Managerial Implications

The findings of this study also provide crucial implications for manufacturing firms to carry out KCPs to acquire a SCA more effectively. First, managers should devote themselves to transforming their firms into learning organizations, which can drive firms to become more efficient at acquiring, integrating, and creating knowledge, thereby producing knowledge continually. A learning organization is an organization that allows its members to adopt effective learning mechanisms to create knowledge and values. Enhancing organizational learning is instrumental in building a learning organization. Managers can advance organizational learning through establishing a learning platform, fostering a learning culture, implementing an incentive mechanism, encouraging employees to acquire knowledge outside the organization, etc. In addition, managers have to promote dynamics and spirals of knowledge creation by taking a leading role in managing the SECI process. Managers can nurture an enabling environment that encourages employees to communicate with each other (to facilitate the exchange of tacit and explicit knowledge within the organization) and provide various types of trainings on learning methods (such as acquisitive learning, experiential learning, and learning by doing) to promote the acquisition, integration, and creation of knowledge by employees and improve the organization's capability in creating knowledge.

Second, the results of this paper indicate that the KCP does not directly strengthen SCA, but can strengthen SCA through the mediation of technological innovation capability. Therefore, the most important practical implication of this paper is that managers should be aware of the importance of the KCP in the link between technological IC and SCA. Our findings indicate that merely concentrating on the KCP may not be sufficient, despite the commonly held view that knowledge is the most important strategic asset that enables firms to develop a SCA [9]. Technological IC seems more important in helping firms obtain a SCA. This is consistent with the research that argues that firms stand to benefit from investing in their technological innovation capabilities for products and processes that generate the most valuable, distinctive, and difficult to imitate strategic assets that allow the firms to achieve superior performance [27]. Thus, while focusing on locating, capturing, transferring, sharing existing knowledge, and creating new knowledge, managers should also concentrate on providing spaces and opportunities for individuals or teams to engage in improving technological IC at the same time.

5.3. Limitations and Future Research

This study has two limitations. First, a convenience sample of manufacturing firms in the Pearl River Delta region of China was used, which limited the generalizability of the findings. China is a vast country in which the development of businesses is subject to local political, cultural, and resource contexts. Future studies should investigate larger samples to further generalize their findings. Second, this study employed cross-sectional data to analyze the influence of knowledge creation on technological innovation capability and SCA, despite the relationship being a dynamic process. Third, concrete outcomes are difficult to observe within a short period of time. Thus, longitudinal research should be conducted in order to investigate knowledge creation and its results in businesses; this would improve the understanding of how knowledge creation contributes to technological innovation capability.

Acknowledgments: This study was supported financially by the Humanities and Social Science Foundation Supported by Ministry of Education in China (16YJC630160), Natural Science Foundation of Guangdong Province in China (2016A030310419), Fund for Science and Technology Projects of Guangdong Province in China (2016B070702001, 2017A070707001), and the Fundamental Research Funds for the Central Universities in China (2017BQ079).

Author Contributions: Chuanpeng Yu and Yenchun Jim Wu conceived and designed the experiments; Chunpei Lin performed the experiments; Chuanpeng Yu and Chunpei Lin analyzed the data; Zhengang Zhang contributed reagents/materials/analysis tools; Chuanpeng Yu wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Survey instrument (with factor loadings).

Item	Loading
Socialization (α = 0.900)	
SOC1: My firm usually adopts cooperative projects across directorates	0.821
SOC2: My firm usually uses apprentices and mentors to transfer knowledge	0.837
SOC3: My firm usually adopts brainstorming retreats or camps	0.870
SOC4: My firm usually adopts employee rotation across areas	0.809
Externalization (α = 0.912)	
EXT1: My firm usually adopts a problem-solving system based on a technology like case-based reasoning	0.837
EXT2: My firm usually adopts groupware and other learn collaboration tools	0.855
EXT3: My firm usually adopts pointers to expertise	0.782
EXT4: My firm usually adopts modeling based on analogies and metaphors	0.818
EXT5: My firm usually captures and transfers experts' knowledge	0.818
Combination (α = 0.855)	
COM1: My firm usually adopts web-based access to data	0.770
COM2: My firm usually uses web pages	0.715
COM3: My firm usually uses databases	0.812
COM4: My firm usually adopts repositories of information, best practices, and lessons learned	0.793
Internalization (α = 0.833)	
INT1: My firm usually adopts on-the-job training	0.851
INT2: My firm usually adopts learning by doing	0.706
INT3: My firm usually adopts learning by observation	0.815
Process innovation capability (α = 0.957)	
Process IC1: My firm is able to create and manage a portfolio of interrelated technologies	0.873
Process IC2: My firm is able to master and absorb the basic and key technologies of business	0.846
Process IC3: My firm continually develops programs to reduce production costs	0.769
Process IC4: My firm has valuable knowledge for innovating manufacturing and technological processes	0.846
Process IC5: My firm has valuable knowledge on the best processes and systems for work organization	0.790
Process IC6: My firm organizes its production efficiently	0.811
Process IC7: My firm assigns resources to the production department efficiently	0.801
Process IC8: My firm is able to maintain a low level of stock without impairing service	0.809
Process IC9: My firm is able to offer environmentally friendly processes	0.847
Process IC10: My firm manages production organization efficiently	0.813
Process IC11: My firm is able to integrate production management activities	0.811
Product innovation capability (α = 0.930)	
Product IC1: My firm is able to replace obsolete products	0.844
Product IC2: My firm is able to extend the range of products	0.868
Product IC3: My firm is able to develop environmentally friendly products	0.864
Product IC4: My firm is able to improve product design	0.835
Product IC5: My firm is able to reduce the time to develop a new product until its launch in the market	0.857

Table A1. *Cont.*

Item	Loading
Sustainable competitive advantage (α = 0.934)	
SCA1: The quality of the products or services that my firm offers is better than that of the competitor's products or services	0.837
SCA2: My firm is more capable of R&D than the competitors	0.801
SCA3: My firm has better managerial capability than the competitors	0.823
SCA4: My firm's profitability is better	0.835
SCA5: The corporate image of my firm is better than that of the competitors	0.856
SCA6: The competitors are difficult to take the place of my firm's competitive advantage	0.877

References

1. Zheng, N.; Wei, Y.; Zhang, Y.; Yang, J. In search of strategic assets through cross-border merger and acquisitions: Evidence from Chinese multinational enterprises in developed economies. *Int. Bus. Rev.* **2016**, *25*, 177–186. [CrossRef]
2. Nonaka, I.; Kodama, M.; Hirose, A.; Kohlbacher, F. Dynamic fractal organizations for promoting knowledge-based transformation—A new paradigm for organizational theory. *Eur. Manag. J.* **2014**, *32*, 137–146. [CrossRef]
3. Parveen, F.; Jaafar, N.I.; Ainin, S. Social media's impact on organizational performance and entrepreneurial orientation in organizations. *Manag. Decis.* **2016**, *54*, 2208–2234. [CrossRef]
4. Penrose, E.T. *The Theory of the Growth of the Firm*; Oxford University Press: Oxford, UK, 2009; ISBN 978-0-19-957384-4.
5. Prahalad, C.K.; Hamel, G. The core competency of the corporation. *Harv. Bus. Rev.* **1990**, *68*, 79–91.
6. Leonard-Barton, D. Core capabilities and core rigidities: A paradox in managing new product development. *Strateg. Manag. J.* **1992**, *13*, 111–125. [CrossRef]
7. Grant, R. Toward a knowledge-based theory of the firm. *Strateg. Manag. J.* **1996**, *17*, 109–122. [CrossRef]
8. Kazadi, K.; Lievens, A.; Mahr, D. Stakeholder co-creation during the innovation process: Identifying capabilities for knowledge creation among multiple stakeholders. *J. Bus. Res.* **2016**, *69*, 525–540. [CrossRef]
9. Huang, K.F.; Dyerson, R.; Wu, L.Y.; Harindranath, G. From temporary competitive advantage to sustainable competitive advantage. *Br. J. Manag.* **2015**, *26*, 617–636. [CrossRef]
10. Centobelli, P.; Cerchione, R.; Esposito, E. Knowledge management in startups: Systematic literature review and future research agenda. *Sustainability* **2017**, *9*, 361. [CrossRef]
11. Ben-Menahem, S.M.; von Krogh, G.; Erden, Z.; Schneider, A. Coordinating knowledge creation in multidisciplinary teams: Evidence from early-stage drug discovery. *Acad. Manag. J.* **2016**, *59*, 1308–1338. [CrossRef]
12. Argyris, C. Single-loop and double-loop models in research on decision making. *Admin. Sci. Q.* **1976**, *21*, 363–375. [CrossRef]
13. Chandrasekaran, A.; Linderman, K. Managing knowledge creation in high-tech R&D projects: A multimethod study. *Decis. Sci.* **2015**, *46*, 267–300. [CrossRef]
14. Peschl, M.F.; Fundneider, T. Designing and enabling spaces for collaborative knowledge creation and innovation: From managing to enabling innovation as socio-epistemological technology. *Comput. Hum. Behav.* **2014**, *37*, 346–359. [CrossRef]
15. Teece, D.J. Nonaka's contribution to the understanding of knowledge creation, codification and capture. In *Towards Organizational Knowledge*; Palgrave Macmillan: London, UK, 2013; pp. 17–23.
16. Nonaka, I.; von Krogh, G. Perspective—Tacit knowledge and knowledge conversion: Controversy and advancement in organizational knowledge creation theory. *Organ. Sci.* **2009**, *20*, 635–652. [CrossRef]
17. Fabrizi, A.; Guarini, G.; Meliciani, V. Public knowledge partnerships in European research projects and knowledge creation across R&D institutional sectors. *Technol. Anal. Strateg. Manag.* **2016**, *28*, 1056–1072. [CrossRef]
18. Tseng, C.Y.; Pai, D.C. Knowledge search, spillover and creation capability in India's pharmaceutical industry. *Technol. Anal. Strateg. Manag.* **2014**, *26*, 207–222. [CrossRef]
19. Galati, F.; Bigliardi, B. Does different NPD project's characteristics lead to the establishment of different NPD networks? A knowledge perspective. *Technol. Anal. Strateg. Manag.* **2017**, *29*, 1196–1209. [CrossRef]

20. Mahr, D.; Lievens, A.; Blazevic, V. The value of customer cocreated knowledge during the innovation process. *J. Prod. Innov. Manag.* **2014**, *31*, 599–615. [CrossRef]
21. Nonaka, I.; Takeuchi, H. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*; Oxford University Press: Oxford, UK, 1995; ISBN 978-0195092691.
22. Schoenherr, T.; Griffith, D.A.; Chandra, A. Knowledge management in supply chains: The role of explicit and tacit knowledge. *J. Bus. Logist.* **2014**, *35*, 121–135. [CrossRef]
23. Park, C.; Vertinsky, I.; Becerra, M. Transfers of tacit vs. explicit knowledge and performance in international joint ventures: The role of age. *Int. Bus. Rev.* **2015**, *24*, 89–101. [CrossRef]
24. Nonaka, I.; Toyama, R.; Konno, N. SECI, *Ba* and leadership: A unified model of dynamic knowledge creation. *Long Range Plan.* **2000**, *33*, 5–34. [CrossRef]
25. Nonaka, I.; Toyama, R. The theory of the knowledge-creating firm: Subjectivity, objectivity and synthesis. *Ind. Corp. Chang.* **2005**, *14*, 419–436. [CrossRef]
26. OECD. *The Measurement of Scientific and Technological Activities: Proposed Guidelines for Collecting and Interpreting Innovation Data: Oslo Manual*, 3rd ed.; OECD Eurostat: Paris, France, 2005.
27. Camisón, C.; Villar-López, A. Organizational innovation as an enabler of technological innovation capabilities and firm performance. *J. Bus. Res.* **2014**, *67*, 2891–2902. [CrossRef]
28. Chatterji, A.K.; Fabrizio, K.R. Using users: When does external knowledge enhance corporate product innovation? *Strateg. Manag. J.* **2014**, *35*, 1427–1445. [CrossRef]
29. Hervas-Oliver, J.L.; Sempere-Ripoll, F.; Boronat-Moll, C. Process innovation strategy in SMEs, organizational innovation and performance: A misleading debate? *Small Bus. Econ.* **2014**, *43*, 873–886. [CrossRef]
30. Lichtenthaler, U. Toward an innovation-based perspective on company performance. *Manag. Decis.* **2016**, *54*, 66–87. [CrossRef]
31. Sears, J.; Hoetker, G. Technological overlap, technological capabilities, and resource recombination in technological acquisitions. *Strateg. Manag. J.* **2014**, *35*, 48–67. [CrossRef]
32. Brunswicker, S.; Vanhaverbeke, W. Open innovation in small and medium-sized enterprises (SMEs): External knowledge sourcing strategies and internal organizational facilitators. *J. Small Bus. Manag.* **2015**, *53*, 1241–1263. [CrossRef]
33. Martín-de Castro, G.; Delgado-Verde, M.; Navas-López, J.E.; Cruz-González, J. The moderating role of innovation culture in the relationship between knowledge assets and product innovation. *Technol. Forecast. Soc. Chang.* **2013**, *80*, 351–363. [CrossRef]
34. Chen, Y.S.; Chang, T.W.; Lin, C.Y.; Lai, P.Y.; Wang, K.H. The influence of proactive green innovation and reactive green innovation on green product development performance: The mediation role of green creativity. *Sustainability* **2016**, *8*, 966. [CrossRef]
35. Cuevas-Rodríguez, G.; Cabello-Medina, C.; Carmona-Lavado, A. Internal and external social capital for radical product innovation: Do they always work well together? *Br. J. Manag.* **2014**, *25*, 266–284. [CrossRef]
36. Tallman, S.; Jenkins, M.; Henry, N.; Pinch, S. Knowledge, clusters, and competitive advantage. *Acad. Manag. Rev.* **2004**, *29*, 258–271. [CrossRef]
37. Ahmad, A.; Bosua, R.; Scheepers, R. Protecting organizational competitive advantage: A knowledge leakage perspective. *Comput. Secur.* **2014**, *42*, 27–39. [CrossRef]
38. Vanpoucke, E.; Vereecke, A.; Wetzels, M. Developing supplier integration capabilities for sustainable competitive advantage: A dynamic capabilities approach. *J. Oper. Manag.* **2014**, *32*, 446–461. [CrossRef]
39. Crescenzi, R.; Nathan, M.; Rodríguez-Pose, A. Do inventors talk to strangers? On proximity and collaborative knowledge creation. *Res. Policy* **2016**, *45*, 177–194. [CrossRef]
40. Mao, H.; Liu, S.; Zhang, J.; Deng, Z. Information technology resource, knowledge management capability, and competitive advantage: The moderating role of resource commitment. *Int. J. Inf. Manag.* **2016**, *36*, 1062–1074. [CrossRef]
41. Urbancova, H. Competitive advantage achievement through innovation and knowledge. *J. Compet.* **2013**, *5*, 82–96. [CrossRef]
42. Piening, E.P.; Salge, T.O. Understanding the antecedents, contingencies, and performance implications of process innovation: A dynamic capabilities perspective. *J. Prod. Innov. Manag.* **2015**, *32*, 80–97. [CrossRef]
43. Carayannis, E.G.; Sindakis, S.; Walter, C. Business model innovation as lever of organizational sustainability. *J. Technol. Transf.* **2015**, *40*, 85–104. [CrossRef]

44. Frohlich, M.T. Techniques for improving response rates in OM survey research. *J. Oper. Manag.* **2002**, *20*, 53–62. [CrossRef]
45. Armstrong, J.; Overton, T. Estimating nonresponse bias in mail surveys. *J. Mark. Res.* **1977**, *14*, 396–402. [CrossRef]
46. Podsakoff, P.M.; MacKenzie, S.B.; Lee, J.Y.; Podsakoff, N.P. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *J. Appl. Psychol.* **2003**, *88*, 879–903. [CrossRef] [PubMed]
47. Sanchez, J.I.; Brock, P. Outcomes of perceived discrimination among hispanic employees: Is diversity management a luxury or a necessity? *Acad. Manag. J.* **1996**, *39*, 704–719. [CrossRef]
48. Sabherwal, R.; Becerra-Fernandez, I. An empirical study of the effect of knowledge management processes at individual, group, and organizational levels. *Decis. Sci.* **2003**, *34*, 225–260. [CrossRef]
49. Camisón, C.; Villar-López, A. An examination of the relationship between manufacturing flexibility and firm performance: The mediating role of innovation. *Int. J. Oper. Prod. Manag.* **2010**, *30*, 853–878. [CrossRef]
50. Menguc, B.; Auh, S. Development and return on execution of product innovation capabilities: The role of organizational structure. *Ind. Mark. Manag.* **2010**, *39*, 820–831. [CrossRef]
51. Wiggins, R.R.; Ruefli, T.W. Sustained competitive advantage: Temporal dynamics and the incidence and persistence of superior economic performance. *Organ. Sci.* **2002**, *13*, 81–105. [CrossRef]
52. Chang, C.H. The influence of corporate environmental ethics on competitive advantage: The mediation role of green innovation. *J. Bus. Ethics* **2011**, *104*, 361–370. [CrossRef]
53. Yu, X.; Chen, Y.; Nguyen, B.; Zhang, W. Ties with government, strategic capability, and organizational ambidexterity: Evidence from China's information communication technology industry. *Inf. Technol. Manag.* **2014**, *15*, 81–98. [CrossRef]
54. Carson, S.J.; Madhok, A.; Wu, T. Uncertainty, opportunism, and governance: The effects of volatility and ambiguity on formal and relational contracting. *Acad. Manag. J.* **2006**, *49*, 1058–1077. [CrossRef]
55. Fornell, C.; Larcker, D.F. Evaluating structural equation models with unobservable variables and measurement error. *J. Mark. Res.* **1981**, *18*, 39–50. [CrossRef]
56. O'Leary-Kelly, S.W.; Vokurka, R.J. The empirical assessment of construct validity. *J. Oper. Manag.* **1998**, *16*, 387–405. [CrossRef]
57. Anderson, J.C. An approach for confirmatory measurement and structural equation modeling of organizational properties. *Manag. Sci.* **1987**, *33*, 525–541. [CrossRef]
58. Kim, W.; Park, J. Examining structural relationships between work engagement, organizational procedural justice, knowledge sharing, and innovative work behavior for sustainable organizations. *Sustainability* **2017**, *9*, 205. [CrossRef]
59. Lazzarini, S.G. Strategizing by the government: Can industrial policy create firm-level competitive advantage? *Strateg. Manag. J.* **2015**, *36*, 97–112. [CrossRef]
60. Kim, N.; Im, S.; Slater, S.F. Impact of knowledge type and strategic orientation on new product creativity and advantage in high-technology firms. *J. Prod. Innov. Manag.* **2013**, *30*, 136–153. [CrossRef]
61. Spraggon, M.; Bodolica, V. Collective tacit knowledge generation through play: Integrating socially distributed cognition and transactive memory systems. *Manag. Decis.* **2017**, *55*, 119–135. [CrossRef]

sustainability

Article

Transformational Training Programs and Quality Orientation of Employees: Does Employees' Loyalty Matter?

Nidal Fawwaz Al Qudah [1,*], Yang Yang [1] and Muhammad Adeel Anjum [1,2]

[1] School of Management, Harbin Institute of Technology, Harbin City 150001, China; yfield@hit.edu.cn
[2] Department of Management Sciences, Balochistan University of Information Technology, Engineering and Management Sciences, Quetta City 87300, Pakistan; muhammad.adeel@buitms.edu.pk
* Correspondence: nidal@stu.hit.edu.cn or naq682002@yahoo.com; Tel.: +86-0451-8641-4009

Received: 31 December 2017; Accepted: 5 February 2018; Published: 9 February 2018

Abstract: Transformational training programs, employee loyalty and quality orientation of employees have been some of the important concerns for both academicians and practitioners for decades. Yet, little is known about their underlying relationship dynamics, especially in the context of higher education institutions. The pivotal aim of this study was to investigate the interplay of transformational training programs, loyalty and quality orientation of employees. For this, a causal model demonstrating the direct and indirect relationships of transformational training programs, employee loyalty and quality orientation was built and tested. Data for this study were collected from 212 ($n = 212$) academics (deans, head of departments and faculty members) from all private sector universities in Amman, Jordan, through a cross sectional survey. Results indicated that both direct and indirect effects of transformational training programs on quality orientation of employees were significant. More specifically, the positive effects that transformational training programs have on quality orientation of employees are through employee loyalty. This finding significantly advances the existing body of knowledge and implies that transformational training programs enhance employees' loyalty which, in turn, escalates employees' orientations towards quality. Hence, it is concluded that the objective of inculcating quality orientation amongst employees cannot be achieved with mere reliance upon transformational training programs. Several contextual factors, such as employee loyalty, should also be focused on and fostered to ensure the effects that training programs have on certain desirable outcomes.

Keywords: transformational training programs-TTP; quality orientation of employees-QOE; employee loyalty-EL; universities; Jordan

1. Introduction

In today's highly dynamic and competitive environment, 'quality orientation' has emerged as an important concern for organizations. Therefore, every organization, despite its nature and scale of operations, strives hard to address the very concern as it is perceived as a 'mantra' of firms' ultimate survival in ruthless competition. This is why, researchers and practitioners devote considerable attention to understand this phenomenon. Quality orientation refers to a set of attitudes and behaviors that affect the quality of interaction between the staff of any organization and its customers and its commitment to continuous improvement during the delivery of customers' perceived quality and to achieve customer satisfaction [1]. However, as a management practice, it points to the conscious efforts of an organization towards achieving high levels of service quality and customer satisfaction. That is why quality orientation plays a fundamental role in improving service quality, service delivery and customer satisfaction [2]. Given this, quality orientation is perceived as a key strategic resource of improving organizational performance [3]. Researchers opine that high-quality orientation coupled with systematic planning and monitoring offers numerous benefits, such as increased productivity and amplified organizational performance [4]. That is why, perhaps, organizations embrace 'quality orientation' as an integral element of their business philosophy [3].

Quality orientation is not only an issue of manufacturing concerns, but, also an important concern for the managers of service sector organizations. Scholars dully recognize the importance of quality orientation for service sector organizations. According to them, quality orientation is an important tool that helps achieve competitive advantage, yet it is one of the least researched topics, especially in the service sector [5]. The service sector, by and large, capitalizes upon its human resource to achieve and sustain competitive advantage. Review of relevant literature reveals that knowledge and skill sets of an organization's employees are key to its performance, competitiveness and advancement. Therefore, firms develop and enhance quality of their human resource through appropriate trainings and development initiatives [6]. However, training programs that are consistent with employees' needs, organizational goals, and business strategy tend to be more successful than those that are not [7]. Studies indicate that effective training programs can significantly affect employee satisfaction and loyalty [8], which are some of the most essential ingredients for a superior service quality and organizational success [9,10].

However, a critical review of substantial body of literature on 'quality orientation' reveals that no attempt has yet been made to ascertain whether or not, the transformational training programs affect quality orientation of employees, and if they do, how? Also, what role does employee loyalty plays in this nexus? This study aims at bridging this knowledge gap by finding out the answer of a major question: 'what effects trainings programs have on quality orientations and what is the mechanism of such effects?' The answer to this question would not only provide new insights on underlying relationship dynamics of training programs, employee loyalty and quality orientation, but would also offer several practical implications that, if acted upon, would help practitioners and policy makers to help achieve the bottom line objectives of organizational effectiveness and sustainability.

2. Literature Review and Conceptual Framework

2.1. Transformational Training Programs (TTP)

Training is a content-based activity, generally away from the workplace, with a coach leading and aiming to modify a person's behavior or attitude [11]. Training programs can be defined as "planned efforts that are aimed at increasing individual skills [6]. Training programs are also referred to as systematic processes of changing the behavior of employees towards achieving organizational goals [12]. Trainings are vital for organizations as they are a source of increasing intellectual capital and escalating employee commitment [13]. Transformational training programs, on the other hand, are a bit different. The objective of transformational training programs is not to change individual behaviors, but to change the way learners think about new knowledge or skills. A plethora of empirical evidence is available on consequences and outcomes of training. McFarlane in his study investigated the consequences of training programs and explored trainees' reactions towards training goals, content, material, trainers, environment, training process and trainees' acquired knowledge and skill sets. He concludes that all these factors play very important role in determining perceived usefulness of training programs [14]. Another study concludes that training programs equip trainees with a determined level of knowledge and skills and shape their behaviors and actions to a desired level [15]. Alawneh demonstrated importance of several contextual factors in determining training transfer [16]. A study worth mentioning here was performed by Kirkpatrick. He identified four levels of training evaluation, namely: (i) reaction criteria; (ii) learning criteria; (iii) behavior criteria; and (iv) results criteria [17].

(1) *Reaction criteria*: Reaction criteria are trainees' perceptions of training programs in organizations [17]. In higher education institutions, the reaction of participants is usually judged

through self-reporting. This method asks trainees to report advantages and disadvantages of training programs [18]. Such criteria is widely used because of its ease [19].

(2) *Learning criteria:* Learning criteria, in higher education institutions, is assessed through learning outcomes and is evaluated by conducting several tests to measure: performance, presentation, and demonstration of learned skills [20]. A variety of assessment techniques, such as speeches and writing samples are used to assess the extent of learning [21].

(3) *Behavioral criteria:* Behavioral criteria measure the performance of trainees on their actual jobs [22]. In higher education institutes, behavioral criteria usually base upon performance indicators such as work-related outcomes. Application of skills and knowledge gained from training programs in research projects is an example of behavioral criteria in higher education institutions [23].

(4) *Results criteria*: Results criteria in organizational settings are measured through different benchmarks, such as: efficiency, productivity and profitability [23]. Results criteria in higher education is measured by assessing the competence level of students [24].

2.2. Employee Loyalty (EL)

Loyalty is a two-way path. If an organization desires its employees to be loyal, they must earn it by creating a stable and challenging workplace [25]. Scholars opine that loyalty is an emotional assurance of employees' ambition to involve and remain determinedly constant and responsible with the organization [26]. Researchers hold different views about loyalty. For some, employee loyalty is an action-oriented approach which deals with the behavior of employees [27]. Whereas, for others, it is the commitment that employees have for their organizations [28]. Martensen and Grønholdt [29] note that the fundamental principle underlying the concept of employee loyalty is emotional attachment [30]. However, some of the researchers identify two basic approaches to determine employee loyalty, namely: the attitudinal approach, and the behavioral approach. From attitudinal perspective, loyalty refers to an individual's psychological inclinations, feelings, identification, attachment or commitment to the organization [31,32]. However, the cognitive nature of attitudinal approach makes the measurement of loyalty difficult and questionable [31,33,34]. Whereas, behavioral approaches view loyalty as an observable phenomenon that is obvious and can be easily materialized in the context of employee-organization relationships [35,36]. Researchers also explain that employee loyalty is a major driving force behind the sustainable development of organizations [37,38].

Cook [39] advanced the body of knowledge on loyalty by introducing a taxonomy of employee loyalty. According to him, employee loyalty can be categorized either as active or passive. First refers to the subjective feelings and desires of employees to stay with an organization. Such subjective feelings and desires arise when employees feel that organizational goals are congruent with their own. Whereas, the latter (passive loyalty) is that state of mind or phenomenon which captures employees' dissatisfaction. A worth discussing fact here is that, despite being dissatisfied, employees do not want to leave organization due to some lucrative benefits that they get (e.g., high wages). If these conditions disappear, employees no longer remain loyal to their organizations. Meschke [40] revamped Cook's taxonomy by introducing the concept of ' tripartite employee loyalty'. According to Meschke, loyalty revolves round three objects: supervisor, working group, and the organization. These objects should be kept in view while investigating the outcomes of employee loyalty. However, an employee's loyalty towards different reference objects may conflict with each other [41]. This point is conceivable as employees do not display the same loyalty towards their supervisor, working group, and organization at the same time. Rather, it is likely that their loyalty towards one or more of these reference objects differs due to their potential outcomes such as: openness to leave, openness to reapply, and openness to change. Hence, a valid and reliable measure of employee loyalty is inevitable [40].

2.3. Quality Orientation of Employees (QOE)

Quality orientation, as a construct is dynamic and abstract in nature. There exist two main perspectives on quality orientation of employees (QOE). First perspective view it as a 'business philosophy', whereas, the latter considers it as a 'managerial practice'. Despite of their differences, both perspectives view QOE as an essential ingredient of organizational success. Proponents of first perspective argue that quality-oriented firms manage and control internal processes to ensure quality products and services to the customers which is ultimate for organizational success [41]. Perhaps, this is why, quality orientation has become an emerging business philosophy as it helps achieving and sustaining competitive advantage [42]. Quality orientation depicts a philosophical commitment of an organization to developing and maintaining a sustainable quality-based competitive advantage leading to increased business performance [43]. Dahlgaard and Mi Dahlgaard-Park explored the aim of this philosophy and found that it could change organizational culture from passive and defensive to proactive and open culture with an open participation of every organizational member [44]. Therefore, it is essential to formulate quality orientation philosophy at the time when organizations begin to look up for competitive superiority through customer satisfaction with its quality products and services [43].

The latter perspective, on the other hand, defines quality orientation as a managerial practice as well as an employee behavior which is oriented towards achieving high levels of service quality and customer satisfaction [5]. This perspective stresses that quality orientation plays fundamental role in service delivery, and that quality orientation behaviors foster service excellence and assure customer satisfaction [3]. Quality orientation of any organization is also linked to a widespread understanding among organizational members about the importance of quality [45,46].

3. Hypotheses Development

Many scholars have investigated the effects of transformational training programs on certain attitudinal outcomes [47]. For instance, a study evaluated the impact of transformational trainings on employees' satisfaction and found them positively associated. Moreover, this study uncovered that the trainees' motivation towards a transformational training program can have significant positive effects on certain attitudinal outcomes (e.g., job satisfaction) [48]. Another study by Sivanathan and colleagues [49] explored how transformational training interventions improve occupational safety. Results of this pre-test/post-test quasi-experiment involving swimming pool supervisors and swim instructors revealed that transformational training interventions significantly improved instructors' perceptions about their supervisors' behaviors regarding safety compliance. Also, it was revealed that the change in instructors' perceptions of their supervisors' behaviors served as a mechanism through which changes in safety behaviors occurred. Duygul and Kublay [50] examined effects of transformational training programs on nursing practices (inventory management). Results indicated significant improvements in the inventory management practices of nurses after attending transformational training programs. Another study by Owoyemi and colleagues [51] explored the relationship between training programs and employees' commitment. Results of this survey of 250 employees revealed a statistically significant and positive relationship between training and employees' commitment. Based on this evidence, we also expect that transformational training programs would be positively related to employee loyalty and their quality orientations. Therefore, we propose that;

Hypothesis 1. *Transformational training programs would be positively associated with quality orientation of employees.*

Hypothesis 2. *Transformational training programs would be positively associated with employee loyalty.*

A study by Antoncic and Antoncic [52] investigated the relationship between employee loyalty and firm's growth. Findings indicate that employee loyalty and firm growth are positively associated. Czyż–Gwiazda [53] tested the interrelationships between business process orientation, maturity level and the level of quality orientation implementation and found that business process orientation, maturity level and the level of quality orientation implementation had strong associations between them. Therefore, we also posit that:

Hypothesis 3. *Employee loyalty would be positively associated with quality orientation.*

As discussed earlier, the pivotal aim of this study is to investigate the interrelationships among transformational training programs, employee loyalty and quality orientation of employees. Though the arguments in support of Hypotheses 1, 2 and 3 clearly demonstrate the connections between major study variables, however, the mechanism through which transformational training programs relate to quality orientation of employees is yet to be conceptualized. We assume that employee loyalty serves as a mechanism that connects transformational training programs and quality orientations. The rationale behind this assumption is that transformational training programs enhance employee loyalty which then induces quality orientations. Therefore, we propose;

Hypothesis 4. *Employee loyalty would mediate the relationship of transformational training programs and quality orientation.*

The dynamics of major study variables are shown in Figure 1.

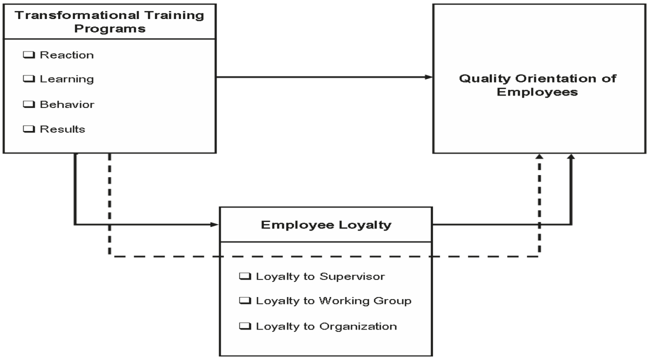

Transformational Training Programs
❑ Reaction
❑ Learning
❑ Behavior
❑ Results
Quality Orientation of Employees
Employee Loyalty
❑ Loyalty to Supervisor
❑ Loyalty to Working Group
❑ Loyalty to Organization

Figure 1. Research Model.

4. Methodology

4.1. Research Design and Sample

Since the purpose of this research was to understand how transformational training programs enhance quality orientation of employees. More specifically, this study intends to empirically investigate the relationship/effect of transformational training programs on quality orientation of employees with the mediating effect of employee loyalty. Correlational design bets serves this purpose, hence, this study adopts a correlational design to investigate the relationships dynamics of transformational training programs, employee loyalty and quality orientation of employees. The population of the current study consisted of academics working in private Jordanian universities in Amman City, the capital of Jordan. Due to cost, time and access constraints, probability sampling was not possible. Therefore, convenience sampling technique was applied to approach respondents. Participants were approached in their native work settings and were given a self-reported questionnaire to fill. All respondents were briefed about the aims of this study and were also informed that their participation was voluntary and that their responses will be analyzed and reported as group data without disclosing their identities. A total of 225 respondents, including deans, heads of departments and faculty members, participated in this study, out of which 13 questionnaires were discarded due to incomplete information (missing responses) and 212 duly filled surveys were retained for further analysis. Table 1 contains the demographic profile of respondents.

Table 1. Demographic Profile.

Variables	Dimensions	Frequency & %
Gender	Male Female	145 (68.4%) 67 (36.6%)
Age	Less than 30 Years 30–39 Years 40–49 Years 50 years and above	17 (08%) 35 (16.5%) 103 (48.6%) 57 (26.9%)
Experience	Less than 5 years 5–9 Years 10–19 Years 20 years and above	11 (5.2%) 97 (45.8%) 61 (28.8%) 41 (20.3%)
Job Titles	Dean Head of Departments Faculty	29 (13.7%) 62 (29.2%) 121 (57.1%)

4.2. Measurement of the Constructs

The items for measuring constructs were drawn from the literature. Specifically, the scales developed and validated by previous researchers were used to gauge respondents' perceptions. Scale items were slightly modified according to the context of present study. 5-point Likert-type scale ranging from 5 (strongly agree) to 1 (strongly disagree) was used to rate scale items. Respondents were asked to indicate levels of their agreement or disagreement with question statements given in the survey. The items to measure transformational training programs were adapted from Al Qudah, et al. [54]; this scale consisted of 13 items and showed a composite reliability of 0.925. Employees' loyalty was measured with the help of an 18 items scale developed and validated by Meschke [40], composite reliability of this scale in our study was 0.972. Quality orientation of employees was measured with the help of a 10 items scale developed by Alrubaiee, et al. [3], this scale showed an overall reliability of 0.899 in present context. Composite and dimension wise reliability coefficients (cronbach alpha) are shown in Table 2.

Table 2. Reliability Analysis.

	Cronbach's Alpha		CR
	No. of Items	Value	
Reaction	3	0.823	0.72766
Learning	4	0.859	0.78089
Behavior	3	0.812	0.86530
Results	3	0.808	0.78046
Transformational Training Programs	13	0.925	-
Loyalty to supervisor	6	0.898	0.83674
Loyalty to working group	6	0.874	0.85623
Loyalty to organization	6	0.887	0.80544
Employee loyalty	18	0.972	-
Quality Orientation of Employees	10	0.899	0.86561

SPSS version 22 was used for data analysis. Structural equation modeling (SEM) technique, using AMOS 22, was applied to test articulated hypotheses. The reliability of all scales was tested once again in order to verify whether or not the transformational training programs, employee loyalty and quality orientation of employees constructs show internal consistency in AMOS. The values of CR (composite reliability) as shown in Table 2 verify that all CR values were higher than the threshold level of 0.7 [55].

5. Results

5.1. Model Fit

As stated above, structural equation modeling (SEM), using AMOS 22, was run to test the hypotheses and to assess the effect and significance level of each path in research model. First of all, an overall model fit was determined by running confirmatory factor analysis (CFA). Several fit indices were calculated to determine the degree to which structural equation model fits the sample data. Obtained values of model fit indices proved the goodness of fit. The values of fit indices, as shown in Table 3 were reported as: $\chi^2/\mathrm{df} = 1.647$, goodness of fit index (GFI) and adjusted goodness of fit index (AGFI) were 0.912 and 0.972 respectively, the normed fit index (NFI) was 0.916 and the Tucker-Lewis index (TLI) was 0.954, the comparative fit index (CFI) was 0.965 and the root mean square error of approximation (RMSEA) was 0.055. All reported values indicated a good fit between theoretical model and data [56–58]. Figure 2 shows the results of structural equation modeling.

Table 3. Overall fit indices of Structural Model with all Constructs.

Model	χ^2	df	χ^2/df	GFI	AGFI	NFI	TLI	CFI	RMSEA
Default model	172.898	105	1.647	0.912	0.927	0.916	0.954	0.965	0.055
Saturated model	0.000	0		1.000		1.000		1.000	
Independence model	2059.518	136	15.144	0.348	0.267	0.000	0.000	0.000	0.259

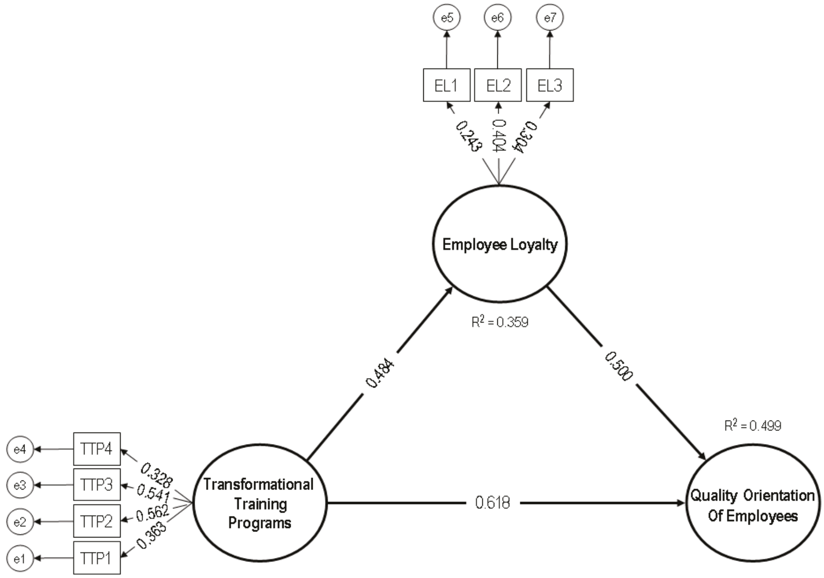


Figure 2. Result of structural equation modeling (SEM).

5.2. Hypotheses Testing

5.2.1. Independent Variable → Dependent Variable

Table 4 represents the effect of independent variable (transformational training programs) on a dependent variable (quality orientation of employees). Results show that transformational training programs can significantly predict quality orientation of employees ($\beta = 0.618$, C.R = 8.704; *p*-value = ***). These evidences provide enough support for H1.

Table 4. Direct affects Testing Result (Independent variable → Dependent variable).

Hypothesis	Regression Weights		Estimate	SE	C.R.	*p* Value	Hypothesis
	From	To					
H1	TTP	QOE	0.618	0.071	8.704	***	Accepted

Note: *** $p < 0.05$.

5.2.2. Independent Variable → Mediating Variable

The effect of transformational training programs on employee loyalty are shown in Table 5. Results show that transformational training programs can significantly predict employee loyalty (β = 0.484, C.R = 6.315, *p*-value = ***) Hence, H2 is supported.

Table 5. Direct affect Testing Result (Independent variable → mediate variable).

Hypothesis	Regression Weights		Estimate	SE	C.R.	*p* Value	Hypothesis
	From	To					
H2	TTP	EL	0.484	0.076	6.315	***	Accepted

Note: *** $p < 0.05$.

5.2.3. Mediating Variable → Dependent Variable

It is clear from Table 6 that the effects of employee loyalty on quality orientation of employees are significant (β = 0.500, C.R = 6.849; *p*-value = ***). Hence, H3 is supported.

Table 6. Direct affect Testing Result (Mediate variable → Dependent variable).

Hypothesis	Regression Weights		Estimate	SE	C.R.	*p* Value	Hypothesis
	From	To					
H3	EL	QOE	0.500	0.073	6.849	***	Accepted

Note: *** $p < 0.05$.

5.2.4. Independent Variable → Mediate Variable → Dependent Variable

The indirect effects of transformational training programs on quality orientation are summarized in Table 7 and Figure 3. As indicated by results, indirect effect of transformational training programs on quality orientation of employees (0.173) are significant which indicates that employee loyalty has significantly mediated the relationship between transformational training programs and quality orientation of employees. Hence, H4 is supported.

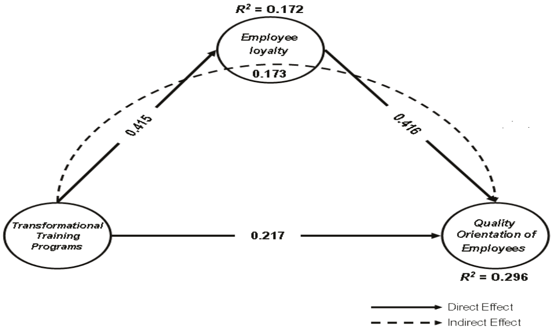


Figure 3. Result of indirect effects.

Table 7. Indirect Effects (Independent variable → mediate variable → Dependent variable). (Mediating effect of Employee loyalty in the relationship between transformational training programs and quality orientation of employees).

Hypothesis	**From**	**Mediation**	**To**	**Direct Effect**			***p* Value**	**Indirect Effect**	**SMC**	***p* Value**	**Results**
				From	**To**	**Value**					
H4	TTP	EL	QOE	TTP	QOE	0.217	***	0.0173	EL = 0.172 QOE = 0.296	*** ***	Mediating
				TTP	EL	0.415	***				
				EL	QOE	0.416	***				

Note: *** $p < 0.05$.

6. Discussion, Implications and Conclusions

6.1. Discussion

This study investigated the interplay of transformational training programs, employees' loyalty and quality orientations. Four hypotheses were articulated for testing. First hypothesis claimed a positive relationship between transformational training programs and quality orientation of employees. Results provided enough empirical evidence in support of this hypothesis. As expected, a significant positive association was found between transformational training programs and employees' quality orientation, which implies that transformational training programs augment attitudinal orientations of employees. These results are consistent with the findings of Barling, et al. [47] and Owoyemi, et al. [51], who also found that training is a significant factor in developing and shaping the desirable attitudes of employees. Hence, it can be asserted that organizations should launch effective training programs to reap the benefits of desirable attitudinal outcomes. We, in hypothesis two, posited that transformational training programs would also be positively associated with employee loyalty. Results also supported this supposition. Results indicated that transformational training programs significantly enhance employee loyalty. The rationale behind this relationship could be the fact that benefits of training programs are twofold, that is, training programs not only improve existing knowledge base and skill sets of employees, but also bring a positive change in employee attitudes. Training opportunities make employees feel that the organization is concerned about them, this feeling in turn results in increased levels of satisfaction and loyalty to the organization. This finding is consistent with that of Sivanathan, et al. [49] who also opined more or less similar. According to them, transformational training programs have significant impacts on the satisfaction and motivation levels of trainees.

Likewise, hypothesis three posited that employee loyalty would positively relates to employees' quality orientations. This hypothesis was also supported. Employee loyalty was found positively associated with quality orientation which means that employees with an attitudinal inclination of loyalty for their firm are more likely to exhibit factorable behaviors geared towards attainment of general goals of the firm such as high-quality service orientation for customers. This finding is consistent with the findings of Antoncic and Antoncic [52] who report a positive relationship between employee loyalty and firm's growth.

Finally, results also indicated that employee loyalty significantly mediated the relationship between transformational training programs and quality orientation of employees (H4). This con notes that employee loyalty serves as mechanism through which transformational training programs are related to quality orientation of employees. In other words, it can be explained as training improves employees' loyalty which, in turn, triggers their orientation towards quality. This finding, to some extent, is consistent with the findings of Ismail, et al. [59] who confirmed that the empowerment being one of the components of loyalty to the organization, mediates the relationship between transformational training and service quality.

6.2. Implications

This study significantly advances the existing body of knowledge in many ways: (i) drawing on the literature, it develops and tests a causal model that redefines how transformational training programs relate to quality orientations especially in middle eastern context; (ii) it also bridges a substantial knowledge gap by explaining the nature and dimensionality of relationships between study variables and, by doing so; (iii) this study highlights empirical utility of training programs in the workplaces. From a practical perspective, this study invokes managers' attention towards training programs and their beneficial concomitants (outcomes) such as: loyalty and quality orientations. The confirmation of the mediating role of employee loyalty also bears important implications for the management. That is, the management of organizations should focus on fostering employee loyalty, as without it the bottom line objective of quality cannot be reached. Moreover, management should also be concerned about training programs as they drive employee loyalty which consequently enhances quality orientation of employees. Building on this, top management of private Jordanian universities may use the framework of this study to develop relevant and effective strategies and tactics to enhance quality orientation of their staff. In addition, the management should endeavor to develop an organizational climate that can promote organizational citizenship behavior and enhance a positive quality orientation of employees. These findings are of particular significance because this study was conducted in a Middle East country (Jordan) which significantly differs from developed nations such as the US, China or European countries in terms of culture, organizational settings and demographic characteristics. Therefore, we believe, this study has made useful contributions to the debate on benefits of transformational training programs for universities in dynamic environments.

However, these findings must be interpreted in light of certain limitations. Since this study was conducted in Jordan, results cannot be generalized to other areas. Moreover, the prime objective of this study was to provide firms with a general guideline on how transformational training programs could enhance employees' loyalty and quality orientation, therefore, it only characterizes the ends instead of the entire continuum. The possibility of social desirability bias during data collection cannot be ruled out, hence, future researchers could substantiate results of this study by adopting a 360 degree approach of data collection. Given its cross sectional nature, this study also limits the possibilities of determining reverse causality, so we recommend longitudinal design in future studies.

6.3. Conclusions

Training programs offer prime opportunities for expanding the knowledge, skills and abilities of employees which are the potent agents of achieving organizational objectives (e.g., improved quality of products and services, customer satisfaction and organizational effectiveness). Hence, every organization, despite its nature and scope, should invest in training. As proven, employee loyalty is another resource which positively contributes towards attainment of organizational goals, therefore, efforts should be made to enhance this viable resource.

Acknowledgments: We appreciate anonymous reviewers, for their valuable feedback and comments during review process.

Author Contributions: Nidal Al Qudah conceived the idea of this study, collected and analyzed relevant data and prepared an initial draft of paper;Yang Yang provided intellectual guidance on the conduct of this study; and Muhammad Adeel Anjum wrote this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Javalgi, R.G.; Whipple, T.W.; Ghosh, A.K.; Young, R.B. Market orientation, strategic flexibility, and performance: Implications for services providers. *J. Serv. Mark.* **2005**, *19*, 212–221. [CrossRef]
2. Chiang, F.F.; Birtch, T.A. Reward climate and its impact on service quality orientation and employee attitudes. *Int. J. Hosp. Manag.* **2011**, *30*, 3–9. [CrossRef]
3. Alrubaiee, L.; Al Zuobi, H.; Abu-Alwafa, R. Exploring the Relationship between Quality Orientation, New Services Development and Organizational Performance. *Am. Acad. Sch. Res. J.* **2013**, *5*, 315.
4. Wang, E.T.; Wei, H.-L. The importance of market orientation, learning orientation, and quality orientation capabilities in TQM: An example from Taiwanese software industry. *Total Q. Manag. Bus. Excell.* **2005**, *16*, 1161–1177. [CrossRef]
5. Marinova, D.; Ye, J.; Singh, J. Do frontline mechanisms matter? Impact of quality and productivity orientations on unit revenue, efficiency, and customer satisfaction. *J. Mark.* **2008**, *72*, 28–45. [CrossRef]
6. Singh, R.; Mohanty, M. Impact of training practices on employee productivity: A comparative study. *Intersci. Manag. Rev.* **2012**, *2*, 87–92.
7. Wexley, K.N. *Developing and Training Human Resources in Organizations*; Harper Collins Publishers: New York, NY, USA, 1991.

8. Hang, N.T.L. *An Investigation of Factors Affecting Employee Satisfaction and Loyalty in the Fire Fighting and Prevention Police of Ho Chi Minh City Vietnam*; University of tampere University of Economics: Ho Chi Minh City, Vietnam, 2016.
9. Jun, M.; Cai, S.; Shin, H. TQM practice in maquiladora: Antecedents of employee satisfaction and loyalty. *J. Oper. Manag.* **2006**, *24*, 791–812. [CrossRef]
10. Hart, D.W.; Thompson, J.A. Untangling employee loyalty: A psychological contract perspective. *Bus. Ethics Q.* **2007**, *17*, 297–323. [CrossRef]
11. Mullins, L. *Management and Organisational Behaviour*, 9th ed.; Prentice Hall: London, UK, 2010.
12. Ivancevich, J.M.; Konopaske, R. *Human Resource Management*, 12th ed.; Prentice Hall: London, UK, 2012.
13. Barrett, A.; O'Connell, P.J. Does training generally work? The returns to in-company training. *ILR Rev.* **2001**, *54*, 647–662. [CrossRef]
14. McFarlane, D.A. Evaluating training programs: The four levels. *J. Appl. Manag. Entrep.* **2006**, *11*, 96.
15. Dullien, T. *Transactional or Transformational Training*; Indian Gaming: Liberty Lake, WA, USA, 2008; pp. 88–119.
16. Alawneh, M. Factors Affecting Training Transfer: Participants Motivation to Transfer Training, Literature Review. In Proceedings of the Academy of Human Resource Development International Research Conference in the Americas, Panama City, FL, USA, 20–24 February 2008.
17. Kirkpatrick, D.L. Invited reaction: Reaction to Holton article. *Hum. Resour. Dev. Q.* **1996**, *7*, 23–25. [CrossRef]
18. Dysvik, A.; Martinsen, Ø.L. The relationship between trainees' evaluation of teaching and trainee performance among Norwegian executive students. *Educ. Psychol.* **2008**, *28*, 747–756. [CrossRef]
19. Arthur, W., Jr.; Tubré, T.; Paul, D.S.; Edens, P.S. Teaching effectiveness: The relationship between reaction and learning evaluation criteria. *Educ. Psychol.* **2003**, *23*, 275–285. [CrossRef]
20. Lievens, F.; Sackett, P.R. The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *J. Appl. Psychol.* **2012**, *97*, 460. [CrossRef] [PubMed]
21. O'Flaherty, J.; Phillips, C. The use of flipped classrooms in higher education: A scoping review. *Internet High. Educ.* **2015**, *25*, 85–95. [CrossRef]
22. Praslova, L. Adaptation of Kirkpatrick's four level model of training criteria to assessment of learning outcomes and program evaluation in higher education. *Educ. Assess. Eval. Account.* **2010**, *22*, 215–225. [CrossRef]
23. Landy, F.J.; Conte, J.M. *Work in the 21st Century, Binder Ready Version: An Introduction to Industrial and Organizational Psychology*; John Wiley & Sons: New York, NY, USA, 2016.
24. Biesta, G. Good education in an age of measurement: On the need to reconnect with the question of purpose in education. *Educ. Assess. Eval. Account.* **2009**, *21*, 33–46. [CrossRef]
25. Kumar, D.; Shekhar, N. Perspectives envisaging employee loyalty: A case analysis. *J. Manag. Res.* **2012**, *12*, 100.
26. Bloemer, J.; Odekerken-Schröder, G. The role of employee relationship proneness in creating employee loyalty. *Int. J. Bank Mark.* **2006**, *24*, 252–264. [CrossRef]
27. Duboff, R.; Heaton, C. Employee loyalty: A key link to value growth. *Strategy Leadersh.* **1999**, *27*, 8–13. [CrossRef]
28. Cunha, M.P. "The Best Place to Be" Managing Control and Employee Loyalty in a Knowledge-Intensive Company. *J. Appl. Behav. Sci.* **2002**, *38*, 481–495. [CrossRef]
29. Martensen, A.; Grønholdt, L. Internal marketing: A study of employee loyalty, its determinants and consequences. *Innov. Mark.* **2006**, *2*, 92–116.
30. Mehta, S.; Singh, T.; Bhakar, S.; Sinha, B. Employee loyalty towards organization—A study of academician. *Int. J. Bus. Manag. Econ. Res.* **2010**, *1*, 98–108.
31. Guillon, O.; Cezanne, C. Employee loyalty and organizational performance: A critical survey. *J. Organ. Chang. Manag.* **2014**, *27*, 839–850. [CrossRef]
32. Yee, R.W.; Yeung, A.C.; Cheng, T.E. An empirical study of employee loyalty, service quality and firm performance in the service industry. *Int. J. Prod. Econ.* **2010**, *124*, 109–120. [CrossRef]
33. Hajdin, M. Employee loyalty: An examination. *J. Bus. Ethics* **2005**, *59*, 259–280. [CrossRef]
34. Coughlan, R. Employee loyalty as adherence to shared moral values. *J. Manag. Issues* **2005**, *17*, 43–57.
35. Rusbult, C.E.; Farrell, D.; Rogers, G.; Mainous, A.G. Impact of exchange variables on exit, voice, loyalty, and neglect: An integrative model of responses to declining job satisfaction. *Acad. Manag. J.* **1988**, *31*, 599–627. [CrossRef]
36. Naus, F.; Van Iterson, A.; Roe, R. Organizational cynicism: Extending the exit, voice, loyalty, and neglect model of employees' responses to adverse conditions in the workplace. *Hum. Relat.* **2007**, *60*, 683–718. [CrossRef]

37. Khuong, M.N.; Linh, V.A.; Duc, V.M. The Effects of Transformational and Ethics-Based Leaderships on Employee's Loyalty towards Marketing Agencies in Ho Chi Minh City, Vietnam. *Int. J. Innov. Manag. Technol.* **2015**, *6*, 158. [CrossRef]
38. Wu, L.; Norman, I. An investigation of job satisfaction, organizational commitment and role conflict and ambiguity in a sample of Chinese undergraduate nursing students. *Nurse Educ. Today* **2006**, *26*, 304–314. [CrossRef] [PubMed]
39. Cook, S. *The Essential Guide to Employee Engagement: Better Business Performance through Staff Satisfaction*; Kogan Page Publishers: London, UK, 2008.
40. Meschke, S. Tripartite Employee Loyalty (TEL): Validation, Measurement and Selected Outcomes of a New Concept. *SSRN Electron. J.* **2016**. [CrossRef]
41. Provis, C. Dirty hands and loyalty in organisational politics. *Bus. Ethics Q.* **2005**, *15*, 283–298. [CrossRef]
42. Menor, L.J.; Roth, A.V. New service development competence and performance: An empirical investigation in retail banking. *Prod. Oper. Manag.* **2008**, *17*, 267–284. [CrossRef]
43. Mehra, S.; Joyal, A.D.; Rhee, M. On adopting quality orientation as an operations philosophy to improve business performance in banking services. *Int. J. Qual. Reliab. Manag.* **2011**, *28*, 951–968. [CrossRef]
44. Dahlgaard, J.J.; Mi Dahlgaard-Park, S. Lean production, six sigma quality, TQM and company culture. *TQM Mag.* **2006**, *18*, 263–281. [CrossRef]
45. Demirbag, M.; Sahadev, S. Exploring the antecedents of quality commitment among employees: An empirical study. *Int. J. Qual. Reliab. Manag.* **2008**, *25*, 494–507. [CrossRef]
46. Maletič, D.; Maletič, M.; Gomišček, B. The impact of quality management orientation on maintenance performance. *Int. J. Prod. Res.* **2014**, *52*, 1744–1754. [CrossRef]
47. Barling, J.; Weber, T.; Kelloway, E.K. Effects of transformational leadership training on attitudinal and financial outcomes: A field experiment. *J. Appl. Psychol.* **1996**, *81*, 827. [CrossRef]
48. Hassan, R.A.; Fuwad, B.A.; Rauf, A.A. Pre-training motivation and the impact of transformational leadership training on satisfaction with trained supervisors: A field experiment. In *Allied Academies International Conference. Academy of Strategic Management. Proceedings*; DreamCatchers Group, LLC: Cullowhee, NC, USA, 2009; p. 35.
49. Sivanathan, N.; Turner, N.; Barling, J. Effects of transformational leadership training on employee safety performance: A quasi-experiment study. In *Academy of Management Proceedings*; Academy of Management: Briarcliff Manor, NY, USA, 2005; pp. N1–N6.
50. Duygulu, S.; Kublay, G. Transformational leadership training programme for charge nurses. *J. Adv. Nurs.* **2011**, *67*, 633–642. [CrossRef] [PubMed]
51. Owoyemi, O.A.; Oyelere, M.; Elegbede, T.; Gbajumo-Sheriff, M. Enhancing employees' commitment to organisation through training. *Int. J. Bus. Manag.* **2011**, *6*, 280.
52. Antoncic, J.A.; Antoncic, B. Employee loyalty and its impact on firm growth. *Int. J. Manag. Inf. Syst.* **2011**, *15*, 81. [CrossRef]
53. Czyż-Gwiazda, E. Business Process Orientation and Quality Orientation Interrelationship-Survey Results. *Wrocław Univ. Econ.* **2015**, 275–291. [CrossRef]
54. Al Qudah, N.F.; Yang, Y.; Li, Z. Perceived Effectiveness of Transformational Training Programs: Dimensions, Measurement and Validation. In Proceedings of the 2017 International Conference on Industrial Engineering, Management Science and Application (ICIMSA), Seoul, Korea, 13–15 June 2017; pp. 1–6.
55. Murtagh, F.; Heck, A. *Multivariate Data Analysis*; Springer: Berlin, Germany, 2012; Volume 131.
56. Bagozzi, R.P.; Yi, Y. On the evaluation of structural equation models. *J. Acad. Mark. Sci.* **1988**, *16*, 74–94. [CrossRef]
57. Bearden, W.O.; Etzel, M.J. Reference group influence on product and brand purchase decisions. *J. Consum. Res.* **1982**, *9*, 183–194. [CrossRef]
58. Marcoulides, G.A.; Schumacker, R.E. *Advanced Structural Equation Modeling: Issues and Techniques*; Psychology Press: New York, NY, USA, 2013.
59. Ismail, A.; Halim, F.A.; Abdullah, D.N.M.A.; Shminan, A.S.; Muda, A.L.A.; Samsudin, S.; Girardi, A. The mediating effect of empowerment in the relationship between transformational leadership and service quality. *Int. J. Bus. Manag.* **2009**, *4*, 3. [CrossRef]

sustainability

MDPI

Article

Visualizing the Academic Discipline of Knowledge Management

Peng Wang [1], Fang-Wei Zhu [1,*], Hao-Yang Song [1], Jian-Hua Hou [2] and Jin-Lan Zhang [1]

[1] Faculty of Management and Economics, Dalian University of Technology, Dalian 116024, China; wangpeng26893@126.com (P.W.); haoyang@mail.dlut.edu.cn (H.-Y.S.); zhangjinlan530@mail.dlut.edu.cn (J.-L.Z.)

[2] Research Center of Science Technology and Society, Dalian University, Dalian 116622, China; hqzhixing@gmail.com

* Correspondence: zhufangwei@mail.dlut.edu.cn; Tel.: +86-411-8470-7746

Received: 29 December 2017; Accepted: 18 February 2018; Published: 2 March 2018

Abstract: The aim of this paper was to evaluate the research status of knowledge management (KM) and identify the characteristics of KM in the literature. We selected and studied in detail 7628 original research articles from the Web of Science from 1974 to 2017. Although many studies have contributed to the evolution of the KM domain, our results showed that a comprehensive bibliometric and visualization investigation was required. The literature on KM has grown rapidly since the 1970s. The United States of America, as the original contributing country, has also internationally collaborated the most in this field of study. The National Cheng Kung University has made the highest number of contributions. The majority of authors contributed a small number of publications. Additionally, the most common category in KM research was management. The main publications for KM research include *Journal of Knowledge Management*, and *Knowledge Management Research & Practice*. A keywords analysis determined that "knowledge sharing", "innovation", "ontology", and "knowledge management" were consistent hotspots in knowledge management research. Through a document co-citation analysis, the intellectual structures of knowledge management were defined, and four emerging trends were identified that focus on new phenomenon, the practice of knowledge management, small and medium enterprises (SMEs) management based on knowledge perspective, innovation and performance, and big data-enabled KM. We also provide eight research questions for future studies. Our results will benefit academics, researchers, and research students who want to rapidly obtain an overview of knowledge management research. This study can also be a starting point for communication between academics and practitioners.

Keywords: bibliometric; knowledge management; keywords analysis; intellectual structure; emerging trends; knowledge mapping

1. Introduction

With the advent of the era of the knowledge economy, knowledge management (KM) has become an important factor for promoting sustainable development of organizations and the economy. KM is also an increasingly important topic in the cross-disciplinary fields of management, computer science, and information science. KM has considerably progressed, attracting attention from researchers, practitioners, and policy-makers [1–3]. KM involves a series of managing activities that mainly concern the adoption, creation, storage, transfer, sharing, and application of knowledge. These activities could be divided into two main macro-processes: knowledge management adoption and knowledge management development [4–7]. The development process includes five phases: creation, storage, transfer, sharing, and application [7–9]. The intellectual antecedents of knowledge management can be traced back to the classical Greek era, which defined the epistemological debate in Western philosophy.

Modern KM research can be traced back to the mid-1970s [10]. Many researchers have contributed to the evolution of knowledge management [10,11]. In the 1980s, some new aspects of knowledge management, including knowledge acquisition, knowledge engineering, and knowledge-based systems, contributed by artificial intelligence research, systematically developed the field of knowledge management [12]. In the 1990s, knowledge management initiatives were flourishing with the help of information technology (IT). KM has helped to address and solve some of the challenges faced by Total Quality Management (TQM) and business process re-engineering [13]. The importance of managing knowledge has become a focus for all types of organizations, as KM is increasingly impacting large companies, SMEs, startups, supply chains, etc. In addition, the development of big data has created new issues for knowledge management [14–19].

Given the depth and breadth of KM practice, the numbers of publications in this field are growing rapidly. Professional journals, such as the *Journal of Knowledge Management*, *Knowledge Management Research & Practice, Academy of Management Review*, *Strategic Management Journal*, *Sloan Management Review*, *Harvard Business Review*, and the *Journal of the Association for Information Science and Technology and Scientometrics*, are now dedicated to different aspects of KM [20,21]. Largely due to the widespread use of KM, efforts have been increasingly invested into tracing the change trajectory of KM research, and its disciplinary characteristics.

However, two main problems remain evident in the existing reviews in the field of KM: some studies draw their conclusions based on subjective judgment, which may create controversies due to the limitations of the researcher's personal knowledge, and the previous qualitative analyses, such as bibliometric and scientometric analyses or systematic reviews, have been limited in terms research scope, timeframe, analytical unit, or focus on specific KM themes [11,18,22]. A bibliometric and visualization perspective of prior publications is lacing. Therefore, a bibliometric and visualizing investigation of the global KM research status is important for understanding the research advances and emerging trends. Unlike previous reviews of KM research, we conducted a bibliometric visualization review and obtained an overall picture of this fast-growing field between 1974 and 2017.

The objectives of this study are as follows. First, we wanted to identify the distribution of KM research including publications over time, countries and territories, institutes, authors, sources, and categories in KM-related research. Second, through co-word analysis of the keywords, we wanted to determine the main research topics. Third, we provided an of KM Intellectual Structure by hiring Citespace. Finally, the ultimate goal of this paper was to identify emerging trends.

To achieve these goals, we posed the following eight questions:

(1) What are the characteristics and growth trends of KM publications?
(2) What are the international collaborating countries that have the most KM research?
(3) Where are the active contributors located?
(4) What are the characteristics of the authorship distribution?
(5) What are the core KM disciplines and journals?
(6) What are the core KM research keywords?
(7) What is the intellectual structure of KM research?
(8) What are the emerging trends in KM research?

Based on the answers to these eight questions, these results obtained in this study benefit academics, researchers, and management students who want to quickly obtain an overview of knowledge management research. Our findings could assist researchers to better understand the current research progress in the KM domain and to identify the bibliometric characters of KM research. Our results about the emerging trends of KM will help researchers choose valuable research topics in the future. In addition, the research results can be a starting point for communication between academics and practitioners.

2. Related Work

Several studies investigated the performance and characteristics of knowledge management, with a wide variety of results. Styhre examined the KM research and found that KM is moving in the progressive direction [23]. Butler analyzed the KM field and suggested that KM could be divided into general, strategy-oriented, information-oriented, human-oriented, and process-oriented perspectives [24]. Lee and Chen visualized the trends in KM with KM data prior to 2006, and determined the 10 most important current research trends in KM [25]. Lee and Chen also revealed the research themes and trends in KM from 1995 to 2010 [26]. Li et al. analyzed the KM research status in China, and obtained some new findings by comparing current with previous KM-related research [27]. Gu found that KM had become an interdisciplinary theory developing on the boundaries of a variety of scientific disciplines [11]. Yogesh et al. provided an overview of 1043 articles for the period of 1974 to 2008, suggesting that KM systems and KM environment were the two most popular topics [28]. Serenko and Bontis ranked the knowledge management and intellectual capital academic journals, and found the top five academic journals in this field [29]. Serenko and Dumay found that the KM discipline is at the pre-science stage and the majority of KM citations exhibited a bimodal citation distribution peak [30]. Serenko applied a meta-analysis technique to integrate the overall findings of KM articles [31]. Akhavan et al. found that the most cited articles in KM were from the United States and the United Kingdom [32]. Considering the literature outlined above, some limits in terms of with the choice of research scope, timeframe, and analytical unit were noted. A bibliometric and visualization perspective of prior publications was also lacking.

Thus, we completed a wider investigation of the challenges faced by KM by profiling a large set of existing KM publications in terms of publication year, author, country, keywords, intelligence structure, and emerging trends. By doing this, we provide a comprehensive investigation of KM research.

3. Materials and Methods

The data used for this study were obtained from the Web of Science Core collection database, a Web-based user interface of *Web of Knowledge* developed by Clarivate Analytics. We adapted the same search strategy used by Lee and Chen to search for papers with the term "knowledge management" in titles, abstracts, or indexing terms [25]. As a result, we obtained 19,393 records prior to the end of 2017. For this study, we considered only articles, because they are the higher ranked scientific contributions. Although the reviews receive a greater number of citations, their scientific contribution is less important, and may introduce considerable noise into our analysis because they often contain too many topics [20,21]. After filtering out the less representative record types, the dataset was reduced to 7628 original research articles that were assumed to be in some way related to KM.

Bibliometric analysis is an effective way to investigate and examine performance in one knowledge domain [33]. Bibliometric analysis can be defined as a statistical method of determining the quantitative features of bibliographic information, literature, articles, and journals. The popularity of bibliometric studies is mainly due to the intrinsic characteristics of the raw data. Among the methodological options for an investigation study, bibliometric approaches have received increasing amounts of attention in various areas of research. Bibliometric studies have been completed for information systems, organizational studies, marketing-related subjects, operations management, and strategic management [34]. These works present an overview of the evolution of the publication years, document types, number of citations, most cited papers, influential authors, institutions, and countries. In other studies, visualization tools were used to provide a map of the bibliometric results. Detecting emerging research trends has been a focus for many researchers [35]. Various methods have been advocated for the purpose of detecting emerging research trends, such as historiography mapping [36,37], document co-citation [38], author co-citation [39], co-word analysis [40], and journal mapping [41].

Bibliometric mapping is usually used to display a structural overview of an academic field or a journal [42]. Some widespread mapping techniques have been designed and developed as computer programs like VOSviewer and Citespace. Compared with other quantitative literature review methods, a bibliometric review is usually used to display the quantitative characteristics of an academic field. Conversely, a systematic review provides an in-depth study and highlights strengths and weaknesses in the literature, evidence research gaps, and identifies appropriate research questions. To achieve objective of this study of investigating and visualizing the global research status in the KM field, we chose the bibliometric and bibliometric mapping method.

In this study, we present a bibliometric profile of KM. In addition, some research tools were used in this study. For example, we used Bibexcel to construct a co-occurrence matrix [43]. Citespace was used for co-citation analysis [44]. Ucinet [45] and Vosviewer [46] were used for social network analysis and visualization and Carrot was used for cluster analysis [47]. Other tools such as Excel were also used for basic statistical analysis and visualization of the bibliometric results. To evaluate the present KM situation, some indicters were used in this paper. For instance, frequency is one of the most commonly used indicators in the bibliometric knowledge domain and is considered the main indicator that highlights the present situation in a research field. Some network indicators were also used in this paper, such as degree centrality and betweenness centrality [45]. The reason for choosing these indicators was that they were also the most commonly used indicators in knowledge network analysis. For the emerging trends analysis, a method was introduced by Chen [44] that combines modularity and a burst index. This method is widely used and has been proven to be able to detect the emerging research trends in other domain. The overall approach and methodology is shown in Figure 1.

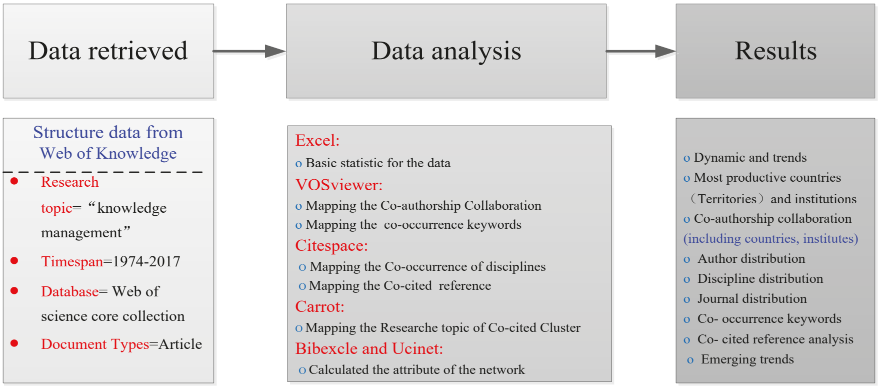

Data retrieved
Data analysis
Results
Structure data from Web of Knowledge
Research topic= "knowledge management"
Timespan=1974-2017
Database= Web of science core collection
Document Types=Article
Excel:
o Basic statistic for the data
VOSviewer:
o Mapping the Co-authorship Collaboration
o Mapping the co-occurrence keywords
Citespace:
o Mapping the Co-occurrence of disciplines
o Mapping the Co-cited reference
Carrot:
o Mapping the Researche topic of Co-cited Cluster
Bibexcle and Ucinet:
o Calculated the attribute of the network
o Dynamic and trends
o Most productive countries (Territories) and institutions
o Co-authorship collaboration
(including countries, institutes)
o Author distribution
o Discipline distribution
o Journal distribution
o Co- occurrence keywords
o Co- cited reference analysis
o Emerging trends

Figure 1. Research methodology.

4. Results

4.1. Distribution by Publication Year

Table 1 displays several characteristics of KM-related publications based on the year of publication. The annual number of articles and countries and the average number of authors and cited references increased significantly during the period of 1974 to 2017. Through checking the published papers over time, only one article was published in 1974, with an increasing number of KM publications after 1999. In 2012, a peak of 588 articles were published. After 2013, the number of publications steadily declined. Each KM publication had an average of 1.7 authors between 1974 and 1998, whereas the number steadily increased to 2.7 for 1999–2017. The annual number of countries participating in KM research also quickly increased from one country in 1974 to 77 in 2011, whereas the average number of cited references declined from 27.2 from 1974–1998 to 21.8 from 1999 to 2017. The correlation between Times Cited (TC) for an article and the length of time since its publication is shown in Table 1. The average length of an article fluctuated slightly, with an overall average of 13.5 pages.

Table 1. Knowledge management (KM) research article characteristics by year from 1974 to 2017.

Publication Year	NP (%) of 7628 Papers	No. Cr (TE)	AV. AU	AV. NR	AV. TC	AV. PG
1974	1 (0.013%)	1	1	37	22	8
1975	4 (0.052%)	1	1	25	1.3	6.3
1976	1 (0.013%)	1	1	48	4	13
1977	1 (0.013%)	1	1	22	6	15
1986	2 (0.026%)	2	1	26	7	18.5
1987	1 (0.013%)	1	1	24	74	12
1988	1 (0.013%)	1	3	30	9	8
1989	4 (0.052%)	2	2	20.8	14.3	11.8
1990	2 (0.026%)	1	2	8	8	5.5
1991	6 (0.079%)	4	2	17.7	3.7	17.3
1992	5 (0.066%)	5	1.8	15.4	1.5	11.2
1993	7 (0.092%)	5	2	27.6	6.6	21.3
1994	7 (0.092%)	7	2	20.6	95	10
1995	12 (0.157%)	4	1.8	22.4	9.8	16.3
1996	15 (0.179%)	7	2.3	28.2	99.1	15.7
1997	38 (0.498%)	9	2	23.7	40.4	12.3
1998	57 (0.747%)	17	2	23.9	61.4	13.4
1999	104 (1.363%)	24	1.9	26.5	54.6	12.0
2000	153 (2.006%)	28	2.0	26.7	40.5	13.5
2001	209 (2.740%)	31	2.3	31.4	45.6	14
2002	290 (3.802%)	40	2.3	25.3	30.3	12
2003	290 (3.802%)	43	2.6	26.2	29.4	13
2004	330 (4.326%)	48	2.8	28.8	24.5	12.6
2005	395 (5.178%)	49	2.7	33.1	28.5	13.9
2006	383 (5.021%)	57	2.6	34.8	26.4	13.7
2007	346 (4.536%)	54	2.6	41.4	24.2	15.4
2008	420 (5.506%)	56	2.6	43.2	19.6	14.1
2009	496 (6.502%)	64	2.6	41.5	21.7	14.0
2010	520 (6.817%))	62	2.7	48.7	18.0	14.5
2011	562 (7.368%)	77	3.0	51.1	15.6	15.1
2012	588 (7.708%)	69	2.8	51.6	10.9	14.8
2013	503 (6.594%)	70	3.2	56.4	10.2	15.4
2014	475 (6.227%)	70	3.0	58.2	8.1	15.0
2015	481 (6.306%)	67	3.2	61.9	4.8	16.0
2016	463 (6.070%)	74	3.1	58.2	1.8	15.4
2017	456 (5.978%)	67	3.1	65.6	0.5	16.4
Total	6285/100%	123	-	-	-	-

Noted: NP = number of publications; No. CR (TE) = number of countries; AV. AU = average number of authors; AV. NR = average numbers of references; AV. TC = average number of Times Cited; AV. PG = average number of pages.

In this study period, the growth in cumulative publications fit an exponential S-shaped function. S-shaped growth is a typical characteristic of a relatively mature stag research field [30]. Figure 2 indicates that KM research areas have entered the mature stage as of 2013.

4.2. Distribution and International Collaboration among Countries (Territories)

A total of 123 countries (territories) participated in KM publication activities from 1974 to 2017. Figure 3 shows the geographical distribution of the important countries (territories). Table 2 ranks the number of articles for each country contributing to KM publications. Notably, an article may be authored by many authors in several different countries. Therefore, the sum of articles published by each country may be larger than the total number of articles. The 1624 institutions in the U.S. published 1763 (25.26%) articles and had the largest number of authored papers. England (territories) was ranked second and Taiwan (territories) ranked third. China contributed 579 (7.6%) articles from 576 institutions and Spain published 553 (7.3%) articles out of 602 institutions.

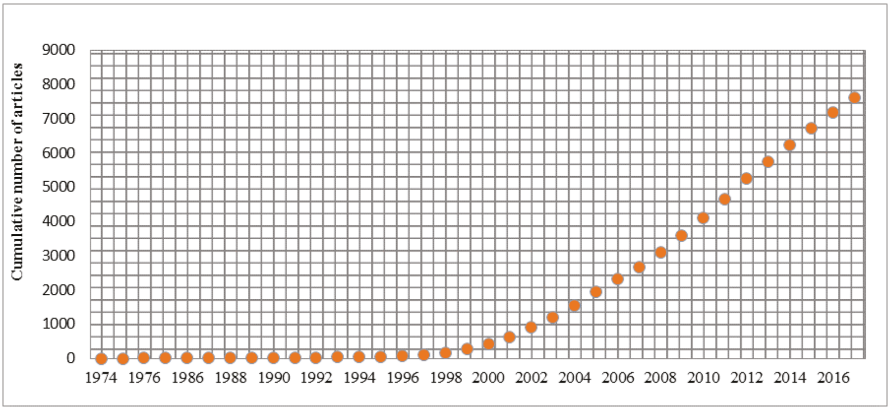


Figure 2. Cumulative growth in knowledge management publications, 1974–2017.

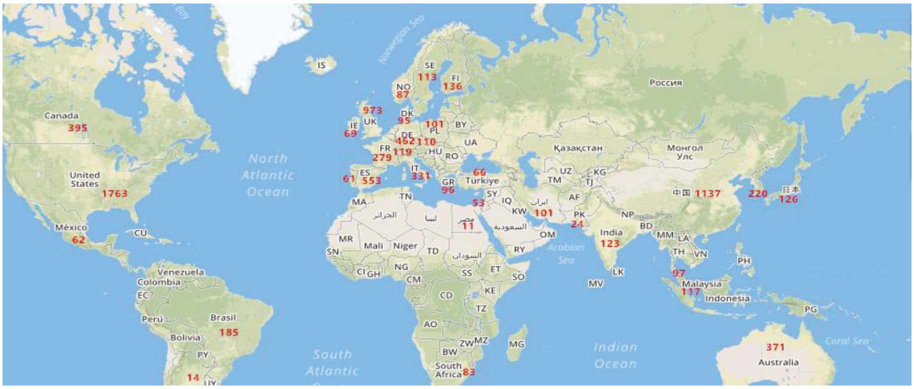


Figure 3. Geographic distribution of KM research articles.

By investigating citations from papers according to country distribution (Table 2), we found U.S.-authored papers were cited by 17,462 articles with 58,283 citations, accounting for 42.2% of all citations. U.S.-authored papers also had the highest average number of citations per article with a frequency of 33.06. The publications from England were next, distantly following the U.S., cited by 13,954 articles with 16,733 (11%) citations. The subsequent countries (territories) include Taiwan, China, and Spain.

International collaboration in science is both a reality and a necessity, and it is in the interest of all nations [48]. A network consisting of nodes with the collaborating countries between 1974 and 2017 is shown in Figure 4. The connection strength that determines the collation frequency between nodes (countries or territories) shows that the U.S. had the closest collaborative relationships with China, Canada, and England. England had the closest collaborative relationships with the U.S., Spain, and China. Taiwan had the closest collaborative relationships with Australia, the U.S., and some Asian countries. Germany had the closest collaborative relationships with European countries, such as Austria, England, and France, and China, the fifth-ranked country, had the closest relationships with the U.S., England, and Germany.

Table 2. Knowledge management (KM) research country (territory) ranked by the number of articles (>100 publications).

Rank	Country (Territory)	No. of Articles (%)	Citations (%)	Average	Citing Articles (%)	Institution
1	U.S.	1763 (23.1%)	58,283 (42.2%)	33.06	17,462 (48.8)	1624
2	England	837 (11 %)	16,733 (12.1%)	19.99	13,953 (18.3)	856
3	Taiwan	595 (7.8%)	10,428 (7.6%)	17.53	8028 (10.5)	233
4	China	579 (7.6%)	8924 (6.5%)	15.41	7461 (9.8)	576
5	Spain	553 (7.3%)	7485 (5.4%)	13.54	6352 (8.3%)	602
6	Germany	462 (6.1%)	4707 (3.4%)	10.19	4385 (5.8%)	735
7	Canada	395 (5.2%)	9937 (5.7%)	20.09	7124 (9.3%)	507
8	Australia	371 (4.9%)	5499 (4.0%)	14.82	5142 (6.7%)	367
9	Italy	331 (4.3%)	3843 (2.8%)	11.61	3564 (4.7%)	471
10	France	279 (3.7%)	3758 (2.7%)	13.47	3729 (4.9%)	533
11	Netherlands	223 (2.9%)	3918 (2.8%)	17.57	3687 (4.8%)	394
12	South Korea	220 (2.9%)	3895 (2.8%)	17.7	3394 (4.5%)	208
13	Brazil	185 (2.4%)	1190 (0.9%)	6.43	1148 (1.5%)	310
14	Finland	136 (1.8%)	1819 (1.3%)	13.38	1680 (2.2%)	130
15	Japan	126 (1.7%)	2540 (1.8%)	20.16	2471 (3.2%)	209
16	India	124 (1.6%)	1194 (0.9%)	9.63	1121 (1.5%)	254
17	Switzerland	119 (1.6%)	1906 (1.4%)	16.02	1846 (2.4%)	308
18	Singapore	117 (1.5%)	3321 (2.4%)	28.38	2985 (3.9%)	159
19	Sweden	113 (1.5%)	2350 (1.7%)	20.80	2246 (2.9%)	267
20	Scotland	111 (1.5%)	2011 (1.5%)	18.12	1936 (2.5%)	157
21	Austria	110 (1.4%)	1170 (0.8%)	10.64	1161 (1.5%)	221
22	Iran	102 (1.4%)	616 (0.4%)	6.04	566 (0.7%)	91
23	Poland	101 (1.3%)	638 (0.5%)	6.32	610 (0.8%)	165

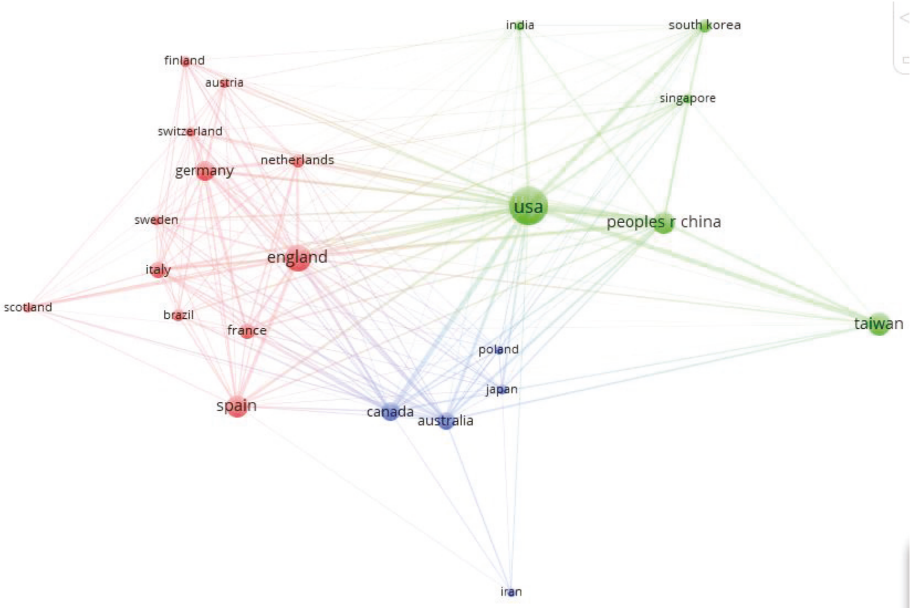


Figure 4. International collaboration network of the top 23 countries in KM research. The network was created using VOSviewer. The thickness of the linking lines between two countries is directly proportional to their collaboration frequency.

Table 3 shows the collaboration frequency distribution of papers from the main nations in the KM field. Tables 2 and 3 indicate that the USA is not only the original contributing country, but also the largest international collaborating country. England is a close second with 377 collaborations compared with the rank of the published number of articles. China and Australia rose in the rankings in terms of international collaboration. However, an opposite trend was observed in Taiwan, ranking fourth. Spain maintained a stable ranking, at fifth place. Table 3 also presents a summary of the Ucinet statistical results of four common parameters of each country: degree centrality, betweenness centrality, effective size, and constraint [33].

Degree centrality is defined as the number of links incident upon a node. It is a count of the number of ties directed to the node. In an international collaboration network, degree centrality often interpreted as a form of popularity or gregariousness. Betweenness centrality, a centrality measure within a graph, quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. In an international collaboration network, the country or institution with a high probability of occurring on a randomly chosen shortest path between two randomly chosen vertices will have high betweenness. From Table 3, the U.S. and Canada had the highest degree centrality, whereas England, Australia, Italy, the Netherlands, Spain, and Sweden were placed second with 21, and China and Switzerland were ranked third.

Betweenness centrality, an indicator for measuring nodes' control capacity over the network, also showed the USA and Canada played an important role in the top 23 international collaboration network. The other two parameters, effective size and constraint, confirmed the important role of the USA, Canada, and England.

Table 3. A social network analysis of the international collaboration network of the top 24 countries.

Country	NO.ICA	NO.ICC	DC	BC
USA	596	77	22	11.705
England	377	68	21	4.339
Taiwan	99	24	14	0.545
China	249	42	20	2.695
Spain	195	68	21	10.032
Germany	165	47	19	2.311
Canada	180	61	22	11.705
Australia	184	49	21	10.973
Italy	114	56	21	4.339
France	122	51	19	2.1
The Netherlands	104	43	21	4.339
South Korea	67	26	14	1.111
Brazil	49	44	19	1.837
Finland	45	29	14	1.703
Japan	46	38	19	6.607
India	42	46	18	0.907
Switzerland	62	52	20	3.231
Singapore	61	34	19	1.889
Sweden	58	48	21	4.339
Scotland	50	30	12	0.182
Austria	47	34	18	2.404
Iran	29	13	5	0
Poland	31	34	18	0.907

Note: NO.ICA = number of international collaboration articles, NO.ICC = number of international collaboration countries, DC = degree centrality, and BC = betweenness centrality.

In a comprehensive view, the collaboration mainly appears in high yield and developed countries. Previous studies indicated cooperation with foreign institutions did not achieve high cited papers. International cooperation does not embody high influence, but it has a very important impact on small and developing countries [29,32]. Therefore, small and developing countries should strengthening international cooperation to improving the publications influence.

4.3. Institution Distribution and Collaboration

A total of 4801 institutions participated in KM-related research, with 66.8% participating only once, 12.2% participating twice, and 22% participating more than twice. The top 25 of the most productive institutions are displayed in Table 4. National Cheng Kung University had the highest number of publications with 82 papers, followed by Hong Kong Polytechnic University with 77 papers, and the City University of Hong Kong ranked third with 56 papers. The subsequent countries include the National University of Singapore and the University of Cambridge. Simultaneously, the cited numbers for each paper are also displayed in Table 4. Harvard University was cited the most with 4263 citations, and the average number of times cited was 121.8. The University of Illinois followed closely with 2685 citations and with an average number of times cited of 63.9. The City University of Hong Kong ranked third with 2435 citations and an average number of times cited of 43.5.

Table 4. The most productive institutions for KM articles.

Rank	Institution	Country	Article	% of 7628	NO.TC	AV.TC
1	National Cheng Kung University	Taiwan	82	1.1%	1419	17.3
2	The Hong Kong Polytechnic University	China	77	1.0%	1542	20
3	City University of Hong Kong	China	56	0.7%	2435	43.5
4	National University of Singapore	Singapore	54	0.7%	2186	40.5
5	University of Cambridge	England	51	0.7%	1553	30.5
6	Universidad de Granada	Spain	48	0.6%	660	13.8
7	National Chiao Tung University	Taiwan	44	0.6%	902	20.5
8	The University of Manchester	England	44	0.6%	857	19.5
9	Nanyang Technological University	Singapore	42	0.6%	518	12.3
10	University of Illinois	USA	42	0.6%	2685	63.9
11	National Taiwan University	Taiwan	41	0.5%	900	22
12	Loughborough University	England	41	0.5%	887	21.6
13	University of Toronto	Canada	41	0.5%	1737	42.4
14	Polytechnic University of Valencia	Spain	39	0.5%	400	10.3
15	Universidad Carlos III de Madrid	Spain	38	0.5%	1032	20.2
16	University of Murcia	Spain	38	0.5%	884	23.3
17	McMaster University	England	37	0.5%	936	25.3
18	National Sun Yat-sen University	Taiwan	37	0.5%	1059	28.7
19	Cranfield University	England	36	0.5%	800	22.2
20	The University of Arizona	USA	36	0.5%	652	18.1
21	University of Warwick	England	36	0.5%	917	25.5
22	Harvard University	USA	35	0.5%	4263	121.8
23	Indiana University	USA	35	0.5%	1353	38.7
24	University of California	USA	35	0.5%	1930	55.1
25a	Korea Advanced Institute of Science & Technology	South Korea	33	0.4%	1654	50.1
25b	National Tsing Hua University	Taiwan	33	0.4%	543	16.5

Note: NO.TC = number of citations, AV.TC = average number of citations.

Then the top 297 institutions with more than or equal to 10 publications were chosen for our collaboration network analysis. The collaboration network map displayed in Figure 5 was created using VOSviewer. In the collaboration analysis, we were concerned about the collaboration frequency between two institutions. In Figure 5, the thickness of the linking lines between the two institutions is directly proportional to their collaboration frequency.

In the network map, the centrality of a node representing an institution is a graph-theoretical property that quantifies the importance of the node's position in a network. Table 5 presents a summary of the statistical results obtained using Ucinet. The statistical results of two common centralization indexes, degree centrality and betweenness centrality, for each institution qualitatively confirms the above findings.

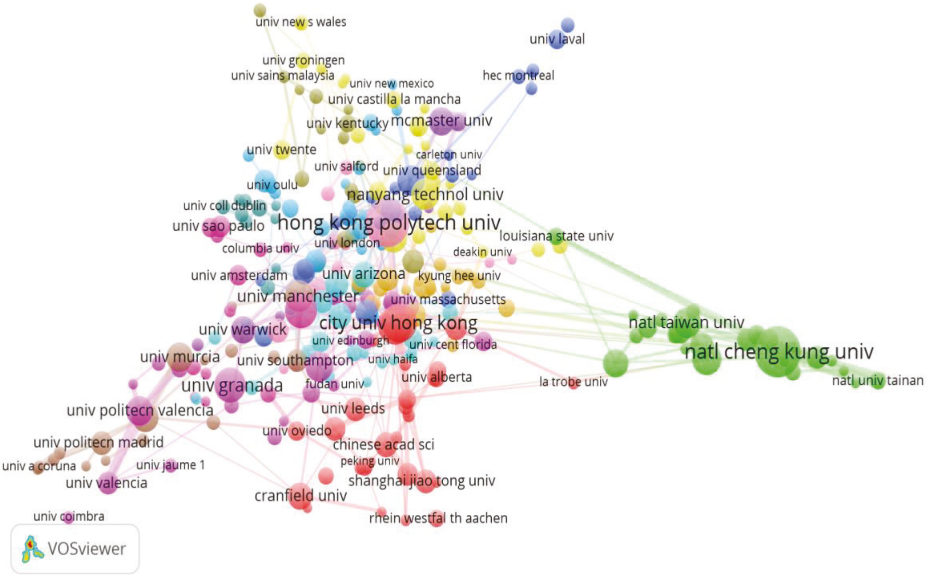


Figure 5. Collaboration network for institutions with more than two articles published contains 791 nodes.

Table 5. A social network analysis of the collaboration network of the top 16 KM research institutions.

Rank	Research Institution	Degree Centrality	Research Institution	Betweenness Centrality
1	City University of Hong Kong	44	City University of Hong Kong	1042.8
2	Aalto University	31	National University of Singapore	710.0
3	Chinese Academy of Sciences	31	University of Cambridge	451.0
4	Boston College	29	Indiana University	399.5
5	Brunel University	29	The Hong Kong Polytechnic University	356.3
6	Carnegie Mellon University	29	Chinese Academy of Sciences	352.0
7	National University of Singapore	29	The University of Arizona	338.7
8	Brigham and Women's Hospital	28	Carnegie Mellon University	334.4
9	Arizona State University	27	Aalto University	331.7
10	Beihang University	27	University of Illinois	331.3
11	Boston University	27	Brunel University	325.3
12	Aston University	26	Arizona State University	308.4
13	Auburn University	25	Boston University	292.1
14	Cardiff University	25	Universitat de Barcelona	289.3
15	BI Norwegian Business School	24	The University of Maryland	274.0
16	Chang Jung Christian University	24	University of Oxford	273.5
17	Chung Hua University	24	Brigham and Women's Hospital	266.0
18	Universities in Asia	23	Beihang University	218.0
19	University of Cambridge	23	Penn State University	216.0
20a	Bar Ilan University	22	National Cheng Kung University	214.2
20b	Indiana University	22	University of Warwick	202.3

From Table 5, City University of Hong Kong was ranked first for degree centrality. Aalto University was second place with a degree centrality value of 710, and the Chinese Academy of Sciences ranked third. Compared with the rank of the betweenness centrality, City University of Hong Kong was not only the first-ranked country in terms of degree centrality, but also had the highest betweenness centrality. The National University of Singapore ranked second and University of Cambridge ranked third.

4.4. Authorship Distribution

The total number of authors who contributed to the output set was 15,380. From 1974 to 2017, the average number of authors per article was 2.8. Table 6 shows the distribution of the number of authors with different numbers of articles. The large majority of authors contributed a very small number of publications, and 12,409 authors had only one article, 1820 authors had two articles, and 578 authors published three articles. The most productive author in the field KM articles was Chen from National Cheng Kung University. The second most productive author was Bontis from McMaster University. The third most productive author is Chen from National Cheng Kung University. Gottschalk and Serenko were ranked fourth, from Lakehead University and BI Norwegian Business School, respectively.

Table 6. The distribution of number of author with different numbers of articles.

NO.AU	1	1	1	2	2	1	2	5	7	5	11	15	19	26	41	41	125	328	14,808
NO.AR	25	21	20	19	18	17	15	14	13	12	11	10	9	8	7	6	5	4	≤3

Figure 6 displays the articles with number of authors by years. An upward trend was observed in the number of authors per article. The output of single-author papers is waning; the rate of single authorship had fallen drastically in KM research. Of the top 100 highly cited papers, single-author papers accounted for 27%. Although previous research indicated a strong positive correlation exists between the number of authors and the number of quotes, the higher the number of authors, the more often they are cited [30]. Additionally, the single-authored paper may be endangered in many fields, but this research still provides the methods and means for advancing research.

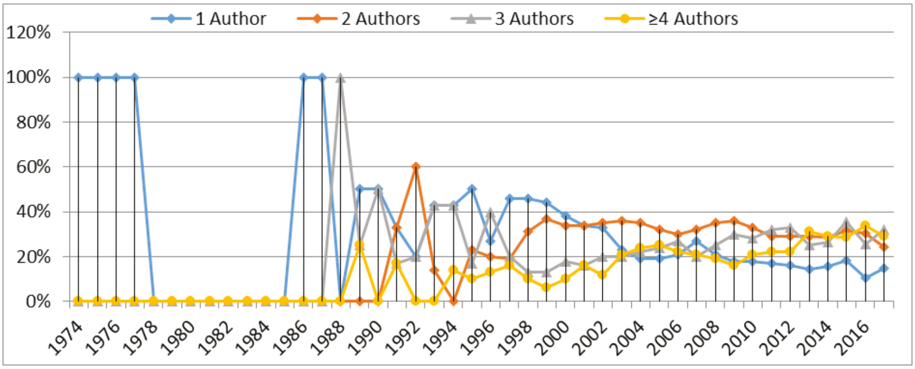


Figure 6. The percentage of articles with different numbers authors by year.

4.5. Distribution of Subject Categories

Table 7 displays the top 30 KM categories ranked in terms of the number of publications. The most common category was Management with 2334 records, followed by business economics with 1723 records, and Computer Science Information Systems with 1349 records.

Figure 7 shows a betweenness centrality network of these categories by using Citespace after being simplified with Minimum Spanning Tree network scaling, which retains the most salient connections. The nodes represent a category in which the number of articles had high betweenness centrality. From Table 8, the centrality of the Engineering, Computer Science, Interdisciplinary Applications, Management, Public, Environmental, and Occupational Health, and Psychology categories are notable. Burst, an indicator used to detect emerging trends, was used to detect emerging KM research subject categories. From Table 8, Computer Science, Theory, and Methods was ranked first with a burst value of 119.2, followed by Computer Science, Artificial Intelligence, and Computer Science. This means that KM research belonging to these three categories has been rapidly increasing in recent years.

Table 7. The top 30 KM categories ranked by the number of publications.

Subject Category	Records	% of Total
Management	2334	30.6%
Information Science Library Science	1723	22.6%
Computer Science Information Systems	1349	17.7%
Computer Science Artificial Intelligence	1050	13.8%
Operations Research Management Science	782	10.2%
Business	704	9.2%
Computer Science Interdisciplinary Applications	589	7.7%
Engineering Industrial	466	6.1%
Computer Science Theory Methods	461	6.0%
Computer Science Software Engineering	392	5.1%
Engineering Electrical Electronic	361	4.7%
Engineering Manufacturing	312	4.1%
Engineering Multidisciplinary	305	4.0%
Engineering Civil	168	2.2%
Education Educational Research	160	2.1%
Computer Science Cybernetics	154	2.0%
Economics	151	2.0%
Medical Informatics	151	2.0%
Health Care Sciences Services	124	1.6%
Environmental Sciences	109	1.4%
Social Sciences Interdisciplinary	97	1.3%
Planning Development	94	1.2%
Telecommunications	93	1.2%
Public Environmental Occupational Health	89	1.2%
Psychology Multidisciplinary	83	1.1%
Ergonomics	81	1.1%
Environmental Studies	74	1.0%
Construction Building Technology	71	0.9%
Automation Control Systems	67	0.9%

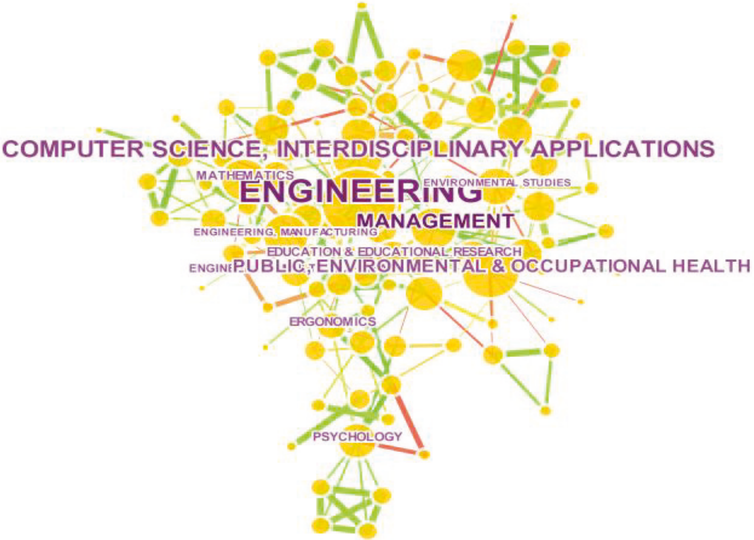

COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS
MATHEMATICS
ENGINEERING
ENVIRONMENTAL STUDIES
MANAGEMENT
ENGINEERING, MANUFACTURING
EDUCATION & EDUCATIONAL RESEARCH
PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH
ERGONOMICS
PSYCHOLOGY

Figure 7. Disciplines involved in KM.

Table 8. The betweenness centrality distribution and burst value of the KM subject.

Rank	Rank by Betweenness Centrality		Rank by Burst	
	Subject	Betweenness Centrality	Subject	Burst
1	Engineering	0.41	Computer Science, Theory, and Methods	84.68
2	Computer Science, Interdisciplinary Applications	0.31	Computer Science, Artificial Intelligence	78.32
3	Management	0.25	Computer Science	62.86
4	Public, Environmental, and Occupational Health	0.23	Computer Science, Software Engineering	30.63
5	Psychology	0.16	Computer Science, Interdisciplinary Applications	16.05
6	Mathematics	0.16	Engineering, Multidisciplinary	14.57
7	Education and Educational Research	0.16	Science and Technology—Other Topics	14.54
8	Ergonomics	0.16	Psychology, Experimental	13.8
9	Engineering, Multidisciplinary	0.14	Computer Science, Information Systems	12.25
10	Engineering, Manufacturing	0.13	Environmental Sciences	12.04
11	Environmental Studies	0.12	Green and Sustainable Science and Technology	11.7
12	Psychology, Applied	0.12	Environmental Studies	10.9
13	Science & Technology—Other Topics	0.11	Mathematical and Computational Biology	8.29
14	Environmental Sciences and Ecology	0.1	Psychology, Multidisciplinary	8.21
15a	Engineering, Chemical	0.1	Psychology	8.17
15b	Agriculture	0.1		

4.6. Journal Distribution

KM research was published in 1558 journals. The top 20 journals are displayed in Table 9. Knowledge management research publications were highly concentrated in these top journals and approximately one-third of the articles were found in these most productive journals. This is a phenomenon that follows Bradford's law and is consistent with observations in other fields. Of these top 20 journals, 1.3% of the 1558 journals had published 2449, or 32.1%, of the 7628 total articles. The major KM research journals include *Journal of Knowledge Management*, *Knowledge Management Research & Practice*, *Lecture Notes in Computer Science*, *Lecture Notes in Artificial Intelligence*, *Expert Systems with Applications*, *International Journal of Technology Management*, *Decision Support Systems*, and *Journal of Universal Computer Science*, with more than 100 articles each.

Table 9. The top 20 knowledge management publication journals.

Rank	Source Title	Records	% of Total
1	Journal of Knowledge Management	418	5.5%
2	Knowledge Management Research & Practice	244	3.2%
3	Lecture Notes in Computer Science	240	3.1%
4	Lecture Notes in Artificial Intelligence	222	2.9%
5	Expert Systems with Applications	198	2.6%
6	International Journal of Technology Management	139	1.8%
7	Decision Support Systems	109	1.4%
8	Journal of Universal Computer Science	104	1.4%
9	International Journal of Information Management	98	1.3%
10	Industrial Management Data Systems	94	1.2%
11	Information Management	64	0.8%
12	Knowledge Based Systems	62	0.8%
13	Journal of Computer Information Systems	61	0.8%
14	Journal of Information Science	60	0.8%
15	Kybernetes	59	0.8%
16	Journal of the American Society for Information Science and Technology	58	0.8%
17	Journal of Business Research	56	0.7%
18	Management Decision	55	0.7%
19	Computers In Human Behavior	54	0.7%
20	International Journal of Production Research	54	0.7%

4.7. Keyword Co-Word Network

Co-word analysis is based on the theory that research fields can be characterized and analyzed based on patterns of keyword usage in publications, which has been successfully used for examining the dynamic evolution of science [41]. Co-word analysis is a content analysis technique that is effective for mapping the strength of the association between keywords in textual data. The network map based on co-word analysis represents the search topics of a specific discipline, which is especially appropriate for describing the development of multidisciplinary fields that combine more complex knowledge. A prior study confirmed the reliability and adequacy of the co-word method for mapping the structure of a scientific field [49], which satisfactorily identified groups of research themes and the process by which fields evolved. In this study, we analyzed a total of 7628 published articles related to KM extracted from the ISI database for the period of 1974 to 2017. After processing, we obtained 13,012 keywords. Most keywords appeared only on one occasion, and only 32 keywords appeared more than 50 times. Table 10 shows the most important keywords ranked by frequency. From Table 10, Knowledge Management, with an occurrence frequency of 3401, was ranked first, followed by keywords Knowledge Sharing, Innovation, Ontology, and Knowledge Management Systems (KMs).

In the introduction, we defined the concept of knowledge management. Here, we introduce other main concepts. Knowledge sharing is an activity through which knowledge, namely information, skills, or expertise, is exchanged among people, friends, families, communities, or organizations. In the KM domain, many studies discussed the different aspects of knowledge sharing. Innovation, consistent with the OECD definition, is defined as a new or significantly improved product (a good or service), process (production or delivery method), marketing method, or managerial method [50]. In the KM field, many studies discussed the relationship between KM and innovation and found that knowledge management plays an important role in innovation. Ontology, a useful technology for KMs or KM identification, storage, and knowledge integration, has also received considerable attention from researchers and practitioners [51]. Knowledge management systems (KMs) can be defined as an information system used to collect, process, and sharing the knowledge, promoting the learning, re-use, and innovation of knowledge, and strengthening the core competence of the organization. Specifically, according to the literature, KMSs are divided into two groups: IT-based tools defined in the literature as KM-Tools, and the organizational practices defined as KM-practices [10,52–54].

Table 10. The most important key words ranked by frequency with more than 25 uses.

Author Keyword	Frequency	Author Keyword	Frequency
knowledge management	3401	case study	83
knowledge sharing	371	semantic web	79
innovation	245	data mining	78
Ontology/ontologies	268	knowledge acquisition	77
knowledge management systems	178	communities of practice	75
knowledge	160	collaboration	73
organizational learning	156	project management	71
knowledge transfer	174	social capital	67
intellectual capital	121	organizational performance	66
knowledge creation	115	absorptive capacity	64
tacit knowledge	98	performance	63
information management	95	competitive advantage	58
information technology	94	management	57
information systems	93	new product development	56
organization culture	87	Web 2.0	56
learning	83	trust	

Then the top 835 keywords with a frequency greater than or equal to five were chosen for our co-occurrence network analysis. The co-word network map displayed in Figure 8 was with VOSviewer. In the co-occurrence keyword analysis, we investigated the co-occurrence frequency of two co-occurrence keywords. The higher the co-occurrence frequency of the two words, the closer the relationship between them, which is represented by the location of the two words. The size of the node represents the frequency of the keyword co-occurrence with other keywords. We drew the following conclusion that Knowledge Sharing has a higher co-occurrence frequency with Innovation, Knowledge Creation and Ontology have a higher co-occurrence frequency with Algorithm and Ontology Change Management, and Knowledge has a higher co-occurrence frequency with Management and Competitive Advantage (Figure 8).

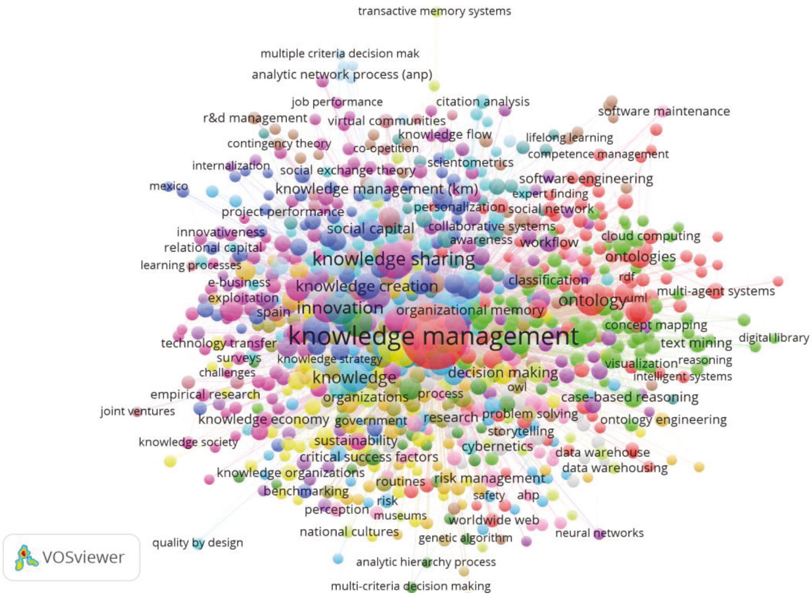


Figure 8. The co-words network of author keywords.

To statistically quantify the importance of each keyword within the co-word network, we used social network analysis. Table 11 presents a summary of the statistical results obtained using the Ucinet too. We ranked the keywords according to degree centrality and Freeman's betweenness centrality. The degree centrality indicates Knowledge Management, Knowledge Sharing, Innovation, Knowledge Transfer, and Organizational learning play an important role in KM research. Betweenness centrality confirmed the degree centrality analysis result, and highlights the keyword ontology.

Table 11. Keywords by degree centrality, betweenness centrality, and effective size.

Keyword	Degree Centrality	Keyword	Betweenness Centrality
knowledge management	804	knowledge management	409,205.3
knowledge sharing	323	knowledge sharing	33,225.5
innovation	248	innovation	17,793.34
knowledge transfer	194	ontology	10,521.03
organizational learning	185	knowledge transfer	10,416.32
knowledge	184	organizational learning	8946.413
ontology	176	knowledge	8879.887
knowledge creation	158	knowledge management systems	6898.824
knowledge management systems	155	knowledge creation	6300.001
learning	148	learning	5458.257

Table 11. *Cont.*

information systems	136	information systems	5039.425
tacit knowledge	134	tacit knowledge	4859.399
collaboration	127	information management	4369.465
information management	122	collaboration	4126.945
information technology	119	information technology	3869.026
intellectual capital	118	intellectual capital	3704.719
knowledge acquisition	114	knowledge management (km)	3635.064
performance	106	semantic web	3409.226
semantic web	104	knowledge acquisition	3283.395
communities of practice	102	communities of practice	2992.594
social capital	100	performance	2686.415
management	98	project management	2570.325
project management	96	data mining	2569.916
case study	95	management	2496.138
organizational culture	95	social capital	2375.459
data mining	89	case study	2216.155
absorptive capacity	88	new product development	2071.063
new product development	87	organizational culture	1976.951
knowledge management (km)	86	internet	1751.147
competitive advantage	83	absorptive capacity	1730.339

4.8. Intellectual Structure of Knowledge Management

Small first introduced the notion of co-citation and used the node-link network to visualize the co-citation relationship of 10 famous particle physics papers. Since then, many studies have created a visualization of co-citation relationships [39]. In a series of subsequent co-citation studies, White and Griffith documented the co-citation analysis principles and applications to map the advance of science, and identified the dynamic intellectual structure of science as a whole, or of particular domains [40]. Researchers later extended the unit of analysis from papers to authors, leading to author co-citation analysis (ACA) [55]. With many self-reflective co-citation research studies, two major types of co-citation analyses, Document Co-Citation Analysis (DCA) and Author Co-Citation Analysis (ACA) of Information Science, were used to visualize the intellectual structure of a whole domain, or of particular fields of study [40]. For this study, we used Document Co-citation Analysis (DCA) to explore the intellectual structure of knowledge management. Citespace, a tool for visualizing the intellectual structure, was used [56].

In this section, an individual network was derived from the 50 most cited articles published in the corresponding time period of two years, which ranging from 1974 to 2017. Then, these networks were merged into a network of 295 co-cited references that form an overview of the evolution of a scientific field over time (Figure 9). To improve the clarity of a visualized evolution network, we used a simplified network using pruning [31]. Here, a topology-based approach instead of a threshold-based approach was chosen for to more extensively consider intrinsic topological properties [56–58]. In this study, pathfinder network scaling instead of minimal spanning trees was used to preserve the chronological growth patterns in the co-citation networks. In Figure 9, the size of a node indicates the number of citations received by the associated reference. Each node is depicted with a series of citation tree-rings across the time frame slices. The structural properties of a node are displayed with a purple ring. The thickness of the purple ring indicates the degree of its betweenness centrality. Table 12 shows the most cited articles with detailed indicators.

From Table 12, the most cited papers by citation counts were during the period of 1995 to 2010. There are two main reasons for this phenomenon. The first is that modern knowledge management rapidly gained in popularity after 2000. The second is that the papers published in recent years need approximately 13–15 years to reach the highest number of citations.

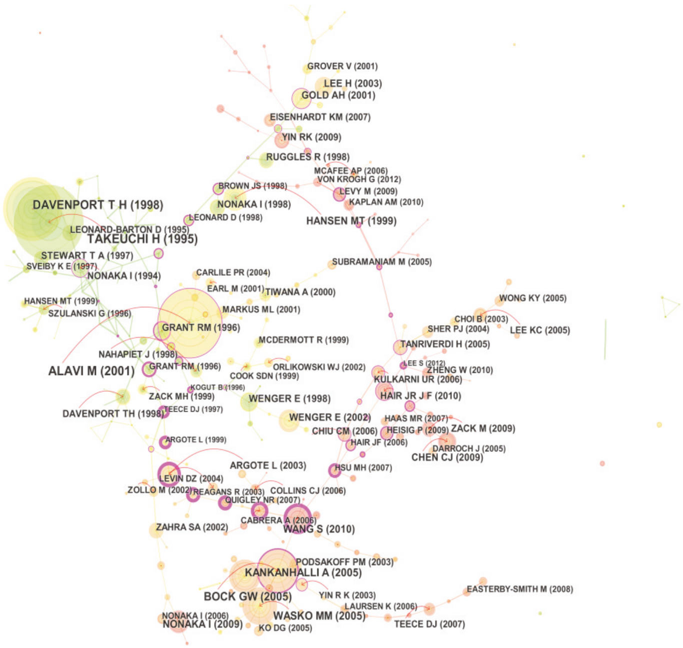


Figure 9. Citations in knowledge management research, shown as a Pathfinder network of cited references.

Table 12. The top 15 most cited papers by citation counts.

Rank	Citation Counts	First Author	Journal (Book)	Year
1	231	Nonaka	Oxford University Press	1995
2	212	Alavi	MIS Quarterly	2001
3	199	Davenport	Harvard Business School Press	1998
4	133	Bock	MIS Quarterly	2005
5	118	Wasko	MIS Quarterly	2005
6	109	Kankanhalli	MIS Quarterly	2005
7	80	Hansen	Harvard Business Review	1999
8	77	Wang	Human Resource Management Review	2010
9	77	Nonaka	Organization Science	2009
10	76	Wenger	Harvard Business School Press	2002
11	70	Wenger	Systems Thinker	1998
12	67	Gold	Journal of Management Information Systems	2001
13	66	Lee	Journal of Management Information Systems	2003
14	65	Chen	Journal of Business Research	2009
15a	63	Argote	Management Science	2003
15b	63	Hair	Prentice-Hall	2010

To further investigate the features of the intellectual structure of KM research, we used cluster mapping of co-citation document networks to complete a visualization analysis of the evolution of the intellectual base in the KM field. Based on the co-citation document networks, we used Citespace to divide the co-citation network into a number of clusters of co-cited references. These references are tightly connected within the same clusters, but loosely connected between different clusters. Table 13 lists 15 major clusters by their size, that is, the number of members in each cluster. Clusters with fewer members tend to be less representative than larger clusters because small clusters are likely to be formed by the citing behavior of a small number of publications.

Table 13. Summary of the largest 15 KM clusters.

ID	Size	Silhouette	Label (TF*IDF)	Label (LLR)	Label (MI)	Mean Year
0	26	0.961	profitability	Asia; call center; case study; modularity; dynamic capability	knowledge acquisition; knowledge creation; knowledge sharing; knowledge transfer; barriers and facilitator	2009
1	20	0.974	profitability	knowledge management system; new product development; organizational knowledge management; corporate strategy; product development	quality; knowledge management system; dimension; customer orientation; information system	2000
2	20	0.991	social constructionist analysis; pseudo-knowledge sharing	technology mediated learning; knowledge sharing; identity; gender; enjoyment	service quality; strategy; model; satisfaction; performance; success	2004
3	18	0.961	empirical analysis	boundary spanning; ERP system; ERP usage; key user; information technology professional	innovation; thinking; managing knowledge; systems thinking; information	2011
4	18	0.975	information sharing; work groups	human resource management; innovation; ultra-peripheral region; manufacturing performance; knowledge management	human capital; human resource management; innovation; start up; human resource management	2007
5	18	0.936	business format; concept	knowledge organization; product development; creation theory; management; community	dynamic capability; process alignment; organizational learning culture; competitive advantage; information technology	2001
6	18	0.93	knowledge management tutorial; people	knowledge management; cognitive congruence; schema; relationship script; resource-based view	resource based view; competitive advantage; firm; epistemology; creativity	1999
7	18	0.93	knowledge management; knowledge assets	organizational impact of knowledge-based system; knowledge engineering; core competency; job quality; knowledge-based system	intellectual capital measurement; knowledge management; intangible assert cognition	1995
8	18	0.835	information technology management; successful knowledge management projects	strategic alliance; knowledge transfer; causal ambiguity; organizational learning; knowledge management	resource based view; competitive advantage; firm; epistemology; creativity	1995
9	17	0.919	biotechnology sector	transitive memory system; team performance; field study; group decision making; coordinating expertise	human capital; human resource management; innovation; start up; corporate	2002
10	15	0.8	information technology management; information technology	Socio technical system; organizational memory; firm; appropriation problem; technological change	management of technology; technological learning; knowledge management; knowledge transfer; strategy	1994
11	15	0.894	impact; innovation	information sharing/withholding; knowledge transfer; reference group; profession; science	knowledge market; dyadic knowledge; knowledge management; knowledge exchange; intangible knowledge	2000
12	14	0.953	social media research; influence	knowledge based view; information system; social software; open innovation; managing knowledge	knowledge acquisition; knowledge creation; knowledge sharing; knowledge transfer; barriers and facilitator	2010
13	13	0.954	fundamental issue; antecedents	information technology; organizational performance; business performance; competitive advantage; research proposition	dynamic capability; process alignment; organizational learning culture; competitive advantage; information technology	2005
14	10	0.949	learning processes; firm level perspective	organizational change; product development; model; transformation; technology	social interaction; organization; innovation performance; manufacturing firm	2006

Note: Clusters are referred to in terms of the labels selected by the TF*IDF, LLR (log-likelihood ratio), and MI (mutual information) methods. In information retrieval, TF*IDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus, whereas those chosen by log-likelihood ratio(LLR) tests and mutual information(MI) tend to reflect a unique aspect of a cluster [44,59].

Cluster #0 was the largest clusters, containing 26 nodes, and the value of the silhouette is 0.961. As the cluster was the largest cluster in the literature co-citation network, the theme of this cluster was relatively fragmented. To obtain more information about Cluster #0, we used Carrot to explain Cluster #0 in more detail. Table 14 outlines Cluster #0 using the lingo algorithm.

Table 14. Details of the largest cluster (Cluster #0).

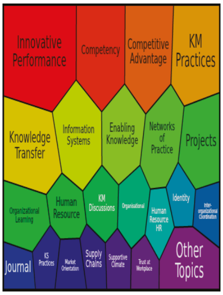
Cluster Details	Cited Articles (Ranked by Citations)			
	Frequency	First Author	Year	Title
 Innovative Performance Competency Competitive Advantage KM Practices Knowledge Transfer Information Systems Enabling Knowledge Networks of Practice Projects Organizational Learning Human Resource KM Discussions Organisational Human Resource HR Identity Inter-organizational Coordination Journal KS Practices Market Orientation Supply Chains Supportive Climate Trust at Workplace Other Topics	65	Chen	2009	Strategic human resource practices and innovation performance: the mediating role of knowledge management capacity
	63	Hair	2010	Multivariate data analysis: A global perspective
	55	Zack	2009	Knowledge management and organizational performance: an exploratory analysis
	43	Heisig	2009	Harmonisation of knowledge management—comparing 160 KM frameworks around the globe
	40	Zheng	2010	Linking organizational culture, structure, strategy, and organizational effectiveness: Mediating role of knowledge management
	38	Hair	2006	Multivariate data analysis: A global perspective
	36	Haas	2007	Different knowledge, different benefits: Toward a productivity perspective on knowledge sharing in organizations
	36	Hsu	2007	Knowledge sharing behavior in virtual communities: The relationship between trust, self-efficacy, and outcome expectations
	35	Darroch	2005	Knowledge management, innovation and firm performance
	33	He	2009	A comparison of purchase decision calculus between potential and repeat customers of an online store

Table 14 shows that the earliest article in Cluster #0, "Knowledge management, innovation and firm performance" [60], mainly described the relationship between knowledge management and firm performance, which was then followed by the studies of Hair [61] and Haas [62]. Ranked by cited frequency, the core members of Cluster #0 represent major milestones in relation to knowledge management in or across organizations, including knowledge performance, competency, knowledge for innovation, and knowledge sharing. The second largest clusters (#1 and #2) both have 20 members and silhouette values of 0.971 and 0.991, respectively. We also used Carrot to explain Cluster #1 in more detail (Table 15). Ranked by cited frequency, the core members of Cluster #1 represent major milestones in relation to knowledge value, including the basic theory of knowledge value for firms, knowledge assets, and knowledge value.

Table 16 details Cluster #2. From Table 16, the most active citation in the cluster was "Behavioral Intention Formation in Knowledge Sharing: Examining the Roles of Extrinsic Motivators, Social-Psychological Factors, and Organizational Climate". The core members of Cluster #2 represent major milestones of knowledge management research from the psychological perspective.

We also sorted the citation curve that includes the betweenness centrality and burst. The betweenness centrality of a node in the network measures the importance of the position of the node in the network. Table 17 shows 10 essential references in the synthesized network with high centrality. These references are important in terms of how they connect individual nodes in the network, and how they connect aggregated groups of nodes, such as co-citation clusters. Four of these nodes are in Cluster #11 and Cluster #4. These works can be seen as landmark works in the context of our broadly defined area of management.

Table 15. Details of the second largest cluster (Cluster #1).

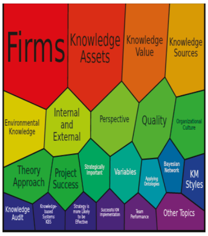
Cluster Details	Cited Articles (Ranked by Number of Citations)			
	Frequency	First Author	Year	Title
 Firms Knowledge Assets Knowledge Value Knowledge Sources Environmental Knowledge Internal and External Perspective Quality Organizational Culture Theory Approach Project Success Strategically Important Variables Applying Ontologies Bayesian Network KM Styles Knowledge Audit Knowledge-based Systems KBS Strategy is more Likely to be Effective Successful KM Implementation Team Performance Other Topics	80	Hansen	1999	What's your strategy for managing knowledge?
	67	Gold	2001	Knowledge Management: An Organizational Capabilities Perspective
	66	Lee	2003	Market Process Reengineering through Electronic Market Systems: Opportunities and Challenges
	61	Nonaka	1998	The concept of 'Ba': building a foundation for knowledge creation.
	57	Ruggles	1998	The state of the notion: knowledge management in practice
	43	Lee	2001	Exploring mediation between environmental and structural attributes: the penetration of communication technologies in manufacturing organizations
	37	Brown	1998	Organizing Knowledge
	31	Liao	2003	Knowledge management technologies and applications-literature review from 1995 to 2002
	30	Dell	1998	If Only We Knew What We Know: Identification and Transfer of Internal Best Practices
	29	Becerra-Fernandez	2001	Organizational Knowledge Management: A Contingency Perspective

Table 16. Details of the second largest cluster (Cluster #2).

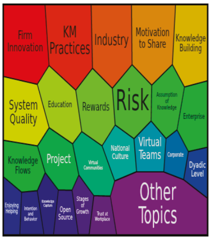
Cluster Details	Cited Articles (Ranked by Number of Citations)			
	Frequency	First Author	Year	Title
 Firm Innovation KM Practices Industry Motivation to Share Knowledge Building System Quality Education Rewards Risk Assumption of Knowledge Enterprise Knowledge Flows Project Virtual Communities National Culture Virtual Teams Corporate Dyadic Level Enjoying Helping Intention and Behavior Knowledge Capture Open Source Stages of Growth Trust at Workplace Other Topics	133	Bock	2005	Behavioral intention formation in knowledge sharing: examining the roles of extrinsic motivators, social-psychological factors, and organizational climate
	118	Wasko	2005	Why should I share? Examining social capital and knowledge contribution in electronic net- works of practice
	109	KankanHalli	2005	Contributing knowledge to electronic knowledge repositories: an empirical investigation
	50	Podsakoff	2003	Common method biases in behavioral research: A critical review of the literature and recommended remedies
	44	Ko	2005	Antecedents of Knowledge Transfer from Consultants to Clients in Enterprise System Implementations
	34	Lin	2007	A stage model of knowledge management: An empirical investigation of process and effectiveness
	32	Garud	2005	Vicious and Virtuous Circles in the Management of Knowledge: The Case of Infosys Technologies
	29	Ardichvili	2003	Motivation and barriers to participation in virtual knowledge-sharing communities of practice
	28	Alavi	2005	An Empirical Examination of the Influence of Organizational Culture on Knowledge Management Practices
	28	Delone	2003	Information System Success: A Ten Years Update

Table 17. Betweenness centrality ranking of the citations.

Centrality	First Author	Year	Source	Cluster ID
0.91	Wang	2010	Human Resource Management Review	4
0.89	Teece	1997	Strategic Management Journal	11
0.84	Cabrera	2006	The International Journal of Human Resource Management	4
0.81	Argote	2003	Management Science	11
0.81	Reagans	2003	Administrative Science Quarterly	11
0.81	Argote	1999	Springer, Berlin	11
0.8	Quigley	2007	Organization Science	11

Table 17. *Cont.*

0.67	Hsu	2007	International Journal of Human-Computer Studies	0
0.48	Grant	1996	Organization Science	11
0.41	Lee	2012	Journal of Knowledge Management	0

A citation burst has two attributes: the intensity of the burst and the length of the burst status. Table 18 lists references with the strongest citation bursts across the entire dataset during the study period of 1974 to 2017. The first article with a strong citation burst is "Working Knowledge: How Organizations Manage What They Know" from Cluster #24. The second-ranked article is "Situated Learning: Legitimate Peripheral Participation" and "Multivariate Data Analysis" is ranked third.

Table 18. Top 15 references with strongest citation bursts.

Strength	Reference	Burst Start Year	Burst End Year
106.699	Spender, J.C.; 1996, Strategic Management Journal	1996	2003
78.583	Grant, R.M.; 1996, Strategic Management Journal	2005	2009
70.1433	Nonaka, I.; 1995, Oxford University Press	1999	2006
30.9065	Alavi, M.; 2001, MIS Quart	2010	2013
29.37	Davenport, T.H.; 1998, Harvard Business School Press	2013	2017
29.1145	Bock, G.W.; 2005, MIS Quart	2005	2002
26.2479	Wang, S.; 2010, Human Resource Management Review	2010	2013
25.7965	Nonaka, I.; 1994, Organization Science	2009	2013
25.3509	Wasko, M.M.; 2005, MIS Quart,	2005	2007
24.9287	Kankanhalli, A.; 2005, MIS Quart	2013	2017
24.7793	Hansen, M.T.; 1999, Harvard Business Review	2000	2006
23.7432	Hair, J.F., Jr.; 2010, Prentice-Hall	2010	2003
22.1144	Wenger, E.; 1998, Cambridge University Press	1998	2005
22.0813	Leonard-Barton, D.; 1995, Harvard Business School Press	2000	2006
21.9779	Stewart, T.A.; 1997, Crown Business	1997	2004

4.9. Emerging Trends

The modularity of a network measures the degree to which nodes in the network can be divided into a number of groups, such that nodes within the same group are connected tighter than the nodes in different groups. The collective intellectual structure of the knowledge of a scientific field can be represented as associated networks of co-cited references. These networks evolve over time. Newly published articles may introduce profound structural variation or have little or no impact on the structure. Figure 10 shows the changes in the modularity of networks during the past 10 years. Each network was constructed based on a two-year sliding window. The number of publications per year increased considerably. The modularity dipped in 2012 and then returned to the previous level. Based on this observation, groundbreaking works plausibly appeared in 2012.

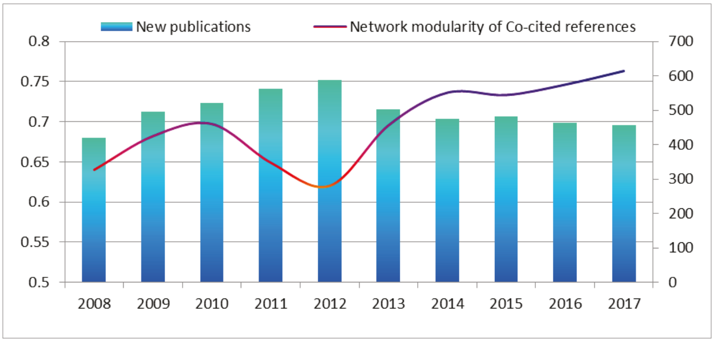


Figure 10. The modularity of the network.

Therefore, we specifically investigated potential emerging trends in 2012, and attempted to explain the significant decrease in the modularity of the network. If the publications in 2012 had a subsequent citation burst, then we expected that the publication played an important role in changing the overall intellectual structure. Ten publications in 2012 were found to have subsequent citation bursts (Table 19). Notably, from Table 19, Krogh [63] and Andreeva [64] were ranked first and second on the list. Both introduced research topics about new phenomena and the practice of knowledge management, and have current citation bursts after 2014. Other articles on the list address other research topics about SMEs management based on knowledge perspective, innovation, performance, and big data. These observations suggest that the modularity change in 2012 is an indication of an emerging trend in these areas.

Table 19. Articles published in 2012 with subsequent citation bursts in descending order of local citation counts.

Reference	Citations	Title	Source	Burst	Duration
Krogh	35	How does social software change knowledge management? Toward a strategic research agenda	Journal of Strategic Information Systems	15.8391	2014–2017
Andreeva	32	Does knowledge management really matter? Linking knowledge management practices competitiveness and economic performance	Journal of Knowledge Management	14.9919	2014–2017
Durst	30	Knowledge management in SMEs: a literature review	Journal of Knowledge Management	13.963	2014–2017
Chen	21	Business intelligence and analytics: from big data to big impact	MIS Quarterly	10.0375	2014–2017
Zhou	18	How knowledge affects radical innovation: Knowledge base, market knowledge acquisition, and internal knowledge sharing	Strategic Management Journal	8.4999	2014–2017
Guthire	14	Reflections and projections: A decade of Intellectual Capital Accounting Research	British Accounting Review	7.8466	2015–2017
Lee	18	An integrated view of knowledge management for performance	Journal of Knowledge Management	7.8046	2014–2017
McAfee	12	Big data: the management revolution.	Harvard Business Review	6.7234	2015–2017
Podsakoff	14	Sources of method bias in social science research and recommendations on how to control it.	Annual Review of Psychology	6.457	2015–2017
Chan	11	An empirical investigation of factors affecting e-collaboration diffusion in SMEs	International Journal of Production Economics	5.6126	2014–2017

Four articles published in 2012 with subsequent citation bursts were review articles. Therefore, we deduced that review articles provide easier access to more citations in a short time period than other types of publications. This is consistent with previous studies [65].

5. Discussion and Conclusions

5.1. Discussion

Considering the limitations imposed by subjective judgment, chosen research scope in terms of time frame, analytical unit, and the lack of visualization perspective of prior publications, our paper comprehensively investigates global knowledge management from 1974 to 2017 to provide a quick overview of KM research. In this study, a coherent comprehensive bibliometric evaluation framework was used to investigate an emerging and promising cross-disciplinary domain, KM. We outlined the key development landscape of KM, including the growth pattern, international collaboration of countries, institutions, author distribution, intellectual structure, and emerging trends. The growth analysis showed that the scientific KM research is emerging as a cross-disciplinary domain among computer science, information science, management, and other research areas. The published KM papers significantly increased since 1991 in an S-shaped pattern, which is consistent with the analysis performed by Styhre [23]. The subsequent country (territory) comparative analysis indicated the U.S., England, Taiwan, and China are the four largest contributors of the published KM literature. Compared with the findings of Gu, Japan and Canada were replaced by Taiwan and China [11]. The scientific research cooperation network analysis indicated that the U.S. is not only the original contributor, but also the largest international collaborating country. England is a close second with 246 international collaborated articles and China ranked third. National Cheng Kung University in Taiwan, Hong Kong Polytechnic University in China, and City University of Hong Kong (China) were the three largest contributors. We observed a decline in single-authored studies and relative stability in studies with two or three authors, and a clear growth trend in multi-authored articles, which is consistent with the analysis of single-authored and multi-authored KM studies [32].

The major publications for knowledge management research include *Journal of Knowledge Management*, *Knowledge Management Research & Practice*, and *Lecture Notes in Computer Science*. These findings agree with prior scientometric research that only highlighted the importance of *Journal of Knowledge Management* [11,23,32].

The visual co-word keyword analysis determined that Knowledge Management, Knowledge Sharing, Innovation, Ontology, KMs, Knowledge Management Systems, and Knowledge are consistent hotspots in KM research. The co-words network analysis showed that the central term Knowledge Management is closely related to the terms Ontology, Organizational Learning, Knowledge Sharing, and Information Technology. These combinations of related issues show that the KM research is focused on knowledge acquisition and sharing to improve knowledge management performance and organization dynamic capacity. This finding supports the conclusion on KM research in business literature as an independent stream, as stated by Akhavan et al. [32].

With the visual co-citation network analysis of references performed with CiteSpace and Carrot, we defined the intellectual structures of knowledge management, and found that four emerging research topics focus on new phenomena and the practice of knowledge management, SMEs management based on knowledge perspective, innovation and performance, and big data-enabled KM.

For new phenomena and the practice of knowledge management, rapid technological changes affect the information and communication technologies that are providing new data mining and predictive analytics solutions. Additionally, the rapid development of social networks, such as Facebook and Twitter [66,67], also influences knowledge management. So, given this context, we can formulate the first research questions (RQ):

RQ1: *How do different emerging technologies change knowledge management?*

Secondly, for SMEs management based on the knowledge perspective, some studies have emphasized the importance of the role of KM in small- and medium-sized enterprises. A consensus conclusion shows that SMEs are starting to make focus on KM practices. However, little research has been completed about the KM of SMEs. Most notably, few empirical studies have been performed on SMEs [6,7,15]. Some academics have focused on SMEs and discussed the KM of SMEs, but some important research issues have been neglected. Given this context, we formulated the next two research questions:

RQ2: *What are the critical difference between SMEs KM and large companies?*
RQ3: *How should the effective development of SMEs KM be promoted?*

Third, innovation and performance based on KM or a KM-based viewpoint is a hot topic in the KM domain. Many studies discussed the relationship between knowledge management and innovation and performance. However, its mechanism is still unclear [21,50]. In additional, most studies focused on large companies. SMEs and startup company innovation and performance research based on the KM perspective should be highlighted. It was then possible to formulate the fourth research question:

RQ4: *How to promote the mechanism research among knowledge management, innovation and performance, not only in large companies but also in SMEs and startups?*

Fourth is big data-enabled KM. The rapid development of big data has created many challenges for KM. Big data can be considered as a knowledge asset, and thus the field of knowledge management gained new momentum with the introduction of big data analytics for knowledge creation [14]. So, we formulated the fifth research question:

RQ5: *How should big data be managed to address the challenges of KM caused by big data?*

For additional studies, we examined the papers' abstracts, and found that most of the literatures on knowledge transfer and knowledge sharing have introduced various measures to promote knowledge sharing, but few were successful in practice. Therefore, we must strengthen the transfer between KM academic research and KM practice. From this, we propose the next research question:

RQ6: *How can the communication between KM academic research and KM practice be strengthened?*

Another research gap was also observed. Previous studies usually focused on the research of knowledge sharing, transfer, and creation, and lacked research on KM failures, such as knowledge hiding and knowledge hoarding [68,69]. Some scholars have begun to focus on this kind of behavior. However, the critical factors leading to these behaviors are still unclear. Therefore, determining the critical factors is an important task for future knowledge management research about negative behavior. From this, we propose the last two research questions:

RQ7: *What are the main behaviors leading to KM failure?*
RQ8: *What are the critical factors leading to KM failure?*

5.2. Implications for Academics and Practitioners

Based on the above proposed research gaps and questions, our results provide guidance and draw implications for future research and practices. For academics, these implications may offer some possible areas or interesting questions for the development of KM.

On the other hand, the findings above have implications for both academics and practitioners. Firstly, the research presented in this paper particularly benefits academics, researchers, and research students wanting to quickly obtain a visualization overview of knowledge management research.

The research topic analysis, which was based on co-keywords, can also be useful for curriculum designing. For example, considering their importance in KM research, knowledge sharing, knowledge and innovation, ontology, knowledge management systems, knowledge transfer, organizational learning, and knowledge creation should be included in curricula for graduate and undergraduate programs about KM.

Based on our findings about the emerging trends in KM, researchers can better understand the development of KM and quickly and efficiently determine valuable research topics for the future. New phenomena and the practice of knowledge management, SMEs management based on knowledge perspective, innovation and performance, and big data-enabled KM are emerging research topics, which should receive more attention from researchers in this field.

Moreover, by identifying the current KM research status, this study provides an opportunity for practitioners and academics to check the extent to which academic research is keeping pace with the KM issues confronted by managers. This may become a starting point for communication between academics and practitioners.

5.3. Limitations and Direction for Future Research

The results from our study should be interpreted in light of several potential limitations due to the research design and the intrinsic drawbacks of bibliometric methods. First, by focusing on two research objectives, we used 7628 original research articles retrieved from the Web of Science core collection for bibliometric analyses, which may be criticized. Although the Web of Science core collection is an effective and good data source for bibliometric analysis, some limitations exist if it is used as a unique database. Future research can address this limitation by expanding the data sources used and merging the data from various databases, like Scopus, Emerging, and PubMed.

Secondly, we mainly used the frequency indicator to outline the present KM situation because frequency is the most commonly used indicator in bibliometric analyses. However, some valuable units may be ignored. Although betweenness centrality and degree centrality were also used to improve our analysis of international collaboration of among countries, distribution and collaboration of institution, and co-word keyword networks, future research is still needed to integrate various indicators.

Lastly, in the intellectual structure analysis section, our study followed the general paradigm of bibliometric research, and did not analyze the epistemology and ontology problems in the articles, which may cause some misunderstanding for readers. This is due to the limited functions of the intrinsic drawbacks of bibliometric analyses. However, we believe that considering the problems about epistemology and ontology in the articles is important and valuable. Therefore, we hope to address up this gap by introducing more methods, like rounded theory method and systematic reviews, in future research.

Acknowledgments: This research is funded by National Natural Science Foundation of China (71372085). The authors are very grateful for the valuable comments and suggestions of the anonymous reviewers, which significantly improved the article.

Author Contributions: Peng Wang designed this research and collected the data set for the experiment. Fang-Wei Zhu analyzed the data to show the validity of this paper. Hao-Yang Song, Jian-Hua Hou and Jin-Lan Zhang wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Spender, J.C. Making knowledge the basis of a dynamic theory of the firm. *Strateg. Manag. J.* **1996**, *17*, 45–62. [CrossRef]
2. Nissen, M.E. Redesigning reengineering through measurement-driven inference. *MIS Q.* **1998**, *22*, 509–534. [CrossRef]
3. Pirró, G.; Mastroianni, C.; Talia, D. A framework for distributed knowledge management: Design and implementation. *Future Gener. Comput. Syst.* **2010**, *26*, 38–49. [CrossRef]
4. Wiig, K.M. Knowledge management: Where did It Come From and Where Will It Go? *Expert Syst. Appl.* **1997**, *13*, 1–14. [CrossRef]
5. Bhatt, G.D. Organizing knowledge in the knowledge development cycle. *J. Knowl. Manag.* **2004**, *1*, 15–26. [CrossRef]
6. Wong, K.Y.; Aspinwall, E. An empirical study of the important factors for knowledge management adoption in the SME sector. *J. Knowl. Manag.* **2005**, *9*, 64–82. [CrossRef]
7. Centobelli, P.; Cerchione, R.; Esposito, E. Knowledge management systems: The hallmark of SMEs. *Knowl. Manag. Res. Pract.* **2017**, *15*, 294–304. [CrossRef]
8. Money, W.; Turner, A. Knowledge Management Information Technology User Acceptance: Assessing the Applicability of the Technology Acceptance Model. In *Knowledge Management in Modern Organizations*; Jennex, M., Ed.; Idea Group Inc.: Calgary, AB, USA, 2007; pp. 233–254.
9. Nikabadi, S.M. Framework for knowledge management processes in supply chain. *Ira. J. Inf. Process. Manag.* **2014**, *28*, 611–642.
10. Alavi, M.; Leidner, D.E. Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Q.* **2001**, *25*, 107–136. [CrossRef]
11. Gu, Y. Global knowledge management research: A bibliometric analysis. *Scientometrics* **2004**, *61*, 171–190. [CrossRef]
12. Henry, N. Bureaucracy, technology, and knowledge management. *Public Adm. Rev.* **1975**, *35*, 572–578. [CrossRef]
13. Barclay, B.R.O.; Murray, P.C. What is knowledge management? *Knowledge Praxis*, 11 May 2009. Available online: http://www.mediaaccess.com/whatis.html (accessed on 5 May 2017).
14. Esposito, C.; Ficco, M.; Palmieri, F.; Castiglione, A. A knowledge-based platform for big data analytics based on publish/subscribe services and stream processing. *Knowl.-Based Syst.* **2015**, *79*, 3–17. [CrossRef]

15. Durst, S.; Edvardsson, I.R. Knowledge management in SMEs: A literature review. *J. Knowl. Manag.* **2012**, *16*, 879–903. [CrossRef]
16. Cerchione, R.; Esposito, E.; Spadaro, M.R. A literature review on knowledge management in SMEs. *Knowl. Manag. Res. Pract.* **2016**, *14*, 169–177. [CrossRef]
17. Cerchione, R.; Esposito, E. A systematic review of supply chain knowledge management research: State of the art and research opportunities. *Int. J. Prod. Econ.* **2016**, *182*, 276–292. [CrossRef]
18. Inkinen, H. Review of empirical research on knowledge management practices and firm performance. *J. Knowl. Manag.* **2016**, *20*, 230–257. [CrossRef]
19. Centobelli, P.; Cerchione, R.; Esposito, E. Knowledge management in startups: Systematic literature review and future research agenda. *Sustainability* **2017**, *9*, 361. [CrossRef]
20. Nordenflycht, A.V. What is a professional service firm? Toward a theory and taxonomy of knowledge-intensive firms. *Acad. Manag. Rev.* **2010**, *35*, 155–174. [CrossRef]
21. Leiponen, A.; Helfat, C.E. Innovation objectives, knowledge sources, and the benefits of breadth. *Strateg. Manag. J.* **2010**, *31*, 224–236. [CrossRef]
22. Ensign, P.C.; Hébert, L. How reputation affects knowledge sharing among colleagues. *MIT Sloan Manag. Rev.* **2010**, *51*, 79–81.
23. Styhre, A. *Understanding Knowledge Management: Critical and Post-Modern Perspectives*; Business School Press: Copenhagen, Denmark, 2003.
24. Butler, F.A.; Stevens, R. Standardized assessment of the content knowledge of English language learners k–12: Current trends and old dilemmas. *Lang. Test.* **2001**, *18*, 409–427.
25. Lee, M.R.; Chen, T.T. *Visualizing Trends in Knowledge Management. Knowledge Science, Engineering and Management*; Springer: Berlin, Germany, 2007; pp. 362–371.
26. Lee, M.R.; Chen, T.T. Revealing research themes and trends in knowledge management: From 1995 to 2010. *Knowl.-Based Syst.* **2012**, *28*, 47–58. [CrossRef]
27. Li, C.; Guo, F.; Zhi, L.; Han, Z.; Liu, F. Knowledge management research status in china from 2006 to 2010: Based on analysis of the degree theses. *Scientometrics* **2013**, *94*, 95–111. [CrossRef]
28. Dwivedi, Y.K.; Venkitachalam, K.; Sharif, A.M.; Al-Karaghouli, W.; Weerakkody, V. Research trends in knowledge management: Analyzing the past and predicting the future. *Inf. Syst. Manag.* **2011**, *28*, 43–56. [CrossRef]
29. Serenko, A.; Bontis, N. Global ranking of knowledge management and intellectual capital academic journals. *J. Knowl. Manag.* **2009**, *13*, 4–15. [CrossRef]
30. Serenko, A.; Dumay, J. Citation classics published in knowledge management journals. Part-i: Articles and their characteristics. *J. Knowl. Manag.* **2015**, *19*, 401–431. [CrossRef]
31. Serenko, A. Meta-analysis of scientometric research of knowledge management: Discovering the identity of the discipline. *J. Knowl. Manag.* **2013**, *17*, 773–812. [CrossRef]
32. Akhavan, P.; Ebrahim, N.A.; Fetrati, M.A.; Pezeshkan, A. Major trends in knowledge management research: A bibliometric study. *Scientometrics* **2016**, *107*, 1–16. [CrossRef]
33. Wallace, D.P.; Fleet, C.V.; Downs, L.J. The research core of the knowledge management literature. *Int. J. Inf. Manag.* **2011**, *31*, 14–20. [CrossRef]
34. Romo-Fernández, L.M.; Guerrero-Bote, V.P.; Moya-Anegón, F. Co-word based thematic analysis of renewable energy (1990–2010). *Scientometrics* **2013**, *97*, 743–765. [CrossRef]
35. Braun, T.; Schubert, A. A quantitative view on the coming of age of inter-disciplinarity in the sciences 1980–1999. *Scientometrics* **2003**, *58*, 183–189. [CrossRef]
36. Rinia, E.J.; Leeuwen, T.N.V.; Vuren, H.G.V.; Raan, A.F.J.V. Comparative analysis of a set of bibliometric indicators and central peer review criteria: Evaluation of condensed matter physics in the netherlands. *Res. Policy* **1998**, *27*, 95–107. [CrossRef]
37. Takeda, Y.; Mae, S.; Kajikawa, Y.; Matsushima, K. Nanobiotechnology as an emerging research domain from nanotechnology: A bibliometric approach. *Scientometrics* **2009**, *80*, 23–38. [CrossRef]
38. Garfield, E. Historiographic mapping of knowledge domains literature. *J. Inf. Sci.* **2004**, *30*, 119–145. [CrossRef]
39. Small, H. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Am. Soc. Inf. Sci.* **1973**, *24*, 265–269. [CrossRef]
40. White, H.D.; Griffith, B.C. Author co-citation: A literature measure of intellectual structure. *J. Am. Soc. Inf. Sci.* **1981**, *32*, 163–171. [CrossRef]

41. Callon, M.; Courtial, J.P.; Laville, F. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics* **1991**, *22*, 155–205. [CrossRef]
42. Leydesdorff, L. Top-down decomposition of the journal citation reportof the social science citation index: Graph-and factor-analytical approaches. *Scientometrics* **2004**, *60*, 159–180. [CrossRef]
43. Persson, O.; Danell, R.; Schneider, J.W. How to use Bibexcel for various types of bibliometric analysis. In *Celebrating Scholarly Communication Studies: A Festschrift for Olle Persson at his 60th Birthday*; Umeå University Library: Umeå, Sweden, 2009; pp. 9–24.
44. Chen, C. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *J. Am. Soc. Inf. Sci. Technol.* **2006**, *57*, 359–377. [CrossRef]
45. Borgatti, S.P.; Everett, M.G.; Freeman, L.C. Ucinet for Windows: Software for Social Network Analysis. 2002. Available online: http://www.citeulike.org/group/11708/article/6031268 (accessed on 20 February 2018).
46. Van Eck, N.J.; Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **2010**, *84*, 523–538. [CrossRef] [PubMed]
47. Cobos, C.; Muñoz-Collazos, H.; Urbano-Muñoz, R.; Mendoza, M.; León, E.; Herrera-Viedma, E. Clustering of web search results based on the cuckoo search algorithm and Balanced Bayesian Information Criterion. *Inf. Sci.* **2014**, *281*, 248–264. [CrossRef]
48. Wagner, C.S.; Leydesdorff, L. Network structure, self-organization, and the growth of international collaboration in science. *Res. Policy* **2005**, *34*, 1608–1618. [CrossRef]
49. Whittaker, J. Creativity and conformity in science: Titles, keywords and co-word analysis. *Soc. Stud. Sci.* **1989**, *19*, 473–496. [CrossRef]
50. Manley, K.; McFallan, S.; Kajewski, S. Relationship between construction firm strategies and innovation outcomes. *J. Constr. Eng. Manag.* **2009**, *135*, 764–771. [CrossRef]
51. Fensel, D. Ontology-based knowledge management. *Computer* **2002**, *35*, 56–59. [CrossRef]
52. Fink, K.; Ploder, C. Knowledge Management Toolkit for SMEs. *Int. J. Knowl. Manag.* **2009**, *5*, 46–60. [CrossRef]
53. Centobelli, P.; Cerchione, R.; Esposito, E. Aligning enterprise knowledge and knowledge management systems to improve efficiency and effectiveness performance: A three-dimensional Fuzzy-based decision support system. *Expert Syst. Appl.* **2018**, *91*, 107–126. [CrossRef]
54. Cerchione, R.; Esposito, E. Using knowledge management systems: A taxonomy of SME strategies. *Int. J. Inf. Manag.* **2017**, *37*, 1551–1562. [CrossRef]
55. Zhang, J.; Chen, C.; Li, J. Visualizing the intellectual structure with paper-reference matrices. *IEEE Trans. Vis. Comput. Graph.* **2009**, *15*, 1153–1160. [CrossRef] [PubMed]
56. Chen, C. Searching for intellectual turning points: Progressive knowledge domain visualization. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5303–5310. [CrossRef] [PubMed]
57. Small, H.; Upham, P. Citation structure of an emerging research area on the verge of application. *Scientometrics* **2009**, *79*, 365–375. [CrossRef]
58. Skuce, D.; Lethbridge, T.C. Code4: A unified system for managing conceptual knowledge. *Int. J. Hum. Comput. Stud.* **1995**, *42*, 413–451. [CrossRef]
59. Chen, C.; Ibekwe-Sanjuan, F.; Hou, J. The structure and dynamics of cocitation clusters: A multiple-perspective co-citation analysis. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 1386–1409. [CrossRef]
60. Darroch, J. Knowledge management, innovation and firm performance. *J. Knowl. Manag.* **2005**, *9*, 101–115. [CrossRef]
61. Hair, J.F.; Black, W.C.; Babin, B.J.; Anderson, R.E. *Multivariate Data Analysis: A Global Perspective*; Prentice Hall: Upper Saddle River, NJ, USA, 2006.
62. Haas, M.R.; Hansen, M.T. Different knowledge, different benefits: Toward a productivity perspective on knowledge sharing in organizations. *Strateg. Manag. J.* **2007**, *28*, 1133–1153. [CrossRef]
63. Krogh, G.V. How does social software change knowledge management? Toward a strategic research agenda. *J. Strateg. Inf. Syst.* **2012**, *21*, 154–164. [CrossRef]
64. Andreeva, T.; Kianto, A. Does knowledge management really matter? Linking knowledge management practices, competitiveness and economic performance. *J. Knowl. Manag.* **2012**, *16*, 617–636. [CrossRef]
65. Marks, M.S.; Marsh, M.C.; Schroer, T.A.; Stevens, T.H. An alarming trend within the biological/biomedical research literature toward the citation of review articles rather than the primary research papers. *Traffic* **2013**, *14*, 1. [CrossRef] [PubMed]

66. Xu, W.W.; Chiu, I.H.; Chen, Y.; Mukherjee, T. Twitter hashtags for health: Applying network and content analyses to understand the health knowledge sharing in a twitter-based community of practice. *Qual. Quant.* **2015**, *49*, 1361–1380. [CrossRef]
67. Pi, S.M.; Chou, C.H.; Liao, H.L. A study of Facebook groups members' knowledge sharing. *Comput. Hum. Behav.* **2013**, *29*, 1971–1979. [CrossRef]
68. Freudenthal, G. The role of shared knowledge in science: The failure of the constructivist programme in the sociology of science. *Soc. Stud. Sci.* **1984**, *14*, 285–295. [CrossRef]
69. Connelly, C.E.; Zweig, D.; Webster, J.; Trougakos, J.P. Knowledge hiding in organizations. *J. Organ. Behav.* **2012**, *33*, 64–88. [CrossRef]

sustainability

MDPI

Article

Internal Social Network, Absorptive Capacity and Innovation: Evidence from New Ventures in China

Wei Shan [1,2], Chu Zhang [1,*] and Jingyi Wang [1]

[1] School of Economics and Management, Beihang University, Beijing 100191, China; shanwei@buaa.edu.cn (W.S.); amanda1249@126.com (J.W.)
[2] Key Laboratory of Complex System Analysis and Management Decision, Ministry of Education, Beijing 100191, China
* Correspondence: zhangchu@buaa.edu.cn

Received: 30 December 2017; Accepted: 3 April 2018; Published: 5 April 2018

Abstract: This research investigates the impact of the internal social network on new venture's innovation by building a comprehensive structural equation modeling (SEM) that integrates three streams of research: internal social network, innovation, and absorptive capacity. Based on a sample of 279 new ventures from China, the current study's results show that absorptive capacity plays a full mediating effect in the relationship of the internal social network and innovation. Particularly, among the skill set of absorptive capacity, a mere skill of knowledge acquisition does not guarantee an enhancement of new venture's innovation. For new ventures to better utilize the social capital generated by the internal network in the process of innovation, they must focus more on the skills of knowledge digestion and knowledge application. The authors further separate the new ventures into two different sub-samples: the new venture supported by mature enterprises (M-type) and the independent new venture (I-type). This study's findings indicate that the effect of the social network on innovation through knowledge digestion is greater in the M-type sample than in the I-type sample; internal social network heterogeneity in general plays a less important role in improving a new venture's innovation than internal social network density, for both M-type and I-type new ventures.

Keywords: new ventures; internal social networks; absorptive capacity; innovation

1. Introduction

During the last decades, innovation has become one of the most critical skills for businesses to enhance their competitiveness and achieve success [1–3]. According to Cefis and Marsili [4], innovation is a main determinant of not only a firm's prosperity, but also its survival. Innovation is closely related to the sustainability of a firm, an industry and even the whole business environment. As Peter Drucker has put it, if an established organization, which in this age necessitating innovation, is not able to innovate, it faces decline and extinction [5]. Innovation, as defined by Thompson [6], is the generation, acceptance, and implementation of new ideas, processes, products, or service. By implementing innovation, firms may be able to create new markets and increase their business values [7], thus maintaining a sustainable competitiveness.

This is especially important for new ventures. Usually, start-up companies do not have a lot of resources at their disposal, or access to big platforms. To survive in today's increasingly competitive environment, new ventures must continually evolve to grow and deliver whatever it is that customers need or want. Innovating could create great value for customers and give those new ventures an edge over their competition, thus helping them grab a market share and create a strong brand identity in the long run. Therefore, how to build sustainable innovation capacity for new ventures has attracted considerable attention from industry, as well as management scholars.

Among the numerous studies and reports on innovation, a general belief was that innovation processes are interactive processes [8]. The "interactive" here actually suggests a "social network" concept. To have an innovation project work out, it requires an active involvement of personnel from different organizational bases [9], such as scientists, designers, marketers, end users, and even external partners. According to Felix et al. [10], network could provide a good cross-functional structure for firms to implement new initiatives that need joint work from all employees and departments, such as social media marketing, innovation and so on. Previous research has greatly emphasized the effect of inter-organizational or external networks on firms' innovation [3,11–15], as external partners might provide a variety of supports to the innovative firm [16]. However, the fact that firms innovate based on their internal structure [17] is somewhat neglected; the generation of new ideas often starts within the firm and the implementation is necessarily controlled by the firm itself. Google is a typical example of an organization that is committed to utilizing "internal social capital", encouraging its employees to innovate from within rather than looking outside for new ideas. Google has attracted and retained a lot of creative talents in doing so, and quickly grew into a monopoly in the online space. The success of Google evidences the fundamental role of an internal social network in boosting innovation.

Numerous studies have provided evidence on how internal resources, such as human resources [18], technological capability [19], and so on, would affect a firm's innovation. Liu, Gong, Zhou and Huang [18], for example, investigated whether, how and when different types of a firm's human resource system jointly influence the employee's creativity. Chen and Huang [20] also indicated that strategic human resource practices have a positive effect on innovation performances via a firm's knowledge management capacity. Srivastava and Gnyawali [21] examined the paradox of technological capabilities, they showed that the benefits of portfolio resources are greater for firms with low technological capability. Bakar and Ahmad [22] sought answers to the question about which of a firm's resources contributes most to product innovation performance (PIP), and concluded that intangible resources are the main drivers of PIP in a Malaysian context. Carnabuci and Diószegi [23] investigated the relationship between internal social network position and an individual's innovative performance in a firm, and they showed that individuals with an innovative cognitive style are most innovative when embedded within a closed network of densely interconnected contacts. Carnabuci and Diószegi's study is more on an individual level, while the current research emphasized the effect of the internal social network at the firm level. Plenty of literature has also contributed in examining internal factors that might affect innovation in an organization from a resource-based view [24,25]. However, as an important firm resource, the internal social network did not receive adequate attention in innovation literature. The authors of the current paper tried to address this research gap by focusing mainly on the relationship of a new venture's internal social network and its innovation.

Another related research stream is the firm's absorptive capacity. There is increasing evidence in the academic literature that merely entering into network relationships does not increase innovation definitely [16]. Literature has studied the mediating role of knowledge management [26], ego–network dynamics [27], dynamic capability [28] and so on in the relationship between social network and innovation. The knowledge-based view believes that innovation is closely related to organizational learning [29], as the transference of knowledge among organizations might provide opportunities for further cooperation, and stimulate the creation of new ideas [9,30]. Absorptive capacity fosters a firm's absorption of technological knowledge and enhances its search for new ways to achieve new competitive advantages, thus is highly related to innovation throughout organizations [31]. Previous research has examined the critical role of absorptive capacity in enhancing a business unit's innovation [32–35], and highlighted the need for strong absorptive capacity in leveraging the shared knowledge for innovation [36,37].

Given the importance of firms' absorptive capacities in innovation, the authors proposed that a new venture's innovation could be jointly affected by absorptive capacity and the internal social network. Previous literature has investigated the joint effect of absorptive capacity with other factors. Nieto and Quevedo [38], for example, demonstrated that the absorptive capacity variable determines innovative effort to a greater extent than technological opportunity and knowledge spillovers. Kotabe, Jiang, and Murray [39] also investigated the effect of absorptive capacity and the social network on innovation, and their results showed that absorptive capacity complements the social network in enhancing both incremental and radical innovations. However, they focused on the external social network with government, while the current study's focus is the internal social network.

The present paper is an attempt to provide a comprehensive framework that integrates these three research streams: internal social network, absorptive capacity, and innovation. By establishing a SEM model, the authors examine the interaction effects between absorptive capacity, internal social network, and their impacts on a new venture's innovation. More specifically, the authors examine the mediating effect of absorptive capacity in the relationship between the internal social network and innovation in new ventures.

In brief, the current research contributes to research in the following aspects. First, this paper focuses especially on the effect of the internal social network/resources on innovation for a new venture. Most previous studies have focused on the impact of the external social network in improving a firm's innovation [3,11–14]. Very little attention has been paid to the role of the internal social network on innovation. This paper examines two dimensions of the internal social network: internal social network density and internal social network heterogeneity. The results show that the internal social network density affects a new venture's innovation more than social network heterogeneity does. Second, scholars have considered different types of innovation [3], different aspects of social networks [14,40], or different types of knowledge [9] in examining the effect of the social network on innovation. However, to the authors' knowledge, no one has compared the results between different types of new ventures. This paper further separates the new ventures into two different types based on the business practices in China: the new venture supported by mature enterprises (M-type) and the independent new venture (I-type). The comparison of the two groups of new ventures shows that the impact of the internal social network on innovation is different across different types of new ventures. The total effects of network heterogeneity on innovation in M-type and I-type samples are 1.1% and 14.6% respectively. In contrast, the total effects of network density on innovation in M-type and I-type samples are 65.3% and 48% respectively. This could provide great insights to different new ventures as how to effectively improve their innovation based on limited resources. Third, the authors incorporate absorptive capacity in examining the relationship between the internal social network and innovation. Numerous studies have focused on the relationships between any two of these three streams [32,41,42], but few have taken all these three streams into consideration at same time. The authors closely investigate the mechanism as to how the internal social network affects a new venture's innovation by taking absorptive capacity into consideration, which provides great insights for new ventures to exploit their internal social capital to the full extent and to achieve sustainability by increasing their innovation. The authors highlight the full mediating effect of absorptive capacity in the relationship between social network and innovation.

The paper is organized as follows: Section 2 presents the theoretical background and establishes the framework for this study; Section 3 is concerned with the methodology; Section 4 describes the results and analyzes the findings; Finally, conclusions are given and further research concepts are identified in Section 5.

2. Theory and Hypotheses

2.1. Social Network and Innovation

Social network refers to a relational set of social actors. It could be the relationships between individuals in a group [43,44], or the relationships between individuals and groups [45], or even the relationships between organizations. The relations in the social network can be either formal or informal. The formal networks usually refer to the relationship established by contract, consanguinity, and so on, whereas the informal networks are mostly set by emotion, friendship and so on. This paper examines the internal social network within an organization. Since formal and informal networks usually coexist within an organization, the authors focus on the interplay between group members in terms of both formal and informal relationships.

Social network is regarded as an important way to derive actual or potential resources [46]. Through the active participation of citizens or members in organizations, social capital as a collective resource can be created [47]. Previous literature has emphasized the positive effect of the social network, particularly the external network, on innovation. Scholars believe that most successful innovators invest in the breadth of accumulated knowledge and absorb information from all kinds of sources, not just external, but also internal [48]. The potential collaboration with external partners might facilitate the interactive learning process among firms [11] and help those firms get new skills [9]. Through social networks firms share knowledge that can improve their capacities for innovation [49,50], which leads to greater levels of both product and process innovations and creates a sustainable competitive advantage for firms [51]. Moreover, the implementation of new ideas greatly depends on coordination with external networks such as business partners, customers and suppliers [41,52–55]. Generally, researchers find that the external network has been positively correlated with innovation and performance of a firm [14,40].

The importance of the social network is especially highlighted when it comes to new ventures, as the social network may provide them with resources that are crucial to their growth, such as funding, information, and even opportunities [56]. Ostgaard and Birley [57] and Adler and Kwon [58] also empirically proved the important role the social network plays in new ventures.

While most of these researchers focus on the external linkages, this paper is trying address the effect of the internal social network on innovation. The resource-based view maintains that firm success is not only determined by external factors but also by its internal characteristics [59,60]. Previous studies have examined the effect of other internal resources on a firm's innovation. Bougrain and Haudeville [61] assess how internal research capacity helps a firm to exploit scientific and technical knowledge, and they find that internal research capacity might enhance the firm's ability to carry its project to success. The internal social network, also as a valuable internal resource, has been overlooked in the innovation research area.

Generally, the effectiveness of social network mechanisms is related to several dimensions, such as network diversity, network size and network density [62,63]. This paper focuses on the two main dimensions that have been extensively discussed in previous literature: the network density and the network heterogeneity. The first is used to illustrate the intensity of relations [64,65]. The second one describes the diversity of the relations.

Thus far, researchers have emphasized the importance of network density in improving a firm's performance. Mehra, Dixon, Brass, and Robertson [66] found that the density of friendship relations within an organizational group was positively related to group performance. Henttonen, Janhonen, and Johanson [67] echoed a similar conclusion. According to Luo [68], even a team structure with fully connected cliques can have a positive impact on performance. There are two main underlying reasons. First, members of highly dense networks are fully connected, which would foster the transference of information and cooperation between members [69,70]. A second reason, as pointed out by Coleman [70], is the high trust level between members that is facilitated by network density. It is easier for firms or groups to have better performances with more people sharing same belief and trusting each other, as proven by Reagans and Zuckerman [62], Zacharatos, Barling, and Iverson [71], and Snape and Redman [72]. To summarize, a high internal social network density would prompt knowledge sharing and improve the innovation of the firm. The authors believe that this general conclusion is also applicable to new ventures.

Hypothesis 1 (H1). Internal social network density has a positive effect on the innovation of a new venture.

There seems to be wide agreement that a diversified network will improve the innovation of a firm [14,42,73–75]. Regarding individuals, previous literature showed that heterogeneous networks positively contribute to individual innovative performance [76,77]. Faced with the increasingly competitive world, firms also need both the width and depth of knowledge for successful innovation nowadays. A diversified network increases the variety of the information, resources, and knowledge accessed [3], which would in turn increase the innovation of a firm. Der Foo, Wong, and Ong state a new venture with higher heterogeneity is more dynamic [78]. Industrial economics supports the idea that heterogeneous structures affect the decision-making or behavior, which in turn affects innovation performance. According to organizational behavior theory, heterogeneity would also affect subsequent innovation activities and innovation processes. Moreover, researchers find that the more diversified the social networks, the higher the innovation level of a firm [14,40]. Therefore, this paper proposes the hypothesis as follows:

H2. Internal social network heterogeneity has a positive effect on the innovation of a new venture.

2.2. Absorptive Capacity and Innovation

The concept of absorptive capacity was first raised by Cohen and Levinthal [33,79]. It refers to the capacity that a firm recognizes the value of new information, assimilates it, and applies it to commercial ends. Absorptive capacity then has been explored in more areas, which led to a review of its definition in academia. Zahra and George [34] expanded the concept and further defined it as “a set of organizational routines and processes by which firms acquire, assimilate, transform and exploit knowledge to produce a dynamic organizational capability.” According to Zahra and George [34], absorptive capacity includes two elements: (a) potential absorptive capacity, meaning the capacity that makes the firm receptive to acquiring and assimilating knowledge/information; and (b) realized absorptive capacity, which refers to a firm’s function of transforming and exploiting the knowledge.

Based on the definition from both Cohen and Levinthal [33,79], Zahra and George [34], this paper summarizes three dimensions of the absorptive capacity: knowledge acquisition, knowledge digestion, and knowledge application. Among the three dimensions, knowledge acquisition and knowledge digestion are consistent with the definition of potential absorptive capacity in Zahra and George’s model. Knowledge application is a generalization of a firm’s capacity in transforming and exploiting knowledge. The current authors believe that these three dimensions are sufficient to characterize a firm’s absorptive capacity.

Absorptive capacity is important in the process of innovation [80]. Many researchers stress that absorptive capacity contributes both directly [81,82] and indirectly [83,84] to innovation performance. Innovation is originated through enterprises’ knowledge circulation (that is a process on inflows and outflows of knowledge) which facilitates the development, and even the commercialization, of internal innovation [85,86]. Innovation practices concern the inter-organizational exchange of knowledge and the inflow of external knowledge into an organization which requires absorptive capability [87,88]. Therefore, the process of creating a new technological knowledge cannot be efficient without a solid absorptive capacity [89]. The innovation process involves the acquisition, dissemination, and use of new knowledge [90–92]. Since absorptive capacity demonstrates a firm’s ability in dealing with knowledge, the chances are high that the absorptive capacity is correlated with a firm’s innovation.

An enterprise with higher absorptive capacity tends to adjust its internal organization to changes in its environment, to explore opportunities, even solutions, and to exploit innovation to meet its needs as well. According to Cohen and Levinthal [33], organizations with a higher level of absorptive capacity are more likely to harness new knowledge to help their innovation. Those firms have a higher ability to create new knowledge from the obtained knowledge. Moreover, Kim [93] notes that the absorptive capacity might help firms acquire information from their external partners and create new ideas from the transferences of knowledge. Evidence from high tech firms of the Pearl River Delta in China also has shown that absorptive capacity has a positive effect on a firm’s innovation [94].

Therefore, a higher absorptive capacity might not only generate more profits from the knowledge it absorbed, but also enhance the firm's innovation [32].

H3. Absorptive capacity has a positive effect on the innovation of a new venture.

Absorptive capacity includes a firm's capacity for acquiring, digesting and applying knowledge. These elements of absorptive capacity play different but complementary roles in enhancing organizational performances [80]. The knowledge acquisition refers to a firm's capability to identify and acquire knowledge that is critical to its operations [34]. Even though knowledge acquisition capability does not generate a direct benefit to firms, it is a prerequisite for a firm to further utilize the knowledge. Knowledge digestion capability enables a firm to analyze, process, interpret and understand the knowledge obtained, which lays a foundation to transform the knowledge into innovation. Therefore, the authors propose the hypotheses below:

H3a. Knowledge acquisition has a positive effect on innovation.

H3b. Knowledge digestion has a positive effect on innovation.

Among the three elements of absorptive capacity, the knowledge application capability is more likely to have a direct effect on a firm's innovation. Knowledge application capability requires knowledge transforming and exploiting skills. The knowledge transforming skill helps firms to develop new blueprints of products with new information or technology, while the exploiting skill can help convert knowledge into new products [95]. Kazanjian, Drazin, and Glynn [96] also observe that firms need the skills of leveraging and recombining knowledge for product line extensions or new product development. Gao et al. [80] proved that firm's innovative activities are more likely to be enhanced when the firm has a higher level of knowledge transformation and exploitation capability. Therefore, the current authors believe that knowledge application skill is beneficial to firms in enhancing their innovation.

H3c. Knowledge application has a positive effect on innovation.

2.3. *Social Network and Absorptive Capacity*

Previous literature supports the theory that both the social network and absorptive capacity might improve the innovation of a firm. An interesting question naturally arises as to how the social network would interact with absorptive capacity in this system. The authors therefore propose the hypotheses on the relationships of social network and absorptive capacity in this section.

Absorptive capacity depends on the level of knowledge in the corporation [33]. As an endowment for firms, the internal social network could provide direct information, knowledge, and complementary resources for firms. Hence, previous studies focused on how the network would help a firm improve its absorptive capacity. The firm's centrality in a network of relationships naturally reflects the resources it can get from the network, for example. Powell et al. [52] find that the greater the firm's centrality, the easier for the firm to acquire information from the network, thus the higher the absorptive capacity. Burt [41] states that the structural holes in the social network will improve the firm's knowledge acquisition. Uzzi [97] believes that the weak ties in the network might incentivize firms to actively extend their networks, from which the firms might have more exposures to heterogeneous information and, thus, a higher ability to get new knowledge. Hein and Rauschnabel [98] propose a conceptual model to illustrate how enterprise social network can help increase knowledge sharing efficiency by adopting new technology. Furthermore, Dyer and Singh [99] and Jung-Erceg, Pandza, Armbruster, and Dreher [100] have emphasized the influences of social capital and inter-firm relationships on a firm's absorptive capacity (knowledge digestion and application) respectively.

This paper focuses on the density and heterogeneity aspects of the network. Network density refers to "the proportion of direct ties in a network relative to the total number possible" (Wikipedia.com). Network density could facilitate the build-up of trust and cooperation while constraining opportunism among members [101–103]. Thus, a dense network is advantageous for knowledge exchange and acquisition [104,105]. Additionally, network ties could also act as a device for screening and interpreting novel information, resulting in an enhanced absorptive capacity of the firm, especially the knowledge acquisition and digestion skills [106]. Even though some researchers claim that a dense network would increase the redundancy of information [42], this redundancy plays a limited role as the key focus here is on finding and absorbing novelty, making considerations of efficiency less of an issue [103,107]. Therefore, network density, no matter internal or external, can help a firm with knowledge absorption and digestion.

Regarding the effect of network density on realized absorptive capacity (knowledge application), previous studies have provided some insights as well. According to Zahra and George [34], network density, or connectedness, might allow units to better transform and exploit new knowledge. By developing the trust and cooperation between members, network density could foster commonality of knowledge and encourage communication in organizations [105,108]. This might improve the efficiency of knowledge exchange and application throughout organizations. Moreover, members of a dense network usually have common goals; therefore, it is less likely to cause conflicts in utilizing knowledge and implementing innovation programs [109]. Thus, the authors propose that:

H4. Internal social network density has a positive effect on the absorptive capacity of a new venture.

H4a. Internal social network density has a positive effect on knowledge acquisition.

H4b. Internal social network density has a positive effect on knowledge digestion.

H4c. Internal social network density has a positive effect on knowledge application.

Network heterogeneity, or network diversity, is another aspect of social relations. It is generally believed that entrepreneurs with more diverse networks will get more support and thus are easier to success [13,110,111]. A similar conclusion is reached regarding the effect of network heterogeneity on absorptive capacity. First and foremost, a diverse network is favorable for firms in acquiring different types of knowledge. Burt [41] has stressed that accesses to these non-redundant contacts is of great benefit to obtain novel information (novelty value). Thus, it is reasonable to propose that network heterogeneity would secure the diversification of the organizational knowledge and facilitate potential absorptive capacity (knowledge acquisition).

Second, diverse knowledge structures support explorative learning and increase the prospect that new external knowledge is related to existing knowledge [112]. The more diverse an individual's external network, the easier it is to find the required knowledge and the more likely an individual will engage in transformation and exploitation activities, and the more likely they are to be exposed to potential new knowledge, which positively affects the recognition of new knowledge [33,113]. People with different backgrounds might have different understandings toward the new information. By sharing their perspectives and views with each other, it is easier for group with higher diversity to assimilate the new knowledge. Meanwhile, the interplay between members with diverse background would augment a firm's capacity for making novel linkages and associations [33]. Overall, the heterogeneity facilitates the digestion of new knowledge that constitutes potential absorptive capacity.

There are different views regarding the impact of heterogeneity on knowledge application in the process of innovation. On the one hand, a diverse internal social network suggests a variety of experiences, expertise and cultures. In most cases, employees with only certain expertise are not sufficiently competent to succeed with a whole project. Therefore, a diverse internal network with different expertise could play complementary roles in transforming the knowledge and applying the knowledge in innovation. Some research, on the other hand, has raised doubts on network heterogeneity. They argue that participation in decision making might have a negative effect on new product development speed due to the difficulty in gaining consensus [114]. A participation of diversified employees might hamper information-processing efficiency and negatively affect knowledge utilization [115]. The authors of this current paper think that there usually is an inspiring and determined leader in a new venture, thus making it less possible to have a consensus problem in decision making. Hence, the authors believe that a diversified internal social network, overall, would be beneficial to knowledge application in the innovation process.

H5. Internal social network heterogeneity has a positive effect on the absorptive capacity of a new venture.

H5a. Internal social network heterogeneity has a positive effect on knowledge acquisition.

H5b. Internal social network heterogeneity has a positive effect on knowledge digestion.

H5c. Internal social network heterogeneity has a positive effect on knowledge application.

Based on the hypotheses above, the authors proposed the following conceptual model (Figure 1).

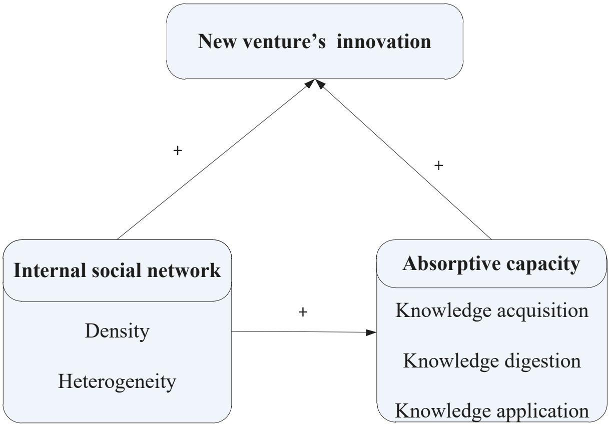


Figure 1. The conceptual model.

3. Methodology

To examine these proposed hypotheses, the authors conducted a survey among people working in new ventures in mainland China. A multi-item questionnaire was developed based on prior literature and the pilot survey. Following the validation of reliability and validity of the scales, data analysis was carried out with Amos.

3.1. Measurement

The survey measurements were composed of three major sections: (a) internal social network; (b) absorptive capacity; and (c) innovation. Altogether, the authors developed six constructs. The scales and items that were adopted for the present survey have been validated in prior empirical studies. Responses were measured on a well-established 5-point Likert scale ranging from "strongly disagree" (1) to "strongly agree" (5). The questionnaire was written in Chinese.

3.1.1. Internal Social Network

Two constructs regarding internal social network were included: internal social network density and internal social network heterogeneity.

The items for network density were adapted from Ke et al. and Peng et al., which included formal and informal contacts between group members [116,117]. Regarding formal contacts, the authors refer to the discussion meetings and coordination in work. Apropos of the informal contacts, the authors mean by informal discussions and personal relationships between the members.

Concerning items of network heterogeneity, the authors adapted the measurements from Zhang, Sun and Wang's research to suit the context of this research [118]. Respondents were asked if members (in their group/company) had large differences in the level of education, if members had different working experience; if members had different ways of thinking; if members had different years of working in this group; if members had differences in working styles, knowledge sets, and experiences. The network heterogeneity was also measured by five 5-point Likert scales anchored by 1 = "Strongly disagree" and 5 = "Strongly agree." The measurement methodology was consistent with Jehn, Northcraft, and Neale [119].

3.1.2. Absorptive Capacity

Absorptive capacity refers to the ability of an enterprise to acquire external knowledge, combine, and assimilate it within the organizational setting [34]. This process enables enterprises to develop less in-house research and development activities [120]. Cohen and Levinthal [33] highlighted three main dimensions of the absorptive capacity. Consistent with Cohen and Levinthal, the authors developed three constructs: knowledge acquisition, knowledge digestion, and knowledge application. The items for these three constructs were adapted from Jansen, Van Den Bosch, and Volberda [121].

3.1.3. Innovation

Innovation has never failed to catch people's attention ever since the 1940s when Schumpeter deemed creative destruction at the heart of economic growth. Innovation has a broad definition. It could be both an outcome and a process. Zaltman, Duncan, and Holbek [122] and Rogers [123] believe innovation is more like a "result". According to their definition, innovation is an idea, practice, or material artifact perceived as new by the relevant unit of adoption. Amabile, Conti, Coon, Lazenby, and Herron [124] also define innovation as the successful implementation of creative ideas within an organization. Since the purpose of this paper is to help new ventures recognize the possible ways to enhance its performance, it must focus not only on the successful result but also the process of innovation. Therefore, the authors borrowed a more general definition of innovation from Thompson [6], that is, the generation, acceptance, and implementation of new ideas, processes, products, or services.

Measures of innovation were adopted from Chen [125] and then adapted to suit the context of the current research. The six items for innovation include: the frequency of new product (or service) development, the responsive feedback to customers' requirements, the improvement in a firm's performance, and the adoption of a new service, new technology, and new methods.

3.2. Procedure and Method

A combination of SPSS 24 and Amos software package 20.0 was used to carry out all the data analyses. A pilot survey was carried out first to assess the clarity of the questionnaire and its suitability to the participants. The questions that might cause ambiguity and confusion based on the feedbacks were rephrased. Furthermore, the authors conducted exploratory factor analysis (EFA) to identify the underlying factor structure with data from the pilot survey. Two items were deleted for innovation construct that caused low measurement quality, resulting in a questionnaire with 6 constructs (Appendix A.1). Subsequently, the data was re-collected using the updated questionnaire. The confirmatory factor analysis (CFA) was conducted to assess the reliability and convergence validity of the measurement model based on the re-collected data. Following the confirmatory factor analysis, SEM analysis was carried out to test the hypothetical model. Last, a sub-group comparison analysis was conducted between M-type and I-type ventures in the extension.

3.3. Sample and Data Collection

The pilot survey was conducted based on a sample of 90 respondents in August 2015 of which 82.222% responded, and 75.556% were valid. Based on the pilot survey, the authors distributed the questionnaires in November 2015 to a sample of 279 respondents. The questionnaires were distributed both online and offline. The online questionnaire was created on wjx.cn, the most popular online survey service provider in China. The authors then sent out the online questionnaire to potential respondents who were working in new ventures via emails, chats, social media, and so on. Offline questionnaires were distributed mainly to part-time MBA students at Beihang University (China). One item was added as the screening question regarding how many years the company the respondent working at has been founded. Responses from companies that existed for more than 5 years were not adopted. The authors also asked the respondents to specify the name of their company. The authors fully erased the respondents' details, dealing with their concern on privacy in distributing the questionnaire. In total, 230 questionnaires were retrieved. The authors excluded responses that: (a) reported same scale for every item; (b) were answered in a certain pattern; (c) had missing data, yielding a usable sample of 202. The survey covered new ventures from five provinces in China: Beijing, Shanghai, Tianjin, Shandong, and Hebei, providing a strong basis to conduct the next step of the analysis.

Table 1 demonstrates the demographic information of the respondents in detail. Among the 202 valid samples, 58.416% were male and 41.584% were female. Most of the respondents were 25–35 years old (75.743%) and were working in new ventures of a knowledge-intensive industry (63.366%); 92.574% of respondents received a higher education in universities (45.049%) or postgraduate schools (47.525%). Two types of new ventures in collecting data were identified: the new ventures supported by mature enterprises (M-type) and the independent new ventures (I-type). 54.555% of respondents were from M-type new ventures and 45.545% were from I-type new ventures. Regarding the size of their working teams, 38.119% of respondents reported a team number of 6–11 and 44.059% reported a team number that was larger than 11.

Table 1. Basic information of the sample (n = 202).

Variable	Category	Numbers	Ratio (%)
Gender	Male	118	58.416
	Female	84	41.584
Age	<25	32	15.842
	25–35	153	75.743
	>35	17	8.416
Education level	College or below	2	0.990
	Bachelor	91	45.049
	Postgraduate	96	47.525
	Ph.D. or above	13	6.436
Industry type	knowledge-intensive	128	63.366
	labor intensive	42	20.792
	capital-intensive	32	15.842
Venture type	M-type	110	54.455
	I-type	92	45.545
Number of team members	2–5	36	17.822
	6–11	77	38.119
	>11	89	44.059

4. Results

4.1. Exploratory Factor Analysis

Exploratory factor analysis (EFA) is to identify the underlying factor structure of a set of observed data. To perform EFA, the authors first conducted validity analysis through KMO (Kaiser-Mayer-Olkin) and Bartlett sphere test. KMO indicates the amount of variance shared among the items designed to measure a latent variable when compared to that shared with the error. Kaiser (1974) recommends accepting values greater than 0.5 as acceptable [126]. Table 2 shows the KMO index >0.5 and the Bartlett test is statistically significant ($p < 0.01$), which indicates that the data have sufficient inherent correlations to perform exploratory factor analysis (EFA).

Table 2. KMO and Bartlett sphere test of the sample.

	KMO	Bartlett
Result	0.654	790.081 ***
Df		300

*** $p < 0.01$.

To understand the factor structure and the measurement quality, a principal component analysis was conducted with varimax rotation, and an evaluation of the eigen values was used to identify the number of factors to retain. Hair et al. (2006) suggests that an item should be removed if (1) the factor loading is lower than 0.5; (2) the item loads in two different factors at same time with both loadings higher than 0.4; and (3) the item does not load in a group to which it belongs [127]. Following this suggestion, two items of innovation construct were dropped. The final results show that all items were included on the correct factor with a load above 0.5. The identified factors correspond to the six constructs presented in the conceptual model. Detailed results of the EFA can be found in Appendix A.2.

The reliability of measurement was also tested through the Cronbach alpha coefficient, which was assessed for each construct (Table 3). As all alpha values are above the recommended threshold of 0.70, the reliability of the data is established.

Table 3. Assessment of reliability.

Construct	Item	Cronbach α	
Social network density	4	0.825	0.749
Social network heterogeneity	5	0.738	
Knowledge acquisition	4	0.718	
Knowledge digestion	4	0.803	
Knowledge application	4	0.841	
Group innovation	4	0.860	

4.2. Measurement Assessment

To verify the factor structure, the measurement model for the constructs was tested using confirmatory factor analysis (CFA) with AMOS 20.0. To ensure that the model was a good fit, several indices were calculated, including Chi-square/degrees of freedom, Goodness of Fit Index (GFI), Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), Incremental Fit Index (IFI) and Root Mean Square of Approximation (RMSEA). The results are summarized in Table 4, which indicates a reasonably good fit [127,128].

Table 4. Summary of the overall fit indices for confirmatory factor analysis (CFA) model. CFI: Comparative Fit Index; IFI: Incremental Fit Index; GFI: Goodness of Fit Index; RMSEA: Root Mean Square of Approximation; TLI: Tucker–Lewis Index.

Fitting Estimation	χ^2/df	CFI	IFI	GFI	RMSEA	TLI
CFA	1.218	0.974	0.974	0.903	0.033	0.967
Recommended value	<3	>0.9	>0.9	>0.9	<0.08	>0.9

Additionally, composite reliability (CR) was also calculated based on the standardized regression weights (factor loadings) by the CFA (Table 5). When the CR exceeds 0.60, it suggests that the measures could consistently represent the latent construct [129]. The results demonstrate the CR values of all the constructs in CFA range from 0.737 to 0.865, which indicates a good internal consistency [127].

Table 5. Reliability and validity of the measurements.

Construct	Item	Factor Loading	1-R2	CR	AVE
Social network density	SND1	0.682	0.534	0.813	0.528
	SND2	0.819	0.329		
	SND3	0.838	0.298		
	SND4	0.524	0.726		
Social network heterogeneity	SNH1	0.563	0.683	0.737	0.362
	SNH2	0.608	0.630		
	SNH3	0.718	0.485		
	SNH4	0.528	0.721		
	SNH5	0.575	0.669		
Knowledge acquisition	KAC1	0.620	0.615	0.751	0.433
	KAC2	0.781	0.390		
	KAC3	0.631	0.601		
	KAC4	0.581	0.663		
Knowledge digestion	KAS1	0.657	0.569	0.800	0.503
	KAS2	0.619	0.617		
	KAS3	0.735	0.460		
	KAS4	0.811	0.342		
Knowledge application	KAP1	0.683	0.533	0.829	0.550
	KAP2	0.834	0.304		
	KAP3	0.801	0.358		
	KAP4	0.630	0.603		
Group innovation	GIP1	0.791	0.374	0.865	0.617
	GIP2	0.837	0.299		
	GIP3	0.758	0.426		
	GIP4	0.753	0.433		

Convergent validity refers to the correspondence or convergence between similar constructs, which can be examined in terms of the factor loadings and average variance extracted (AVE) score (Table 5). The analysis is usually regarded as acceptable if factor loading estimates and average variance extracted (AVE) values are higher than 0.5 [127,130]. The AVE scores for constructs social network density, knowledge digestion, knowledge application, and innovation were 0.528, 0.503, 0.550, and 0.617 respectively, all of which met the threshold condition. The AVE scores of knowledge acquisition and social network heterogeneity did not reach 0.5. Deleting these two constructs might have made the statistical results look better; however, the authors still decided to keep the two constructs due to the following reasons: (a) even though the AVE values for social heterogeneity and knowledge acquisition were a little bit lower than 0.5, the results showed that all the indicators of knowledge acquisition and social network heterogeneity loaded significantly ($p < 0.001$) and substantially (all factor loadings >0.5) onto their constructs in this model. Moreover, the AVE score of knowledge acquisition was quite close to the threshold (0.433). Previous literature has documented similar treatment to this by keeping the constructs [131]; (b) the authors might have missed out on some insightful conclusions about the social network and absorptive capacity if the authors deleted the two constructs. The authors were then able to show the different roles heterogeneity and density play in affecting a venture's innovation by considering the heterogeneity in the current model. Additionally, by considering the knowledge acquisition dimension, this study can provide full insights to new ventures as to how absorptive capacity plays a role in the relationship of an internal social network and innovation.

Furthermore, given the two constructs were strongly supported in the literature (see Section 2), the authors decided to keep the constructs.

To recapitulate, the confirmatory factor analysis indicated a satisfactory level of reliability and validity overall, as well as an adequate model fit, meaning that the model was valid. Therefore, it was appropriate to conduct the next step, structural model analysis.

4.3. Structural Model Assessment

The authors used structural equation modeling (SEM) with software AMOS 20.0 in this paper to test all the hypotheses. Introduced in the 1960s and 1970s, SEM has been extensively used in sociology, psychology, and other social sciences [132]. SEM includes a diverse set of mathematical models, computer algorithms, and statistical methods that fit networks of constructs to data (Wikipedia.com). There are two main reasons why the authors adopted SEM in this paper. One of the reasons is that SEM is able specify relationships between unobserved constructs (or latent variables) from observable variables [133]. There are latent variables, like network density, in this paper that could not be measured directly. SEM could help deal with those latent variables and test the hypotheses using the observed data. Next, SEM is considered as one of the most significant techniques in understanding multiple relationships [134]. In the proposed hypothetical model, there are both direct relationships and indirect relationships. SEM allows the authors to test the overall theory as well as specific relationships between observed relationships.

4.3.1. Conceptual Model Analysis

The conceptual model and hypotheses were evaluated by structural model analysis with maximum likelihood estimation. The overall model fit was assessed in terms of same fit indices as above. According to the results, the Chi-square/degrees of freedom had a value of 1.305, which is below the recommended criteria of 3. The statistics for CFI, IFI and TLI are 0.962, 0.963, and 0.954 respectively. The value of RSMEA is 0.039. These indices demonstrate that the model is a good fit with the observed data.

The statistical significances of path coefficients were also examined. Figure 2 depicts the standardized regression coefficient of each path together with its significance. According to the squared multiple correlation coefficients (R^2), 58% of the variance in innovation can be explained by the hypothetical model.

Continuing from the result, it was found that internal social network density had a significantly positive effect on the three constructs of absorptive capacity (knowledge acquisition, knowledge digestion, and knowledge application), but no significant effect on innovation directly. Thus, **H1** is not supported, and **H4** (**H4a**–**H4c**) is supported. Regarding internal social network heterogeneity, there was no evidence showing that internal social network heterogeneity had a significantly direct effect on innovation either. Therefore, **H2** is not supported. Concerning the three constructs of absorptive capacity, only the relationship between internal social network heterogeneity and knowledge acquisition was confirmed to be significant. Thus, **H5b** is supported, but **H5a** and **H5c** failed to be supported. Moreover, knowledge digestion and knowledge application positively predicted the innovation. Accordingly, **H3b** and **H3c** are supported, while H3a is not supported.

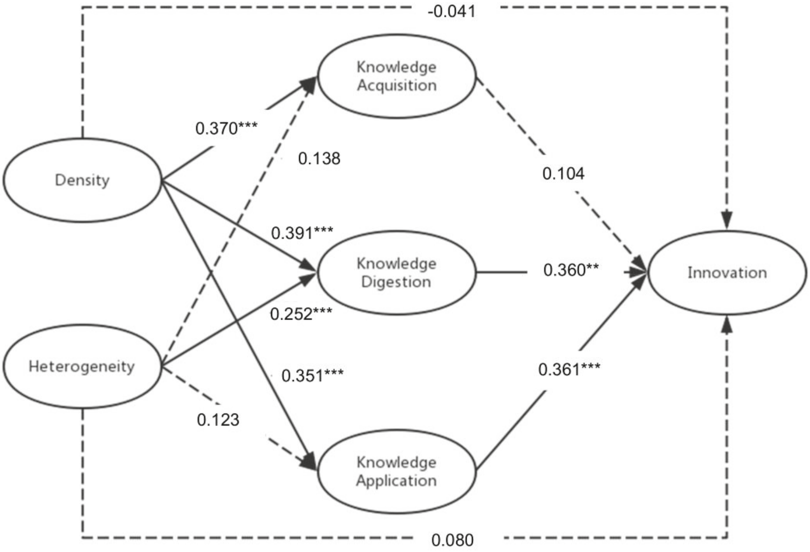


Figure 2. Standardized regression results of the structural model. (Dotted lines indicate non-supported paths, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

4.3.2. Modified Model

Based on the result of the conceptual model (Figure 2), the authors reevaluated the hypotheses. The model was modified and reexamined by deleting the insignificant paths one by one. The fit indices indicated an overall good fit for the modified model, as summarized in Table 6. The specific standardized weight estimates are illustrated in Figure 3, all of which are significantly positive. Furthermore, 58% of the variances in innovation are explained by the modified model.

Table 6. Summary of the overall fit indices for hypothetical model and modified model.

Fitting estimation	χ^2/df	CFI	IFI	GFI	RMSEA	TLI
Conceptual model	1.305	0.962	0.963	0.892	0.039	0.954
Modified model	1.336	0.957	0.958	0.889	0.041	0.949
Recommended value	<3	>0.9	>0.9	>0.9	<0.08	>0.9

Absorptive capacity played an important role as a mediator in the relationship between internal social network and innovation. That is to say, internal social network did not have a direct impact on innovation but did indirectly affect innovation via absorptive capacity. The authors highlighted three dimensions of absorptive capacity in this paper. Figure 3 demonstrates that the results of the modified model are consistent with the conclusions stated above: (1) the heterogeneity of the internal network would positively affect a new venture's skills in knowledge digestion (β = 0.259 **), while the density of the internal network is crucial in enhancing a new venture's skill in knowledge acquisition (β = 0.392 ***), knowledge digestion (β = 0.392 ***) and knowledge application (β = 0.403 ***); (2) regarding the absorptive capacity, what really matters in its relationship with innovation are knowledge digestion dimension (β = 0.387 ***) and knowledge application dimension (β = 0.435 ***).

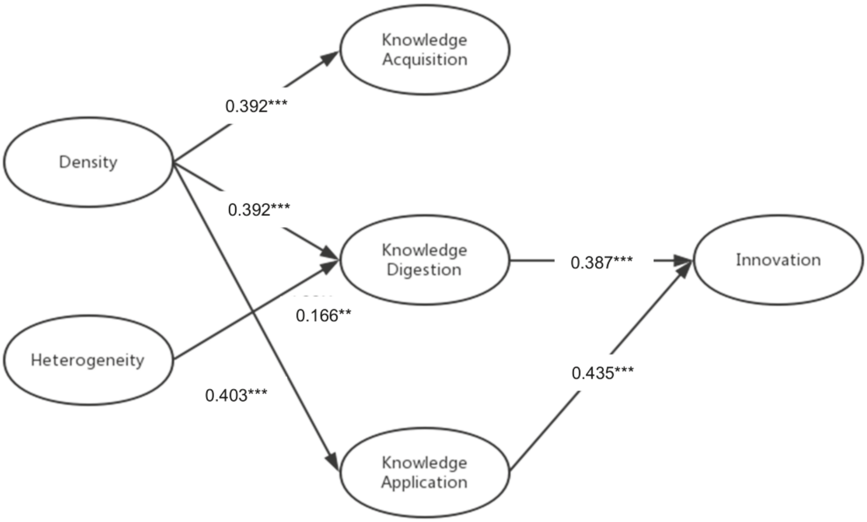


Figure 3. Standardized regression result of the modified model (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

Overall, the standardized (indirect) effect of social network density and heterogeneity on innovation are 0.538 (density) and 0.081 (heterogeneity) respectively, which means that the internal social network will significantly improve a new venture's innovation.

4.3.3. Discussion and Managerial Insights

Both internal network heterogeneity and density would have a positive effect on a firm's performance in innovation. This result is consistent with Reagans and Zuckerman [62], and Pelled [135]. However, the current results further indicate that the direct effect of the social network on innovation is considerably weaker and statistically insignificant (Figure 2). According to Salazer et al. [16], merely entering into network relationships does not guarantee the organization an increase in its innovation. Social network, in essence, is a type of resource for the organization. Only by utilizing this "resource" through a certain mechanism, can firms achieve better performance in innovation. Results in the current study show that the only way for the internal social network to influence innovation is through a firm's absorptive capacity. Thus, the direct effect in hypotheses **H1**–**H2** is not supported. Kotabe, Jiang, and Murray's study confirmed absorptive capacity complements the external network in enhancing innovation [39]. The current study further emphasizes the importance of absorptive capacity as a mediator in the interaction between the internal social network and innovation.

Second, among the three dimensions of absorptive capacity, the knowledge digestion and knowledge application skills of a new venture would have significantly positive effects on its innovation (**H3b** and **H3c** are supported). Even though knowledge acquisition skill might lay a foundation for further utilization of knowledge, it does not necessarily increase a new venture's innovation. The key characteristic of innovation is to "create" new ideas from the transferences of knowledge and information. Knowledge acquisition skill does not help firms "generate" new ideas, it only helps firms get the necessary knowledge/information to be further transformed. Therefore, it might not be too surprising to see that **H3a** is not significantly supported. However, the conclusion adds importance to the knowledge digestion and application skills for innovation: new ventures should not just be a "receptor" of new knowledge, but also be an active "user" of the knowledge in the process of innovation.

Last, as expected in the hypotheses, internal social network density would significantly improve the new venture's absorptive capacity ($p < 0.01$). The result of this study confirmed the previous research on network density and absorptive capacity [34,106]. Regarding the impact of heterogeneity on absorptive capacity, network heterogeneity might be able to secure the diversification of the organizational knowledge, which is of help (but not guaranteed) in increasing a firm's knowledge acquisition skill. Meanwhile, a diverse internal network with different expertise could play complementary roles in transforming and applying the knowledge. However, previous literature has debated whether heterogeneity would increase or hamper information-processing efficiency and knowledge utilization. Atuahene-Gima and Cardinal believe that heterogeneity might make it difficult to reach a consensus, thus a participation of diversified employees might hamper information-processing efficiency and negatively affect knowledge utilization [114,115]. The current authors' results show that the heterogeneity of the internal social network could positively increase a new venture's knowledge acquisition and knowledge application skills. However, the result is not statistically significant. The cause for the result could be the interaction between the positive effect and negative effect of the heterogeneity. Regardless, the current paper's result suggests that it does not require new ventures to maintain a diversity of employee backgrounds in order to have better knowledge application skill. Rather than putting unnecessary investment to maintaining network diversity, it is better for new ventures to build a dense team in which members work closely with each other.

4.4. Extension: Different Types of New Ventures

New ventures differ from each other in many aspects, such as initial funding, business resources and so on. The external linkages are sometimes important as well, since the support or resources from outside might greatly improve their efficiency in innovation. In the context of China, the authors separated the new venture companies into two different types: new ventures supported by mature enterprises (M-type new venture) and independent new venture (I-type new venture). The M-type new venture is common in China, as there are many venture capital (VC) funds by those large and mature companies. One prime example is the Innovation Foundation Project established by Tencent (qq.com). Not only can those new ventures get financial support, but they are also provided with vast technical or managerial resources, which are unavailable to those I-type new ventures.

The authors identified 102 responses from M-type ventures and 90 from I-type ventures of all the 202 responses. To test the generality of the current framework and check the effect of the social network on innovation for those two different types of companies, the authors conducted a group comparison study by estimating the modified structural model with data from two sub-samples. The values of RSMEA for the pooled sample and both sub-samples were 0.046, 0.062 (M-type), and 0.071 (I-type) respectively. It shows that the structural model in this paper fit well with the pooled sample and both sub-samples. Regarding the venture type difference, the model accounted 71.6% of variance in innovation for the M-type venture, and 43.9% for the I-type ventures, respectively.

Table 7 presents the result of M-type and I-type ventures, respectively. It was found that most of the path coefficients were significant for both sub-samples. The only exception in the M-type sample was the path from network heterogeneity to knowledge digestion. In contrast, the network heterogeneity would positively affect knowledge digestion in the I-type sample ($\beta = 0.294$ **). Noteworthy is the path from knowledge application to innovation, which was significant in the M-type sample ($\beta = 0.622$ ***) but insignificant in the I-type sample.

To test the effect of venture type, the authors followed the process proposed by Keil et al. [136]. Table 7 also gives the subgroup analysis result. T value was used to detect the differences in specific paths between M-type and I-type ventures.

Table 7 details that the effect of the social network on innovation through knowledge digestion was greater in those I-type new ventures (**H3b**, **H4b**, and **H5b**). One possible reason is that I-type companies did not have to deal with the big supporter company, thus the I-type ventures might have less bureaucracy and their employees were usually more active. Their quickness in assimilating knowledge and spotting new opportunities was especially amplified by the density and heterogeneity of their networks.

Table 7. Standardized path coefficients of the two types of new ventures.

Assumption	M-Type	I-Type	t-Value
Knowledge digestion → Innovation (H3b)	0.303 ***	0.577 ***	1.449
Knowledge Application → Innovation (H3c)	0.622 ***	0.105	2.523 **
Network density → Knowledge acquisition (H4a)	0.450 ***	0.348 **	0.262
Network density → Knowledge digestion (H4b)	0.374 **	0.430 **	0.184
Network density → Knowledge application (H4c)	0.464 ***	0.279 **	0.666
Network heterogeneity → Knowledge digestion (H5b)	0.015	0.294 **	0.677

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

The M-type startups, on the contrary, had more advantages in resources. With the help of large companies, M-types were more efficient in implementing the new ideas, which would in turn accelerate their innovation. Providing they were good at transforming and exploiting the current knowledge set into generating new ideas (knowledge application), the innovation performance would be improved. Therefore, the effect of knowledge application skill on innovation was much larger and more significant among M-type ventures (**H3c**).

This finding is interesting, as it is generally believed that the more support for those new ventures, the better performances in innovation they might have. However, the current study shows that the effects of their knowledge digestion skills on innovation were actually weaker. To have greater success, those ventures have to keep reflecting and avoid slowness in spotting new market trends. For those independent ventures, the lacking of resources might limit them from implementing innovative projects. Therefore, those ventures should be actively involved in seeking external supports once they have an innovative idea. Regarding VC funds, the implication was that they should not step in a new venture's operation too early. The best way for them to support a new venture is to provide more resources to new ventures in the implementation phase of innovative projects.

Still another interesting finding emerged from this study. Network heterogeneity in general played a less important role in improving a new venture's innovation than network density. Network heterogeneity affected innovation through knowledge digestion (Figure 3); however, in the M-type sample, the path from network heterogeneity to knowledge digestion was not significant. The total effects of network heterogeneity on innovation in M-type and I-type samples were 1.1% and 14.6%, respectively. In contrast, the total effects of network density on innovation in M-type and I-type samples were 65.3% and 48% respectively. The example of wechat—an instant message application—might help illustrate this result. Wechat was supported by Tencent from the very beginning. The heterogeneity of their group members was relatively low for most of its founding members were from same company, Tencent. However, this did not stop the success of wechat. The founding group still made it the most popular instant message application by working intensively with each other. Moreover, they kept adding new features into their product on a continuous basis. Therefore, this indicates that it is more useful for new ventures to build a dense team in which members work closely with each rather than focusing on recruiting unnecessarily heterogeneous members, especially for M-type ventures.

5. Conclusions

The paper investigates the impacts of the social network, particularly the internal social network, on new ventures' innovation. Using a comprehensive framework, the current paper integrated the stream of social network, innovation, and absorptive capacity. The authors distinguished this paper from others by highlighting the full mediating effect of absorptive capacity in the relationship of the social network and innovation. The study also extended previous research by comparing the effect of social networks in the different types of new ventures. The main conclusions are shown as follows:

An internal social network can improve the innovation of a new venture through its absorptive capacity, that is, the absorptive capacity (or firm's learning skills) acts as a mediator in the relationship of internal social network and innovation. The communication within the organization is a profoundly social and interactive process, which enables the transference of knowledge and creation of new ideas. A higher network density indicates a closer connection between the members [69,70], which would greatly improve the acquisition, digestion [104,105], and application of knowledge [109] (absorptive capacity). Concurrently, the heterogeneity of the internal social network would equip the organization with a variety of knowledge, leading to a better grasp of new knowledge (knowledge digestion) [112].

The current research further shows that, of the three dimensions in absorptive capacity, the skills of knowledge digestion and application seem to be strong determinants of innovation in new ventures. The knowledge acquisition skill, per se, has little influence on innovation. What really matters is the process of internalizing the new knowledge and creating new ideas. These findings provide insights for new ventures regarding how to make the best use of their internal social networks to improve their absorptive capacities, as well as their innovation.

A deeper understanding of the effect of the internal social network on innovation can be gained by comparing the two different types of new ventures. Based on whether the new venture has external support, the authors classify new ventures into two types: the new venture supported by mature enterprises (M-type) and the independent startups (I-type). The authors show that the effect of the social network on innovation through knowledge digestion skill is greater in those I-type new ventures. This finding is especially important to those VC funds who expect high innovation performance from the ventures they support, as it indicates that it is not always best to support new ventures throughout the whole process of innovation. The best way for them to support a new venture is to provide resources to it in its late phase of the innovation process, essentially the implementation phase of innovative projects.

The research contributes by examining the interactive effects of the internal social network and absorptive capacity on innovation for new ventures. However, the paper still has certain limitations that further research might try to overcome. First, the research focuses on Chinese new ventures. Although the approach is appropriate, it is still too early to arrive at a general conclusion that might be applied to all countries. Therefore, further research across the world or in other countries is strongly recommended. Second, the data has an acceptable but not perfect level of performance on KMO index (EFA) and factor loadings (CFA); future research might try to increase the sample size or develop a new questionnaire to have a better statistical result. Third, the questionnaires in the current research were randomly distributed to different industries. As the economy develops, some emerging industries are demonstrating new characteristics that differ from traditional sectors. The authors believe that the different characteristics of these sectors deserve attention in future research. Fourth, the changes in business settings might bring new insights to this research. Specifically, technological progress, such as internal Web 2.0 tools and augmented reality, is offering tremendous potential in terms of internal collaboration and knowledge management [98]. Via the display of situationally relevant information in the most comprehensible manner, those technologies might greatly enhance firms' absorptive capacities, which could further improve the role of the absorptive capacity in the relationship between the internal social network and innovation. Therefore, it would be interesting to verify that in the next step. The authors believe there are other factors that might affect innovation in new ventures. As innovation is becoming more and more important, continued exploration of the potential factors under innovation would be insightful.

Acknowledgments: The authors are very grateful for the insightful comments and suggestions of the anonymous reviewers and the editor, which have helped to significantly improve this article. Furthermore, this research was supported by the National Natural Science Foundation of China (No. 71371025), and Beijing Natural Science Foundation (No. 9182010).

Author Contributions: Wei Shan and Chu Zhang conceived and designed the model and experiments; Jingyi Wang performed the experiments; Wei Shan, Chu Zhang and Jingyi Wang analyzed the data; Wei Shan and Chu Zhang wrote the paper. All authors read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1 Questionnaire

Internal social network

Internal social network density

- Members have discussion meetings on a regular basis (SND1).
- Members often communicate and collaborate on work (SND2).
- Members often discuss issues with each other (SND3).
- Members get along with each other (SND4).

Internal social network heterogeneity

- Members have large differences in the level of education (SNH1).
- Members have different working experience (SNH2).
- Members have different ways of thinking (SNH3).
- Members have different years of working in this group (SNH4).
- Members have difference in working styles, knowledge sets, and experiences (SNH5).

Absorptive capacity

Knowledge acquisition

- Members frequently collect industrial technological information and managerial insights in a detailed manner. (KAC1)
- Members often evaluate technology and management information obtained externally (KAC2).
- Members communicate with other people to acquire new knowledge on a regular basis (KAC3).
- Members often compare the difference between obtained and existing technology and management knowledge (KAC4).

Knowledge digestion

- Members can store up new technology and knowledge (KAS1).
- Members can quickly catch (grasp) the new technology and business mode from outside (KAS2).
- Members can quickly understand and analyze the change of market demand (KAS3).
- Members can quickly spot new opportunities in current business environment (KAS4).

Knowledge application

- Members have strong ability in adapting the newly absorbed knowledge to products/operations. (KAP1).
- Members often utilize new technology and ideas in the development process of new products (KAP2).
- Members often integrate the new technology with new ideas to develop new products or increase the efficiency of operations (KAP3).
- Teams/members have been rewarded for applying new technology or new managerial approach (KAP4).

Innovation performance

- Team/company often uses new products and service (GIP1).
- Team/company often introduces new technology to improve workflow (GIP2).
- Team/company often uses new methods to improve product performance (GIP3).
- Team/company often makes use of new methods to improve group performance (GIP4).
- Team/company often develops new products and service that can be accepted by the market (GIP5).
- Team/company can change the service project and method based on the customers demand (GIP6).

Note: GIP5 and GIP6 were deleted in our pilot survey.

Appendix A.2 EFA Analysis Result of Pilot Study

1. Result of explanatory factor analysis (EFA) (Original Questionnaire).

Rotated Composition Matrix							
	Components						
	1	2	3	4	5	6	7
KAP 3	0.834	0.154	0.143	−0.017	−0.031	−0.123	0.105
KAP 2	0.825	0.105	0.066	0.210	−0.083	−0.070	0.137
KAP 1	0.763	0.228	0.121	−0.016	0.180	−0.094	0.053
KAP 4	0.701	0.176	0.136	0.066	−0.025	−0.077	−0.369
GIP 3	0.097	0.888	0.153	0.043	−0.040	−0.072	−0.008
GIP 2	0.216	0.798	0.176	−0.085	−0.016	−0.183	0.103
GIP 4	0.269	0.653	0.331	0.151	−0.135	−0.214	0.011
GIP 1	0.365	0.637	0.093	−0.045	0.030	−0.214	0.314
GIP 6	0.004	0.557	0.188	−0.205	0.069	0.299	0.427
SND 3	0.211	0.149	0.846	0.056	0.049	−0.018	0.098
SND 2	0.034	0.198	0.828	−0.079	0.085	0.141	0.060
SND 1	0.036	0.119	0.825	0.151	−0.138	−0.051	0.089
SND 4	0.195	0.180	0.598	−0.121	0.190	0.156	−0.159
KAS 3	−0.031	−0.019	0.170	0.800	−0.098	−0.040	0.259
KAS 2	0.114	−0.038	−0.077	0.796	−0.059	0.138	−0.249
KAS 4	−0.002	−0.091	0.071	0.793	0.195	−0.121	0.207
KAS 1	0.121	0.109	−0.110	0.743	−0.209	0.034	−0.190
SNH 5	−0.112	0.054	−0.052	−0.004	0.739	−0.030	−0.025
SNH 4	−0.156	−0.033	0.151	−0.185	0.707	0.016	−0.213
SNH 3	0.018	−0.224	0.184	−0.067	0.704	−0.096	−0.024
SNH 1	0.139	0.096	−0.062	0.081	0.680	0.166	0.055
SNH 2	0.411	−0.055	−0.080	−0.051	0.628	0.062	0.258
KAC 2	−0.039	−0.057	0.048	0.131	0.017	0.786	0.059
KAC 3	−0.139	−0.104	0.280	0.019	−0.005	0.744	−0.048
KAC 4	−0.053	−0.201	0.111	−0.154	0.169	0.666	0.039
KAC 1	−0.089	−0.005	−0.212	−0.009	−0.074	0.647	0.042
GIP 5	0.105	0.303	0.084	0.066	−0.070	0.085	0.784

Extraction Method: principal component analysis.

Total variance explained.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.488	20.325	20.325	5.488	20.325	20.325
2	3.267	12.101	32.426	3.267	12.101	32.426
3	2.680	9.926	42.353	2.680	9.926	42.353
4	2.484	9.201	51.553	2.484	9.201	51.553
5	1.776	6.577	58.131	1.776	6.577	58.131
6	1.612	5.969	64.100	1.612	5.969	64.100
7	1.047	3.878	67.978	1.047	3.878	67.978
8	0.989	3.663	71.641			
9	0.899	3.329	74.970			
10	0.845	3.130	78.099			
11	0.772	2.861	80.960			
12	0.608	2.252	83.212			
13	0.538	1.993	85.205			
14	0.525	1.944	87.149			
15	0.489	1.812	88.961			
16	0.436	1.614	90.575			
17	0.395	1.462	92.037			
18	0.351	1.300	93.337			
19	0.323	1.197	94.533			
20	0.291	1.077	95.610			
21	0.265	0.982	96.593			
22	0.215	0.797	97.389			
23	0.191	0.708	98.098			
24	0.172	0.635	98.733			
25	0.128	0.474	99.207			
26	0.117	0.435	99.642			
27	0.097	0.358	100.000			

2. Result of EFA analysis after deleting two items in innovation.

Rotated Composition Matrix						
	Components					
	1	2	3	4	5	6
KAP 3	0.830	0.133	0.191	0.001	−0.013	−0.124
KAP 2	0.801	0.042	0.209	0.216	−0.050	−0.026
KAP 1	0.738	0.107	0.299	−0.017	0.203	−0.056
KAP 4	0.729	0.161	0.058	0.048	−0.054	−0.131
SND 2	0.035	0.836	0.178	−0.066	0.075	0.120
SND 3	0.187	0.827	0.221	0.070	0.058	0.004
SND 1	0.010	0.800	0.209	0.159	−0.129	−0.014
SND 4	0.234	0.646	0.023	−0.127	0.152	0.070
GIP 3	0.090	0.170	0.854	0.014	−0.046	−0.040
GIP 2	0.187	0.167	0.848	−0.099	−0.002	−0.120
GIP 1	0.350	0.094	0.692	−0.045	0.053	−0.175
GIP 4	0.251	0.323	0.681	0.135	−0.130	−0.173
KAS 3	−0.052	0.139	0.091	0.818	−0.076	−0.002
KAS 4	−0.015	0.049	−0.022	0.813	0.209	−0.110
KAS 2	0.134	−0.065	−0.123	0.782	−0.082	0.103
KAS 1	0.144	−0.096	0.030	0.727	−0.231	0.005
SNH 5	−0.129	−0.042	0.054	−0.015	0.737	−0.018
SNH 3	0.027	0.204	−0.270	−0.061	0.691	−0.143
SNH 4	−0.164	0.162	−0.080	−0.194	0.689	0.002
SNH 1	0.125	−0.047	0.094	0.073	0.685	0.175
SNH 2	0.390	−0.082	0.015	−0.035	0.656	0.080
KAC 2	−0.053	0.054	−0.041	0.124	0.031	0.818
KAC 3	−0.118	0.314	−0.198	0.015	−0.020	0.704
KAC 1	−0.104	−0.210	0.009	−0.020	−0.057	0.683
KAC 4	−0.047	0.133	−0.239	−0.151	0.170	0.648

Extraction Method: principal component analysis.

Total variance explained.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.214	20.858	20.858	5.214	20.858	20.858
2	3.174	12.695	33.553	3.174	12.695	33.553
3	2.661	10.644	44.196	2.661	10.644	44.196
4	2.375	9.500	53.697	2.375	9.500	53.697
5	1.696	6.784	60.481	1.696	6.784	60.481
6	1.344	5.376	65.857	1.344	5.376	65.857
7	0.982	3.929	69.786			
8	0.903	3.614	73.400			
9	0.861	3.443	76.842			
10	0.803	3.211	80.053			
11	0.638	2.550	82.603			
12	0.584	2.337	84.941			
13	0.512	2.049	86.990			
14	0.483	1.932	88.922			
15	0.461	1.844	90.766			
16	0.352	1.407	92.173			
17	0.336	1.343	93.516			
18	0.322	1.288	94.805			
19	0.280	1.120	95.925			
20	0.256	1.023	96.948			
21	0.212	0.850	97.798			
22	0.186	0.744	98.542			
23	0.135	0.541	99.082			
24	0.131	0.522	99.605			
25	0.099	0.395	100.000			

References

1. Ahuja, G.; Lampert, C.M.; Tandon, V. Moving beyond Schumpeter: Management research on the determinants of technological innovation. *Acad. Manag. Ann.* **2008**, *2*, 1–98. [CrossRef]
2. McCraw, T.K. *Prophet of Innovation: Joseph Schumpeter and Creative Destruction*; Harvard University Press: Cambridge, MA, USA, 2007.
3. Kim, Y.; Lui, S.S. The impacts of external network and business group on innovation: Do the types of innovation matter? *J. Bus. Res.* **2015**, *68*, 1964–1973. [CrossRef]
4. Cefis, E.; Marsili, O. Survivor: The role of innovation in firms' survival. *Res. Pol.* **2006**, *35*, 626–641. [CrossRef]
5. Drucker, P.F. *Innovation and Entrepreneurship: Practice and Principles*; Harper & Row: New York, NY, USA, 1985.
6. Thompson, V.A. Bureaucracy and innovation. *Adm. Sci. Q.* **1965**, *10*, 1–20. [CrossRef]
7. Soto-Acosta, P.; Loukis, E.; Colomo-Palacios, R.; Lytras, M.D. An empirical research of the effect of internet-based innovation on business value. *Afr. J. Bus. Manag.* **2010**, *4*, 4096–4105.
8. Edquist, C. *Systems of Innovation: Technologies, Institutions, and Organizations*; Pinter Publishers/Cassell Academic: London, UK, 1997.
9. Pérez-Luño, A.; Medina, C.C.; Lavado, A.C.; Rodríguez, G.C. How social capital and knowledge affect innovation. *J. Bus. Res.* **2011**, *64*, 1369–1376. [CrossRef]
10. Felix, R.; Rauschnabel, P.A.; Hinsch, C. Elements of strategic social media marketing: A holistic framework. *J. Bus. Res.* **2017**, *70*, 118–126. [CrossRef]
11. Capaldo, A. Network structure and innovation: The leveraging of a dual network as a distinctive relational capability. *Strat. Manag. J.* **2007**, *28*, 585–608. [CrossRef]
12. Shu, S.T.; Wong, V.; Lee, N. The effects of external linkages on new product innovativeness: An examination of moderating and mediating influences. *J. Strat. Mark.* **2005**, *13*, 199–218. [CrossRef]
13. Uzzi, B. Social structure and competition in interfirm networks: The paradox of embeddedness. *Adm. Sci. Q.* **1997**, *42*, 35–67. [CrossRef]

14. Laursen, K.; Salter, A. Open for innovation: The role of openness in explaining innovation performance among UK manufacturing firms. *Strat. Manag. J.* **2006**, *27*, 131–150. [CrossRef]
15. Jiang, Y.; Chun, W.; Yang, Y. The Effects of External Relations Network on Low-Carbon Technology Innovation: Based on the Study of Knowledge Absorptive Capacity. *Sustainability* **2018**, *10*, 155. [CrossRef]
16. Salazar, A.; Gonzalez, J.M.H.; Duysters, G.; Sabidussi, A.; Allen, M. The value for innovation of inter-firm networks and forming alliances: A meta-analytic model of indirect effects. *Comput. Hum. Behav.* **2016**, *64*, 285–298. [CrossRef]
17. Oerlemans, L.A.; Meeus, M.T.; Boekema, F.W. Do networks matter for innovation? The usefulness of the economic network approach in analysing innovation. *Tijdschrift Voor Economische En Sociale Geografie* **1998**, *89*, 298–309. [CrossRef]
18. Liu, D.; Gong, Y.; Zhou, J. Human resource systems, employee creativity, and firm innovation: The moderating role of firm ownership. *Acad. Manag. J.* **2017**, *60*, 1164–1188. [CrossRef]
19. Zhang, X.; Chen, H.; Wang, W.; Ordóñez de Pablos, P. What is the role of IT in innovation? A bibliometric analysis of research development in IT innovation. *Behav. Inf. Technol.* **2016**, *35*, 1130–1143. [CrossRef]
20. Chen, C.J.; Huang, J.W. Strategic human resource practices and innovation performance: The mediating role of knowledge management capacity. *J. Bus. Res.* **2009**, *62*, 104–114. [CrossRef]
21. Srivastava, M.K.; Gnyawali, D.R. When do relational resources matter? Leveraging portfolio technological resources for breakthrough innovation. *Acad. Manag. J.* **2011**, *54*, 797–810. [CrossRef]
22. Bakar, L.J.A.; Ahmad, H. Assessing the relationship between firm resources and product innovation performance: A resource-based view. *Bus. Process Manag. J.* **2010**, *16*, 420–435. [CrossRef]
23. Carnabuci, G.; Diószegi, B. Social networks, cognitive style, and innovative performance: A contingency perspective. *Acad. Manag. J.* **2015**, *58*, 881–905. [CrossRef]
24. Tarafdar, M.; Gordon, S.R. Understanding the influence of information systems competencies on process innovation: A resource-based view. *J. Strateg. Inf. Syst.* **2007**, *16*, 353–392. [CrossRef]
25. Sok, P.; O'Cass, A. Achieving superior innovation-based performance outcomes in SMEs through innovation resource–capability complementarity. *Ind. Mark. Manag.* **2011**, *40*, 1285–1293. [CrossRef]
26. Huang, J.W.; Li, Y.H. The mediating effect of knowledge management on social interaction and innovation performance. *Int. J. Manpow.* **2009**, *30*, 285–301. [CrossRef]
27. Yan, Y.; Guan, J. Social capital, exploitative and exploratory innovations: The mediating roles of ego-network dynamics. *Technol. Forecast. Soc. Chang.* **2018**, *126*, 244–258. [CrossRef]
28. Lin, H.; Zeng, S.; Liu, H.; Li, C. Bridging the gaps or fecklessness? A moderated mediating examination of intermediaries' effects on corporate innovation. *Technovation* **2018**. [CrossRef]
29. Calantone, R.J.; Cavusgil, S.T.; Zhao, Y. Learning orientation, firm innovation capability, and firm performance. *Ind. Mark. Manag.* **2002**, *31*, 515–524. [CrossRef]
30. Nielsen, B.B. The role of knowledge embeddedness in the creation of synergies in strategic alliances. *J. Bus. Res.* **2005**, *58*, 1194–1204. [CrossRef]
31. García-Sánchez, E.; García-Morales, V.J.; Martín-Rojas, R. Influence of Technological Assets on Organizational Performance through Absorptive Capacity, Organizational Innovation and Internal Labour Flexibility. *Sustainability* **2018**, *10*, 770. [CrossRef]
32. Tsai, W. Knowledge transfer in intraorganizational networks: Effects of network position and absorptive capacity on business unit innovation and performance. *Acad. Manag. J.* **2001**, *44*, 996–1004. [CrossRef]
33. Cohen, W.M.; Levinthal, D.A. Absorptive capacity: A new perspective on learning and innovation. *Adm. Sci. Q.* **1990**, *35*, 128–152. [CrossRef]
34. Zahra, S.A.; George, G. Absorptive capacity: A review, reconceptualization, and extension. *Acad. Manag. Rev.* **2002**, *27*, 185–203. [CrossRef]
35. Ethiraj, S.K.; Kale, P.; Krishnan, M.S.; Singh, J.V. Where do capabilities come from and how do they matter? A study in the software services industry. *Strateg. Manag. J.* **2005**, *26*, 25–45. [CrossRef]
36. Sivadas, E.; Dwyer, F.R. An examination of organizational factors influencing new product success in internal and alliance-based processes. *J. Mark.* **2000**, *64*, 31–49. [CrossRef]
37. Mowery, D.C.; Oxley, J.E.; Silverman, B.S. Strategic alliances and interfirm knowledge transfer. *Strat. Manag. J.* **1996**, *17*, 77–91. [CrossRef]
38. Nieto, M.; Quevedo, P. Absorptive capacity, technological opportunity, knowledge spillovers, and innovative effort. *Technovation* **2005**, *25*, 1141–1157. [CrossRef]

39. Kotabe, M.; Jiang, C.X.; Murray, J.Y. Examining the complementary effect of political networking capability with absorptive capacity on the innovative performance of emerging-market firms. *J. Manag.* **2017**, *43*, 1131–1156. [CrossRef]
40. Caloghirou, Y.; Kastelli, I.; Tsakanikas, A. Internal capabilities and external knowledge sources: Complements or substitutes for innovative performance? *Technovation* **2004**, *24*, 29–39. [CrossRef]
41. Burt, R.S. *Structural Holes: The Social Construction of Competition*; Harvard University Press: Cambridge, MA, USA, 1992.
42. Al-Laham, A.; Amburgey, T.L.; Baden-Fuller, C. Who is my partner and how do we dance? Technological collaboration and patenting speed in US biotechnology. *Br. J. Manag.* **2010**, *21*, 789–807. [CrossRef]
43. Mitchell, J.C. *Social Networks in Urban Situations: Analyses of Personal Relationships in Central African Towns*; Manchester University Press: Manchester, UK, 1969.
44. Gottlieb, B.H. Social networks and social support: An overview of research, practice, and policy implications. *Health Educ. Q.* **1985**, *12*, 5–22. [CrossRef] [PubMed]
45. Huang, Y. Social class, social network and mental happiness. *Taiwan. Sociol. Assoc.* **1998**, *21*, 171–210.
46. Nahapiet, J.; Ghoshal, S. Social capital, intellectual capital, and the organizational advantage. *Acad. Manag. Rev.* **1998**, *23*, 242–266. [CrossRef]
47. Putnam, R.D. *Bowling alone: The Collapse and Revival of American Community*; Simon and Schuster: New York, NY, USA, 2001.
48. Murovec, N.; Prodan, I. Absorptive capacity, its determinants, and influence on innovation output: Cross-cultural validation of the structural model. *Technovation* **2009**, *29*, 859–872. [CrossRef]
49. Belso-Martínez, J.A.; Xavier Molina-Morales, F.; Mas-Verdu, F. Perceived usefulness of innovation programs for high-tech and low-tech firms. *Manag. Decis.* **2013**, *51*, 1190–1206. [CrossRef]
50. Belso-Martinez, J.A.; Molina-Morales, F.X.; Mas-Verdu, F. Combining effects of internal resources, entrepreneur characteristics and KIS on new firms. *J. Bus. Res.* **2013**, *66*, 2079–2089. [CrossRef]
51. Hervas-Oliver, J.L.; Albors-Garrigos, J.; de-Miguel, B.; Hidalgo, A. The role of a firm’s absorptive capacity and the technology transfer process in clusters: How effective are technology centres in low-tech clusters? *Entrep. Reg. Dev.* **2012**, *24*, 523–559. [CrossRef]
52. Powell, W.W.; Koput, K.W.; Smith-Doerr, L. Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Adm. Sci. Q.* **1996**, *4*, 116–145. [CrossRef]
53. Boudreau, K.; Lakhani, K. How to manage outside innovation. *MIT Sloan Manag. Rev.* **2009**, *50*, 69.
54. Chesbrough, H.W. *Open Innovation: The New Imperative for Creating and Profiting from Technology*; Harvard Business Press: Brighton, MA, USA, 2006.
55. Dyer, J.H.; Hatch, N.W. Using supplier networks to learn faster. *MIT Sloan Manag. Rev.* **2004**, *45*, 57.
56. Keil, T.; Maula, M.; Schildt, H.; Zahra, S.A. The effect of governance modes and relatedness of external business development activities on innovative performance. *Strat. Manag. J.* **2008**, *29*, 895–907. [CrossRef]
57. Burt, R.S. Structural holes and good ideas. *Am. J. Sociol.* **2004**, *110*, 349–399. [CrossRef]
58. Ostgaard, T.A.; Birley, S. New venture growth and personal networks. *J. Bus. Res.* **1996**, *36*, 37–50. [CrossRef]
59. Adler, P.S.; Kwon, S.W. Social capital: Prospects for a new concept. *Acad. Manag. Rev.* **2002**, *27*, 17–40. [CrossRef]
60. Prahalad, C.K.; Hamel, G. The core competence of the corporation. *Harv. Bus. Rev.* **1990**, *68*, 79–91. [CrossRef]
61. Padgett, R.C.; Galan, J.I. The effect of R&D intensity on corporate social responsibility. *J. Bus. Ethics* **2010**, *93*, 407–418. [CrossRef]
62. Reagans, R.; Zuckerman, E.W. Networks, diversity, and productivity: The social capital of corporate R&D teams. *Org. Sci.* **2001**, *12*, 502–517. [CrossRef]
63. Koka, B.R.; Prescott, J.E. Strategic alliances as social capital: A multidimensional view. *Strateg. Manag. J.* **2002**, *23*, 795–816. [CrossRef]
64. Bougrain, F.; Haudeville, B. Innovation, collaboration and SMEs internal research capacities. *Res. Pol.* **2002**, *31*, 735–747. [CrossRef]
65. Henttonen, K.; Janhonen, M.; Johanson, J.E. Internal social networks in work teams: Structure, knowledge sharing and performance. *Int. J. Manpow.* **2013**, *34*, 616–634. [CrossRef]
66. Wong, S.S. Task knowledge overlap and knowledge variety: The role of advice network structures and impact on group effectiveness. *J. Org. Behav.* **2008**, *29*, 591–614. [CrossRef]

67. Mehra, A.; Dixon, A.L.; Brass, D.J.; Robertson, B. The social network ties of group leaders: Implications for group performance and leader reputation. *Org. Sci.* **2006**, *17*, 64–79. [CrossRef]
68. Luo, J.D. Social network structure and performance of improvement teams. *Int. J. Bus. Perform. Manag.* **2005**, *7*, 208–223. [CrossRef]
69. Burt, R.S. The social capital of structural holes. In *The New Economic Sociology: Developments in an Emerging Field*; Russell Sage Foundation: New York, NY, USA, 2002.
70. Coleman, J.S. *Foundations of Social Theory*; Belknap Press: Cambridge, MA, USA, 1990.
71. Zacharatos, A.; Barling, J.; Iverson, R.D. High-performance work systems and occupational safety. *J. Appl. Psychol.* **2005**, *90*, 77–93. [CrossRef] [PubMed]
72. Snape, E.; Redman, T. HRM practices, organizational citizenship behaviour, and performance: A multi-level analysis. *J. Manag. Stud.* **2010**, *47*, 1219–1247. [CrossRef]
73. Duysters, G.; Lokshin, B. Determinants of alliance portfolio complexity and its effect on innovative performance of companies. *J. Prod. Innov. Manag.* **2011**, *28*, 570–585. [CrossRef]
74. Leiponen, A.; Helfat, C.E. Innovation objectives, knowledge sources, and the benefits of breadth. *Strat. Manag. J.* **2010**, *31*, 224–236. [CrossRef]
75. Un, C.A.; Cuervo-Cazurra, A.; Asakawa, K. R&D collaborations and product innovation. *J. Prod. Innov. Manag.* **2010**, *27*, 673–689. [CrossRef]
76. Baron, R.A. Opportunity recognition as pattern recognition: How entrepreneurs "connect the dots" to identify new business opportunities. *Acad. Manag. Perspect.* **2006**, *20*, 104–119. [CrossRef]
77. Cross, R.; Cummings, J.N. Tie and network correlates of individual performance in knowledge-intensive work. *Acad. Manag. J.* **2004**, *47*, 928–937. [CrossRef]
78. Der Foo, M.; Wong, P.K.; Ong, A. Do others think you have a viable business idea? Team diversity and judges' evaluation of ideas in a business plan competition. *J. Bus. Vent.* **2005**, *20*, 385–402. [CrossRef]
79. Cohen, W.M.; Levinthal, D.A. Innovation and learning: The two faces of R & D. *Econ. J.* **1989**, *99*, 569–596. [CrossRef]
80. Gao, S.; Xu, K.; Yang, J. Managerial ties, absorptive capacity, and innovation. *Asia Pac. J. Manag.* **2008**, *25*, 395–412. [CrossRef]
81. Lichtenthaler, U.; Lichtenthaler, E. A capability-based framework for open innovation: Complementing absorptive capacity. *J. Manag. Stud.* **2009**, *46*, 1315–1338. [CrossRef]
82. Scuotto, V.; Del Giudice, M.; Carayannis, E.G. The effect of social networking sites and absorptive capacity on SMES'innovation performance. *J. Technol. Transf.* **2017**, *42*, 409–424. [CrossRef]
83. Lane, P.J.; Koka, B.R.; Pathak, S. The reification of absorptive capacity: A critical review and rejuvenation of the construct. *Acad. Manag. Rev.* **2006**, *31*, 833–863. [CrossRef]
84. Lowik, S.; Kraaijenbrink, J.; Groen, A.J. Antecedents and effects of individual absorptive capacity: A micro-foundational perspective on open innovation. *J. Knowl. Manag.* **2017**, *21*, 1319–1341. [CrossRef]
85. Nonaka, I.; Takeuchi, H. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*; Oxford University Press: New York, NY, USA, 1995.
86. Chesbrough, H.; Crowther, A.K. Beyond high tech: Early adopters of open innovation in other industries. *R&D Manag.* **2006**, *36*, 229–236. [CrossRef]
87. Lichtenthaler, U. Open innovation: Past research, current debates, and future directions. *Acad. Manag. Perspect.* **2011**, *25*, 75–93. [CrossRef]
88. West, J.; Bogers, M. Open innovation: Current status and research opportunities. *Innovation* **2017**, *19*, 43–50. [CrossRef]
89. Bharati, P.; Zhang, C.; Chaudhury, A. Social media assimilation in firms: Investigating the roles of absorptive capacity and institutional pressures. *Inf. Syst. Front.* **2014**, *16*, 257–272. [CrossRef]
90. Damanpour, F. Organizational innovation: A meta-analysis of effects of determinants and moderators. *Acad. Manag. J.* **1991**, *34*, 555–590. [CrossRef]
91. Moorman, C.; Miner, A.S. Organizational improvisation and organizational memory. *Acad. Manag. Rev.* **1998**, *23*, 698–723. [CrossRef]
92. Verona, G. A resource-based view of product development. *Acad. Manag. Rev.* **1999**, *24*, 132–142. [CrossRef]
93. Kim, L. Crisis construction and organizational learning: Capability building in catching-up at Hyundai Motor. *Org. Sci.* **1998**, *9*, 506–521. [CrossRef]

94. Jian, Z.Q.; Wu, L.Z.; Huang, J. The impact of absorptive, knowledge integration on the organizational innovation and organizational performance. *Sci. Res. Manag.* **2008**, *29*, 80–86.
95. Kogut, B.; Zander, U. What firms do? Coordination, identity, and learning. *Org. Sci.* **1996**, *7*, 502–518. [CrossRef]
96. Kazanjian, R.K.; Drazin, R.; Glynn, M.A. Implementing strategies for corporate entrepreneurship: A knowledge-based perspective. *Strateg. Entrepr.* **2002**, 173–199.
97. Uzzi, B. The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *Am. Soc. Rev.* **1996**, 674–698. [CrossRef]
98. Hein, D.W.; Rauschnabel, P.A. Augmented reality smart glasses and knowledge management: A conceptual framework for enterprise social networks. In *Enterprise Social Networks*; Springer Gabler: Wiesbaden, Germany, 2016.
99. Dyer, J.H.; Singh, H. The relational view: Cooperative strategy and sources of interorganizational competitive advantage. *Acad. Manag. Rev.* **1998**, *23*, 660–679. [CrossRef]
100. Jung-Erceg, P.; Pandza, K.; Armbruster, H.; Dreher, C. Absorptive capacity in European manufacturing: A Delphi study. *Ind. Manag. Data Syst.* **2007**, *107*, 37–51. [CrossRef]
101. Gulati, R. Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances. *Acad. Manag. J.* **1995**, *38*, 85–112. [CrossRef]
102. Gulati, R. Social structure and alliance formation patterns: A longitudinal analysis. *Adm. Sci. Q.* **1995**, 619–652. [CrossRef]
103. Hagedoorn, J.; Duysters, G. Learning in dynamic inter-firm networks: The efficacy of multiple contacts. *Org. Stud.* **2002**, *23*, 525–548. [CrossRef]
104. Jaworski, B.J.; Kohli, A.K. Market orientation: Antecedents and consequences. *J. Mark.* **1993**, 53–70. [CrossRef]
105. Rowley, T.; Behrens, D.; Krackhardt, D. Redundant governance structures: An analysis of structural and relational embeddedness in the steel and semiconductor industries. *Strat. Manag. J.* **2000**, 369–386. [CrossRef]
106. Gilsing, V.; Nooteboom, B.; Vanhaverbeke, W.; Duysters, G.; Van den Oord, A. Network embeddedness and the exploration of novel technologies: Technological distance, betweenness centrality and density. *Res. Pol.* **2008**, *37*, 1717–1731. [CrossRef]
107. Gilsing, V.; Nooteboom, B. Density and strength of ties in innovation networks: An analysis of multimedia and biotechnology. *Eur. Manag. Rev.* **2005**, *2*, 179–197. [CrossRef]
108. Galunic, D.C.; Rodan, S. Resource recombinations in the firm: Knowledge structures and the potential for Schumpeterian innovation. *Strat. Manag. J.* **1998**, 1193–1201. [CrossRef]
109. Rindfleisch, A.; Moorman, C. The acquisition and utilization of information in new product alliances: A strength-of-ties perspective. *J. Mark.* **2001**, *65*, 1–18. [CrossRef]
110. Uzzi, B. Embeddedness in the making of financial capital: How social relations and networks benefit firms seeking financing. *Am. Soc. Rev.* **1999**, 481–505. [CrossRef]
111. Witt, P. Entrepreneurs' networks and the success of start-ups. *Entrepr. Reg. Dev.* **2004**, *16*, 391–412. [CrossRef]
112. McGrath, R.G. Exploratory learning, innovative capacity, and managerial oversight. *Acad. Manag. J.* **2001**, *44*, 118–131. [CrossRef]
113. Tushman, M.L. Special boundary roles in the innovation process. *Adm. Sci. Q.* **1997**, *22*, 587–605. [CrossRef]
114. Atuahene-Gima, K. The effects of centrifugal and centripetal forces on product development speed and quality: How does problem solving matter? *Acad. Manag. J.* **2003**, *46*, 359–373. [CrossRef]
115. Cardinal, L.B. Technological innovation in the pharmaceutical industry: The use of organizational control in managing research and development. *Org. Sci.* **2001**, *12*, 19–36. [CrossRef]
116. Ke, J.L.; Shi, J.T.; Sun, J.M. Dimensions' developing and structure's testing of team social capital. *Stud. Sci. Sci.* **2007**, *25*, 935–940.
117. Peng, W.; Zhou, H.L.; Fu, Z.P. Effect mechanism of team internal social network on team innovation performance: An empirical study based on R&D team in enterprises. *Sci. Res. Manag.* **2013**, *34*, 135–142.
118. Zhang, X.E.; Sun, Z.H.; Wang, B. An empirical analysis of the effect of heterogeneity within entrepreneurial team on entrepreneurial performance: A case of 264 enterprises in seven provinces and cities. *East China Econ. Manag.* **2013**, *27*, 112–115.
119. Jehn, K.A.; Northcraft, G.B.; Neale, M.A. Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Adm. Sci. Q.* **1999**, *44*, 741–763. [CrossRef]

120. Nicotra, M.; Romano, M.; Del Giudice, M. The evolution dynamic of a cluster knowledge network: The role of firms' absorptive capacity. *J. Knowl. Econ.* **2014**, *5*, 240–264. [CrossRef]
121. Jansen, J.J.; Van Den Bosch, F.A.; Volberda, H.W. Managing potential and realized absorptive capacity: How do organizational antecedents matter? *Acad. Manag. J.* **2005**, *48*, 999–1015. [CrossRef]
122. Zaltman, G.; Duncan, R.; Holbek, J. *Innovations and Organizations*; Wiley: New York, NY, USA, 1973.
123. Rogers, E.M. *Diffusion of Innovations*, 4th ed.; Simon and Schuster: New York, NY, USA, 2010.
124. Amabile, T.M.; Conti, R.; Coon, H.; Lazenby, J.; Herron, M. Assessing the work environment for creativity. *Acad. Manag. J.* **1996**, *39*, 1154–1184. [CrossRef]
125. Chen, G.H. *A Study on the Impact of Enterprise R&D Team's Informal Network Structure on the Performance of Product Innovation*; Renmin University: Beijing, China, 2008.
126. Kaiser, H.F.; Rice, J.; Little, J.; Mark, I. Educational and psychological measurement. *Am. Psychol. Assoc.* **1974**, *34*, 111–117. [CrossRef]
127. Hair, J.F.; Black, W.C.; Babin, B.J.; Anderson, R.E. *Multivariate Data Analysis*, 7th ed.; Pearson Prentice Hall: Upper Saddle River, NJ, USA, 2006.
128. Hu, L.; Bentler, P.M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Model.* **1999**, *6*, 1–55. [CrossRef]
129. Bagozzi, R.P.; Yi, Y. On the evaluation of structural equation models. *J. Acad. Market. Sci.* **1988**, *16*, 74–94. [CrossRef]
130. Al-Somali, S.A.; Gholami, R.; Clegg, B. An investigation into the acceptance of online banking in Saudi Arabia. *Technovation* **2009**, *29*, 130–141. [CrossRef]
131. Kim, M.J.; Chung, N.; Lee, C.K. The effect of perceived trust on electronic commerce: Shopping online for tourism products and services in South Korea. *Tour. Manag.* **2011**, *32*, 256–265. [CrossRef]
132. Kaplan, D. *Structural Equation Modeling: Foundations and Extensions*, 2nd ed.; Sage Publications: New York, NY, USA, 2008.
133. Hancock, G.R. Fortune cookies, measurement error, and experimental design. *J. Mod. Appl. Stat. Methods* **2003**, *2*, 293–305. [CrossRef]
134. Chahal, H.; Kaur Sahi, G.; Rani, A. Moderating role of perceived risk in credit card usage and experience link. *J. Indian Bus. Res.* **2014**, *6*, 286–308. [CrossRef]
135. Pelled, L.H. Demographic diversity, conflict, and work group outcomes: An intervening process theory. *Org. Sci.* **1996**, *7*, 615–631. [CrossRef]
136. Keil, M.; Tan, B.C.Y.; Wei, K.K.; Saarinen, T.; Tuunainen, V.; Wassenaar, A. A cross-cultural study on escalation of commitment behavior in software projects. *MIS Q.* **2000**, *24*, 299–325. [CrossRef]

sustainability MDPI

Article

Crowdsourcing Analysis of Twitter Data on Climate Change: Paid Workers vs. Volunteers

Andrei P. Kirilenko [1,*], Travis Desell [2], Hany Kim [3] and Svetlana Stepchenkova [1]

[1] The Department of Tourism, Recreation and Sport Management, University of Florida, P.O. Box 118208, Gainesville, FL 32611-8208, USA; svetlana.step@ufl.edu
[2] The Department of Computer Science, University of North Dakota, Streibel Hall, 3950 Campus Road Stop 9015, Grand Forks, ND 58202-9015, USA; tdesell@cs.und.edu
[3] The Department of Business Administration and Tourism and Hospitality Management, Mount Saint Vincent University, 166 Bedford Highway, Halifax, NS B3M 2J6, Canada; Hany.Kim@msvu.ca
* Correspondence: andrei.kirilenko@ufl.edu; Tel.: +1-352-294-1648

Received: 25 September 2017; Accepted: 30 October 2017; Published: 3 November 2017

Abstract: Web based crowdsourcing has become an important method of environmental data processing. Two alternatives are widely used today by researchers in various fields: paid data processing mediated by for-profit businesses such as Amazon's Mechanical Turk, and volunteer data processing conducted by amateur citizen-scientists. While the first option delivers results much faster, it is not quite clear how it compares with volunteer processing in terms of quality. This study compares volunteer and paid processing of social media data originating from climate change discussions on Twitter. The same sample of Twitter messages discussing climate change was offered for processing to the volunteer workers through the Climate Tweet project, and to the paid workers through the Amazon MTurk platform. We found that paid crowdsourcing required the employment of a high redundancy data processing design to obtain quality that was comparable with volunteered processing. Among the methods applied to improve data processing accuracy, limiting the geographical locations of the paid workers appeared the most productive. Conversely, we did not find significant geographical differences in the accuracy of data processed by volunteer workers. We suggest that the main driver of the found pattern is the differences in familiarity of the paid workers with the research topic.

Keywords: citizen-scientist; climate change; crowdsourcing; MTurk; social networks; Twitter

1. Introduction

Development and applications of climate change policies require their acceptance and support by the public. The traditional method of measuring public perceptions of climate change relied on surveys, such as the Climate Change in the American Mind [1]. Recent development of social media, however, offered new, unobtrusive opportunities for measuring public perceptions of climate change worldwide. In this new line of research, Twitter, the fourth most popular social networking site, is the medium most frequently used in research. EBSCO Academic Search Primer database contains 23 journal articles with the words "Twitter" and "Climate change" in the abstract as compared to Facebook (the most popular social media site, 14 journal articles), YouTube (second most popular social media site, 4 journal articles), Instagram (third most popular social media site, 1 journal article), and Reddit (fifth most popular social media site, no journal articles).

Very few of these papers, however, utilized the "Big Data" advantage of social media, exploring the content of the large corpora of Twitter messages. Kirilenko and Stepchenkova [2] used a 1-million sample of tweets to research geographical variations in climate change discourse worldwide. Cody et al. [3] analyzed 1.5 million tweets containing the words "climate" to explore temporal changes in sentiment (described in the paper as "a tool to measure happiness") expressed by the people in relation to climate change. Yang et al. [4] researched the effect of climate and seasonality on depressed mood using automated content analysis of 600 million tweets. Holmberg and Hellsten [5] studied 250 thousand tweets to identify gender differences in climate change communication. Leas et al. [6] analyzed the impact of a celebrity speaking on climate change on social media discussion. Kirilenko et al. [7] and Sisco et al. [8] analyzed the impact of extreme weather events on attention to climate change in social media.

The scarcity of "Big Data" research on climate change perceptions expressed in social media is related to the challenges in content analysis of large volumes of texts. Classification of social media messages requires considerable monetary and time investments, which easily become prohibitive when large datasets are processed. Even when machine learning methods are used, supervised classification still requires a manually classified sample that serves for both algorithm training and for groundtruthing. One popular solution to this research bottleneck is to break the work into small, manageable, easily understandable tasks and then to use the Internet to outsource processing of each task to amateur scientists (referred to as "workers") acting as volunteers or contractors. This method was popularized by Howe [9] as "crowdsourcing". The most famous crowdsourcing effort is probably the Galaxy Zoo project targeted at the classification of imagery of over one million galaxies collected by the Sloan Digital Sky Survey [10], which so far has produced over 50 million classifications [11,12].

The challenge of outsourcing data processing to untrained workers (either volunteered or paid) is associated with quality control. While a significant body of literature studied the quality of paid crowdsourcing (mostly Amazon's Mechanical Turk; further "MTurk"), and few papers dealt with the quality of volunteered crowdsourcing, we are aware of only two studies that attempted to compare the performance of paid and volunteer workers processing the same data in crowdsourced projects. A research completed by Mao et al. [13] investigated the performance of volunteer and paid crowd workers in exoplanet detection through analysis of the planet transit light curves. A set of light curves was offered to volunteer citizen-scientists through the crowdsourcing platform Planet Hunters (www.planethunters.org). A visually similar interface was built as a set of the Amazon MTurk tasks and offered to the paid workers. Overall, the performance of the paid workers was the same or only slightly below that of the volunteer workers, which might be partially related to the high hourly earnings of the paid workers of $4.8–5.6/h compared to the mean earning of below $2/h in paid crowdsourcing projects, as reported by Ross et al. [14]. The authors also noticed that the unpaid citizen scientists were spending almost twice the amount of time on each task as compared to the paid workers.

In a similar study design, Redi and Povoa [15] compared the performance of volunteer participants recruited via Facebook and paid crowdsourcing workers in the estimation of the aesthetic appeal of photographs processed with various filters. The authors found that volunteered work returned a higher correlation between the mean image ratings obtained through crowdsourcing and in a lab experiment, which demonstrated better reliability from volunteered crowdsourcing. They also reported a smaller number of unreliable volunteer workers; however, the volunteers tended to leave the work unfinished more frequently.

Scientific research in high visibility fields is likely to appeal to citizen scientists, making volunteer crowdsourcing a viable alternative to paid workers and ensuring the return of supposedly better quality data. Both of the abovementioned studies that compare the volunteer and paid platforms, however, dealt with same type of data (images) and returned somewhat inconsistent outcomes. We are not aware of a similar comparison done for the textual data from social media. The exploding popularity of social networks led to an ever increasing number of publications using social media data to study public discourse in relation to various natural and/or socio-economic phenomena. Among the available social networking sites, Twitter is one of the most frequently researched, with over 4000 publications on Twitter listed at the Thomson-Reuters' Web of Science.

The goal of our study was to compare the quality of volunteer and paid workers' classification of Twitter messages (tweets) on climate change and provide recommendations on quality control. The study was a part of a larger project on studying the geographical patterns of public perceptions of climate change worldwide [2]. This article is organized as follows. In the second section, we provide a brief review of research dealing with crowdsourcing, with an emphasis on quality control. Then, we present our data and methodology. The results are presented in Section four. The fifth section contains a thorough discussion of our results. Finally, Section six provides a brief conclusion, study limitations, and recommendations for further research.

2. Crowdsourcing and Quality Control Issues

2.1. Crowdsourcing in Scientific Research

Multiple authors used crowdsourcing in their research of climate change impacts. Muller et al. [16] reviewed 29 crowdsourcing projects related to climate change that involved volunteer citizen-scientists engaged in data collection and/or processing. The crowdsourced applications included measurements of snow, rainfall, and other weather data, reporting severe weather outbreaks, recording air quality data, estimating the length of plane contrails (important contributors to warming troposphere), classification of the satellite imagery of tropical cyclones, digitizing weather records found in the 19th century ship logs and many others. Other climate researchers used paid contractor workers. For example, Olteanu et al. [17] used crowdsourcing to process data on climate change coverage in mainstream news and online. Samsel et al. [18] used massive crowd processing of color schemes for digital mapping of ocean salinity change, related to climate change. When data volume is too large for manual processing, even when crowdsourcing is involved e.g., due to costs, crowdsourcing may be used to process a sample of data, which can further be used for training and validation of machine learning algorithms. Thus, Yzaguirre et al. [19] used crowdsourcing to validate their text mining application for extraction of past environmental disaster events in news archives. Paid crowdsourcing platforms are also frequently employed for collecting public opinion data. Ranney and Clark [20] used volunteer and paid online participants to collect data on knowledge about climate change. Attari [21] researched peoples' perceptions on their water use and found over two-times underestimation of consumed water.

Multiple factors promote the growing popularity of crowdsourcing. Between those, the most important is probably its speed and cost. When data processing is easily parallelized, large volumes of data can be processed quickly by breaking data analysis into small, easily comprehended micro-tasks that do not require special training and are then solved by hundreds of citizen-scientists. On one hand, multiple studies reported hourly wages of the crowd workers well below the minimum hourly wage (e.g., [14]). The crowd workers are regarded as independent contractors, which generally frees their employers from tax and legal obligations, reducing costs even further. On the other hand, projects that are deemed socially important appeal to volunteer labor and can be completed at no monetary cost at all—for example, the abovementioned Sloan Digital Sky Survey attracted over 150,000 volunteer workers [11].

Many crowdsourcing platforms are available on the Web (for a review, see [22]). Among those platforms specialized on outsourcing the micro-tasks, the Amazon's Mechanical Turk platform is probably the most popular one, having over 0.5 million registered workers ("Turkers") from 190 countries [23] The demographics of the Turkers and an introductory guide to conducting a crowdsourcing research on MTurk platform was published by Mason and Suri [24].

2.2. Quality Issues in Crowdsourcing

It has been repeatedly demonstrated that complex problems normally requiring advanced technical training can be solved by crowdsourcing; examples include civil engineering [25], bioinformatics [26], astronomy [12] and many others. Furthermore, under certain conditions, even generating novel ideas and innovations can be crowdsourced with results comparable to those obtained from experts [27]. Crowdsourced results may, however, be unreliable due to the following factors:

- Instrumental errors arising from complex data pre- and post-processing, which involves multiple third-party platforms used to prepare data for processing, send tasks to workers, collect processing results, and finally, join the processed data.
- Involuntary errors by human raters, e.g., due to insufficiently clear instructions and workers' cognitive limitations.
- Deliberately poor performance of the human raters. A worker may vandalize the survey and provide wrong data, may try to maximize the number of tasks processed per time unit for monetary or other benefits, may provide incorrect information regarding its geographical location, or may lack motivation [28].

The last item in this list of potential error sources has attracted a lot of attention from practitioners, due to its high potential to render project results unusable. A widely cited experiment that consisted of rating Wikipedia articles by Amazon Mechanical Turk workers demonstrated only a marginally significant correlation between the crowdsourced and experts' ratings [29]. However, the same study showed that simple changes in task design aimed at discouraging workers' cheating, increased the median time spent by a worker to complete one task from 1.5 to 4.1 min, decreased the percentage of unusable classifications from 49% to 6%, and noticeably improved correlation with expert classification.

On the other hand, it seems only logical to suppose that deliberately poor performance should not occur with voluntary participants, as they do not seek maximization of monetary gain from their work. Indeed, a study of motivations of volunteer workers in a crowdsourced scientific project on galaxy image classification [30] found that the primary motivation was seeking to contribute to original scientific research (39.8% of the respondents), followed by an interest in the scientific discipline (12.4%) and discovery (10.4%). Other motivations supplied by participants may, however, contribute towards a "cutting corners" behavior: they include a desire to complete more tasks than other participants, seeking fame for discoveries, and completing a homework assignment [30].

The data quality problem is typically resolved by heavily redundant designs where a single task is assigned to multiple workers; the "true" classification value is then defined as the majority vote (the mode). The required redundancy, however, increases costs while reducing the benefits of using crowdsourcing. Since the highest threat to reliability of paid crowdsourcing results come from a small but highly active group of workers trying to game the system [29], there is an incentive to identify the poorly performing workers and exclude their results from further consideration.

Multiple methods have been suggested to reduce the impact of this group on study results; for an overview of quality control methods in crowdsourced solutions, see [31]. Rouse et al. [32] demonstrated that an improvement in accuracy can be obtained simply by asking the workers if they were attentive in completing the task, and giving them an option to remove their data from consideration. A commonly used solution is to employ a worker reputation system with assigning tasks to workers with approval ratings above a certain pre-set level [33]. Another set of methods of identification and expulsion of the unethical workers is based on a set of indices measuring (1) agreement with the expert "golden standard" data; (2) agreement with the other workers; (3) agreement with the attention check questions and (4) an amount of effort estimated from the task completion time [34]. The "golden standard" is a subset of data that is processed by experts in the field; an important condition is that a lay person should be able to process this data easily and unambitiously. The agreement-based indices target identification of outlier workers or weigh the contributions by worker's deviation from the mean [35]. The attention check and language comprehension questions are verifiable questions [29] that do not require factual knowledge [36]; the results obtained from the workers failing to answer the attention questions correctly should be discarded. Finally, the average time to complete a single task is used to identify low-quality workers presumably spending a lesser amount of time per task [34].

3. Data and Methodology

Twitter data was originally collected for a project on online discussions of climate change, and early results were covered in [2]. Software was developed to systematically poll the Twitter social networking site for the terms “climate change” and “global warming”, which resulted in over 2 million tweets collected; after filtering as described in [7], this dataset was reduced to 1.3 million georeferenced tweets. Out of this database, 600 tweets in English published within the 2012–2014 period were randomly selected for further processing.

The research design was similar to [13]. The same data were offered for processing to the volunteer workers through the Climate Tweet project based on the Citizen Scientist platform hosted at the University of North Dakota [37] and to the paid workers through the Amazon MTurk platform. To follow the best crowdsourcing practices, we used only the best paid workers, defined as those with at least 95% Human Intelligence Task (HIT) approval rating—see [33] for a detailed explanation. Note that the HIT approval rating is a worker’s work quality measure, calculated as a fraction of his/her completed tasks that were approved by requesters.

As a motivation, the volunteer workers were provided with an explanation of the scientific importance of the project; additionally, the screen names of the best workers were published on the project’s login page. The paid workers were provided with monetary compensation of $0.40 for classification of a single bundle (HIT) of 20 tweets. Taking into account the mean processing time discussed in the next section, the mean hourly earnings of a paid worker was $2.03, which is slightly above the average Amazon MTurk earning of just below $2/h: cf. mean earnings of $1.58/h for an Indian, vs. $2.30/h for a U.S. worker [14].

The quality comparisons of data produced by volunteer and paid workers were conducted along the two dimensions: expressed attitudes to the phenomenon of climate change in a processed tweet as well as topics raised. First, workers were asked to evaluate the attitude towards climate change expressed in a tweet using a 5-point scale [−2, 2]:

−2: extremely negative attitude, denial, skepticism (“*Man made GLOBAL WARMING HOAX EXPOSED*”);

−1: denying climate change (“*UN admits there has been NO global warming for the last 16 years*!”), or denying that climate change is a problem, or that it is man-made (“*Sunning on my porch in December. Global warming ain’t so bad*”);

0: neutral, unknown (“*A new article on climate change is published in a newspaper*”);

1: accepting that climate change exists, and/or is man-made, and/or can be a problem (“*How’s planet Earth doing? Take a look at the signs of climate change here*”);

2: extremely supportive of the idea of climate change (“*Global warming? It’s like earth having a Sauna!*”).

Second, the workers were asked to classify the same tweet into up to three of the following 10 topics, unified into broader themes:

- Global warming phenomenon: (1) drivers of climate change, (2) science of climate change, and (3) denial and skepticism;
- Climate change impacts: (4) extreme events, (5) unusual weather, (6) environmental changes, and (7) society and economics;
- Adaptation and mitigation: (8) politics and (9) ethical concerns, and
- (10) Unknown.

For exact questions, refer to Appendix A.

While the task formulation offered to the workers was the same on both platforms, with very similar visual survey layout, the work flow was different due to specifics of the paid and unpaid work organization and differences in the platforms. The paid workers were offered classification tasks in 20-tweet packets; for redundancy, each task was offered multiple times to different workers, so that each tweet was processed by multiple MTurk workers (min = 20, mean = 26, max = 48). Tweets were offered to volunteer workers individually, and each tweet was processed by a fewer number of workers (min = 6, mean = 14, max = 21). For further analysis, we selected only those tweets that were processed by at least nine workers on both platforms, which reduced the number of tweets from 600 to 579. The final classification was produced by the "majority consensus" method, i.e., for each tweet, its "true" classification was decided based on which topical category, or attitude, received the largest number of "votes" from the workers [31].

For groundtruthing purposes, the "Gold Standard" tweets were selected based on [34]. For this, 579 tweets were screened by the first author, who was a climate scientist, and tweets that most easily and transparently could be classified into one of the classification categories were selected (103 in total). e.g., the tweet "*What happened to global warming? It's cold as **** outside*" clearly falls into the category "denial and skepticism" with a negative attitude towards climate change. The selected 103 tweets classified by the experts will be further referred to as the "Expert processed" (E) dataset. The same tweets processed by the paid and volunteer workers will be referred to as P and V datasets, respectively.

The study, therefore, followed the best practices of research employing MTurk workers [38]: (1) utilizing workers' qualifications in task assignment; (2) creating a "Gold standard" expert-processed dataset; (3) using redundancy, and (4) using a majority consensus to adjudicate results. The abovementioned best practices (2)–(4) were also employed with respect to the data produced by volunteers; however the authors were unable to apply the best practice (1) controlling qualification of the volunteered workers.

4. Results

4.1. Descriptive Statistics

The processing of the whole pool of 579 tweets was done by 127 volunteers and 574 paid workers; on average, each volunteer processed 65 tweets, while each paid worker processed only 26. For paid workers, the mean processing time of a 20-tweet task was 11.8 min (35 s/tweet). Few raters spent a very short time per tweet (min = 5 s/tweet), indicating potential cheating behavior. We do not have processing time for the volunteer workers due to the software platform limitations.

Classification results differed for volunteer and paid workers with the former tending to classify tweets into the fewer number of topics. The matched pair two-tail t-test found significant differences between the number of categories in the V (mean = 1.64) and P (mean = 1.83) datasets ($p = 3.3 \times 10^{-30}$). We also found a better interrater agreement between the volunteer workers for all topical classifications: c.f. 75% percentage agreement for V vs. 81% for P. Note that this percentage agreement is inflated by the agreement by chance; Fleiss' generalized kappa adapted by Uebersax [39] for the unequal number of raters per subject (see also [40]) showed that in fact the interrater agreement was poor. Nevertheless, it also showed a better agreement for V raters (mean kappa = 0.24) vs. P raters (mean kappa = 0.14).

For attitude classification, the paid workers demonstrated a tendency to use extreme values of −2 and +2 more frequently than the volunteers; 25.8% of all tweets were rated as extremely positive or extremely negative, vs. only 11.1% for volunteers. Similarly to topic classification, the interrater agreement was higher for V as compared to P raters as measured by percentage agreement (86% for V vs. 78% for P) and generalized kappa (0.22 for V vs. 0.13 for P). While manually examining the P and V datasets, we observed lower quality of paid worker's classification of more difficult content. For example, the tweet "*GW is fact but Sandy is hardly proof. Poor logic . . . Sandy confirms the obvious impact of global warming ....*" was (correctly) classified as having a positive attitude towards existence of global warming by 83% of the volunteer workers vs. 56% of paid workers. Further, only one out of 17 (6%) of the volunteer raters classified the attitude as negative vs. 20% of the paid workers.

4.2. Crowdsourced vs. Expert Classification Quality

Tweet classification was validated by comparison with expert classification (dataset E of 103 tweets). We found consistently better performance from the volunteer workers, as exhibited by a higher correlation between the V and E datasets (mean r = 0.40), as compared to the P and E datasets (mean r = 0.29)—see Table 1. The difference was statistically significant ($p < 0.05$). Similarly, the mean Sørensen–Dice distance between the topic classification vectors was lower for the V vs. E datasets (0.47), as compared to the P vs. E datasets (0.36).

Table 1. Pearson's correlation between the crowdsourced volunteer (V) and paid (P) worker and expert (E) classification of the tweets. The columns represent classification of the topics 1–10 found in Section 3 and of the attitude (A).

Comparison	1	2	3	4	5	6	7	8	9	10	A
V vs. E	0.17 *	0.41 ‡	0.57 ‡	0.40 ‡	0.34 ‡	0.57 ‡	0.46	0.31 ‡	0.32 ‡	0.40 ‡	0.46 ‡
P vs. E	0.13	0.24 ‡	0.39 ‡	0.36 ‡	0.24 †	0.37 ‡	0.34 ‡	0.21 †	0.21 †	0.39 ‡	0.33 ‡

* $p < 0.05$; † $p < 0.01$; ‡ $p < 0.05$.

The majority consensus method to extract the "true" classification from redundant ratings provided equally high quality results for both paid and volunteer workers, with an accuracy (fraction of matches with the E dataset) of ~0.8 for topic, and ~0.7 for sentiment classification (Table 2). The acceptable "realistic" agreement between human coders as measured by an accuracy, coefficient may vary between 0.70 and 0.79 [41], as evidenced by e.g., an Amazon MTurks data analysis [42]. Overall, we conclude that the lower work quality of an average individual paid worker is mitigated by quality control based on massive redundancy, so that using volunteer workers has no data quality benefits over paid workers.

Table 2. Matching expert (E) and majority consensus of volunteer (V) and paid (P) worker classifications (classification accuracy) for the full dataset and the subsample used for groundtruthing. Note higher redundancy rate for P workers as compared to V workers (every tweet was independently processed by 26 and 14 workers on average, respectively). Refer to Figure 1 to compare classification accuracy for the same redundancy rate.

Comparison	Matching Topics	Matching Attitudes	Opposite Attitudes
V vs. P (full dataset)	0.73	0.65	0.01
V vs. P (groundtruthing dataset)	0.75	0.68	0.05
V vs. E (groundtruthing dataset)	0.80	0.70	0.04
P vs. E (groundtruthing dataset)	0.79	0.67	0.03

To estimate the effect of crowdsourcing redundancy on classification quality, we repeatedly reduced the redundancy level in V and P datasets by limiting the maximum number of classifications of a single tweet. The maximum redundancy level for tweets was reduced from 19 for the V dataset and 30 for the P dataset, down to zero. In effect, this emulated the designs in which each tweet was analyzed by a regressive number of workers. To estimate uncertainty arising from a variability in the workers' quality, we performed 10 permutations, each time removing a respective number of randomly selected classifications. The results (Figure 1) showed the quality of majority consensus classification falling faster for the paid as opposed to volunteer workers: e.g., a 70% match between a crowdsourced and expert classification was on average achieved by 12 paid workers vs. just four volunteers.

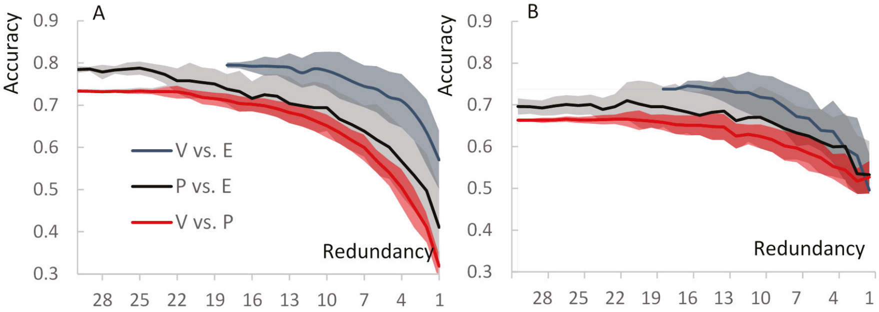


Figure 1. A fraction of matching classifications of tweets' topics (**A**) and attitude (**B**) as a function of crowdsourcing redundancy. The expert (E), majority consensus volunteer (V), and paid (P) worker datasets are being compared. Arial boundaries show the best and the worst estimates and solid lines show the mean estimates (see the text for explanation).

4.3. Geographical Variability

We used each worker's computer internet protocol (IP) address to determine the worker's country of residence. For paid and volunteer crowdsourcing alike, the majority of workers and the majority of completed tasks originated from the U.S., but the overall geographical distributions were very dissimilar (Table 3). Almost 95% of all paid workers and over 95% of their completed tasks came from just two countries, the U.S. and India, with the next country, the U.K., contributing to less than 0.5% of completed tasks. As opposed to that, 95% of volunteer workers came from 16 countries, and 95% of tasks were completed in eight countries. The highest percentage of completed tasks came from the U.S. (64%), followed by the U.K. (13%), with India representing only 1% of completed tasks (Figure 2).

Table 3. Geographical distribution of volunteer (N_v = 127) and paid (N_p = 574) workers and their completed tasks (N_v = 8198 and N_p = 14860) as a percentage of the total. The top 10 countries included into the table represent 78% of volunteer workers and 97% of their completed tasks. For paid workers, the table represents 97% of workers and 97% of completed tasks.

Country	Volunteer Workers		Paid Workers	
	Tasks	Raters	Tasks	Raters
U.S.	64.4	17.1	75.7	76.4
U.K.	13.0	12.4	0.4	0.5
Australia	6.2	10.6	0.1	0.2
Canada	5.6	8.8	0.3	0.3
Indonesia	2.9	7.1	0.0	0.2
Germany	1.2	4.9	0.0	0.2
Ireland	1.1	5.4	0.0	0.2
India	1.0	4.2	20.6	18.2
France	0.8	3.9	0.0	0.2
Brazil	0.7	3.6	0.0	0.2

In terms of data quality, we found significant geographical differences in the P dataset (Table 4), e.g., the crowdsourced topic classification matched the expert one in 80% of the U.S. subsample, but only in 22% of the India subsample. Interestingly, we did not find a similar effect for the V dataset (Table 4). Manual examination of data originating from the IPs located in India showed multiple misclassifications. For example, a tweet "*Global warming is a lie!!! Proof: Step outside!!! Brrrrr!*" was mapped as a climate change impact on weather, on environment and on society. Similarly, a tweet "*The End of an Illusion or no global warming* . . . " was misclassified as global warming drivers, science and impacts on weather. On average, one tweet was classified into 2.2 categories in the India subset vs. 1.78 categories in the U.S. subset.

Exclusion of the rating from the IPs originating in India provided an improvement in classification quality and, consequently, allows to greatly reduce redundancy level (Figure 3). For example, when the India subset was excluded from the data, reducing the dataset by 21%, a 70% match between a crowdsourced and expert classification, was achieved on average by six paid workers vs. 12 paid workers required for the entire dataset.

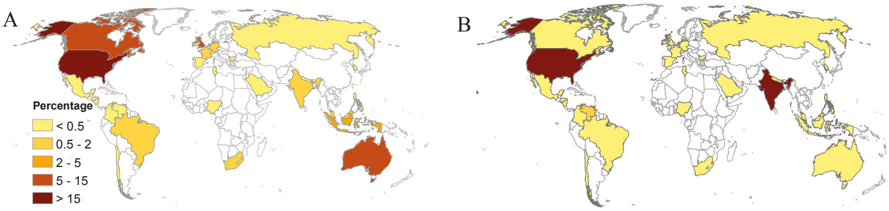


Figure 2. Percentage of classified tweets for the volunteered (**A**) and paid (**B**) workers.

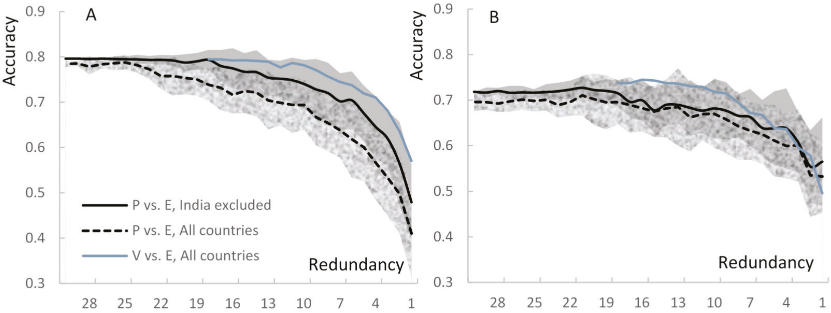


Figure 3. A fraction of matching classifications of tweets' topics (**A**) and attitude (**B**) as a function of crowdsourcing redundancy. The entire expert (E), majority consensus volunteer (V), and paid (P) worker datasets are being compared to subsets of data that excludes India (see the text for explanation).

Table 4. Fraction of matching majority consensus volunteer (V) and paid (P) worker classifications with an expert classification, the two top paid crowdsourcing countries (U.S. and India), and for all other countries. The India V sample for volunteer classification was too small (<1%) to allow comparisons. The individual samples for countries other than India and the US are too small to allow a comparison.

Comparison	Matching Topics		Matching Attitude		Opposite Attitudes	
	V	P	V	P	V	P
U.S.	0.83	0.80	0.72	0.74	0.03	0.05
India		0.22		0.54		0.15
Other countries	0.72	0.47	0.78	0.47	0.03	0.11

5. Discussion

Amazon MTurk best practice guide [38] recommends redundant data processing as a tool to improve the accuracy of the obtained results. However, employing massively redundant research design is costly; therefore, some researchers used the majority consensus method with as few as three redundant ratings. Snow et al. [43] found that the expert-quality evaluation is already achieved with

N = 4 classifications per item. However, we found that for harder-to-process questions related to science, much higher redundancy (N $\gg$ 10) is required for paid workers.

Despite the fact that only the best quality Amazon MTurk workers were selected (HIT approval rating $\geq$ 95%), the performance of the paid workers was still inferior to the performance of volunteers. Consequently, we found that for a particular task, the same accuracy level can be achieved with 12 paid workers as with only four volunteers. The associated cost increase may be prohibitive for many scientific projects, which makes volunteer crowdsourcing an attractive alternative. The downside of volunteer crowdsourcing is that it requires a much longer time to complete the project. In our case, Amazon MTurk processing was completed in five days, with most of the time taken up with validation of the already processed data. The Citizen Scientist platform processing took one year; on average, ~600 tweets per month were processed. We also found that an interaction between the scientists and volunteers was required to keep the public interested in donating their time to the project.

To extract "true" classification from the large redundant pool returned by the crowdsourced workers, we used the most popular and simple method of majority vote. Multiple algorithms have been proposed to reduce the "noise" originating from workers' inaccuracy e.g., [44,45] and others. Application of data cleaning methods based on assigning dataset-specific quality rating to each of the paid workers helps to reduce required redundancy. For example, Dawid and Skene [35] suggest that workers should be weighted based on the deviation of their scores from the mean; the contribution from low quality workers should then be discarded or used with a lesser weight. Ipeirotis et al. [46] demonstrate that separation of workers' error rates into true errors and systematic bias leads to significant improvement of classification, and suggests that as far as each worker processes a large number of assignments (at least 20), the redundancy can be kept to at five iterations per task without significant quality deterioration. In practice, we found that the majority (90%) of the paid workers accepted just one or two task bundles (20–40 tweets), which made these quality control methods marginally applicable. This difference in the number of tweets classified by the paid workers and volunteers may also partially explain the difference in work quality. Indeed, assuming that a higher number of samples processed by a worker leads to better training and hence better quality on the subsequent tasks, volunteer workers would outperform paid ones.

The task completion time presumably measures each worker's thoughtfulness and hence may be another measure of work quality; indeed, Snow et al. [43] found that the per-hour pay encouraged the workers to spend twice as much time processing each task, and returned more accurate results as opposed to the per-task pay. However, we did not find a significant correlation between the task completion time (min 101 s, median 571 s, mean 708 s, max 6093 s) and accuracy. We also noticed that the performance of the fastest and the slowest workers tended to be poor.

Another quality management strategy is to utilize a worker reputation system to employ only the workers with approval ratings above a certain pre-set level [33]; commonly, a 90–95% rating is used. We, however, speculate that workers' reputation might not be a very reliable indicator of their performance. Proliferation of the online rating system means that the workers have become highly motivated in the protection of their online reputation. In a handful of cases, we had to reject incomplete tasks; subsequently, we received complaints and threats to blacklist us as bad requesters. Given the time and effort required to follow up requests from unsatisfied workers and a low cost of individual tasks, there is a strong incentive to avoid a dispute and comply with workers' requests, which thus artificially boosts the approval ratings of workers.

The workers participating in our study were on average earning ~$2/h, which is similar to average MTurk earnings. It is possible that a higher pay rate would return better quality results; however, Gillick and Liu [47] hypothesized that lower compensation might attract the workers less interested in monetary rewards and hence spend more time per task. Having read the online discussion of the MTurk workers, we also noticed that they associate an unusually high pay rate with possible fraud and recommend abstaining from taking such HITs.

In our study, similarly to other research [48], the overwhelming majority (95%) of paid workers came from the U.S. and India. This is not surprising, since the Amazon MTurk workers from other countries are unable to transfer their earnings to a bank account [49]. We found that discarding results from workers outside the U.S. significantly improved data quality and hence reduced the required redundancy of the design; we did not find a similar effect for the volunteer workers. Note that the geographical distribution of the volunteer and paid workers was very different; the volunteer workers came predominantly from the countries with an active public discussion of climate change on Twitter and a high level of Twitter penetration. For example, the daily number of English language tweets originating from the U.S. is ~30 times higher than those for India, but this number is only three times higher than those from the U.K. [2]. We therefore speculate that the main reason for the low quality of India data was insufficient familiarity of the workers with climate change discourse in general. Consequently, geographical worker selection may be an important factor to consider in order to improve the quality of results.

6. Conclusions

The purpose of this research was to compare the quality of volunteer and paid workers' classification of Twitter messages on climate change. We found lower accuracy of data returned by paid crowdsourced workers as compared with volunteer workers, while the latter required significantly longer time to complete. Consequently, a similar accuracy of processed data was achieved with paid workers only with a higher design redundancy; this caused expenses to be high. While conventional methods of accuracy improvement were largely unsuccessful due to the long-tail distribution of processed tasks per worker, limiting the workers' pool to those located in the U.S. significantly improved paid workers' data quality, making it only slightly lower than the volunteers' performance. Therefore, geographical location is an important factor for worker selection. We suggest the consideration of limiting the workers' pool to those countries where the research topic is actively discussed by the public in study designs.

The study has several limitations that might have an impact on its generalizability. While climate change is a world-wide discussed issue, the framing of its various topical aspects could differ depending on the country, thus potentially affecting classifications by the raters. At the same time, it is speculated that topical aspects with little differences in framing could yield lesser geographical differences in processing quality. Another limitation concerns the usage of the simplest, but also the most common "majority filter" for error correction; more advanced methods of error correction might return more precise results. Finally, despite our efforts to make the online interface for paid and volunteer workers as similar as possible, the differences in technical configuration between crowdsourcing platforms prevented us from designing a completely identical interface for the two web sites. These limitations should be addressed in further research.

Acknowledgments: Publication of this article was funded by the University of Florida Open Access Publishing Fund. The data used in this study were collected by the University of North Dakota Center for Regional Climate Studies supported by the NSF ND EPSCoR grant.

Author Contributions: A.P.K. and S.S. conceived and designed the study and wrote the paper; A.P.K. analyzed the data; T.D. and H.K. organized the volunteered and paid crowdsourcing processing, accordingly.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Coding Instructions

Note: Assume that "global warming" and "climate change" (further—CC) are synonyms.

Note: examples are written in *Italics*

Read a tweet. What is the attitude expressed towards CC? Use the −2 −1 0 1 2 codes:

0: neutral, unknown (*A new article on CC is published in a newspaper*) (*He talked about CC*)

1: accepting that CC exists and/or is man-made and/or can be a problem (*How's planet Earth doing? Take a look at the signs of climate change here*)

2: extremely supportive of the idea of CC (*Global warming? It's like earth having a Sauna!!*). Think of code 2 as though it is code 1 plus a strong emotional component and/or a call for action

−1: denying CC (*UN admits there has been NO global warming for the last 16 years*!) or denying that CC is a problem or that it is man-made (*Sunning on my porch in December. Global warming ain't so bad*.)

−2: extremely negative attitude, denial, skepticism ("*Climate change*" *LOL*) (*Man made GLOBAL WARMING HOAX EXPOSED*). Think of code −2 as though it is code −1 plus a strong emotional component.

Classify each tweet using ten categories below. If you think that a tweet belongs to multiple categories, you may use **up to three categories**. If you cannot find any suitable category, leave the cells empty. The categories are in bold.

Categories of GLOBAL WARMING PHENOMENON

1. **Drivers** of CC. Examples:
   - *Greenhouse gases (Carbon Dioxide, Methane, Nitrous Oxide*, etc.*)*
   - *Oil, gas, and coal*
2. **Science**. Examples:
   - *The scientists found that climate is in fact cooling*
   - *IPCC said that the temperature will be up by 4 degrees C*
3. **Denial**, skepticism, Conspiracy Theory. Examples:
   - *Scientists are lying to the public*

Categories of IMPACTS OF CLIMATE CHANGE

4. **Extreme** events. Examples:
   - *Hurricane Sandy, flooding, snowstorm*
5. **Weather** is unusual. Examples:
   - *Hot or cold weather*
   - *Too wet or too dry*
   - *Heavy Snowfall*
6. **Environment**. Examples:
   - *Acid rain, smog, pollution*
   - *Deforestation, coral reef bleaching*
   - *Pests, infections, wildfires*
7. **Society** and Economics. Examples:
   - *Agriculture is threatened*
   - *Sea rising will threaten small island nations*
   - *Poor people are at risk*
   - *Property loss, Insurance*

Categories of ADAPTATION AND MITIGATION

8. **Politics**. Examples:

    - *Conservatives, liberals, elections*
    - *Carbon tax; It is too expensive to control CC*
    - *Treaties, Kyoto Protocol, WTO, UN, UNEP*

9. **Ethics**, moral, responsibility. Examples:

    - *We need to fight Global Warming*
    - *We need to give this planet to the next generation*
    - *God gave us the planet to take care of*

UNKNOWN category

10. **Unknown**, jokes, irrelevant, hard to classify. Examples:

    - *Global warming is cool OMG a paradox*
    - *This guy is so hot its global warming*

References

1. Leiserowitz, A.; Maibach, E.W.; Roser-Renouf, C.; Rosenthal, S.; Cutler, M. *Climate Change in the American Mind: May 2017*; Yale Program on Climate Change Communication; Yale University and George Mason University: New Haven, CT, USA, 2017.
2. Kirilenko, A.P.; Stepchenkova, S.O. Public microblogging on climate change: One year of Twitter worldwide. *Glob. Environ. Chang.* **2014**, *26*, 171–182. [CrossRef]
3. Cody, E.M.; Reagan, A.J.; Mitchell, L.; Dodds, P.S.; Danforth, C.M. Climate Change Sentiment on Twitter: An Unsolicited Public Opinion Poll. *PLoS ONE* **2015**, *10*, e0136092. [CrossRef] [PubMed]
4. Yang, W.; Mu, L.; Shen, Y. Effect of climate and seasonality on depressed mood among twitter users. *Appl. Geogr.* **2015**, *63*, 184–191. [CrossRef]
5. Holmberg, K.; Hellsten, I. Gender differences in the climate change communication on Twitter. *Int. Res.* **2015**, *25*, 811–828. [CrossRef]
6. Leas, E.C.; Althouse, B.M.; Dredze, M.; Obradovich, N.; Fowler, J.H.; Noar, S.M.; Allem, J.-P.; Ayers, J.W. Big Data Sensors of Organic Advocacy: The Case of Leonardo DiCaprio and Climate Change. *PLoS ONE* **2016**, *11*, e0159885. [CrossRef] [PubMed]
7. Kirilenko, A.P.; Molodtsova, T.; Stepchenkova, S.O. People as sensors: Mass media and local temperature influence climate change discussion on Twitter. *Glob. Environ. Chang.* **2015**, *30*, 92–100. [CrossRef]
8. Sisco, M.; Bosetti, V.; Weber, E. When do extreme weather events generate attention to climate change? *Clim. Chang.* **2017**, *143*, 227–241. [CrossRef]
9. Howe, J. The rise of crowdsourcing. *Wired Mag.* **2006**, *14*, 1–4.
10. Clery, D. Galaxy Zoo volunteers share pain and glory of research. *Science* **2011**, *333*, 173–175. [CrossRef] [PubMed]
11. Galaxy Zoo. Available online: https://www.galaxyzoo.org/ (accessed on 25 December 2016).
12. Lintott, C.; Schawinski, K.; Bamford, S.; Slosar, A.; Land, K.; Thomas, D.; Edmondson, E.; Masters, K.; Nichol, R.C.; Raddick, M.J.; et al. Galaxy Zoo 1: Data release of morphological classifications for nearly 900,000 galaxies. *Mon. Not. R. Astron. Soc.* **2011**, *410*, 166–178. [CrossRef]
13. Mao, A.; Kamar, E.; Chen, Y.; Horvitz, E.; Schwamb, M.E.; Lintott, C.J.; Smith, A.M. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing, Palm Springs, CA, USA, 7–9 November 2013.
14. Ross, J.; Irani, L.; Silberman, M.; Zaldivar, A.; Tomlinson, B. Who are the crowdworkers? Shifting demographics in mechanical Turk. In Proceedings of the CHI'10 Extended Abstracts on Human Factors in Computing Systems, Atlanta, GA, USA, 10–15 April 2001; ACM: New York, NY, USA, 2010; pp. 2863–2872.

15. Redi, J.; Povoa, I. Crowdsourcing for Rating Image Aesthetic Appeal: Better a Paid or a Volunteer Crowd? In Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia, Orlando, FL, USA, 7 November 2014; ACM: New York, NY, USA, 2014; pp. 25–30.
16. Muller, C.L.; Chapman, L.; Johnston, S.; Kidd, C.; Illingworth, S.; Foody, G.; Overeem, A.; Leigh, R.R. Crowdsourcing for climate and atmospheric sciences: Current status and future potential. *Int. J. Climatol.* **2015**, *35*, 3185–3203. [CrossRef]
17. Olteanu, A.; Castillo, C.; Diakopoulos, N.; Aberer, K. Comparing Events Coverage in Online News and Social Media: The Case of Climate Change. In Proceedings of the Ninth International AAAI Conference on Web and Social Media, Oxford, UK, 26–29 May 2015.
18. Samsel, F.; Klaassen, S.; Petersen, M.; Turton, T.L.; Abram, G.; Rogers, D.H.; Ahrens, J. Interactive Colormapping: Enabling Multiple Data Range and Detailed Views of Ocean Salinity. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA'16), San Jose, CA, USA, 7–12 May 2016; ACM: New York, NY, USA, 2016; pp. 700–709.
19. Yzaguirre, A.; Warren, R.; Smit, M. Detecting Environmental Disasters in Digital News Archives. In Proceedings of the 2015 IEEE International Conference on Big Data, Santa Clara, CA, USA, 29 October–1 November 2015; pp. 2027–2035.
20. Ranney, M.A.; Clark, D. Climate Change Conceptual Change: Scientific Information Can Transform Attitudes. *Top. Cogn. Sci.* **2016**, *8*, 49–75. [CrossRef] [PubMed]
21. Attari, S.Z. Perceptions of water use. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 5129–5134. [CrossRef] [PubMed]
22. Vukovic, M. Crowdsourcing for Enterprises. In Proceedings of the 2009 Congress on Services-I, Los Angeles, CA, USA, 6–10 July 2009; pp. 686–692.
23. Overview of Mechanical Turk—Amazon Mechanical Turk. Available online: http://docs.aws.amazon.com/AWSMechTurk/latest/RequesterUI/OverviewofMturk.html (accessed on 28 December 2016).
24. Mason, W.; Suri, S. Conducting behavioral research on Amazon's Mechanical Turk. *Behav. Res. Methods* **2012**, *44*, 1–23. [CrossRef] [PubMed]
25. Staffelbach, M.; Sempolinski, P.; Kijewski-Correa, T.; Thain, D.; Wei, D.; Kareem, A.; Madey, G. Lessons Learned from Crowdsourcing Complex Engineering Tasks. *PLoS ONE* **2015**, *10*, e0134978. [CrossRef] [PubMed]
26. Kawrykow, A.; Roumanis, G.; Kam, A.; Kwak, D.; Leung, C.; Wu, C.; Zarour, E.; Sarmenta, L.; Blanchette, M.; Waldispühl, J.; et al. Phylo: A citizen science approach for improving multiple sequence alignment. *PLoS ONE* **2012**, *7*, e31362. [CrossRef] [PubMed]
27. Poetz, M.K.; Schreier, M. The value of crowdsourcing: can users really compete with professionals in generating new product ideas? *J. Prod. Innov. Manag.* **2012**, *29*, 245–256. [CrossRef]
28. Chandler, J.; Paolacci, G.; Mueller, P. Risks and rewards of crowdsourcing marketplaces. In *Handbook of Human Computation*; Springer: New York, NY, USA, 2013; pp. 377–392.
29. Kittur, A.; Chi, E.H.; Suh, B. Crowdsourcing User Studies with Mechanical Turk. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy, 5–10 April 2008; ACM: New York, NY, USA, 2008; pp. 453–456.
30. Raddick, M.J.; Bracey, G.; Gay, P.L.; Lintott, C.J.; Cardamone, C.; Murray, P.; Schawinski, K.; Szalay, A.S.; Vandenberg, J. Galaxy Zoo: Motivations of Citizen Scientists. Available online: http://arxiv.org/ftp/arxiv/papers/1303/1303.6886.pdf (accessed on 27 October 2017).
31. Allahbakhsh, M.; Benatallah, B.; Ignjatovic, A.; Motahari-Nezhad, H.R.; Bertino, E.; Dustdar, S. Quality control in crowdsourcing systems. *IEEE Int. Comput.* **2013**, *17*, 76–81. [CrossRef]
32. Rouse, S.V. A reliability analysis of Mechanical Turk data. *Comp. Hum. Behav.* **2015**, *43*, 304–307. [CrossRef]
33. Peer, E.; Vosgerau, J.; Acquisti, A. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behav. Res. Methods* **2014**, *46*, 1023–1031. [CrossRef] [PubMed]
34. Eickhoff, C.; de Vries, A.P. Increasing cheat robustness of crowdsourcing tasks. *Inf. Retr.* **2013**, *16*, 121–137. [CrossRef]
35. Dawid, A.P.; Skene, A.M. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Appl. Stat.* **1979**, *28*, 20–28. [CrossRef]
36. Goodman, J.K.; Cryder, C.E.; Cheema, A. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *J. Behav. Decis. Mak.* **2013**, *26*, 213–224. [CrossRef]
37. Climate Tweets. Available online: http://csgrid.org/csg/climate/ (accessed on 25 December 2016).

38. Amazon Mechanical Turk Requester Best Practices Guide. Available online: https://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf (accessed on 29 December 2016).
39. Uebersax, J.S. A design-independent method for measuring the reliability of psychiatric diagnosis. *J. Psychiatr. Res.* **1982**, *17*, 335–342. [CrossRef]
40. Gwet, K.L. *Handbook of Inter-Rater Reliability. The Definitive Guide to Measuring the Extent of Agreement among Raters*, 4th ed.; Advanced Analytics, LLC: Gaithersburg, MD, USA, 2014.
41. Donkor, B. Sentiment Analysis: Why It's Never 100% Accurate. 2014. Available online: https://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf (accessed on 29 December 2016).
42. Ogneva, M. How companies can use sentiment analysis to improve their business. Mashable, 19 April 2010. Available online: https://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf (accessed on 29 December 2016).
43. Snow, R.; O'Connor, B.; Jurafsky, D.; Ng, A.Y. Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; Association for Computational Linguistics: Stroudsburg, PA, USA, 2008; pp. 254–263.
44. Welinder, P.; Branson, S.; Perona, P.; Belongie, S.J. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*; NIPS: Vancouver, BC, Canada, 2010; pp. 2424–2432.
45. Whitehill, J.; Wu, T.; Bergsma, J.; Movellan, J.R.; Ruvolo, P.L. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*; NIPS: Vancouver, BC, Canada, 2009; pp. 2035–2043.
46. Ipeirotis, P.G.; Provost, F.; Wang, J. Quality Management on Amazon Mechanical Turk. In Proceedings of the ACM SIGKDD Workshop on Human Computation, Washington, DC, USA, 25 July 2010; ACM: New York, NY, USA, 2010; pp. 64–67.
47. Gillick, D.; Liu, Y. Non-Expert Evaluation of Summarization Systems is Risky. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, CA, USA, 6 June 2010; Association for Computational Linguistics: Stroudsburg, PA, USA, 2010; pp. 148–151.
48. Paolacci, G.; Chandler, J.; Ipeirotis, P.G. Running experiments on amazon mechanical Turk. *Judgm. Decis. Mak.* **2010**, *5*, 411–419.
49. Amazon Mechanical Turk. Available online: https://www.mturk.com/mturk/help?helpPage=worker#how_paid (accessed on 30 December 2016).

sustainability

MDPI

Article

Developing a Methodology of Structuring and Layering Technological Information in Patent Documents through Natural Language Processing

Taeyeoun Roh, Yujin Jeong and Byungun Yoon *

Department of Industrial & Systems Engineering, School of Engineering, Dongguk University, 26, Pil-dong 3-ga, Chung-gu, Seoul 100-715, Korea; arongi1320@naver.com (T.R.); withss501@naver.com (Y.J.)
* Correspondence: postman3@dongguk.edu

Received: 15 September 2017; Accepted: 13 November 2017; Published: 17 November 2017

Abstract: Since patents contain various types of objective technological information, they are used to identify the characteristics of technology fields. Text mining in patent analysis is employed in various fields such as trend analysis and technology classification, and knowledge flow among technologies. However, since keyword-based text mining has the limitation whereby, when screening useful keywords, it frequently omits meaningful keywords, analyzers therefore need to repeat the careful scrutiny of the derived keywords to clarify the meaning of keywords. In this research, we structure meaningful keyword sets related to technological information from patent documents; then we layer the keywords, depending on the level of information. This research involves two steps. First, the characteristics of technological information are analyzed by reviewing the patent law and investigating the description of patent documents. Second, the technological information is structured by considering the information types, and the keywords in each type are layered through natural language processing. Consequently, the structured and layered keyword set does not omit useful keywords and the analyzer can easily understand the meaning of each keyword.

Keywords: text mining; NLP; technological information; patent analysis; text structure

1. Introduction

A patent is objective and proven technological information, which contains one or more unique technical features that cannot be duplicated in other patents. Patents have been used as essential information for effective management strategies [1,2]. Structured data such as the issue date and the number of citations, as well as unstructured data such as summaries and claims included in the patent, can be useful in analyzing competitive markets and technology trends [3]. In many studies on structured data, the features or trends of technology development have been obtained by using patent search and international patent classification (IPC) [4–6]. Currently, considerable research is underway that includes unstructured data rather than only structured data to produce more meaningful results. In order to analyze a large number of patents, it is necessary to analyze the contents of unstructured information such as titles, summaries, and claims [7]. In the most commonly used keyword-oriented analysis, not all keywords are extracted, but keywords are analyzed according to specific criteria [4] and the linkage between technological fields [8].

Keyword-oriented text mining cannot accurately convey the meanings of individual words. Therefore, analysis on tagging parts of speech through natural language processing (NLP) is utilized to recognize information at the sentence level rather than at the keyword level. For example, subject–action–object (SAO) methodology is a representative method that analyzes the subject, verb, and object in a sentence as one structure. The SAO methodology interprets an object as a problem, and the subject and action as a way to solve problems. The SAO methodology can derive more detailed meaning by reflecting the semantic relations among keywords compared to the traditional keyword analysis methods. Using SAO methodology, patent analysis has been conducted from a more semantic point of view, such as for patent infringement and similarity of technology [9] and technology roadmaps [10].

The text described in the patent document must contain various types of technological information, such as functions, components, and operating methods, in accordance with the patent law. However, the existing keyword-based analysis and the SAO methodology analyze the structure of extracted keywords based on the frequency of occurrence and part-of-speech, respectively. Thus, various types of technological information pertinent to the patent are often missing from the documents, and a secondary analysis is necessary. In addition, the existing keyword-based analysis and the SAO methodology have limitations since the inherent information of each patent such as the functions, components, and operating methods cannot be deduced.

Therefore, the purpose of this study is to derive a set of keywords that contain technological information of patented texts without omission in order to overcome the limitations of the text mining used in existing patent analysis. To achieve this purpose, in this research, information about phrases that form parts of speech is used, in addition to the dictionary meaning of words through Natural Language Processing (NLP). The technological information included in the patent exists in a formulated description form based on the part-of-speech, and the sentences described in the title, summary, and claims can be structured according to the type of technological information. In addition, words included in phrases, and words modifying phrases can be identified by using information on phrases. Keywords can then be selected according to the quality of information using the importance of words in a sentence rather than the part-of-speech and frequency. The keywords are structured according to the type of technological information and it is possible to understand the meaning of the keywords without secondary interpretation of the keyword set by using the hierarchical keyword set based on the importance of the word. Various levels of information can be selectively extracted from the detailed information. Since the keyword set obtained through this study can be used to analyze a large number of patents without further use of expert opinion or searching the technical field, this methodology can support the process of searching new fields for technology development. In addition, since the technological information of the keyword set is presented for each type, it is possible to classify the technology without additional clustering in terms of the function and the component.

In Section 2, we discuss the limitations of the existing methods through a literature review on existing patent analysis and text mining. We also present the specific research objectives and explain NLP, a key technology for achieving the objectives of this research. In Section 3, we present the basic concept and the detailed process of the proposed approach. Section 4 shows the application of the proposed methodology to a real case, the user interface field, to derive and verify the results. In Section 5, the limitations and areas for future research are discussed.

2. Background

2.1. Patent Analysis

Patent analysis is performed to understand the nature of the technology and industry such as detailed properties of technology and industrial trends [11]. In addition, patent analysis can extract various information in the patent, through classification, visualization, and clustering analysis [7]. The information contained in the patent can be divided into structured information and unstructured information. Structured information includes the information quantified in the patent database such as patent number, registration date, number of citation, and number of claims. In addition, the technical classification codes defined differently for each country are included to indicate the technical field of a patent. Unstructured information includes information described in the text such as the title, summary, and claims.

Lee et al. [12] and Altuntas et al. [8] analyzed the technology convergence, innovation, and relationship among technologies by investigating the citation network between technology fields. Jeong et al. [4] and Su [13] presented the trends of technology development among structured information by analyzing the number of patents registered each year. Based on IPC codes, Kim [14] discovered core technology in environmental ecology based on data envelopment analysis (DEA) and association rule mining. In addition, Kang et al. [15] proposed a convergence index to explore promising convergence technologies using structured information. Yun [5], Lim et al. [6] and Niemann et al. [16] evaluated the importance of the technology and patenting patterns by using the number of citations.

Unstructured information mainly focuses on the analysis of text information by converting text information into quantitative information using text mining. Noh et al. [17] selected keywords that have the highest text mining efficiency among the titles, abstracts, and claims. Huang et al. [18] and Lee et al. [19] selected the important keywords shown in the summary and claims, and searched for patents and technologies with high similarity using IPC codes. Lee et al. [20] evaluated the novelty of patents by using the similarity between major keywords. Ko et al. [21] analyzed the degree of technology convergence in the technology field. Lee and Sohn [22] identified shale gas development by analyzing the abstracts of patents.

As mentioned above, existing studies have utilized structured information and unstructured information to interpret various types of technological information. However, most of the studies analyze patents by integrating keywords that are shown in terms of technology, and the keywords are not interpreted in terms of the respective patents. In addition, although the patent has specific technological information related to features, functions, methods, components etc., it might only reflect certain areas of configuration at the keyword level. Thus, in such approaches, a secondary analysis should be performed in order to clearly present the meaning of keywords.

In order to overcome the limitations of existing research, this study defines the type of technological information in advance and extracts the unique information of each patent by structuring the information according to the type of technological information. We then propose a method that can interpret both the patent information of one patent as well as the technological field.

2.2. Text Mining

Text mining extracts meaningful information from unstructured data, and is utilized in many research fields because it can express a large amount of text. Text mining can be divided into keyword-based analysis and word-based analysis [23]. Keyword-based analysis methods are based on the frequency of occurrence, such as the method of using the Term Frequency-Inverse Document Frequency (*TF-IDF*) value which is an index to judge word importance in the document, Latent Semantic Analysis (LSA), and Latent Dirichlet Allocation (LDA) [24]. SAO is well known as a representative method in the word-based analysis method.

The method of using the occurrence frequency involves determining the number of times that the keyword appears in the analysis target [25]. Since the *TF-IDF* value is calculated by using the number of keywords appearing in the document, it is the relative value of the degree of importance of words. In the method of using the occurrence frequency and the *TF-IDF* value, parts of the keywords are selected and analyzed by evaluating the importance of the keyword without using all the keywords appearing in the document. Kim et al. [26] and Min et al. [27] selected future promising areas by analyzing the time series of words appearing in papers, news, and policy research reports. Choi et al. [2] predicted promising technologies by investigating the network of keywords. Kim et al. [28] used this approach to create a patent development map in technology fields using the patent keywords.

LSA is a method used to understand the linkage relationship between unknown documents based on the occurrence frequency of keywords. A matrix of the occurrence frequency of keywords in each document is constructed and the similarities between documents are compared, using singular value decomposition. The advantage of this method is that multiple keywords other than a single keyword can be compared together and their latent relationship is discovered. Ghazizadeh et al. [29] proposed a future service direction by clustering users' complaints through LSA. LDA is an analytical method that can derive topics from texts and extract keywords corresponding to each topic [30]. Since the relation between topics and keywords can be derived using the probability of the keywords for each subject based on the Dirichlet probability distribution, it is possible to derive a set of keywords that reflects the meaning contained in the word in comparison with other text mining methodologies. Based on the topic modeling methodology, Kwon et al. [31] analyzed social impacts of emerging technology using LSA. Park and Song [32] used the abstract of a paper to understand the research trends based on LDA. Jang et al. [33] discovered technology opportunity in heterogeneous technology fields using LDA. Jin et al. [34] used Twitter data to select topics and search for issue changes through the network analysis of thematic keywords. Gao and Eldin [35] interpreted topics as one cluster and derived independent meaning for each topic using only the relationship between words in the topic. Furthermore, Guo et al. [36] compared various topic modeling methods and detected user interests in microblog platform.

SAO analysis interprets subject-action-object as a structure, in contrast to analyzing the meaning of one word in existing text mining. In this case, the verb and object are interpreted as a way to solve the subject, and are analyzed by focusing on the function of each word. Park et al. [9] analyzed patent infringement and technology similarity, and Yoon and Kim [23] selected the promising technology and applied it to technical planning. Wang et al. [24] utilized the SAO structure to construct the seven layers of the technology roadmap.

In order to increase the accuracy of the analysis, various studies are conducted to extract more meaningful and accurate keywords. Lee and Kim [37] and Rose et al. [38] suggested keyword extracting methodology based on *TF-IDF* index and the number of frequency. Hulth [39] used natural language processing to extract keyword based on the linguistic meaning of the words.

However, previous text mining based on keyword extraction has four limitations. First, many meaningful keywords are excluded without considering the meaning and function of each word. This process does not reflect the characteristics of patents that have different independent information, as shown in Figure 1. Second, in order to analyze the results of a large amount of patent analysis, further analysis is required to more specifically confirm the keyword and the patent. Therefore, this approach is limited since a large amount of patent analysis hinders the ability to make efficient and quick decisions. Third, keyword extracting methodology based on topic modeling has another problem. Since they are aimed at constructing a group of words and analyzing the words in each group, they are not aimed at the meaning of each word itself. Fourth, since SAO analysis extracting constructed information by noun and verb, other parts of speech cannot be extracted.

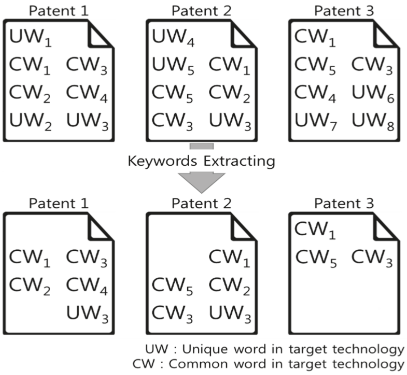


Figure 1. Errors in extracting keyword.

In order to overcome this problem, in this study, we classify the technological information of the patent according to the description form, and extract the information after structuring the information of the patent terms of technological information. This makes it possible to extract the technological information of the patent without omission of useful keyword and specific part of speech. In addition, meaning of extracted keyword in one patent is clarified, showing which technological information it contained. Furthermore, groups of the patent keywords can be representing technology fields as previous approaches do.

2.3. Natural Language Processing (NLP)

Natural Language Processing (NLP) refers to the process of converting a natural language described by humans into a machine language understood by a computer. Besides the dictionary meaning of the words in the sentence, NLP provides various types of information through the existing learning or stored algorithm. Among these types of information, part of speech (POS), which tags the part of speech of each word, can be used to check the grammatical meaning of each word in the sentence, and can decompose a sentence into several phrases and clauses. In addition, one complicated sentence can be decomposed into plural complete sentences according to a predefined sentence decomposition standard.

NLP has mainly been used to classify sentence patterns based on parts of speech or to extract sentence features. Nasukawa and Yi [40] and Yi et al. [41] proposed an algorithm to judge positive and negative statements, and Jin and Xiong [42] suggested an algorithm to classify the types of sentences and translate Chinese into English. Yang and Seo [43] classified the type of claims on the basis of the part of speech and extracted the information on the keyword type.

In this study, not only the meanings of keywords but also the information on phrases are utilized through NLP. In addition, this research classifies the sentences according to the type of technological information they provide by using the form of phrases and the parts of speech that each sentence contains. Then, the level of the information of the keyword is classified using the degree of influence of each keyword in the sentence, and the keyword is extracted by layering the keywords in a structure.

3. Methodology

3.1. Basic Concept

The main purpose of this research is extracting useful keywords from patents and interpreting them in terms of technological information. The main limitation of the previous studies is that they do not aim at the meaning of the keyword itself. Thus, they omit useful keywords, and extracted keywords cannot be interpreted easily. In order to solve this limitation, we construct a set of keywords by structuring the sentences included in a patent by using descriptive information and layering the keywords in order to clearly interpret the meaning of extracted keywords without missing information unique to each patent. In this study, as shown in Figure 2, the technological information of patents is structured and layered by two parallel processes of technological information analysis, patent text structuring/layering. First, in the technological information analysis stage, the analysis of existing studies and various patents is incorporated to classify the technology types, and each sentence is structured according to each type of description and the parts of speech. In this case, when specific words such as pronouns and parts of speech appear, we define pointing words as an indicator that identifies the technical information contained in each sentence. Using pointing words, we can determine the type of technological information contained in a sentence when a certain word appears in the sentence, depending on the meaning of the pointing words. Then, the extracting rule is defined, which is a method of extracting keywords in a hierarchical manner based on the degree of importance of each keyword in technological information. Second, text in patent is preprocessed to apply an extracting rule which is results of the first step, the information of the part of speech and phrase in patented text is tagged through the NLP process to apply pointing words and extracting rule defined in the technological information analysis stage. We then classify and structure the types of technological information of sentences through pointing words. In addition, technological information in patents is layered using NLP. As results of the structuring & layering process, each type of technological information is derived in specific meaning, and information is layered by their importance and impacts.

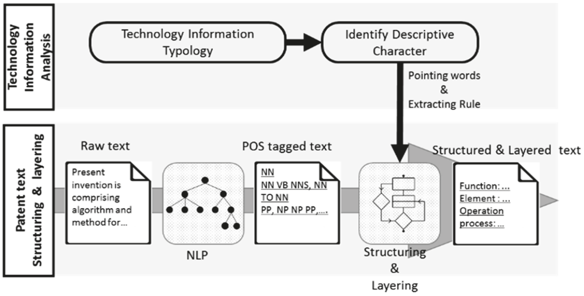

Technology Information Analysis
Technology Information Typology
Identify Descriptive Character
Pointing words & Extracting Rule
Patent text Structuring & layering
Raw text
Present invention is comprising algorithm and method for...
NLP
POS tagged text
NN
NN VB NNS, NN
TO NN
PP, NP NP PP,....
Structuring & Layering
Structured & Layered text
Function: ...
Element : ...
Operation process: ...

Figure 2. Research concept.

3.2. Overall Process

3.2.1. Technological Information Analysis

(1) Definition of technological information type

A patent describes the structure, method, function, substance, or combination of these, in order to protect the invention via the Patent Act that stipulates that a person having ordinary knowledge in the technical field can easily carry out the invention. In addition, the application for the patent should summarize the above matters, and the claims shall be clearly and concisely stated in the Act [44]. The patent also has a claim that specifies the type of technological information it includes. Claims are the most important factor in securing legal protection and uniqueness by explicitly claiming the patent owner's legal rights. In addition, according to 35 U.S.C § 112 [45], 'a component may be expressed as means or steps for performing a specific function other than a description of a material or an act'. Accordingly, when writing the patent specification, the uniqueness of the patent is secured by describing its various unique components.

In accordance with these regulations, unlike general documentation, patents must describe specific technological information. Therefore, all statements described in the patent can be interpreted to include specific technological information. In previous research, the keywords were extracted first and the meaning of each keyword was interpreted from expert opinions and prior research in the technical field, in terms of technological information such as keywords related to function and keywords related to the application field of the patent. However, in this study, we need to clarify the meaning of the keyword from a technological information perspective without additional analysis by structuring and extracting the patent information. In this process, the type of technological information in the patent is defined by referring to the international technology classification, the domestic technology industry classification system, and the existing research that interprets the technological information of the patent from various viewpoints.

(2) Identification of description type according to technological information type

Technological information has different descriptions depending on the form. For example, the function of a patent is described in the form of a verb or a gerund, and a component is expressed in the form of a noun. In this way, the form described by the type of technological information is understood and only the desired information is extracted according to the type of technological information. Thus, this research defines the type of technological information, and uses pointing words to extract only specific information. The pointing word is an indicator that can identify which sentences have technical information and which contents of technical information are included in the sentences. By using pointing words, the sentences appearing in the title, type, and claim can be structured according to the type of technological information.

For example, interface, and system, which are nouns described in the title of patent number US8570295 [46] shown in Table 1, are nouns commonly used in the patent itself or as nouns referring to the whole patent. The first sentence in the abstract of the document in which nouns are subjects is the explanation of the patent. In this case, 'interface' and 'system' can be defined as pointing words that each word explains the patent itself. In other words, if 'interface' or 'system' appeared in a sentence, the sentence contains technological information about the patent itself. It is to be understood that the words 'include' and 'comprise' described in the abstract and claim parts both mean 'inclusion', while the following words, 'layer' and 'fluid channel' correspond to a component. In this case, if 'include' and 'comprise' occur, they can be defined as pointing words in which the component is described. If two defined pointing words are used, it is a sentence that contains technological information that describes the patent itself when it contains only 'interface', and 'system'. If it contains both 'include' and 'comprise', it can be classified into sentences expressing technological information about elements.

Table 1. Index of patent US8570295.

	US8570295
Title	User interface system
Abstract	The user interface system of the preferred embodiment includes: ...
Claim	A user interface comprising: a layer comprising ..., a fluid channel, and a tactile surface, ...

The sentences that are classified according to the type of technological information and included as pointing words have technological information for the different parts of speech. In addition, the technological information is not a single word, but contains a more detailed description in terms of phrases. Therefore, in this study, a 'depth' concept is defined and used as a measure of the degree of importance of keywords in each technological information.

Figure 3 shows the result of tagging information of phrases and the parts of speech in the sentence, "sensors recognize gesture moving horizontally" through NLP. Assuming that 'sensors' is one of the constituent elements of the patent, it can be seen that the most important information is the 'sensors', which provide additional information on horizontal movement, which is a feature of the sensors and gestures. The information in the Depth 1 can be defined as 'sensors' and the keywords in the Depth 2 become 'sensors' and 'recognize'. Thus, the keywords of Depth n + 1 can be defined as a set of keywords appearing at a level of the nth and keywords appearing at a level of n + 1th. Therefore, the extraction rule that extracts layered keywords using the depth concept is defined in the sentence classified by types of technological information.

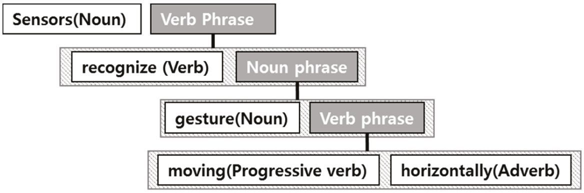


Figure 3. Part of speech (POS) tagged sentence.

3.2.2. Patent Structuring and Layering

The sentences in the title, summary, and claims are converted into a tagged form of words in parts of speech and phrases through the NLP process. The sentences described in the title, summary, and claims are classified and structured according to the types of technological information given by the pointing words defined through the patent analysis and the existing prior studies. The technological information is then layered according to depth, using the extracting rule defined for each type of technological information.

(1) Structuring technological information in patent documents

We use pointing words defined on the basis of descriptive type of technological information. The sentences in the title, summary, and claims are structured according to technology types based on whether or not pointing words appear. As described in Section 3.2.1, the type of technological information is not classified by only one pointing word, but by multiple pointing words and the texts included in the titles, abstracts, and claims are structured according to the types of technological information in a sentence, as shown in Figure 4. For example, if a sentence has a noun that refers to a patent, and a verb that means inclusion, it includes the elements that the patent itself contains. On the other hand, when a sentence has a noun that refers to the patent as well as a verb that is related to inclusion, it can be seen that one of the components of the patent contains elements other than the patent itself. In a sentence, a noun that is not the name of a patent can be considered as a component of the patent.

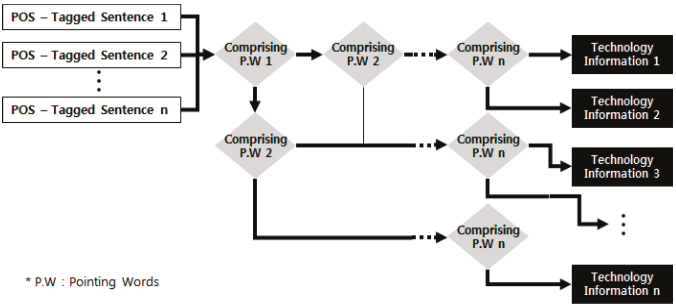


Figure 4. Text structuring based on pointing words.

A patent is characterized by a long, complex sentence, rather than a short sentence. If the pointing words appear once in a complex sentence, errors will most likely occur. Therefore, in this study, using the depth concept throughout the sentence, the error can be prevented only by including the lowest depth, that is, the pointing words as the key information, as shown in Figure 4.

(2) Layering keywords by technological information

Keywords are structured depth using the extraction rule defined for each structured sentence according to technology types. For example, the sentence, "device includes: a sensor recognizes a gesture moving horizontally", is classified as a sentence containing a component, and indicates that the noun is technological information corresponding to a component. Therefore, the phrase containing "a sensor" can be extracted as technological information and layered, as shown in Figure 5. In other words, a lot of information is divided into various layers, enabling to choose how many information extract to interpret a patent. For example, if we want formal information about a sentence, we choose Depth 1 and interpret a main topic of the sentence. If we want more information than Depth 1, we can analyze Depth 2 and Depth 3. Each depth means that Depth 2 contains information about the main topic and their verb, and Depth 3 contains more information than Depth 2, which is an object of the verb.

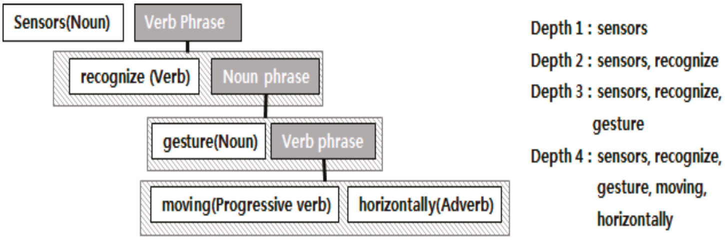


Figure 5. Layered sentence by depth.

3.3. Efficiency Verification

The verification process consists of two stages. First, we validate the efficiency of the patent-level keyword sets by extracting the technological information for each type of patent quantitatively and qualitatively. Then, by analyzing one patent through structured and layered keywords, we confirm whether the level of patent analysis has been analyzed, not the level of technology field analysis. First, we compare the average *TF-IDF* value of a set of keywords to verify the efficiency of the set of keywords derived from a higher technology level in order to verify the set of keywords by a quantitative method. For the qualitative verification, we verify the importance of the keywords obtained based on the technical classification system of the technical field and the advice of experts and verify the efficiency of the method by judging whether the meaningful keywords are present in the existing method and the proposed method. The *TF-IDF* value indicates the degree of importance of each keyword in the document. As the *TF-IDF* value increases, the keyword in the document becomes more important. In this paper, we evaluate the importance of keywords in a set of keywords, rather than the importance of the keywords in the documents. The formula for obtaining the *TF-IDF* value is as follows.

$$TF = \frac{keyword\ frequency\ \in keyword\ set}{all\ word\ freqyency\ \in keyword\ set} \tag{1}$$

$$IDF = \log_2 \frac{number\ of\ patent}{number\ of\ patent\ containing\ keyword} \tag{2}$$

$$TF - IDF = TF \times IDF \tag{3}$$

Therefore, when the average *TF-IDF* value of the keyword set is high, it can be interpreted that the importance of the keywords belonging to the keyword set is high. The average *TF-IDF* values of a set of keywords obtained by the keywords extracted through the conventional method and the proposed method are compared. A comparison of two sets of keywords suggests that a set of keywords with a high *TF-IDF* value contains important keywords in each patent.

This research evaluates the quality of the keywords obtained by the proposed method based on the classification system of the existing technical field and the consultation of experts. Although the keywords evaluated to be meaningful keywords are not revealed by the existing methods, the result shows that the keywords appear when extracted through the proposed method, confirming that important keywords are extracted in the interpretation of patents and technical fields. Finally, it is confirmed that the patent can be obtained by structuring and layering the information at the level of one patent rather than at the level of technology, based on the keywords of the technical field extracted from one patent.

4. Data Analysis

Based on the suggested framework, the extracting rule is defined by both literature review and linguistic analysis about texts included in patent documents. Then, the defined rule is applied to a target technology and then technological information is structured and layered. Finally, the results are verified in qualitative and quantitative ways.

4.1. Selection of Analysis Target

The user interface (UI) field was used in this study in order to select various types of technological information in patents. UI is a technology that enables communication between people and machines. It is a technology field that connects many smart devices and smart contents with which we are in contact [47]. UI technology is attracting attention as 'human-centered' technology, leading the ICT market [48]. Even if products have the same function, a user can select a product that provides the high efficiency and convenience of the UI. Thus, the UI technology should be considered as one of core factors in developing a user-oriented product. The UI field is used in various technologies and industries, and can be divided into detailed fields according to the method in which information is provided to the user and the input method. In the method of providing information, a user-experience interface is provided as a method of using all five senses instead of using only sight. As a method of inputting information, the intention of the user can be input through various methods such as a touch screen, motion recognition, biometric signal recognition etc., other than a keyboard and a mouse input method. In addition to its features, it has various applications such as personal computers, navigation, and medical devices. Therefore, it can be seen that the UI field has various keywords according to the type of technological information for the purpose of this study.

In this study, the patents of UI are analyzed, and keywords are structured and layered according to types of technological information. In order to collect the patents in the UI field, a patent database of United States Patent and Trademark Office (USPTO) was applied. A total of 500 patents are collected by using the search expression (1) "User Interface" in the title and summary from 2011 to 2015, excluding sentences that have particular symbols such as "", [], and <> that could cause an error in the NLP process.

4.2. Technical Analysis

(1) Definition of types of technological information

In this study, four types of technological information were defined by reviewing the patent classification, industry classification technology, and information in the patent in previous studies. First, patents are classified by international standards of the IPC code that have the five classification criteria for a patent, including industry types, the patent form, component, functions, and detailed features. Industrial technical classification has been based on the type of industry, application, function, and operation method.

In terms of the use of various technological information, Huang et al. [17] utilized the SAO structure, interpreting the action as patented functions and the object as the objectives of patents. Lee et al. [18] classified patents by analyzing patent claims in terms of the patent rights. Kim and Choi [49] demonstrated that the meaning of each keyword has a greater effect on the technical classification than positions of keywords such as title, summary, and claims by analyzing the effect of keywords on classification based on Japanese patents.

From literature reviews on technological information, types of technological information can be defined as functions, applications, components, and operation methods. Functions refer to a key feature of the patent. The application is an object where a patent is performed or operated, referring to the areas in which the patent is carried out. Components refer to a generic name of various types of components, such as the physical and functional components included in the patent, and can be divided into non-legal components and legal components. The method of operation refers to the specific method for achieving a specific function described in the patent.

The description of information types proposed in this study can be classified by the patent laws, enabling classification of all sentences contained in the patent document. According to the patent law, the subject of sentences should be a noun that is previously defined because every word in the patent documents must be clearly explained to the subject. A patent has specific nouns that indicate the type of patent, such as a device or method, or general nouns such as 'handle' and 'glasses' in the title of the patent to identify the patent. A sentence that has such nouns as a subject explains a patent itself. Thus, all sentences that have a subject which is a noun to indicate a patent, explain the functions and applications of the patent.

In terms of components, non-legal and legal components are clarified by a verb that includes the meaning of the components in the patent claims and summary. Since the simple identification of components is not able to ensure the uniqueness of each patent, the features of each component in the patent document are defined to ensure uniqueness. Therefore, if the subject of a sentence is a component, it can be said that the sentence describes methods of operation for implementing a key feature of the patent.

Finally, a sentence has a subject that is a noun, indicating the patent itself or its functions. A sentence that explains a function deals with the detailed explanation of the critical features and operating methods. However, according to patent law, a patent should provide novelty compared with the functions of existing patents; therefore, it is extremely unusual to describe the detailed features through the new sentence. Therefore, a sentence in which a subject is a function can be interpreted as unique information for providing an operating method of the patent as well.

(2) Identifying narrative forms of technological information

The technological information described in the patent has a typical description form according to its type. When a sentence is described as a particular form, it can be classified as a sentence that contains the information of the description. By extracting the technological information that is described in the form of classified sentences, it is possible to extract each type of technological information. This research defines pointing words in order to classify the sentences, considering the type of technological information. Pointing words are words that can determine a type of information technology, which is included in the sentence based on the occurrence or absence of the specific words. Pointing words consist of nouns that represent a patent (Representing Noun; RN), verbs that have common meaning (General Verb; GV), verbs that appear in front of components (Component Verb; CV), and nouns that mean components of patents (Component Noun; CN).

RN is a noun related to the form of patents such as device, method, system, algorithm, program, apparatus, and invention. Since its frequency is quite high, yet the *TF-IDF* values are very low, the keyword has information that has little impact on the meaning of the whole patent. Further, in addition to the form of a noun related to the patent, patents may be briefly expressed as a noun, such as 'display' and 'handle'. Since these nouns are in the lowest level in the title of the patent, the terms related to the form of the patent and nouns that are in the lowest level of the noun can be defined as RN. Although RNs that can be obtained in two ways have little significance in the patent, they explain the patent itself. A sentence in which RNs appear can be interpreted as a description of the patent itself. GV consists of verbs such as 'provide', 'suggest', 'mean', 'relate', and 'describe' that have high frequency, yet the *TF-IDF* values are very low. GV are verbs that, when interpreted in Korean, are not interpreted as a specific meaning, and a general description is given such as "the patent provides (or proposes, means) the ~~", before describing the key information. Thus, since a GV does not give special information in the sentence, the subsequent phrases and nouns contain more key features. CV refers to verbs such as 'consist', 'include', 'compose', 'form', and 'involve' that have high frequency yet very low *TF-IDF* values. Nouns and noun phrases that are contained behind the aforementioned verbs can be construed as a component of the patent. CN can be interpreted as a component of patents because it is a noun or noun phrase appearing after CV. Although CN is not the key factor of the patent, the CN can implement the key features or functional components of the patent due to the interaction between the CNs. Therefore, a sentence in which the CN is a subject can be interpreted as the method used. The type of technological information can be determined by using the aforementioned four pointing words included in their position in the title, summary, and claims. In addition, since the technological information has a particular part of speech according to its type, extracting rules can be defined to extract the desired technological information by utilizing NLP as shown in Table 2.

Table 2. Described form and extracting rule of technological information.

Described Form						Extracting Rule
Technological Information Type	Described Item	Pointing Words				
		RN	GV	CV	CN	
Function	Title	O	X	X	X	Verb, Gerund phrase, Verb phrase with preposition
	Abstract	O	X	X	X	Verb phrase, Gerund phrase
	Abstract	O	O	X	X	Gerund phrase with preposition, supine
Object	Title, Abstract	O	O	X	X	Noun phrase with preposition
Unauthorized Component	Abstract	O	X	X	X	Noun phrase, gerund phrase, supine after GV
Authorized Component	Claim	O	X	O	X	Noun phrase, Gerund phrase, supine after 'Comprising: ' with 'a, an'
Method of operation	Abstract, Claim	X	X	X	O	Verb phrase in taking a component as a subject
	Abstract	O	X	X	O	Verb phrase after 'the + RN or CN'

Statements containing information about the function of the patent can be classified into two types. The first type is a sentence that includes RN in the title and summary, but does not contain a CV and GV. In such a case, the extracted verb or gerund phrase can be interpreted as a function. Another form contains the RN, GV, CV, and a preposition before a verb phrase or 'to' infinitive. In such a case, the extracted verb or gerund phrase can be interpreted as a function used to extract the verb phrase or 'to' infinitive subsequent to prepositions as its adverbial usage.

A sentence that expresses the application of the patent includes the RN in the title and summary, and the noun extracted after the preposition can be interpreted as an application. In terms of components, the description forms of legal and non-legal components differ. First, a sentence containing a legal constituent element includes all the sentences described in the claim. It is possible to extract noun phrases, gerund phrases, and 'to' infinitive phrases including the articles 'a/an' described after the word "comprising" and to interpret them as a legal constituent in noun usage. Sentences containing non-legal components appear in the abstract with RN and CV. It is possible to extract noun phrases, gerund phrases, and 'to' infinitive phrases including 'a/an' described after CV, and interpret them as non-legal constituents according to noun usage.

The sentence containing the method of operation appears in two forms based on the presence or absence of the article. First, the verb phrase can be interpreted as the method of operation in the sentence that has the CN, rather than the RN, as the subject. Second, if "The + RN" is a subject of a sentence, it can be interpreted as an operation method.

4.3. Patent Structuring and Layering

The texts included in the title, summary, and claim according to the sentence are separated and each sentence is tagged with POS. Then, based on the tagged information, the text is structured according to the type of technological information, and the structured information is layered according to its depth.

4.3.1. Structuring Technological Information according to the Types

Sentences that extracting rules can be applied to are explored in the title, summary, and claims of patents. Although the title and claims are described only in one sentence, a summary generally has multiple sentences. This paper analyzes 500 sentences with titles and claims, respectively, and 951 sentences with a summary. The patent documents therefore contain a total of 1951 sentences. The number of sentences classified by type of technological information by the pointing words is shown in Table 3.

Table 3. Sentence Ratio.

Technological Information	Number of Sentences	Ratio
Function	411	21%
Object	484	9%
M.O *	64	3%
Function, Object	550	28%
Function, Component	207	11%
Function, Object, Component	271	14%
Object, M.O	53	3%
None	221	11%

* Method of operation.

4.3.2. Keyword Layering by Technological Information

In the title, summary, and claims, sentences are structured according to function, application, object, component, and operation method, considering the depth in which information is extracted at a desired level. Depth 1 contains the keywords that can be obtained at the highest level. Functions include verbs, nouns for applications, nouns or gerunds for components, and verbs for operating methods. The lower depth comprises words that can further describe the words in Depth 1. Table 4 shows the results of extracting the keywords by changing the level of depth from Depth 1 that can obtain only keywords to Depth 4 that can obtain additional information.

Table 4. The number of extracted keyword.

	All	Depth			
		1	2	3	4
Total	4243	1044	1442	1581	1656
Function		550	699	809	874
Object		282	389	436	471
U.C *		454	535	596	643
A.C **		415	966	4063	1117
M.O ***		203	251	290	322

* Unauthorized component; ** Authorized component; *** Method of operation.

4.4. Verification

In order to verify all the keywords obtained from the technical field in this study, we quantitatively and qualitatively compared our proposed method with the method of extracting the top 20% of the existing *TF-IDF* values. In addition, we analyzed the function, application object, component, and operation method of the patent by using the structured keyword which could not be obtained by the conventional method.

4.4.1. Keyword Set Verification

We compared the average *TF-IDF* of the keyword set using the existing keyword extraction method and the structuring and layering (S & L) proposed in this study while changing the ratio of the keywords extracted from each keyword set. Figure 6 shows that as the number of keywords increases, the average value of *TF-IDF* decreases. In addition, it can be confirmed that the S & L method has a higher average value of *TF-IDF* than the conventional method. In this case, as the depth decreases, the average *TF-IDF* value increases. Therefore, it can be said that the keyword with low depth has information that is unique to each patent. On the other hand, as the depth increases, the *TF-IDF* value becomes lower. Therefore, as a keyword that has higher depth is extracted, each keyword appears in various patents rather than in one patent.

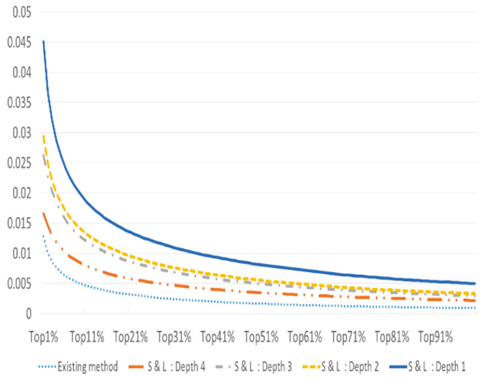


Figure 6. Average of Term Frequency-Inverse Document Frequency (*TF-IDF*) index per extracted ratio.

Table 5 shows the results obtained by comparing a set of keywords through the existing method and S & L. It demonstrates that it is more useful to extract the Depth 1 and 2 keywords using S & L than the method that extracts only the *TF-IDF* values corresponding to the upper 20% of the existing keywords. In addition, the rate that the keyword obtained through S & L is extracted in the method which depends on the *TF-IDF* value is calculated in order to determine whether the problem in which significant keywords are excluded has been resolved. Table 6 shows that only 45.31% of the extracted keywords are extracted from the top 20% of the extracted keywords. It can be seen that the existing method extracts only part of the significant keywords extracted through S & L.

Table 5. Average of *TF-IDF* index per extracted ratio.

Extracting Method	Ratio	Average of *TF-IDF*
Existing Method	20%	0.003228
Existing Method	100%	0.000917
S & L: Depth 1	100%	0.004991
S & L: Depth 2	100%	0.003322
S & L: Depth 3	100%	0.002959
S & L: Depth 4	100%	0.002166

Table 6. Extracted keyword ratio in existing method.

Extracting Ratio	Extracted Keyword (Ratio)	
	S & L: Depth 1	**S & L: Depth 4**
20%	475 (45.31%)	849 (51.27%)
40%	699 (66.95%)	1624 (98.07%)
60%	829 (79.12%)	1654 (99.88%)
80%	910 (87.16%)	1656 (100%)
100%	1044 (100%)	1656 (100%)

4.4.2. Keyword Verification

Through the keyword set verification, it can be confirmed that S & L is a set of keywords that is quantitatively superior to the existing keyword extraction method. For the qualitative analysis of the extracted keywords, valid keywords were selected from S & L's classified functions, applicable object, non-legal constituent element, legal constituent element, and operation method based on expert's advice and related research [6,50]. At this time, since the rank of extracted keywords differs according to the depth, the depth that can select the most significant keyword according to technological information type is derived. Table 7 shows that the main functions of the user interface technology are represented by display, detect, and control, and their applications are graphical display, gesture, information, and application. Elements constituting each patent include functions used to perform detect, control, and receive, as well as physical components such as display, processor, surface, and sensor. In addition, it can be seen that the main functions of the patent are implemented through the actions of detect, associate, integrate, and configure.

Table 7. Extracted keyword index per technological information.

Rank	S & L: Depth 1				S & L: Depth 4					
	Function		**Object**		**Unauthorized Component**		**Authorized Component**		**Method of Operation**	
	Words	**Ranked in E.M ***	**Words**	**Ranked in E.M**	**Words**	**Ranked in E.M**	**Words**	**Ranked in E.M**	**Words**	**Ranked in E.M**
1	Display	14	Graphical	47	Display	14	Display	14	Area	7
2	Detect	340	Display	14	Detect	340	Professor	125	Display	14
3	Control	12	Gesture	11	Surface	10	Detect	340	Associate	**1689**
4	Configure	**1691**	Information	16	Control	12	Configure	**1691**	Detect	340
5	Determine	356	Data	**2691**	Receive	218	Select	340	Integrate	**1555**
6	Manage	**3691**	Application	1	Touch	4	Receive	218	Configure	**1691**
7	receive	218	location	40	sensor	76	control	12	Gesture	11

* Existing Method.

In the existing keyword extraction method, the upper 20% of the *TF-IDF* value is extracted. The total number of keywords to be analyzed is 4243, and when 20% is extracted, only the top 849 keywords are extracted and other keywords are excluded. Table 7 shows that the main functions of the patents in the User Interface field are display, detect, control, configure, and manage. At this time, it can be seen that "configure" and "manage" have *TF-IDF* values as high as the 1691th and 3691th keywords in the entire keyword set, and thus cannot be extracted by the conventional method. In the same way, it can be seen that important keywords in technological information are not extracted by the conventional method.

4.4.3. Verification of Patent Interpretation Method

Keyword extraction through S & L can be used not only in the entire patent data set, but also when extracting keywords from a single patent. Table 8 shows the results obtained by applying US857029 with the information level of Depth 4 in the S & L approach. Analysis of the results in

Table 8 shows that this patent has the function of inputting keywords in various ways and is used in computers. In addition, it consists of display function and the physical components of the keyboard. Legally, the way that users type letters and text on the keyboard is protected. A function of inputting characters by sensing various touches of the keyboard is implemented.

Table 8. Keyword index of US8570295.

US8570295	
Title	Touch screen device, method, and graphical user interface for inserting a character from an alternate keyboard
Function	insert character alternate keyboard
Object	computer-implemented method use
Unauthorized component	display soft keyboard
Authorize component	plurality character-insertion key select soft keyboard different single move break text input area correspond
Method of operation	key select soft keyboard different contact detect response Movement lift off character

5. Conclusions

In this study, the sentence was structured according to types of technological information by analyzing the descriptive form of technological information described in the title, summary, and claims. Each keyword is then layered based on the degree of importance in the technological information, and the technological information is structured according to type. The core keywords and a set of keywords layered by keywords are then obtained. We confirmed that the keyword set extracted by the level based on the depth from the sentence that is classified according to the descriptive information provides more significant keywords than a set using only part of the existing keyword based on the average *TF-IDF* value of the keywords in a quantitative manner. In addition, when extracting some keywords, it is possible to determine whether or not the keywords included in the verified set of keywords are detected with a significant set of keywords, so that the proposed method can extract the meaningful keywords without missing them.

Based on the preliminary research on the UI field and the consultation of experts, it is found that the proposed method can extract the keywords that are important in function, application object, non-legal component, component, and operation method, but not extracted through the existing method. It is verified that the same word can be interpreted differently depending on the type of technological information. In addition, it is possible to analyze each patent according to the type of technological information by extracting the keywords of each patent without defining the characteristics of patents based on extracted keywords from the technical field level. Thus, we can confirm that the proposed approach can derive a more detailed level of information than the existing text mining techniques.

The proposed methodology has three theoretical contributions. First, keywords of patents are extracted by linguistic criteria. Most previous keyword extraction methodology is based on their meaning or the number of occurrence. However, this research suggests that another criterion can be used to extract meaningful keywords in patent. Second, the proposed methodology can be used not only to derive the characteristics of the technology field but also to derive the characteristics of each patent according to the type of technological information. Third, it is advantageous to extract more significant keywords using fewer keywords by applying the depth after classifying the sentences in terms of analysis efficiency, because the method does not screen valid keywords after extracting all keywords. In addition, according to the intention of a researcher, it is possible to obtain a flexible set of keywords according to the purpose of analysis by varying the depth.

The proposed methodology can be utilized in industry fields by various ways. First, when the technical field has a lack of prior knowledge, or when it is difficult to interpret the extracted keywords in the field of fusion technology, the meaning of each keyword can be clarified and the characteristics of the technical field can be derived. Second, when presenting specific technology opportunities such as development purpose, target, development direction in technology, and product development process, time series analysis can be applied to type-specific technological information to explore more detailed technology opportunities. Third, using features that are structured according to types, technological information can be extracted from the patent level, and the technology can be classified based on the new application target, functions, and components by comparing patents. Moreover, in order to analyze the possibility of patent infringement, the technological information of the two patents to be analyzed can be compared from various viewpoints.

However, this research has two limitations. First, the results of the research are greatly affected by the quality of NLP. The patent documents were structured in this study by relying on the parts of speech tagged by the NLP. In other words, if parts of speech tagging process do not work well, each sentence in patents cannot be structured, technological information cannot be extracted in the structured form. Second, the suggested methodology cannot be applied in all technology domains. In the case of the User Interface technology that analyzed in this research, the technological information of the four types of technology, function, application object, component, and operation method is evenly included. However, excellent results cannot be obtained in a technical field that requires technological information other than the four types, for example, when a patent is expressed through an algorithm or a chemical formula.

In order to overcome the limitations of this study, it is necessary to utilize a better quality NLP, such as NLP using deep learning. In addition, in the analysis of other technical fields, various types of technological information need to be defined according to technical fields such as operation sequence and interaction in addition to the four types of technological information through sufficient literature survey. Moreover, the proposed approach needs to be generalized in any types of documents and technologies by reflecting the unique characteristics of documents and technologies.

Acknowledgments: This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2017R1D1A1A09000758).

Author Contributions: T.R. designed the study, outlined the methodology, conducted the data analysis, and wrote the manuscript. Y.J. interpreted the results, and wrote the manuscript. B.Y. implemented the research, designed the study, outlined the methodology, and helped draft the paper. All authors have read and approved the final manuscript.

References

1. Ernst, H. Patent information for strategic technology management. *World Pat. Inf.* **2003**, *25*, 233–242. [CrossRef]
2. Choi, J.; Kim, H.; Im, N. Keyword Network Analysis for Technology Forecasting. *J. Intell. Inf. Syst.* **2011**, *17*, 227–240.
3. Liu, S.J.; Shyu, J. Strategic planning for technology development with patent analysis. *Int. J. Technol. Manag.* **1997**, *13*, 661–680. [CrossRef]
4. Jeong, E.S.; Kim, Y.G.; Lee, S.C.; Kim, Y.T.; Chang, Y.B. Identifying Emerging Free Technologies by PCT Patent Analysis. *J. Korea Inst. Electron. Commun. Sci.* **2014**, *9*, 111–122. [CrossRef]
5. Yun, J.H. Patent Information Analysis: Tools for Systematic R & D Planning. *Ind. Eng. Mag.* **2011**, *18*, 23–28.
6. Lim, C.; Yun, D.; Park, I.; Park, G.; Koh, S.; Yoon, B. Exploring Prospective Research Areas in UI/UX through the Analysis of Patents. *Korean Manag. Sci. Rev.* **2015**, *32*, 1–18. [CrossRef]
7. Tseng, Y.H.; Lin, C.J.; Lin, Y.I. Text mining techniques for patent analysis. *Inf. Process. Manag.* **2007**, *43*, 1216–1247. [CrossRef]
8. Altuntas, S.; Dereli, T.; Kusiak, A. Analysis of patent documents with weighted association rules. *Technol. Forecast. Soc. Chang.* **2015**, *92*, 249–262. [CrossRef]
9. Park, H.; Yoon, J.; Kim, K. Identifying patent infringement using SAO based semantic technological similarities. *Scientometrics* **2011**, *90*, 515–529. [CrossRef]

10. Wang, X.; Qiu, P.; Zhu, D.; Mitkova, L.; Lei, M.; Porter, A.L. Identification of technology development trends based on subject–action–object analysis: The case of dye-sensitized solar cells. *Technol. Forecast. Soc. Chang.* **2015**, *98*, 24–46. [CrossRef]
11. Campbell, R.S. Patent trends as a technological forecasting tool. *World Pat. Inf.* **1983**, *5*, 137–143. [CrossRef]
12. Lee, C.; Kang, B.; Shin, J. Novelty-focused patent mapping for technology opportunity analysis. *Technol. Forecast. Soc. Chang.* **2015**, *90*, 355–365. [CrossRef]
13. Su, H. Global Interdependence of Collaborative R&D-Typology and Association of International Co-Patenting. *Sustainability* **2017**, *9*, 541.
14. Kim, C. A patent analysis method for identifying core technologies: Data mining and multi-criteria decision making approach. *J. Korea Saf. Manag. Sci.* **2014**, *16*, 213–220. [CrossRef]
15. Kang, H.J.; Um, M.J.; Kim, D.M. A study on forecast of the promising fusion technology by US patent analysis. *J. Technol. Innov.* **2006**, *14*, 93–116.
16. Niemann, H.; Moehrle, M.G.; Frischkorn, J. Use of a new patent text-mining and visualization method for identifying patenting patterns over time: Concept, method and test application. *Technol. Forecast. Soc. Chang.* **2017**, *115*, 210–220. [CrossRef]
17. Noh, H.; Jo, Y.; Lee, S. Keyword selection and processing strategy for applying text mining to patent analysis. *Exp. Syst. Appl.* **2015**, *42*, 4348–4360. [CrossRef]
18. Huang, L.; Shang, L.; Wang, K.; Porter, A.L.; Zhang, Y. Identifying target for technology mergers and acquisitions using patent information and semantic analysis. In Proceedings of the 2015 Portland International Conference on Management of Engineering and Technology (PICMET), Portland, OR, USA, 2–6 August 2015.
19. Lee, C.; Song, B.; Park, Y. How to assess patent infringement risks: A semantic patent claim analysis using dependency relationships. *Technol. Anal. Strateg. Manag.* **2013**, *25*, 23–38. [CrossRef]
20. Lee, W.S.; Han, E.J.; Sohn, S.Y. Predicting the pattern of technology convergence using big-data technology on large-scale triadic patents. *Technol. Forecast. Soc. Chang.* **2015**, *100*, 317–329. [CrossRef]
21. Ko, N.; Yoon, J.; Seo, W. Analyzing interdisciplinarity of technology fusion using knowledge flows of patents. *Exp. Syst. Appl.* **2014**, *41*, 1955–1963. [CrossRef]
22. Lee, W.J.; Sohn, S.Y. Patent analysis to identify shale gas development in China and the United States. *Energy Policy* **2014**, *74*, 111–115. [CrossRef]
23. Yoon, J.; Kim, K. Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. *Scientometrics* **2011**, *88*, 213–228.
24. Wang, H.; Cheng, D.; Chen, C.; Wu, Y.; Lo, C.; Lin, H. A Novel Real-Time Speech Summarizer System for the Learning of Sustainability. *Sustainability* **2015**, *7*, 3885–3899. [CrossRef]
25. Guo, X.; Sun, H.; Zhou, T.; Wang, L.; Qu, Z.; Zang, J. SAW Classification Algorithm for Chinese Text Classification. *Sustainability* **2015**, *7*, 2338–2352. [CrossRef]
26. Kim, H.J.; Jo, N.O.; Shin, K.S. Text Mining-Based Emerging Trend Analysis for the Aviation Industry. *J. Intell. Inf. Syst.* **2015**, *21*, 65–82. [CrossRef]
27. Min, K.Y.; Kim, H.T.; Ji, Y.G. A Pilot Study on Applying Text Mining Tools to Analyzing Steel Industry Trends: A Case Study of the Steel Industry for the Company "P". *J. Soc. e-Bus. Stud.* **2014**, *19*, 51–64. [CrossRef]
28. Kim, M.; Park, Y.; Yoon, J. Generating patent development maps for technology monitoring using semantic patent-topic analysis. *Comput. Ind. Eng.* **2016**, *98*, 289–299. [CrossRef]
29. Ghazizadeh, M.; McDonald, A.D.; Lee, J.D. Text Mining to Decipher Free-Response Consumer Complaints Insights From the NHTSA Vehicle Owner's Complaint Database. *Hum. Factors J. Hum. Factors Ergon. Soc.* **2014**, *56*, 1189–1203. [CrossRef] [PubMed]
30. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
31. Kwon, H.; Kim, J.; Park, Y. Applying LSA text mining technique in envisioning social impacts of emerging technologies: The case of drone technology. *Technovation* **2017**, *60–61*, 15–28. [CrossRef]
32. Park, J.; Song, M. A study on the Research Trends in Library & Information Science in Korea using Topic Modeling. *J. Korean Soc. Inf. Manag.* **2013**, *30*, 7–32.
33. Jang, H.; Roh, T.; Yoon, B. User needs-based technology opportunities in heterogeneous fields using opinion mining and patent analysis. *J. Korean Inst. Ind. Eng.* **2017**, *43*, 39–48. [CrossRef]
34. Jin, S.A.; Heo, G.E.; Jeong, Y.K.; Song, M. Topic-Network based Topic Shift Detection on Twitter. *J. Korean Soc. Inf. Manag.* **2013**, *30*, 285–302. [CrossRef]

35. Gao, L.; Eldin, N. Employers' Expectations: A Probabilistic Text Mining Model. *Procedia Eng.* **2014**, *85*, 175–182. [CrossRef]
36. Guo, H.; Chen, Y. User interest detecting by text mining technology for microblog platform. *Arab. J. Sci. Eng.* **2016**, *41*, 3177–3186. [CrossRef]
37. Lee, S.; Kim, H.J. News keyword extraction for topic tracking. In Proceedings of the Fourth International Conference on Networked Computing and Advanced Information Management, NCM'08, Gyeongju, Korea, 2–4 September 2008; pp. 554–559.
38. Rose, S.; Engel, D.; Cramer, N.; Cowley, W. Automatic Keyword Extraction from Individual Documents. 2010. Available online: https://pdfs.semanticscholar.org/5a58/00deb6461b3d022c8465e5286908de9f8d4e.pdf (accessed on 16 November 2017).
39. Hulth, A. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2003; pp. 216–223.
40. Nasukawa, T.; Yi, J. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture*; ACM: New York, NY, USA, 2003; pp. 70–77.
41. Yi, J.; Nasukawa, T.; Bunescu, R.; Niblack, W. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In Proceedings of the Third IEEE International Conference on Data Minning, Melbourne, FL, USA, 22 November 2003; pp. 427–434.
42. Jin, Y.; Xiong, W. A sentence degeneration model and its application in Chinese-English patent machine translation. In Proceedings of the 2011 7th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), Tokushima, Japan, 27–29 November 2011; pp. 421–424.
43. Yang, S.Y.; Soo, V.W. Extract conceptual graphs from plain texts in patent claims. *Eng. Appl. Artif. Intell.* **2012**, *25*, 874–887. [CrossRef]
44. Korea Patent Law. 2016. Available online: www.kipo.go.kr (accessed on 6 November 2017).
45. United States Patent ACT, 35 U.S.C 112 (2011). Available online: www.uspto.gov (accessed on 6 November 2017).
46. Cisesla, M.C.; Yairi, M.B. User Interface System. U.S. Patent 85,702,795, 29 October 2013.
47. Kim, S.J.; Cho, D.E. Technology trends for UX/UI of smart Contents. *Korea Contents Assoc. Rev.* **2016**, *14*, 29–33. [CrossRef]
48. Sohn, K.S. Technology convergence's present and future. *Ind. Eng. Mag.* **2012**, *19*, 28–33.
49. Kim, J.H.; Choi, K.S. Patent document categorization based on semantic structural information. *Inf. Process. Manag.* **2007**, *43*, 1200–1215. [CrossRef]
50. Park, I.; Park, G.; Yoon, B.; Koh, S. Exploring Promising Technology in ICT Sector Using Patent Network and Promising Index Based on Patent Information. *ETRI J.* **2016**, *38*, 405–415. [CrossRef]

sustainability

Article

Data Governance Taxonomy: Cloud versus Non-Cloud

Majid Al-Ruithe *, Elhadj Benkhelifa * and Khawar Hameed

Cloud Computing and Applications Research Lab, School of Computing and Digital Technologies, Staffordshire University, Stoke-on-Trent ST4 2DE, UK; khawar.hameed@staffs.ac.uk

* Correspondence: majid.al-ruithe@research.staffs.ac.uk (M.A.-R.); e.benkhelifa@staffs.ac.uk (E.B.); Tel.: +966-598-343-504 (M.A.-R.); +44-791-640-6720 (E.B.)

Received: 12 October 2017; Accepted: 14 December 2017; Published: 2 January 2018

Abstract: Forward-thinking organisations believe that the only way to solve the data problem is the implementation of effective data governance. Attempts to govern data have failed before, as they were driven by information technology, and affected by rigid processes and fragmented activities carried out on a system-by-system basis. Until very recently, governance has been mostly informal, with very ambiguous and generic regulations, in siloes around specific enterprise repositories, lacking structure and the wider support of the organisation. Despite its highly recognised importance, the area of data governance is still underdeveloped and under-researched. Consequently, there is a need to advance research in data governance in order to deepen practice. Currently, in the area of data governance, research consists mostly of descriptive literature reviews. The analysis of literature further emphasises the need to build a standardised strategy for data governance. This task can be a very complex one and needs to be accomplished in stages. Therefore, as a first and necessary stage, a taxonomy approach to define the different attributes of data governance is expected to make a valuable contribution to knowledge, helping researchers and decision makers to understand the most important factors that need to be considered when implementing a data governance strategy for cloud computing services. In addition to the proposed taxonomy, the paper clarifies the concepts of data governance in contracts with other governance domains.

Keywords: data governance; cloud computing; cloud data governance; taxonomy; systematic review; holistic

1. Introduction

We are accustomed to the concepts of information technology (IT) governance [1] and corporate governance [2]. The term "governance", in general, refers to the way an organisation ensures that strategies are set, monitored, and achieved [3]. As IT has become the backbone of every organisation, by definition, IT governance becomes an integral part of any business strategy, and falls under corporate governance. Historically, data emerged out of disparate legacy transactional systems. Then, data was seen as a by-product of running the business, and had little value beyond the transaction and the application that processed it, hence data was not treated as a valuable shared asset. This continued until the early 1990s, when the value of data started to take another trend beyond transactions. Business decisions and processes increasingly started to be driven by data and data analysis. Further investment in data management was the approach taken to tackle the increasing volume, velocity, and variety of data, such as complex data repositories, data warehouses, Enterprise Resource Planning (ERP), and Customer Relationship Management (CRMs) [4]. Data links became very complex and shared amongst multiple systems, and the need to provide a single point of reference in order to simplify daily functions became crucial, which gave birth to master data management [5].

Data complexity and volume continue to explode; businesses have grown more sophisticated in their use of data, which drives new demands that require different ways to combine, manipulate, store, and present information. Forward-thinking companies recognised that data management solutions alone are becoming very expensive and are unable to cope with business realities, and the data problem must be solved in a different way [6]. During this time, the notion of data governance started to take a different direction, a more important one. Attempts to govern data failed before, as they were driven by IT, and affected by rigid processes and fragmented activities carried out on a system-by-system basis. Until very recently, governance has been mostly informal, in siloes around specific enterprise repositories, lacking structure and the wider support of the organisation. Despite its recognised high importance, data governance is still an under-researched area and less practised in industry [7,8]. Researchers differ in their definitions of data governance. The governance concept can be understood in different contexts, for instance, corporate governance, information governance, IT governance, and data governance; Wende [9] and Chao [10] argue that data governance and IT governance need to follow corporate governance principles.

To achieve successful data governance, organisations need a strategy framework that can be easily implemented in accordance with the needs and resources of information [11,12]. A good data governance framework can also help organisations to create a clear mission, achieve clarity, increase confidence in using organisational data, establish accountabilities, maintain scope and focus, and define measurable successes [11,13]. To facilitate data governance, Seiner [14] argues that organisations must design a data governance model of role responsibilities to identify people who have a level of accountability to define, produce, and use data in the organisation. Along similar lines, some authors in the literature argues that organisations should obtain responsibility for data from the information technology (IT) department, with the participation and commitment of IT staff, business management, and senior-level executive sponsorship in the organization [15]. Experts in this field show that where organisations do not implement data governance, the chaos is not as obvious, but the indicators are glaring, including dirty, redundant, and inconsistent data; inability to integrate; poor performance; terrible availability; little accountability; users who are increasingly dissatisfied with IT performance; and a general feeling that things are out of control [16]. The first efforts to create a framework for data governance were published in 2007 [9,17].

The emergence of cloud computing is a recent development in technology. The National Institute of Standards and Technology (NIST) [18] defined cloud computing as "*a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction*". The cloud computing model enhances availability, and is composed of five essential characteristics, four deployment models, and three service models [19]. The essential characteristics of cloud computing include on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service [20]. The cloud deployment models are the private, public, hybrid, and community models [21]. In addition, cloud computing includes three service delivery models, which are: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) [22]. Cloud computing offers potential benefits to public and private organisations by making IT services available as a commodity [23,24]. The generally claimed benefits of cloud computing include: cost efficiency, unlimited storage, backup and recovery, automatic software integration, easy access to information, quick deployment, easier scale of services, and delivery of new services [25]. Furthermore, other benefits include: optimised server utilisation, dynamic scalability, and minimised life cycle development of new applications. However, cloud computing is still not widely adopted due to many factors, mostly concerning the moving of business data to be handled by a third party [6], where, in addition to the cloud consumer and provider, there are other actors: the cloud auditor, cloud broker, and cloud carrier [26]. Therefore, loss of control of data, security and privacy of data, data quality and assurance, data stewardship, etc. can all be cited as real concerns of adopting the cloud computing business model [27]. Data lock-in is another potential risk, where cloud customers can face difficulties in extracting their data from the cloud [28]. Cloud consumers can also suffer from operational and regulatory challenges, as organisations transfer their data to third parties for storage and processing [29]. In addition, it may be difficult for the consumers to check the data handling practices of the cloud provider or any of the other involved actors [23,30,31]. The cloud computing model is expected to be a highly disruptive technology, and the adoption of its services will, therefore, require even more rigorous data governance strategies and programmes, which may be more complex, but are necessary.

The general consensus among authors is that data governance refers to the entirety of decision rights and responsibilities concerning the management of data assets in organisations. This definition does not, however, provide equal prominence for data governance within the cloud computing technology context. Therefore, this deficit calls for in-depth understanding of data governance and cloud computing. This trend contributes to changes in the data governance strategy in the organisation, such as the organisation's structure and regulations, people, technology, processes, roles, and responsibilities. This is one of the great challenges facing organisations today when they move their data to cloud computing environments, particularly regarding how cloud technology affects data governance. The authors' general observation reveals that the area of data governance in general is under-researched and not widely practised by organisations, let alone when it is concerned with cloud computing, where research is in its infancy and far from reaching maturity.

This forms the main motivation behind this paper, which attempts to provide the readers with a holistic view of data governance for both cloud and non-cloud computing, using a taxonomy approach. The contribution of this paper is unprecedented, with this taxonomy expected to be very valuable in developing coherent frameworks and programmes of Data Governance for both cloud and non-cloud computing. One main question has been considering to formulate the results in this study which is following: what is the main factor that require to develop the data governance for non-cloud and cloud computing?

The remainder of the article is structured in seven sections. The next section discusses what data governance is, and why it is important, followed by a section reviewing the literature on data governance. A subsequent section presents the relationship between data governance and other governance domains. Following this, the data governance taxonomy section presents a holistic taxonomy for data governance for cloud and non-cloud. The final Section presents the conclusions, limitations of research and future work.

2. What Is Data Governance and Why Is It Important?

It is important, before developing a holistic taxonomy, to define the context of data governance. Often, researchers and practitioners confuse data governance and data management. The definition of data management provided by the Data Management Association (DAMA) is: "*data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets*" [12]. Data management in general focuses on the defining of the data element, how it is stored, structured, and moved. Although there is no official standard definition of data governance, to provide clarity, we refer to the most cited definitions offered by some important organisations and specialists.

According to the Data Governance Institute (DGI), data governance is "a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods" [32]. The IT Encyclopedia defines data governance as: "*the overall management of the availability, usability, integrity, and security of the data employed in an enterprise. A sound data governance program includes a governing body or council, a defined set of procedures, and a plan to execute those procedures*" [7]. DAMA, on the other hand, defines data governance as: "*the exercise of authority, control and shared decision-making (planning, monitoring and enforcement) over the management of data assets*" [33]. According to DAMA, data governance is, therefore, high-level planning and control over data management [33]. Wende [9] have also argued that data governance is different from data management, that data governance complements data management, but does not replace it. Ladley [34] defined data governance as "a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods". Weber [7] suggested that data governance "specifies the framework for decision rights and accountabilities to encourage desirable behaviour in the use of data. To promote desirable behaviour, data governance develops and implements corporate-wide data policies, guidelines, and standards that are consistent with the organization's mission, strategy, values, norms, and culture." More recently, "Non-Invasive", a book by Seiner in 2014, defines data governance as "the formal execution and enforcement of authority over the management of data and data related assets" [14].

Some other researchers or practitioners seem also to confuse IT governance and data governance. IT governance is a much more mature area, with the first publications on the topic released about four decades ago [35], while data governance is still under-researched. Organisations with mature IT governance practices tend to have a stronger alignment between IT and business [36], and the author argues that organisations should gain the responsibility for data from the IT department. Besides IT governance, data governance also has a significant role in aligning the organisation's business. Data governance can be used to solve an assortment of business issues related to data and information [16]. Otto [37] argued that a data governance model helps organisations to structure and document the accountabilities for their data quality. Some authors have explicitly demonstrated that data governance is different from IT governance in principle and practice [9,24]. In principle, data governance is designed for the governance of data assets, while IT governance makes decisions about IT investments, the IT application portfolio, and the IT projects portfolio. In practice, IT governance is designed primarily around an organisation's hardware and applications, not its data.

Al Rifai M. et al. [30] argues that enterprise-wide data strategy and governance are important for organisations, and are required to achieve competitive advantage. In addition, all existing sources have hitherto only addressed data governance. The fact that organisations need to take many aspects into consideration when implementing data governance has been neglected so far [9,16,38]. Moreover, some researchers show that organisations which do not implement effective data governance can quickly lose any competitive advantage [14,39]. Seiner [14] illustrated that working without a proper data governance programme is analogous to an organisation allowing each department and each employee to develop, for instance, their own financial chart of accounts. Data governance in any organisation requires the involvement and commitment of all staff, with full sponsorship by the management and senior-level executive sponsorship [40].

Recently, many organisations have become aware of the increasing importance of governing their data to ensure the confidentiality, integrity, quality, and availability of customer data [41,42]. Currently, there is no single approach for the implementation of a data governance programme for all organisations [3]. Good data governance can help organisations to create a clear mission, achieve clarity, increase confidence in using organisational data, establish accountabilities, maintain scope and focus, and define measurable successes [33,43]. Moreover, many authors have suggested that developing effective data governance will lead to many benefits for organisations. These benefits are: enabling more effective decision-making, reducing operational friction, and protecting the needs of data stakeholders as central to a governance programme [44,45]. In addition, other benefits include: training of management and staff to adopt common approaches to data issues, build standard, repeatable processes, reducing costs and increasing effectiveness through coordination of efforts, and ensuring the transparency of processes [16,17,37].

3. Review of the Literature on Data Governance

An up-to-date literature review has been undertaken to help us and the readers understand the research landscape in data governance. This review will be instrumental in developing the aforementioned taxonomy. The review followed the systematic literature review protocol, defined by [46], with customised search strings, a study selection process, and inclusion and exclusion criteria. The search was conducted in the following libraries and databases: Google Scholar, Staffordshire e-resources Libraries, Saudi Digital Library, and the British Library (Ethos). The term "data governance" was used in this search, but we also tried a combination of keywords in order to test for synonyms used in the literature and to cover all relevant publications. The following search strings were also used, "data governance organization", "governance data", "data governance in cloud computing", "data governance for cloud computing", and "cloud data governance". All these search strings were combined by using the Boolean "OR" operator as follows: ((data governance) OR (data governance organization) OR (governance data) OR (data governance in cloud computing) OR (data governance for cloud computing) OR (cloud data governance)).

The search covered the period between 2000 and 2017. The study selection process was based on four stages, and only 52 records on data governance, which meet the criteria and fall within the scope of the study, were attained for the final review. Table 1 provides a summary of these 52 papers, categorised by academic- and practice-oriented contributions for cloud and non-cloud computing.

Table 1. Categorisation of the resultant records on data governance.

Nature of Contribution	Format	References
Academic	Papers in journals and conference proceedings, books, working reports and theses	Non-cloud: [6,7,9,11–14,30,33,34,37,44,47–59]. Cloud Computing: [60–62].
Practice-oriented	Publications by industry associations, software vendors and analysts	Non-cloud: [38,39,63–65], [50,52,66–111]. Cloud Computing: [41,42,53,70–72].

Out of the retained 52 records, only five records were reported in academic literature on data governance for cloud services. All reported research agrees that only a few organisations have addressed data governance, and only partially. Additionally, all reported academic literature stated that data governance is one of the key components for any enterprise cloud; they also described some issues related to moving data to the cloud outside the organisation's premises, such as security, data migration and interoperability. Felici et al. [60] focused more on one aspect of data governance, accountability, where they proposed an accountability model for data stewardship in the cloud, which explains data governance in terms of accountability attributes and cloud-mediated interactions between actors. This model consists of accountability attributes, accountability practices and accountability mechanisms. Tountopoulos [73] focused on addressing interoperability requirements relating to the protection of personal and confidential data for cloud data governance. They also categorised the accountability taxonomy, composed of seven main roles, which are: cloud subject, cloud customer, cloud provider, cloud carrier, cloud broker, cloud auditor, and cloud supervisory authority. Figure 1 shows the numbers of published research on data governance in the last 10 years, following a systematic review.

Cloud data governance has also been overlooked by industry. Cloud Security Alliance, Trustworthy Computing Group, and Microsoft Corporation are regarded as the recognised leaders in this area. The Cloud Security Alliance cloud data governance working group currently focuses on the data protection aspect, with an aim to propose a data governance framework to ensure the availability, integrity, privacy, and overall security of data in different cloud models; this is far from being realised [74]. Trustworthy Computing Group and Microsoft Corporation describe the basic elements of a data governance initiative for privacy, confidentiality, and compliance, and provide guides to help organisations embark on this path [41]. According to a MeriTalk report in 2014, only 44% of IT professionals in the federal government believe their agencies have mature data governance practices in the cloud. This report also suggests that about 56% of agencies are currently in the process of implementing data stewardship or data governance programmes [75].

Evaluating the existing work on data governance for traditional IT and cloud computing reveals that it is still very limited, lacking standards and unified definitions, hence a taxonomy approach to classify different aspects and attributes of data governance will be a highly valuable contribution at this stage.

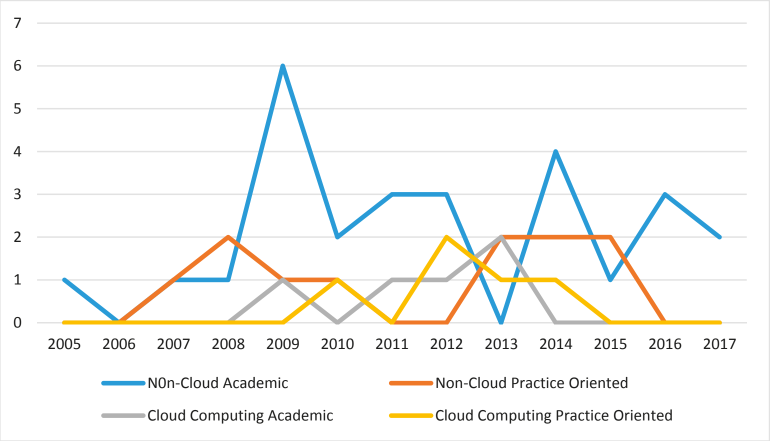


Figure 1. Number of published research on data governance in the last 10 years.

4. Data Governance and Other Governance Domains

With the emergence of new governance domains—to name but the most relevant ones, Corporate Governance, IT Governance, Information Governance, and, more recently, Cloud Computing Governance—it is easy to confuse them, something we have observed in the literature, where authors have interchanged these governance domains as if they are the same thing. It is important, therefore, to differentiate between these domains, and more important to define how they are linked to each other, particularly with respect to data governance. Figure 2 is a simplified view of the interrelations between these domains.

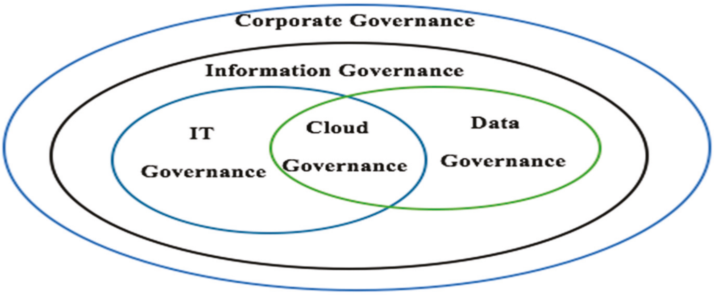


Figure 2. The interrelations between governance domains.

Corporate governance has become important, as effective governance ensures that the business environment is fair and transparent, and that companies can be held accountable for their actions [76].

In contrast, weak corporate governance leads to waste, mismanagement and corruption. According to the Organization for Economic Cooperation and Development (OECD), corporate governance is "*a set of relationships between a company's management, its board, its shareholders, and other stakeholders, corporate governance also provides the structure through which the objectives of the company are set, and the means of attaining the objectives and monitoring performance are determined*" [77].

In recent years, IT has been the backbone of every business [78]. As a result, the concept of IT governance has become more important for organisations. IT governance, similarly to corporate governance, is the process of establishing authority, responsibilities, and communication, along with policies, standards, control mechanisms and measurements to enable the fulfilment of defined roles and responsibilities [79]. Thus, corporate governance can provide a starting point in the definition of IT governance [7]. According to Herbst et al. (2013), IT governance is defined as "*procedures and policies established in order to assure that the IT system of an organization sustains its goals and strategies*" [80]. It is pertinent, however, to note that there is a difference between IT governance and IT functions; this difference is not just about the centralisation or decentralisation of IT structures, but also that it is not the sole responsibility of the CIO [81].

The term "information governance" was introduced by Donaldson and Walker (2004) as a framework to support the work of the National Health Society in the USA. Unfortunately, many organisations have not yet established a clear distinction between information governance and IT governance [82]. Information governance can be viewed as a subset of corporate governance, with the main objectives being to improve the effectiveness and speed of decisions and processes, to reduce the costs and risks to the business or organisation, and to make maximum use of information in terms of value creation [83]. Gartner defines information governance as "*the specification of decision rights and an accountability framework to ensure appropriate behaviour in the valuation, creation, storage, use, archiving and deletion of information*" [84]. The information governance approach focuses on controlling information that is generated by IT and office systems, or their output, but does focus on detailed IT or data capture and quality processes.

Cloud governance is a new term in the IT field; however, it has not been given a clear definition yet [85]. Microsoft defines cloud governance as "*defining policies around managing the factors: availability, security, privacy, location of cloud services and compliance and tracking for enforcing the policies at run time when the applications are running*" [86]. The core of cloud governance revolves around the relationships between provider and consumer, across different business models [87]. The business model should define the way in which an offer is made and how it is consumed. To function at all cloud levels (IaaS, PaaS and SaaS), the business model should be devoid of the type of resources involved.

The literature reported different views on what drives what within these governance domains; in our research, we argue that data governance should be the key driver for all other governance domains, sitting at the heart of everything. The most debated relationship among these governance domains has been that of information governance and data governance, where numerous schools of thought, including the Data Governance Institute, have consistently used information and data governance interchangeably, connoting the understanding that the two terms mean the same thing. A very recent paper, published only in 2016, as part of the proceedings of the 28th Annual Conference of the Southern African Institute of Management Scientists, presented a systematic analysis to prove that data governance is indeed a prerequisite for information governance, and hence the argument was extended to state that data governance must become an ingrained part of both corporate governance and IT governance [88]. Figure 3 provides an illustration of the advocated hierarchy of these governance domains, showing also the difference between management and governance.

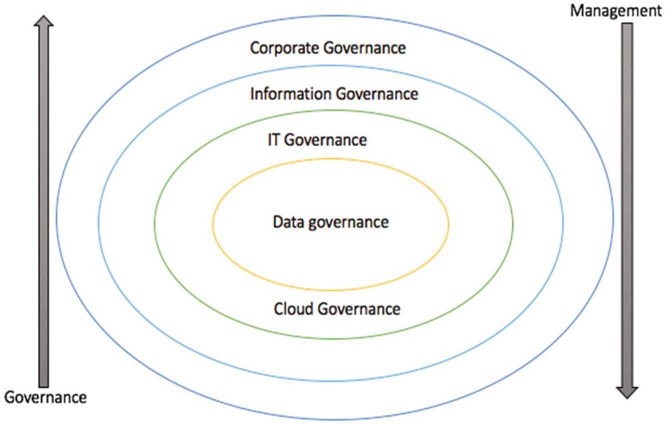


Figure 3. The hierarchy for the difference between management and governance.

5. Data Governance Taxonomy

To construct a holistic taxonomy, we must determine the key dimensions of data governance. This adopted dimension-based approach allows for the categories in the taxonomy to be broken down into discrete areas. A dimension-based approach allows more flexibility in placing content into various nodes, represented by the dimension to which they belong. In the context of data governance, this approach will allow users to manage data governance content more efficiently. Successfully achieving this could be a potentially complex process, and consequently requires more investigative effort and the involvement of different stakeholders. Therefore, the taxonomy for data governance was developed following exploratory and qualitative research, where the method employed was merrily based on a combination of analysing the relevant knowledge in the public domain, resulting from the above described systematic literature review (Section 3) and following the analytic theory [89].

The analytic theory has been useful in understanding the data governance aspects of traditional IT and cloud technology. Sein M. et al. [89] state that "*the analytic theory is used to describe or classify specific dimensions or characteristics of individuals, groups, situations, or events by summarizing the commonalities found in discrete observations. Frameworks, classification schema and taxonomies are numerous in IS*". The analytic theory has been chosen as a concept for this study to identify data governance dimensions for the cloud services. To use analytic theory in making data governance dimensions, we follow three steps. Firstly, understanding the state of the art of data governance for traditional IT and the cloud. Secondly, identifying specific dimensions or characteristics of data governance and cloud computing. Finally, developing the key data governance dimensions for cloud computing, based on the definitions of data governance and factors presented in the literature review, which will construct the desired taxonomy. The adopted approach is considered expedient in expounding a sound theoretical foundation for the study. This approach is used to contextualise the research, for which authors chose the contents that were relevant for the study and how these were employed in order to reach a scientific conclusion. Such an approach is considered essential, following a set of processes or procedures in undergoing a systematic review, which can be verified or validated scientifically.

To the best of the authors' knowledge, and following the aforementioned research approach, there is no published research that defines the key dimensions of data governance for cloud computing. In contrast, for traditional IT (non-cloud), there is some reported research, albeit not much. As illustrated above, data governance for non-cloud and cloud, although showing some similarities at a higher level, differs significantly in details, in addition to some new factors related only to cloud technology. Figure 4 shows the two main classes of data governance, considered as sub-taxonomies: data governance for non-cloud computing, referred to herein as traditional data governance, and data governance for cloud computing, referred to herein as cloud data governance.

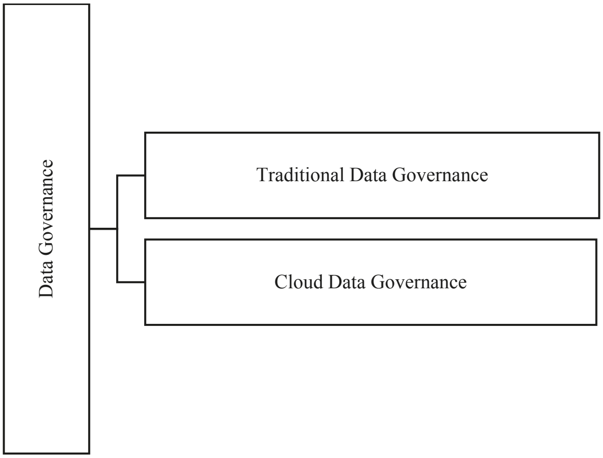


Figure 4. Two main blocks of the data governance taxonomy.

5.1. Traditional Data Governance

As shown in the systematic literature review above, the literature on traditional data governance is still considered insufficient. Some authors expressed their subjective views on aspects of data governance; this subjectivity is driven by the fact that there is no single approach to implementing standard data governance for all types of organisations [4]. This means each organisation's approach to data governance could be different. It is, therefore, very difficult to capture all the different views; instead, after further analysis of the relevant literature, we could identify common aspects of data governance which most authors seem to agree upon. Therefore, traditional data governance could be classified into three main categories: people and organisational bodies, policy, and technology, as shown in the simplified taxonomy below (Figure 5). This is followed by extended descriptions and classification of each aspect.

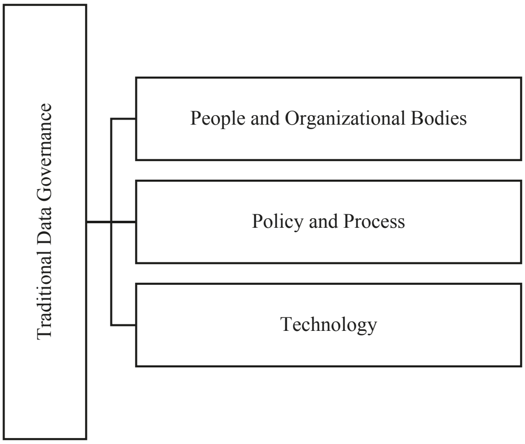


Figure 5. Traditional data governance taxonomy.

5.1.1. People and Organisational Bodies

Data governance will influence the mix of data stakeholders involved in data-related decisions and actions in an organisation, as well as the amount of effort required of each stakeholder. Therefore, in traditional data governance, the people and organisational bodies play important parts when organisations implement data governance for their business [90]. The element of people and organisational bodies in data governance can be defined as any individual or group that could affect or be affected by the data under discussion. People in traditional governance have many tasks, including authority, data stewardship, business rules, collaboration, accountability and culture attitude [91]. The people and organisational bodies element, in the context of traditional data governance, could include the following: data governance office, data governance council, executive sponsorship, chief information officer (CIO), data management committee, compliance committee and data stewards; each has specific roles and responsibilities within their organisations. Figure 6 below summarises the most important aspects of this class of traditional data governance, as agreed by most reported literature.

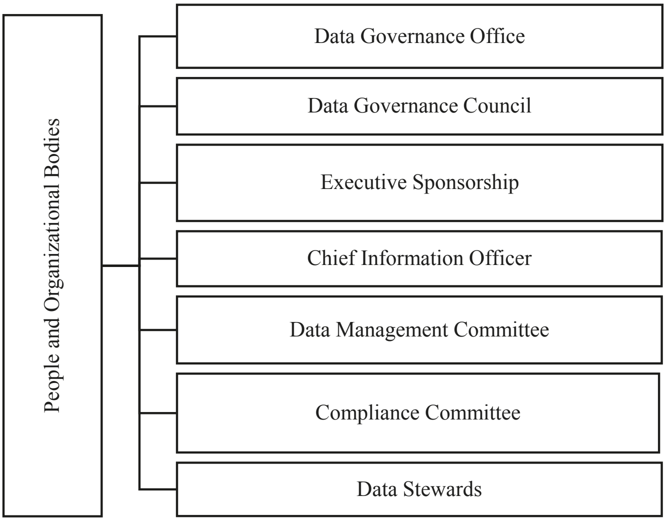


Figure 6. People and organisational bodies taxonomy in traditional data governance.

5.1.2. Policy and Process

Data governance policy is a set of measurable acts and rules for a set of data management functions in order to ensure the benefit of a business process [92]. Regarding data governance processes, they describe the methods used to govern data; these processes should be standardised, documented and repeatable. According to IBM Institute [69], data governance policies and processes should be crafted to support regulatory and compliance requirements for data management functions. The policy and process aspects in traditional data governance could include principles, policies, standards and process, as displayed in Figure 7.

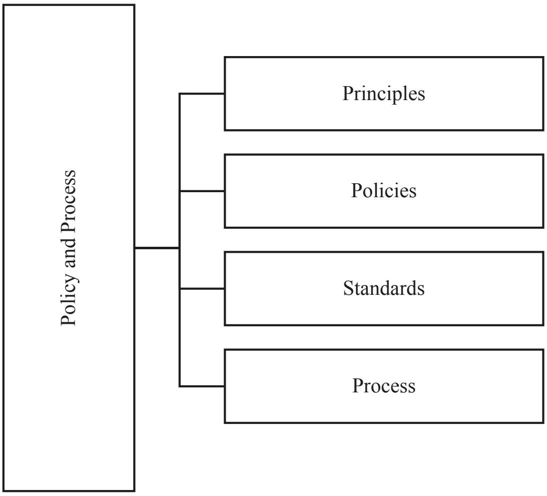


Figure 7. Policy and process elements in traditional data governance.

5.1.3. Technology

Technology is an integral factor for data governance; it is through technology that we can ensure automation and enforce and control data governance policies. However, the role of technology comes after an approved data governance policy and process. Technology in the context of data governance represents the engineering methods that are responsible for reflecting its policies and practice in a measurable way. Therefore, a fit-for-purpose plan for using technical tools to support data governance polices, within the context of roles, responsibilities, and accountabilities, must be established [4,66]. The simplest forms of technology reported for traditional data governance could include hardware, software and monitoring tools, as depicted in Figure 8.

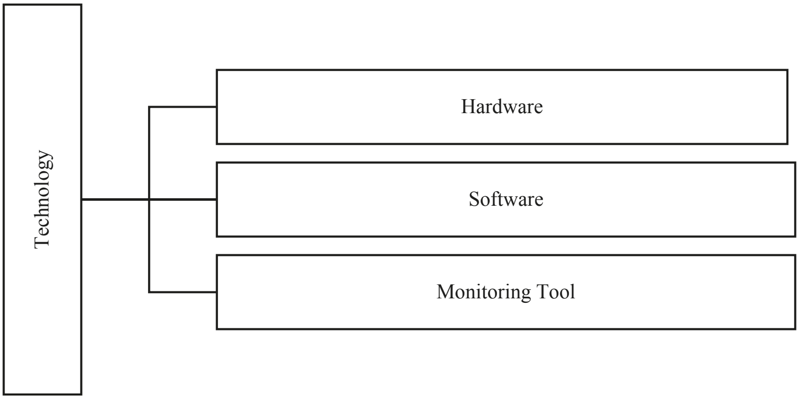


Figure 8. The technology elements of traditional data governance.

5.2. Cloud Data Governance

The impediment to the wider adoption of the cloud computing model has been linked primarily to aspects related to the data governance environment [42,53,60]. While security seems to be the most cited challenge to cloud adoption, Farrell [93] shows that 41% of the security problems in the cloud are related to governance and legal issues. Data governance is considered to be one of the most important aspects of cloud governance [25]. Data governance programmes, built for on-premises IT infrastructure, cannot be deployed for cloud infrastructure and service provisioning, which would require completely new requirements, design and implementation [53,93]. Undoubtedly, the area of cloud data governance is becoming a topic of the coming decades [60,73], although it is still under-researched by both academia and industry, due to its novelty [7,9]. As discussed above, data governance is still underdeveloped and under-practised, even for traditional IT infrastructures, let alone cloud computing environments [4,94]. This is evidenced by the results of the systematic literature review discussed above, where only 11 records discussing data governance for cloud computing were reported. Governance in the cloud needs to understand, moderate and regulate the relationships between different cloud actors or stakeholders in terms of roles and responsibilities [24]. Data governance is meant to classify and assign responsibilities, communication, labelling and policies [57]. There are few studies reporting on data governance for the cloud services. Almost all existing work on data governance for cloud computing focuses on accountability and interoperability [57,60]. Accountability could be addressed at different levels, technological, regulatory and organisational [95].

There is a strong consensus that cloud computing will lead to change in the strategy of traditional data governance in organisations [96]. Cloud data governance is the main focus area in this research, where the aim is to construct a taxonomy that represents the different classifications of this domain. To recall, to the best of the authors' knowledge, this is the first such attempt reported, following the most comprehensive and up-to-date literature review. Figure 9 is a high-level taxonomy of cloud data governance, compiled from the analysis of relevant literature, identified from the systematic literature review. The subsequent sections contain further description of every sub-class.

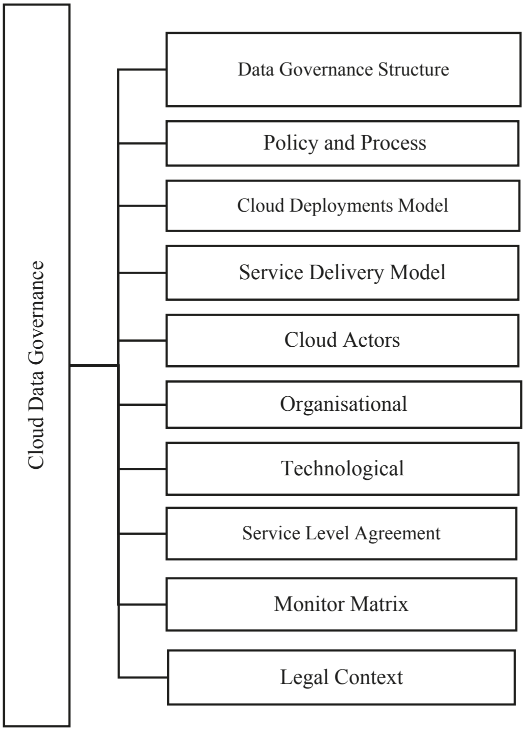


Figure 9. A Cloud Data Governance Taxonomy.

5.2.1. Data Governance Structure

Designing a data governance structure is an important factor in ensuring that requisite roles and responsibilities are addressed throughout the enterprise at the right organisational levels [13]. Several common data governance roles have been identified in existing studies, including the following: executive sponsorship, data management committee, compliance committee, data stewardship team, cloud manager, cloud provider member, IT member and legal member [9,97]. These roles must collaborate to formulate data governance bodies. Figure 10 shows an example of a typical cloud data governance structure.

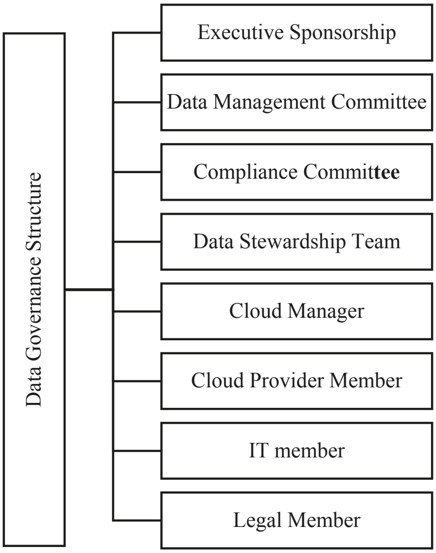


Figure 10. Cloud data governance structure.

5.2.2. Data Governance Function

This refers to master activities for data governance, including functions which data governance teams must take into account when implementing data governance programmes [98]. Establishing consistent policies, standards, and operating processes to ensure the accuracy, availability, and security of data should be part of the data governance strategy, as well as defining the organisation's data assets [3,37]. Therefore, the data governance team must define all data governance policies that address cloud consumers' concerns. The data governance functions can support organisations to make cloud service decisions, such as the geographic distribution of data stored, processed, and in transit; regulatory requirements; data management requirements; and audit policies [99]. Effective data governance in cloud computing requires transparency and accountability, which leads to appropriate decisions that foster trust and assurance for cloud consumers [100]. The outcomes from data governance function activities include standard, procedure, compliance, transformation, integration, management, auditability, transparency, policies, principles and processes. This is considered the master dimension for data governance, but it must comply with other dimensions to develop effective data governance. Figure 11 shows the cloud data governance function and its concerns for cloud computing.

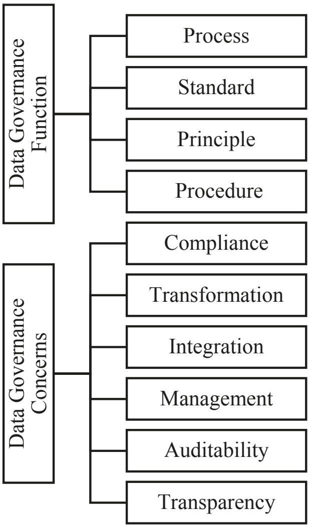


Figure 11. Cloud data governance function and its concerns for cloud computing.

5.2.3. Cloud Deployment Model

This is an important factor to consider in data governance. There are primarily four cloud deployment models, which differ in their provisions; these are the public, private, hybrid and community cloud deployment models. To address data governance, the level of risk and complexity of each cloud deployment must be taken into consideration [18]. According to [110] the implementation of data governance varies greatly, based on the adopted cloud deployment. Figure 12 shows cloud deployment model types to be considered when implementing a cloud data governance programme.

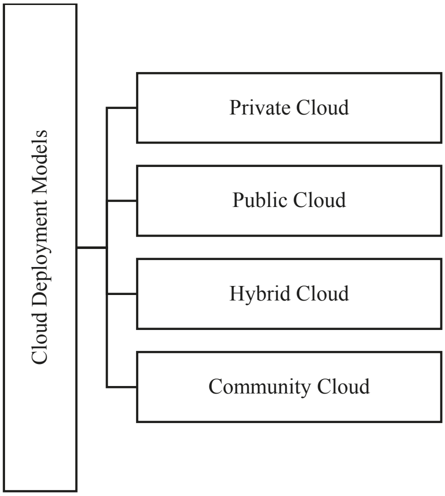


Figure 12. Cloud deployment model types for cloud data governance.

5.2.4. Cloud Service Delivery Model

Cloud services can be categorised into three delivery models: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) [101]. Depending on the model, the cloud consumer will have a differing level of control over their data [61] and each model will require a different approach to data governance and management. Therefore, the data governance teams must consider all the characteristics of the service delivery model and define appropriate policies to enforce control roles and responsibilities. Figure 13 shows the cloud service delivery model to be considered when implementing cloud data governance.

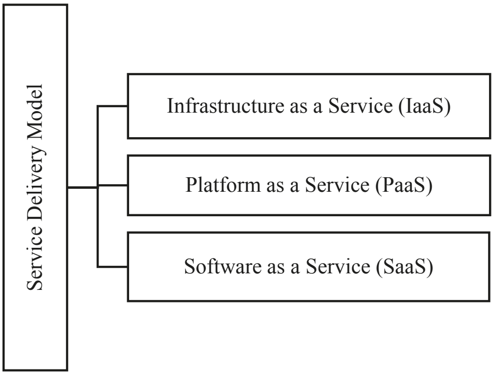

Service Delivery Model
Infrastructure as a Service (IaaS)
Platform as a Service (PaaS)
Software as a Service (SaaS)

Figure 13. Cloud service delivery model for cloud data governance.

5.2.5. Cloud Actors

The actors are also a critical factor in defining cloud data governance. "Cloud actors" refers to individuals or organisations that participate in processes or transactions, and/or perform tasks in the cloud computing environment. NIST's cloud computing reference architecture distinguishes five major actors: the cloud consumer, the cloud provider, the cloud auditor, the cloud carrier and the cloud broker [18]. Each cloud actor has special roles and responsibilities in any one cloud provision, so a data governance programme must clearly define the roles and responsibilities for all cloud actors [102]. Figure 14 shows the cloud actors in cloud data governance.

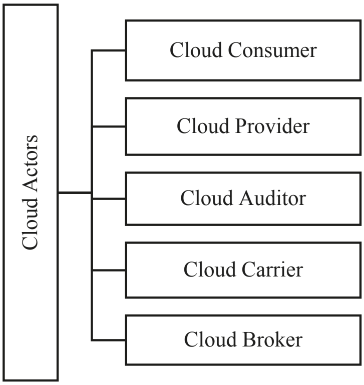


Figure 14. Cloud actors in cloud data governance.

5.2.6. Service Level Agreement (SLA)

One key issue for the cloud consumer is the provision of governance for data which they no longer directly control [103]. Contractual barriers increase between cloud actors. An SLA is an agreement that serves as the foundation of expectation for services between the cloud consumer and the provider [100]. The agreement states what services will be provided, how they will be provided, and what happens if expectations are not met; therefore, an SLA is pivotal in data governance. Thus, the cloud consumer and provider must negotiate all aspects of data governance before developing the SLA. As a result, these agreements are in place to protect both parties. Before evaluating any cloud SLA, cloud consumers must first develop a strong business case for the cloud services, with data governance level policies and requirements and a strategy for their cloud computing environment. The SLA should contain a set of guidelines and policies to assist client organisations in defining governance plans for data which they may choose to move to a cloud provider [104]. These must comply with legal and regulatory requirements. All of these policies can be negotiable between the cloud consumer and cloud provider, to identify the target level of data governance before establishing the contract. The SLA for cloud data governance includes data governance functions; data governance requirements, roles and responsibilities; and data governance metrics and tools. Figure 15 shows the SLA elements for cloud data governance.

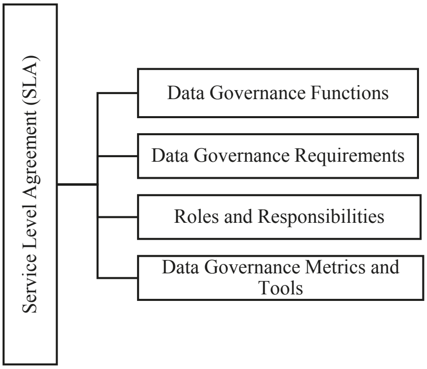

Service Level Agreement (SLA)
Data Governance Functions
Data Governance Requirements
Roles and Responsibilities
Data Governance Metrics and Tools

Figure 15. Service Level Agreement (SLA) elements for cloud data governance.

5.2.7. Organisational Context

Data governance is a major mechanism for establishing control over an organisation's data assets and enhancing their business value [105]. It is also a critical element of implementing a sustainable data management capability, which addresses enterprise information needs and reporting requirements. Organisational factors are important for data governance to be successful [8]. Data governance requires change management in the organisation, in addition to the participation and commitment of IT staff, business management and senior-level executive sponsorship in organisations [37]. Moreover, top management support is considered to be the critical success factor in implementing data governance [61]. Staff in organisations need to learn data governance functions, demanding top management support to enhance the organisation's staff skillset. The organisational context means defining all internal factors that organisations must consider when they manage risks [14]. There are three perspectives for organisational context: the strategic, tactical and operational perspectives. Data governance for cloud computing services should comply with these perspectives. The organisational context for cloud data governance includes organisation charts, organisation vision and mission, organisation strategy, the business model, decision-making processes, training plan, communication plan and change management plan. Figure 16 shows the organisational context elements of cloud data governance.

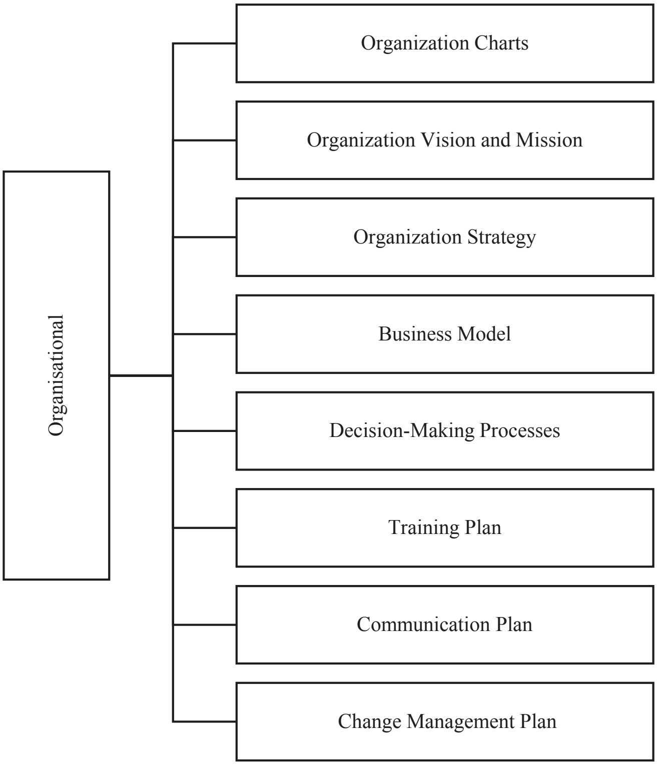


Figure 16. Organisational context of cloud data governance.

5.2.8. Technical Context

Technology is also a key element for data governance success [8]. The technical context represents the issues related to data which will affect the decision of cloud computing adoption and data governance implementation for cloud computing services [106]. Therefore, a lack of technology is considered to be a barrier to successful data governance. Technical factors encapsulate data management issues that affect organisations' strategies, such as security, privacy, quality and integrity. Therefore, it is incumbent upon organisations implementing data governance to assess all technological characteristics available in their organisation, in order to effectively implement data governance. The technical issues that could have an impact on the implementation of data governance for cloud services include availability, reliability, security, privacy, quality, compatibility, ownership, auditing, integrity, data lock-in and performance [106,107]. Figure 17 displays the technological context elements of cloud data governance.

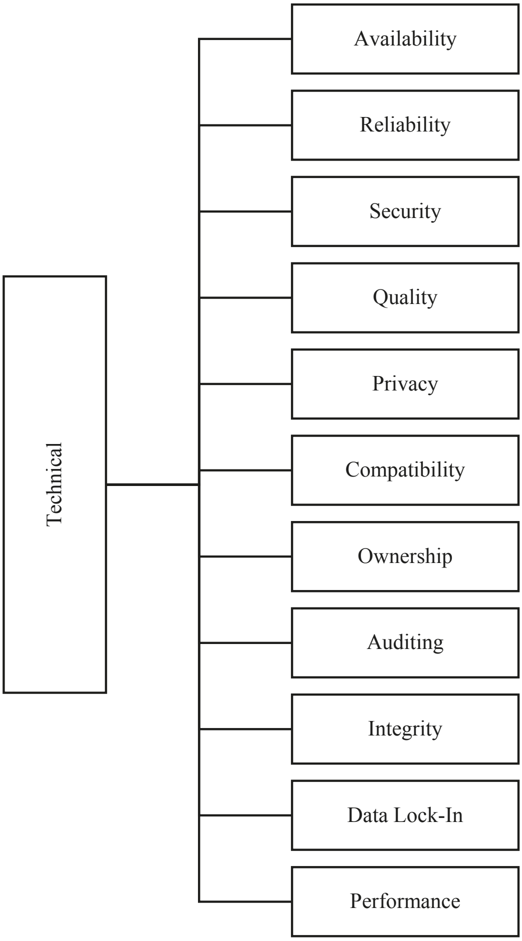


Figure 17. Technical context of cloud data governance.

5.2.9. Legal Context

The legal aspect in this context determines the external and internal laws and regulations related to the data which might affect an organisation's intent to adopt cloud technology [106], which can in turn affect the implementation of an adequate data governance programme for cloud computing services. Therefore, the data governance teams must understand what is implied about data in all relevant contracts before implementing a data governance strategy. Failure to comply with the law when dealing with confidential data erodes trust, which can seriously damage the view of the top management of an organisation regarding the trustworthiness of the cloud provider services [108]. The legal context for cloud data governance includes the Data Protection Act 1998, change of control act and cloud regulations. Figure 18 shows the legal context of cloud data governance.

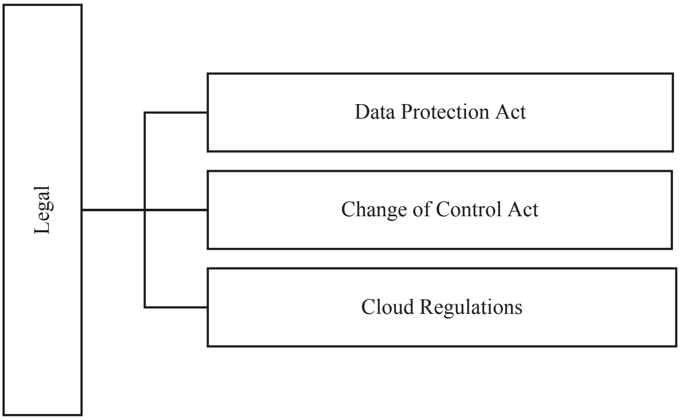


Figure 18. Legal context of cloud data governance.

5.2.10. Monitor Matrix

The monitor matrix in data governance is the exercise of authority, control and shared decision-making over the management of data assets [41]. Measuring and monitoring supports ongoing data governance efforts to ensure that all incoming and existing data meets business rules [109]. By adding a monitoring component to the data governance programme, data quality efforts are enhanced, which in turn renders data much more reliable [109]. Moreover, continuous monitoring ensures compliance with SLAs and the set requirements defined in the data governance strategy [42]. The data governance monitor matrix for cloud computing services includes the cloud control matrix, KPIs and a monitoring tool. Figure 19 shows the elements of the monitor matrix for cloud data governance.

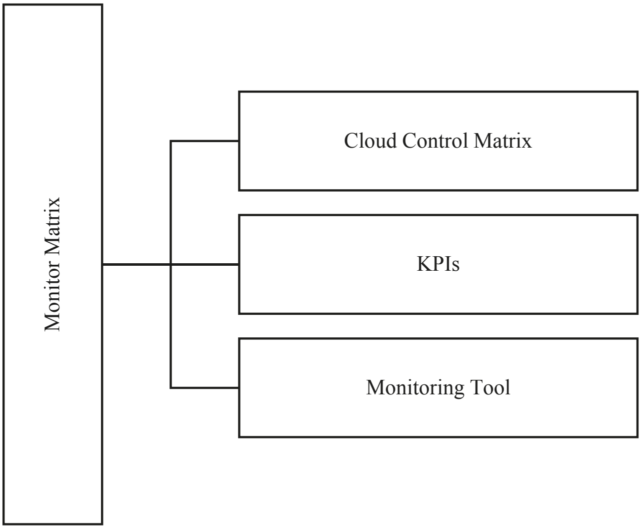


Figure 19. Monitor matrix elements for cloud data governance.

Figure 20 highlights the overall taxonomies of data governance for cloud and non-cloud.

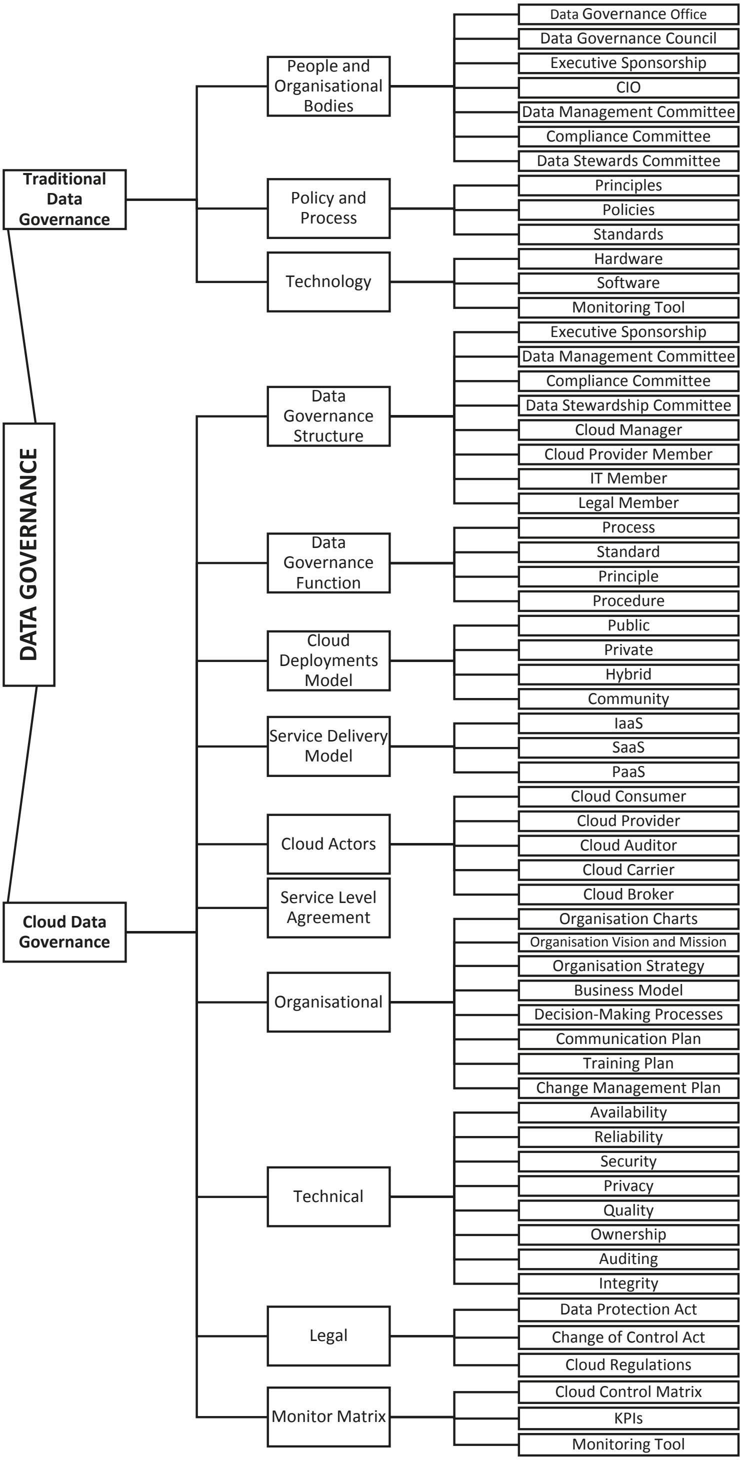


Figure 20. The overall taxonomies of data governance for cloud and non-cloud.

6. Conclusions

Data management solutions alone are becoming very expensive and are unable to cope with the reality of everlasting data complexity. Businesses have grown more sophisticated in their use of data, which drives new demands, requiring different ways to handle this data. Forward-thinking organisations believe that the only way to solve the data problem will be the implementation of effective data governance. With the absence of sufficient literature on data governance in general, and specifically for the cloud paradigm, this paper presents a useful contribution to the relevant research communities. In this paper, we proposed taxonomies for data governance, for both non-cloud and cloud computing networks. A holistic taxonomy that combines all different taxonomies is depicted in Figure 20. These taxonomies are supported by the results of a systematic literature review (SLR), which offers a structured, methodical, and rigorous approach to the understanding of the state of the art of research in data governance. The objective of the study is to provide a credible intellectual guide for upcoming researchers in data governance, to help them identify areas in data governance research where they can make the most impact.

However, this study presents a taxonomy of data governance development requirements for non-cloud and the cloud environments; thus, it does not cover a taxonomy of operational data governance risks that attempts to identify and organize the sources of operational data governance risk. Moreover, this paper is the first of its type, to the best of the authors' knowledge, to cover cloud data governance taxonomy; this presents another limitation, which is related to the lack of relevant literature in this subject domain. The literature shows that most of the existing studies focus on a survey of data governance for non-cloud environments, whilst only three sources in the literature focused on accountability of data governance in cloud computing environments.

Due to the lack of research in this subject area, future work will focus of validation of the proposed taxonomies with specialists from both academia and practitioners. Further research can investigate the application of the proposed taxonomies, especially for cloud data governance, in real world case scenarios. The presented research in this paper shows the lack of research in cloud data governance, which creates an urge for the need to develop a holistic framework for cloud data governance strategy, which highlights the main pillars, processes and attributes to design more specific data governance program. The proposed taxonomies are expected to play an instrumental role in developing such a framework.

Author Contributions: All authors have contributed in this paper by research, scoping writing, and/or reviewing of assigned sections.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nfuka, E.; Rusu, L. Critical Success Factors for Effective IT Governance in the Public Sector Organizations in a Developing Country: The Case of Tanzania. In Proceedings of the ECIS 2010, 18th European Conference on Information Systems, Pretoria, South Africa, 7–9 June 2010.
2. Salami, O.L.; Johl, S.K.; Ibrahim, M.Y. Holistic Approach to Corporate Governance: A Conceptual Framework. *Glob. Bus. Manag. Res* **2014**, *6*, 251.
3. Weber, K.; Otto, B.; Osterle, H. One Size Does Not Fit All—A Contingency Approach to Data Governance. *ACM J. Data Inf. Qual.* **2009**, *1*, 4. [CrossRef]
4. Begg, C.; Caira, T. Exploring the SME Quandary: Data Governance in Practise in the Small to Medium-Sized Enterprise Sector. *Electron. J. Inf. Syst. Eval.* **2012**, *15*, 3–13.
5. Buffenoir, E.; Bourdon, I. Managing extended organizations and data governance. *Adv. Intell. Syst. Comput.* **2013**, *205*, 135–145.
6. Niemi, E. Designing a Data Governance Framework. In Proceedings of the IRIS Conference, At Oslo, Norway, 18 August 2011; Volume 14.
7. Rouse, M. Data governance definition. Available online: www.whatis.techtarget.com (accessed on 9 April 2017).

8. Al-Ruithe, M.; Benkhelifa, E.; Hameed, K. Key dimensions for cloud data governance. In Proceedings of the FiCloud 2016, The IEEE 4th International Conference on Future Internet of Things and Cloud, Vienna, Austria, 22–24 August 2016; pp. 379–386.
9. Wende, K. A Model for Data Governance—Organising Accountabilities for Data Quality Management. In *Proceedings of the 18th Australasian Conference on Information Systems*; University of Southern Queensland: Toowoomba, Australia, 2007; pp. 417–425.
10. Chao, L. (Ed.) Cloud Computing for Teaching and Learning: Strategies for Design and Implementation: Strategies for Design and Implementation. IGI Global, 2012. Available online: https://books.google.com.hk/books?hl=zh-TW&lr=&id=PKWeBQAAQBAJ&oi=fnd&pg=PR1&dq=Cloud+computing+for+teaching+and+learning:+strategies+for+design+and+implementation:+strategies+for+design+and+implementation.+IGI+Global&ots=K2qgWXdeuQ&sig=3MkVNY_ATWYVjYNuthdn6EPAl3g&redir_esc=y#v=onepage&q=Cloud%20computing%20for%20teaching%20and%20learning%3A%20strategies%20for%20design%20and%20implementation%3A%20strategies%20for%20design%20and%20implementation.%20IGI%20Global&f=false (accessed on 1 December 2017).
11. Fu, X.; Wojak, A.; Neagu, D.; Ridley, M.; Kim, T. Data governance in predictive toxicology: A review. *J. Cheminform.* **2001**, *3*, 24. [CrossRef] [PubMed]
12. Prasetyo, H.N.; Surendro, K. Designing a data governance model based on soft system methodology (SSM) in organization. *J. Theor. Appl. Inf. Technol.* **2015**, *78*, 46–52.
13. Panian, Z. Some Practical Experiences in Data Governance. *World Acad. Sci. Eng. Technol.* **2010**, *62*, 939–946.
14. Seiner, R.S. *Non-Invasive Data Governance*, 1st ed.; Technics Publications: New York, NY, USA, 2014.
15. Russom, P. *Data Governance Strategies: Helping Your Organization Comply, Transform, and Integrate*; The Data Warehousing Institute: Los Angeles, CA, USA, 2008.
16. Kamioka, T.; Luo, X.; Tapanainen, T. An Empirical Investigation of Data Governance: The Role of Accountabilities. In Proceedings of the 20th Pacific Asia Conference on Information Systems (PACIS 2016), Chiayi, Taiwan, 27 June–1 July 2016.
17. Poor, M. *Applying Aspects of Data Governance from the Private Sector to Public Higher Education*; University of Pregon: Eugene, OR, USA, 2011; Volume 1277, p. 125.
18. Mell, P.; Grance, T. *The NIST Definition of Cloud Computing Recommendations of the National Institute of Standards and Technology*; NIST Special Publ.: Gaithersburg, MD, USA, 2011; Volume 145, p. 7.
19. Almarabeh, T.; Majdalawi, Y.K.; Mohammad, H. Cloud Computing of E-Government. *Commun. Netw.* **2016**, *8*, 1–8. [CrossRef]
20. Kshetri, N. Cloud computing in developing economies. *IEEE Comput.* **2012**, *43*, 47–55. [CrossRef]
21. Al-Ruithe, M.; Benkhelifa, E.; Hameed, K. Current State of Cloud Computing Adoption—An Empirical Study in Major Public Sector Organizations of Saudi Arabia (KSA). *Procedia Comput. Sci.* **2017**, *110*, 378–385. [CrossRef]
22. Bojanova, I.; Samba, A. Analysis of cloud computing delivery architecture models. In Proceedings of the 2011 IEEE Workshops of International Conference on Advanced Information Networking and Applications (WAINA), Singapore, 22–25 March 2011; pp. 453–458.
23. Forell, T.; Milojicic, D.; Talwar, V. Cloud Management: Challenges and Opportunities. In Proceedings of the 2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), Shanghai, China, 16–20 May 2011; pp. 881–889.
24. Al-Ruithe, M.; Benkhelifa, E.; Hameed, K. A conceptual framework for designing data governance for cloud computing. *Procedia Comput. Sci.* **2016**, *94*, 160–167. [CrossRef]
25. Ko, R.K.L.; Jagadpramana, P.; Mowbray, M.; Pearson, S.; Kirchberg, M.; Liang, Q.; Lee, B.S. TrustCloud: A framework for accountability and trust in cloud computing. In Proceedings of the 2011 IEEE World Congress on Rvices (Services), Washington, DC, USA, 4–9 July 2011; pp. 584–588.
26. Bumpus, W. *NIST Cloud Computing Standards Roadmap*; NIST: Gaithersburg, MD, USA, 2010; pp. 1–3.
27. Ramachandra, G.; Iftikhar, M.; Khan, F.A. A Comprehensive Survey on Security in Cloud Computing. *Procedia Comput. Sci.* **2017**, *110*, 465–472. [CrossRef]
28. Sirimovu, J.U.N.; Artins, O.P. *A Decision Framework to Mitigate Vendor Lock-in Risks in Cloud (SaaS Category) Migration*; Bournemouth University: Poole, UK, 2017.
29. Jennings, B.; Stadler, R. Resource Management in Clouds: Survey and Research Challenges. *J. Netw. Syst. Manag.* **2013**, *23*, 567–619. [CrossRef]

30. Rifaie, M.; Alhajj, R.; Ridley, M. Data governance strategy: A key issue in building enterprise data warehouse. In Proceedings of the iiWAS '09, 11th International Conference on Information Integration and Web-Based Applications & Services, Kuala Lumpur, Malaysia, 14–16 December 2009; pp. 587–591.
31. Neela, K.L.; Kavitha, V. A Survey on Security Issues and Vulnerabilities on Cloud Computing. *Int. J. Comput. Sci. Eng. Technol.* **2013**, *4*, 855–860.
32. Thomas, G. *The DGI Data Governance Framework*; Data Gov. Institute: Orlando, FL, USA, 2006; Volume 20.
33. Cheong, L.K.; Chang, V. The Need for Data Governance: A Case Study. In Proceedings of the 18th Australasian Conference on Information System, Toowoomba, Australia, 5–7 December 2007; Volume 100, pp. 999–1008.
34. Ladley, J. *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program*; Newnes: Boston, MA, USA, 2012.
35. Verhoef, C. Quantifying the effects of IT-governance rules. *Sci. Comput. Program.* **2007**, *67*, 247–277. [CrossRef]
36. De Haes, W.V.G.S. Practiices in IT Governance and Business/IT alignment. *Inf. Syst. Control* **2008**, *2*, 1–6.
37. Otto, B. A Morphology of the Organisation of Data Governance. In Proceedings of the Conference 19th European Conference on Information Systems (ECIS 2011), Helsinki, Finland, 9–11 June 2011; p. 272.
38. HIMSS Clinical & Business Intelligence Committee. *A Roadmap to Effective Data Governance: How to Navigate Five Common Obstacles*; HIMSS Clinical & Business Intelligence Committee: Chicago, IL, USA, 2015.
39. Guillory, K. The 4 Reasons Data Governance Fails. Available online: http://www.noah-consulting.com/experience/papers/4%20Reasons%20Data%20Governance%20Fails%20-%20Guillory.pdf (accessed on 1 December 2017).
40. Héroux, S.; Fortin, A. The influence of IT governance, IT competence and IT-business alignment on innovation. In Proceedings of the 2016 Canadian Academic Accounting Association (CAAA) Annual Conference, Centre St. John's, NL, USA, 2–4 June 2016; pp. 1–36.
41. Salido, J.; Manager, S.P.; Group, T.C.; Corporation, M.; Cavit, D. A Guide to Data Governance for Privacy, Confidentiality, and Compliance. *Microsoft Trust. Comput.* **2010**, *6*, 17.
42. Cloud Security Alliance. *Cloud Data Governance Research Sponsorship*; Cloud Security Alliance: Seattle, WA, USA, 2012.
43. Adelman, S. Without a Data Governance Strategy. *DM Rev.* **2008**, *18*, 32.
44. Otto, B. Data governance. *Bus. Inf. Syst. Eng.* **2011**, *3*, 241–244. [CrossRef]
45. Hallikas, J. *Data Governance and Automated Marketing—A Case Study of Expected Benefits of Organizing Data Governance in an ICT Company*; University of Helsinki: Helsinki, Finland, 2015; pp. 1–89.
46. Kitchenham, B.; Charters, S. Guidelines for performing Systematic Literature Reviews in Software Engineering. *Engineering* **2007**, *2*, 1051.
47. Buffenoir, E.; Bourdon, I. Reconciling complex organizations and data management: The Panopticon paradigm. *arXiv*, 2012.
48. Badrakhan, B.B. Drive toward Data Governance. Available online: http://www.ewweb.com/e-biz/drive-toward-data-governance (accessed on 1 December 2017).
49. Weber, K.; Cheong, L.; Otto, B.; Chang, V. Organising Accountabilities for Data Quality Management-A Data Governance Case Study. In Proceedings of the Conference DW2008: Synergies through Integration and Information Logistics, St Gallen, Switzerland, 27 October 2008; pp. 347–359.
50. Office, D.G. *The State of New Jersey Data Governance Framework Strategic Plan*; New Jersey University: Jersey City, NJ, USA, 2013.
51. Neff, A.; Schosser, M.; Zelt, S.; Uebernickel, F.; Brenner, W. Explicating performance impacts of it governance and data governance in multi-business organisations. In Proceedings of the 24th Australasian Conference on Information Systems (ACIS), Melbourne, Australia, 4–6 December 2013.
52. Kunzinger, F.; Corporation, H.; Haines, P.; Consulting, N.; Schneider, S.; Solutions, V. Delivering a Data Governance Strategy that Meets Business Objectives. In Proceedings of the 14th International Conference on Petroleum Data Integration, Data & Information Management, Houston, TX, USA, 17–19 May 2010.
53. Mary, B.; Mccarthy, P.; Hill, S. Cloud Adoption Points to IT Risk and Data Governance Challenges. Available online: https://www.in.kpmg.com/SecureData/ACI/Files/cloudadoptiondaaprilmay2011.pdf (accessed on 1 December 2017).

54. Soares, S. *The IBM Data Governance Unified Process: Driving Business Value with IBM Software and Best Practices*; Mc Press: Chicago, IL, USA, 2010; p. 153.
55. Allen, C.; Jardins, T.R.D.; Heider, A.; Lyman, K.A.; McWilliams, L.; Rein, A.L.; Schachter, A.A.; Singh, R.; Sorondo, B.; Topper, J.; et al. Data governance and data sharing agreements for community-wide health information exchange: lessons from the beacon communities. *EGEMS* **2014**, *2*, 1057. [CrossRef] [PubMed]
56. Nunn, S. Driving Compliance through Data Governance. *J. AHIMA* **2009**, *80*, 50–51. [PubMed]
57. Imhanwa, S.; Greenhill, A.; Owrak, A. Designing Data Governance Structure: An Organizational Perspective. *GSTF J. Comput.* **2013**, *4*, 1–10.
58. Bhansali, N. *Data Governance: Creating Value from Information Assets*; Auerbach Publications: Boca Raton, FL, USA, 2014.
59. Sarsfield, S. *Data Governance Imperative*; IT Governance Publishing: Cambridgeshaire, UK, 2009.
60. Felici, M.; Koulouris, T.; Pearson, S. Accountability for Data Governance in Cloud Ecosystems. In Proceedings of the 2013 IEEE 5th International Conference on Loud Computing Technology and Science (Cloudcom), Bristol, UK, 2–5 December 2013; pp. 327–332.
61. Groß, S.; Schill, A. Towards user centric data governance and control in the cloud. In Proceedings of the International Workshop on Open Problems in Network Security (iNetSec), Lucerne, Switzerland, 9 June 2011; pp. 132–144.
62. Wendy, Y. Is data governance in cloud computing still a mirage or do we have a vision we can trust. *Softw. World* **2011**, *42*, 15.
63. Mustimuhw Information Solutions Inc. *Data Governance Framework: Framework and Associated Tools*; Mustimuhw Information Solutions Inc.: Duncan, BC, Canada, 2015.
64. Best Practices in Enterprise Data Governance. Available online: https://www.sas.com/content/dam/SAS/en_ca/doc/other1/best-practices-enterprise-data-governance-106538.pdf (accessed on 1 December 2017).
65. Russom, P. Data Governance strateGies. *Bus. Intell. J.* **2008**, *13*, 13–15.
66. Thomas, G. How to Use the DGI Data Governance Framework to Configure Your Program. Data Gov. Inst. Available online: www.DataGovernance.com (accessed on 23 June 2016).
67. Australian Institute of Health and Welfare. *AIHW Data Governance Framework 2014 (AIHW)*; Australian Institute of Health and Welfare: Canberra, Australia, 2014.
68. Loshin, D. *Operationalizing Data Governance through Data Policy Management*; Knowledge Integrity, Inc.: Washington, DC, USA, 2010.
69. Adler, S. *The IBM Data Governance Council Maturity Model: Building a Roadmap for Effective Data Governance*; IBM Corporation: Somers, NY, USA, 2007.
70. Salido, J. Data Governance for Privacy, Confidentiality and Compliance: A Holistic Approach. *ISACA J.* **2010**, *6*, 1–7.
71. Hunter, L. Tools for Cloud Accountability: A4Cloud Tutorial. 2015. Available online: http://www.a4cloud.eu/node/362 (accessed on 4 November 2015).
72. Solutions, C. *Data Governance in the Cloud*; Cloud Industry Forum: York Road Maidenhead, UK, 2013.
73. Tountopoulos, V.; Athens Technology Center. The Problem of Cloud Data Governance. Available online: http://www.cspforum.eu/uploads/Csp2014Presentations/Track_13/The%20problem%20of%20cloud%20data%20governance.pdf (accessed on 4 November 2015).
74. Cloud Security Alliance, Cloud Data Governance Working Group, 2015. Available online: https://cloudsecurityalliance.org/group/cloud-data-governance/ (accessed on 21 May 2015).
75. Alexandria, V. Despite Data Governance Efforts, Eighty-Nine Percent of Federal IT Professionals Are Apprehensive about Migrating IT Services to the Cloud, 2014. Available online: http://www.businesswire.com/news/home/20140909005167/en/Data-Governance-Efforts-Eighty-Nine-Percent-Federal-Professionals#.VeV27Jrovcc (accessed on 12 July 2015).
76. Youssef, A. Exploring Cloud Computing Services and Applications. *J. Emerg. Trends Comput.* **2012**, *3*, 838–847.
77. Government, A. The National Cloud Computing Strategy. *Natl. Broadband Netw.* **2013**, *2013*, 36.
78. Preittigun, A.; Chantatub, W. A Comparison between IT Governance Research and Concepts in COBIT 5. *Int. J. Res. Manag. Technol.* **2012**, *2*, 581–590.
79. Lee, S.U.; Zhu, L.; Jeffery, R.; Group, A.P. Data Governance for Platform Ecosystems: Critical Factors and the State of Practice. *arXiv*, 2017.

80. Herbst, N.R.; Kounev, S.; Reussner, R. Elasticity in Cloud Computing: What It Is, and What It Is Not. In Proceedings of the 10th International Conference on Autonomic Computing, San Jose, CA, USA, 26–28 June 2013; pp. 23–27.
81. Debreceny, R.S.; Gray, G.L. IT Governance and Process Maturity: A Multinational Field Study. *J. Inf. Syst.* **2013**, *27*, 157–188. [CrossRef]
82. Kooper, E.; Maes, M.R.; Lindgreen, R. Information Governance as a Holistic Approach to Managing and Leveraging Information Prepared for IBM Corporation. *Int. J. Inf. Manag.* **2011**, *31*, 1–27.
83. Williams, P.A.H. Information governance: a model for security in medical practice. *J. Digit. Forensics Secur. Law* **2007**, *2*, 57–74. [CrossRef]
84. Gartner, Information Governance. 2016. Available online: http://www.gartner.com (accessed on 17 May 2017).
85. Woldu, L. *Cloud Governance Model and Security Solutions for Cloud Service Providers*; Metropolia Ammattikorkeakoulu: Helsinki, Finland, 2013.
86. Saidah, A.S.; Abdelbaki, N. A new cloud computing governance framework. In Proceedings of the CLOSER 2014, International Conference Cloud Computing Services Science, Barcelona, Spain, 3–5 April 2014; pp. 671–678.
87. Kofi, J.; Kwame, K. Who 'owns' the cloud? An empirical study of cloud governance in cloud computing in ghana. In Proceedings of the 28th European Regional Conference of the International Telecommunications Society (ITS), Passau, Germany, 30 July–2 August 2017.
88. Olaitan, O.; Herselman, M.; Wayi, N. Taxonomy of literature to justify data governance as a prerequisite for information governance. In Proceedings of the 28th Annual Conference of the Southern African Institute of Management Scientists (SAIMS), Pretoria, South Africa, 4–7 September 2016.
89. Sein, M.K.; Henfridsson, O.; Rossi, M. Research essay action design research. *MIS Q.* **2011**, *30*, 611–642.
90. Jansen, W.; Grance, T. Guidelines on Security and Privacy in Public Cloud Computing. Available online: http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-144.pdf (accessed on 1 December 2017).
91. Grant, O.I. *Oklahoma Interoperability Grant Project Oklahoma Interoperability Grant Data Roadmap*; US Department of Health and Human Services: Washington, DC, USA, 2013.
92. Bell, R. *Institutional Data Governance Policy*; Vanderbilt University and Medical Centre: Nashville, Tennessee, 2014; pp. 1–12.
93. Farrell, R. Securing the Cloud—Governance, Risk, and Compliance Issues Reign Supreme. *Inf. Secur. J. Glob. Perspect.* **2010**, *19*, 310–319. [CrossRef]
94. Wende, K. Data Governance Defining Accountabilities for Data Quality Management. In Proceedings of the Italian Workshop on Information Systems (SIWIS 2007, Side Event of ECIS 2007), Carisolo, Italy, 9–14 February 2007.
95. Theoharidou, M.; Papanikolaou, N.; Pearson, S.; Gritzalis, D. Privacy risk, security, accountability in the cloud. In Proceedings of the 2013 IEEE 5th International Conference on Cloud Computing Technology and Science (CloudCom), Bristol, UK, 2–5 December 2013; pp. 177–184.
96. Trivedi, H. Cloud Adoption Model for Governments and Large Enterprises. Master's Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2013.
97. Weber, R.; Iruka, I. *Best Practices in Data Governance and Management for Early Care and Education: Supporting Effective Quality Rating and Improvement Systems*; U.S. Department of Health and Human Services: Washington, DC, USA, 2014.
98. Power, D.; Street, W. Sponsored by All the Ingredients for Success: Data Governance, Data Quality and Master Data Management. *Hub Solut. Des.* **2013**, *2043*, 1–20.
99. Cloud Standards Customer Council (CSCC). *Security for Cloud Computing 10 Steps to Ensure Success*; Cloud Standards Customer Council: Needham, MA, USA, 2012; pp. 1–35.
100. Cloud Standards Customer Council. *Practical Guide to Cloud Service Level Agreements Version 1.0*; Cloud Standards Customer Council: Needham, MA, USA, 2012; pp. 1–44.
101. Bulla, C.M.; Bhojannavar, S.S.; Danawade, V.M. Cloud Computing: Research Activities and Challenges. *Int. J. Emerg. Trends Technol. Comput. Sci.* **2013**, *2*, 206–214.
102. Badger, L.; Grance, T.; Corner, R.P.; Voas, J. *Cloud Computing Synopsis and Recommendations*; NIST Publications: Gaithersburg, MD, USA, 2011.
103. Chawngsangpuii, R.; Das, R.K. A challenge for security and service level agreement in cloud computing. *Int. J. Res. Eng. Technol.* **2014**, 2319–2322. [CrossRef]

104. Cochran, M.; Witman, P.D. Governance and service level agreement issues in a cloud computing environment computing environment. *J. Inf. Technol. Manag.* **2011**, *22*, 41–55.
105. Goals, S.; Dyche, J.; Levy, E. *Data Governance: Getting It Right!* GFT: Stuttgart, Germany, 2015; pp. 1–3.
106. Alkhater, N.; Wills, G.; Walters, R. Factors Influencing an Organisation's Intention to Adopt Cloud Computing in Saudi Arabia. In Proceedings of the 2014 IEEE 6th International Conference on Loud Computing Technology and Science (CloudCom), Singapore, 15–18 December 2014; pp. 1040–1044.
107. Khajeh-Hosseini, A.; Sommerville, I.; Sriram, I. Research Challenges for Enterprise Cloud Computing. *arXiv*, 2010.
108. Confidential, W.S.; Reserved, A.R. *Holistic Approach to Key Challenges Unstructured Data Governance Holistic Approach to Key Challenges*; WhiteBox: Schwyz, Switzerland, 2012.
109. Van der, L.M. *Measuring Data Governance*; Leiden University: Leiden, The Nederland, 2015; p. 89.
110. Cloud Standards Customer Council. Security for Cloud Computing Ten Steps to Ensure Success. Available online: http://www.cloud-council.org/deliverables/CSCC-Security-for-Cloud-Computing-10-Steps-to-Ensure-Success.pdf (accessed on 14 August 2016).
111. Brett. *Data Governance Best Practices and Trends within South African Companies*; Glue Data: Cape Town, South African, 2009.

sustainability

Article

A Hierarchical Feature Extraction Model for Multi-Label Mechanical Patent Classification

Jie Hu [1], Shaobo Li [1,2,*], Jianjun Hu [2,3] and Guanci Yang [1]

[1] Key Laboratory of Advanced Manufacturing Technology of Ministry of Education, Guizhou University, Guiyang 550025, China; jason.houu@gmail.com (J.H.); gcyang@gzu.edu.cn (G.Y.)

[2] School of Mechanical Engineering, Guizhou University, Guiyang 550025, China; jianjunh@cse.sc.edu

[3] Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA

* Correspondence: lishaobo@gzu.edu.cn

Received: 30 November 2017; Accepted: 15 January 2018; Published: 16 January 2018

Abstract: Various studies have focused on feature extraction methods for automatic patent classification in recent years. However, most of these approaches are based on the knowledge from experts in related domains. Here we propose a hierarchical feature extraction model (HFEM) for multi-label mechanical patent classification, which is able to capture both local features of phrases as well as global and temporal semantics. First, a *n*-gram feature extractor based on convolutional neural networks (CNNs) is designed to extract salient local lexical-level features. Next, a long dependency feature extraction model based on the bidirectional long–short-term memory (BiLSTM) neural network model is proposed to capture sequential correlations from higher-level sequence representations. Then the HFEM algorithm and its hierarchical feature extraction architecture are detailed. We establish the training, validation and test datasets, containing 72,532, 18,133, and 2679 mechanical patent documents, respectively, and then check the performance of HFEMs. Finally, we compared the results of the proposed HFEM and three other single neural network models, namely CNN, long–short-term memory (LSTM), and BiLSTM. The experimental results indicate that our proposed HFEM outperforms the other compared models in both precision and recall.

Keywords: text feature extraction; patent analysis; hybrid neural networks; mechanical patent classification

1. Introduction

The World Intellectual Property Organization (WIPO) developed the International Patent Classification (IPC) as a standard taxonomy to classify patents and their applications. According to the report from the WIPO's intellectual property statistics data [1], the current number of worldwide patent applications is rapidly increasing. When an enormous number of patent applications come to the local patent office, it could be a nightmare for the patent examiners. Thus, patent automatic classification (PAC) tasks have drawn much research interest, with many conferences and campaigns hosted around this topic [1,2]. A PAC system is designed for classifying patents into corresponding categories. When a patent application is submitted to a patent office, a search for previous inventions in the field is required, which can be done by retrieving related patents using the classification labels of the submitted patent. The result of this retrieval procedure can be used to decide whether a patent should be granted or not. The patent classification procedure is still time-consuming and labor-intensive work, even for experienced patent examiners, due to the extremely complicated patent language and the hierarchical classification scheme.

In order to find relevant prior arts easier and allow patent examiners pay more attention to the patent innovation content, a PAC system is highly demanded. Significant efforts have been made in many previous studies [3–7]. Many researchers have made contributions to this topic from different perspectives. Some of them focused on the patent text representation [8,9], trying to find the best solution to represent the patent text, while some of them were dedicated to designing the most effective classification algorithms [3,4,7]. Besides this, some others worked on extraction semantic features from patent text [10–12]. Moreover, some researchers tried to identify which parts in a patent document can provide more representative information for classification tasks [5,13]. Almost all these studies highly rely on hand-crafted feature engineering, hence researchers have to design sophisticated feature extractors to extract features from patent documents to achieve competitive performance in the PAC system.

Previous studies showed that distributed representation has great potential to represent texts from both semantic and syntactic perspectives without any external domain knowledge [14,15]. Meanwhile, convolutional neural networks (CNN) can capture salient local lexical-level features and bidirectional long–short-term memory (BiLSTM) can learn long-term dependencies from sequences of higher-level representations in the patent text [16,17].

This paper presents a hybrid hierarchical feature extraction model (HFEM) for multi-label mechanical patent classification. HFEM employs a continuous bag-of-words (CBOW) algorithm to map words in the patent text into word embeddings, which can well represent each word from both syntactic and semantic perspectives with low dimensional vectors. Our algorithm adopts CNN and BiLSTM to capture local lexical-level and long dependency sentence-level features, and uses a supervised feature learning scheme to automatically extract features from patent documents without any expert knowledge.

The main contributions of this paper are summarized as follows:

- A novel hybrid hierarchical feature extraction model (HFEM) for multi-label mechanical patent classification is introduced, which applies deep learning algorithms to patent feature extraction and classification.
- A CNN-based *n*-gram feature extractor is proposed to automatically extract features from a lengthy patent text full of technical and legal terminologies. A long dependency feature extraction model based on bidirectional LSTM is proposed to uncover sequential correlations from higher-level sequence representations.
- We compared HFEM with CNN, LSTM, and BiLSTM. It is shown that HFEM outperforms other compared models in terms of precision, recall and the weighted harmonic mean of precision and recall (F1) scores.

The remainder of this paper is organized as follows. Section 2 presents some related works. Section 3 is devoted to the description of the feature extractor based on CNN and depicts the HFEM architecture. In Section 4, we present the designed datasets and the performance metrics in the experiments. In Section 5, we first define detailed hyper-parameters of the HFEM models, then conduct a series of classification experiments to determine the best variant of the HFEM algorithm. Section 6 presents the comprehensive experiments and analysis. In Section 7, we draw conclusions and present future study directions.

2. Related Works

2.1. Feature Extraction from Text

Previous approaches to represent patent text in the PAC systems of related studies can be roughly classified into two categories: statistical based and semantic based. The bag-of-words (BOW) model [8,18] is a typical, statistically-based text representation approach, which is almost always used in patent analysis studies [1,18]. After stemming, filtering and stop-word removal, the BOW represents each document by the words' occurrences, ignoring their ordering and grammar in the original document. The empirical results [3,19] showed that phrases (*n*-gram) contain more information than single word schemes and could lead to better performance. However, longer phrases may result in the curse of dimensionality issue. For example, in Web 1T five-gram [20], Google Inc. (Mountain View, CA, USA) provides the dataset with its length ranging from unigrams to five-grams. The other variation of BOW is term frequency–inverse document frequency (TF-IDF), which proposes to reduce the dimensions of BOW and increases the weight of words which are relevant to the current document. TF-IDF is often use in patent classification as a text feature extractor [4,5]. However, the BOW discards a large amount of the information from the original document, such as position in the text, semantics, and co-occurrences in different documents. Therefore, it only useful as a lexical level feature.

To address these issues, some scholars used syntactic- and semantic-based approaches to alleviate these problems [9,18,21]. Experienced experts extract representative terms from documents, identify term patterns, and use these patterns find semantic relationships between these terms. WordNet often works as a lexical resource for semantic relation establishment and polysemy-based filters. A semantic-based approach can bring a lot of benefit to the PAC system, but it relies highly on domain knowledge from human experts.

The emerging word embedding encoding approach has shown its capability to capture important syntactic and semantic regularities and identify text contents and subsets of the content. Distributed representation is developed based on the distributional hypothesis, which states that words that appear in the same contexts share similar semantic meaning. That also means words that occur in similar contexts may have similar embeddings. Dongwen et al. [22] proposed a method using the skip-gram algorithm to extract semantic features. When combined with Support Vector Machine for Multivariate Performance Measures (SVM^{perf}), their algorithm achieved state-of-art performance in a Chinese sentiment classification task. Xu et al. [23] designed a document classification framework based on word embedding and conducted a series of experiments on a biomedical documents classification task, which leveraged the semantic features generated by the word embedding approach, achieving highly competitive results. Kuang [15] proposed two algorithms based on the CBOW model and evaluated word embeddings learned from these proposed algorithms for two healthcare-related datasets. The results showed that the proposed algorithms improved accuracy by more than 9% compared to existing techniques.

2.2. Patent Classification

Patent classification is mainly based on the IPC taxonomy, which is a hierarchical structure consisting of sections, classes, subclasses, groups and subgroups. At each sublevel of the hierarchy, the number of categories is multiplied by about 10. As a result, the IPC contains approximately 72,000 categories. Meanwhile, as patent documents are often lengthy and full of technical and legal terminologies, it becomes a highly impracticable and almost impossible task for people who are not from this domain to design a PAC system. Nevertheless, previous methods often come with an elaborately-designed, hand-crafted features extractor to achieve decent performance. The ultimate purpose of a PAC system is to find the best candidate IPC labels, as a patent examiner does. However, the complicated classification taxonomy system and the difficulty of analyzing lengthy patents full of technical and legal terminologies often make it a challenge to develop a high performance PAC system.

A wide variety of algorithms have been proposed for PAC systems. SVM classifiers, parse network of winnows (SNoW), Bayesian classifiers, and neural network classifiers have been investigated for this task. J. Stutzk [5] treated PAC as a multi-label hierarchical classification problem and employed *k* Nearest Neighbors (*k*-NN) and a one-versus-rest SVM to classify patent data with additional geospatial data. From the results, they concluded that coverage error could be significantly decreased, and the application classification system can be improved by incorporating the home addresses of the inventors. Verberne [24] conducted a series of classification experiments with the linguistic classification system (LCS) based on Naive Bayes, Winnow and SVM^{light} in the context of Conference and Labs of the Evaluation Forum Intellectual Property (CLEF-IP) 2011. They found that adding full descriptions to abstracts gives an improvement for classifying documents at the subclass level. Finally, they reached the classification precision score of 74.43%. Lim [6] applied a multi-label Naive Bayes classifier to classify 564,793 registered patents from Korea at the IPC subclass level, using titles, abstracts, claims, technical fields and backgrounds narrative text in the patent as their model's inputs. They found that when taking advantage of all narrative text, they achieved the highest classification precision at 68.31%. Li, Tate et al. [25] proposed a two-layered feed-forward neural network and employed the Levenberg–Marquardt algorithm to train the network for 1948 patent documents from United States Patent Classification (USPC) 360/324. The authors used a stemming approach, the Brown Corpus, to handle most irregular words. They achieved an accuracy of 73.38% and 77.12% on two category sets, respectively.

2.3. Deep Learning in Text Feature Extraction

Recently an increasing number of studies employed deep learning models for text classification, and achieved dominant performance on massive data processing. A CNN is a class of deep, multilayer, feed-forward and back-propagation artificial neural networks [26]. When comparing to standard multilayer neural networks with the same number of hidden units, it takes less time to train and has fewer parameters. A CNN is comprised of one or more convolutional layers and subsampling layers. Each convolutional layer consists of a set of neurons with learnable weights and biases. Each neuron performs dot operations with some inputs and optionally follows it with a non-linearity mapping. The architecture of CNN is designed to learn complex, high-dimensional, nonlinear mapping from large collections of examples [27]. CNN models have been widely used and achieved top performance in computer vision, image classification [28], speech recognition, and Natural Language Processing (NLP) [16], due to their capability of capturing local correlations of spatial or temporal structures. For text modeling, CNN uses a series of convolutional filters on nearby words to extract *n*-gram features at different positions of text. Xiang et al. [29] built a character-level CNN model for several large-scale datasets to show that it could achieve state-of-the-art or competitive results. They treated texts as a kind of raw signal at the character level and applied one-dimensional CNN to them. In order to show CNN's advantage, they also constructed several large-scale datasets. The results show that CNN is an effective method for text classification, especially for large-scale datasets.

Another popular deep learning architecture is recurrent neural networks (RNN), which is designed to handle sequence data and capture long-term dependencies. RNN is able to propagate historical information via a chain-like neural network architecture, which makes them a natural choice for processing sequential data. RNN has been successfully applied to a variety of problems: speech recognition, language modeling, translation, image captioning, etc. While processing sequential data, it looks at the current input x_t as well as the previous outputs of hidden state h_{t-1} at each time step. Unfortunately, it would be a disaster for standard RNN when the gap between two-time steps becomes too large, leading to vanishing/exploding gradients. LSTM [30] is explicitly designed to address this issue, which consists of three-point wise multiplication gates aiming at controlling the ratio of information to forget and to store in the cell states. Gates are a way to optionally let information through. A bidirectional LSTM network is a variation of LSTM which consist of two separate LSTM networks to store the context in both directions; one forward network reading the input sequence from left to right and one backward reading the sequence from right to left. The forward network accumulates any sequence context to the left of each position in the sequence, and the backward net accumulates the sequence context to the right of each position. After processing a sequence in both directions, the outputs of the separate networks are used to compute the final output using the weights and biases from the output neurons. Kiperwasser et al. [17] proposed an approach for feature extraction for dependency parsing based on a BiLSTM encoder. They trained BiLSTM jointly with the parser objective. The results demonstrate the effectiveness of proposed parsers, when compared to the state-of-the-art accuracies on English and Chinese. Ying et al. [16] combined CNN and BiLSTM to extract Chinese events from unstructured data. They employed CNN and LSTM to capture both lexical-level and sentence-level features. The proposed method achieved competitive performance in several aspects on the Automatic Content Extraction (ACE) 2005 dataset.

3. Deep-Learning-Based Hierarchical Feature Extraction Model for Patent Classification

3.1. N-Gram Feature Extraction Based on CNN

Convolutional neural networks were originally applied to computer vision to capture local features [27]. CNN architectures have gradually shown effectiveness in various NLP tasks, and have been used for feature extraction in previous studies, which show that they have the capability of capturing features by themselves [31]. Due to the extremely complicated patent language and hierarchical classification scheme, many previous studies have had to design sophisticated feature extractors to achieve competitive performance for the task. While patent documents are often lengthy and full of technical and legal terminology, it has become a highly impracticable and almost impossible task for people who are not from this domain. Therefore, we employ a CNN-based model (see Figure 1) to extract *n*-gram features from patent texts.

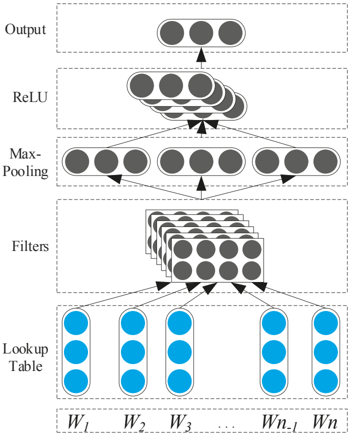


Figure 1. Convolutional neural network with multiple convolutional filters for *n*-gram feature extraction.

We transform each input text into the concatenation of all word vectors, each of which is a word embedding that captures the syntactic and semantic meanings of the word. In this way, we can represent the input text as a vector sequence $V = \{v_1, v_2, \ldots, v_n\}$. The vector sequence V can be converted into a matrix $T \in R^{s \times d}$, where d is the dimension of the word embedding and s is the length of the text. After encoding the input text, we use a convolutional layer to extract the local features, then apply max-pooling and a non-linear layer to merge all local features into a global representation.

Specifically, the convolutional layer extracts local features by sliding continuously window shaped filters with full rows of the matrix T. The width l of the filters is the same as width d of the word embedding. The height h of filters is a number of adjacent rows. Empirical research demonstrates that sliding filters over 2–5 words at a time could achieve strong performance [31]. The filters slide over matrix A and perform convolutional operations. Let $T[i : j]$ denote sub-matrix of T from row i to row j; w_i denotes the i-th filter. Formally, the output of the convolutional layer for i-th filter is computed as:

$$o_i = \mathrm{T}[i : i + h - 1] \otimes w_i \tag{1}$$

$$c_i = f(o_i + b) \tag{2}$$

where $\otimes$ is element-wise multiplication, c_i is the feature learned by the i-th filter, b is the bias, and f is the activation function that can be sigmoid, tangent, etc. In our case, we chose Rectified Linear

Unit ReLU as the nonlinear activation function. After that, we combine all local features map c via a max-pooling function. The max-pooling function applies to each feature map c_i to reduce the dimensionality and capture the highest value from the features. For n filters, the generated n feature maps can be treated as the input of BiLSTM,

$$W = \{c_1, c_2, \ldots, c_n\} \tag{3}$$

Here, comma represents column vector concatenation and c_i is the feature map generated with the i-th filter.

3.2. Long Dependency Feature Extraction Based on Bidirectional LSTM

The LSTM has a series of repeating modules of a neural network for each time step as in standard RNN. At each time step, the cell state c_t (old hidden state h_{t-1}, the input at the current time step x_t) is controlled by a set of gates, the forget gate f_t, the input gate i_t, and the output gate o_t. These gates using previous hidden state h_{t-1} and current input x_i to jointly make decide how to update the current memory cell c_t and the current hidden state h_t. The LSTM transition functions are defined as follows:

Input gates:

$$i_t = \sigma_g(W_i \otimes [h_{t-1}, x_t] + b_i) \tag{4}$$

Forget gates:

$$f_t = \sigma_g\left(W_f \otimes [h_{t-1}, x_t] + b_f\right) \tag{5}$$

tput gates:

$$o_t = \sigma_c(W_o \otimes [h_{t-1}, x_t] + b_o) \tag{6}$$

Cell states:

$$c_t = f_t \ \otimes c_{t-1} + i_t \otimes q_t \tag{7}$$

Cell outputs:

$$h_t = o_t \otimes \sigma_c(c_t \ ) \tag{8}$$

Here, σ_g is the logistic sigmoid function $f(x) = \frac{1}{1+e^{-x}}$, that has an output in [0, 1], σ_c denotes a hyperbolic tangent function, and $\otimes$ denotes the element-wise multiplication.

The LSTM is designed for learning long-term dependencies of time-series data, and it is especially true in the case of bidirectional long–short-term memory networks (BiLSTM), since BiLSTM enables us to classify each element in a sequence while using information from that element's past and future. Figure 2 shows the architecture of BiLSTM. Therefore, we choose BiLSTM to stack to the convolution layer to learn such dependencies in the sequence of higher-level features.

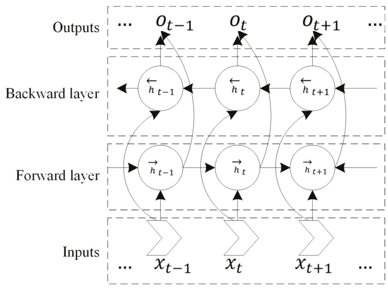


Figure 2. The architecture of HFEM.

3.3. The Architecture of the Hierarchical Feature Extraction Model and Algorithm

Based on the analyses above, a hybrid neural network model based on convolutional and long short-term memory neural networks, is proposed. The architecture of hierarchical feature extraction model (HFEM) is shown in Figure 3, and the algorithm can be detailed as follows:

Input: Narrative text in patent documents.

Output: Probabilities of IPC labels for each patent document.

(1) Split the document into four sections, keep the top 150 words of each section.

(2) Initialize the text with pertained word embedding by looking up the word embedding lookup table, then each patent document is represented by four matrices with dimension 150 × 100.

(3) The four matrices are fed in to four independent CNN channels, each channel applies 128 filters with the dimensions of 3 × 100. The convolutional operation converts four input channels into four feature maps with the dimension of 148 × 128.

(4) Concatenation, maximum, average, and summation strategies are employed to join the feature maps. After four concatenation operations, four kinds of feature maps are obtained with the dimension of 592 × 128, 148 × 128, 148 × 128, and 148 × 128 respectively.

(5) The four feature maps are fed into four BiLSTM networks with 128 forward and backward LSTM neurons. After the BiLSTM network, each feature map is reduced to a matrix of 1 × 256.

(6) Sigmoid function is utilized to calculate the feature vector's probabilities for each label.

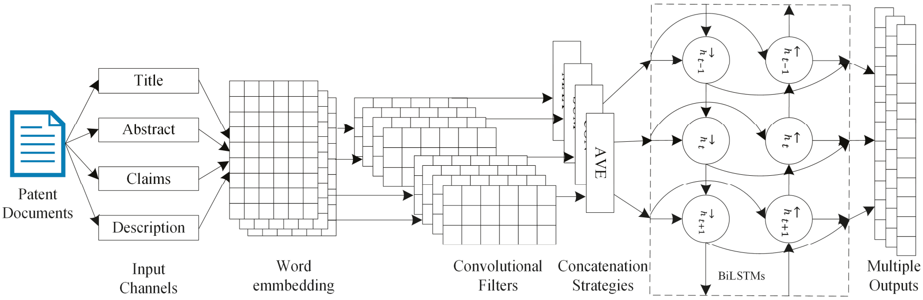


Figure 3. The architecture of HFEM.

Since each patent document consists of multiple narrative text sections, the classification model should take various sections into account. We apply a CNN to extract *n*-gram features from mechanical patent documents with consecutive window filters. After that, we combine all the local *n*-gram feature maps extracted from different sections, via four kinds of concatenation strategies to concatenate local features into global ones. Especially, we did not employ a max-pooling layer to the convolutional network due to the fact that a max-pooling operation will break the continuous sequence organization of selected features. However, the BiLSTM is explicitly designed for sequence data. We do not apply pooling after the convolution operations since we stack the BiLSTM on top of the CNN.

After the CNN layer, the four channel inputs have been converted into four feature maps. We use W_{title}, $W_{abstract}$, W_{claims} and $W_{description}$ to denote the feature maps from the four input channels, respectively. Then we employ concatenation, maximum, average and summation strategies to concatenate features into global ones.

$$W_{CON} = W_{title} \oplus W_{abstract} \oplus W_{claims} \oplus W_{description} \tag{9}$$

Here, $\oplus$ denotes the matrix concatenation operation and W_{CON} denotes the result after concatenation operation, so the shape of the W_{CON} matrix will be four times that of the feature map.

$$W_{MAX} = MAX(W_{title}, W_{abstract}, W_{claims}, W_{description}) \tag{10}$$

where the $MAX()$ operation selects the maximum value from each feature map.

$$W_{AVE} = AVE(W_{title}, W_{abstract}, W_{claims}, W_{description}) \tag{11}$$

where the $AVE()$ conducts the sum operation of each feature map first, then averages the value.

$$W_{SUM} = W_{title} + W_{abstract} + W_{claims} + W_{description} \tag{12}$$

Here, $+$ denotes the summation operation and W_{SUM} denotes the sum of each feature map. Then, the W_{CON}, W_{MAX}, W_{AVE} and W_{SUM} channel features are fed into the BiLSTM jointly. Our approach is different from those methods that use multi-layer CNNs and train the CNN and LSTM separately. We treat the model as an entire network and train the CNN and BiLSTM layers simultaneously. We adopt Adaptive Moment Estimation (ADAM) [32] to minimize the objective function to solve the optimization problem. For the training procedure, we randomly feed the model with a batch of training set until the results converge.

4. Datasets and Evaluation Metrics

In order to check the performance of the proposed algorithm, we established the training, validation and test datasets, all of which are subsets of the CLEF-IP 2011 dataset [33]. The CLEF-IP 2011 dataset consists of more than 2.6 million patent documents from the European Patent Office (EPO) and 400,000 patent documents from the World Intellectual Property Organization (WIPO), representing of 1.35 million (each patent may consist of multiple patent documents) patents filed between 1978 and 2009. These 1.35 million patents contain three kinds olanguage contents, namely English, German, and French.

Generally speaking, a patent document includes bibliographic information, a title, document number, issued date, patent type, classification information, a list of inventors, a list of applicant companies or individuals, abstract, claims section, and a full-text description. More specifically, the title of a patent indicates the name of the patent; the abstract part gives a brief technical description of the innovation; the patent type explains the patent type, and the classification part presents one or multiple class labels. The claim section's main function is to protect the inventors' rights. The description section describes the process, the machine, manufacture, composition of matter, or improvement invented, a brief summary and the background of the invention, the detailed description, and a brief description of its application. The documents also contain meta-information on the assignee, date of application, inventor, etc.

In our experiments, we extracted records from the CLEF-IP 2011 dataset that contain at least one IPC-R classification label, which belongs to section F and the title, abstract, claims, and description textual content in English, namely M-CLEF. Figure 4 shows the yearly distribution of the patent document quantities in the M-CLEF dataset. In the IPC hierarchic taxonomy, section F includes patent applications ranging from the mechanical engineering, lighting, heating and weapons fields. All different document versions for a single patent are merged into a single document with fields updated from its latest versions. After data cleaning, the final M-CLEF dataset consists of 107,302 patent documents. In order to approach a realistic patent classification scenario, we split the dataset into training, validation, and test datasets based on time, so that patents in the training and validation datasets have timestamps earlier than those in the test datasets. More specifically, we used the patents published during 2006 to 2008 as test data and randomly split the rest of patents (all before 2006) into 80% and 20%, as training and validation datasets, respectively. Table 1 shows the detailed numbers for each dataset used in our experiments. We conduct a series of experiments on the multi-label patent classification task. On average, each patent in the training set has 1.4 classification labels at the subclass level.

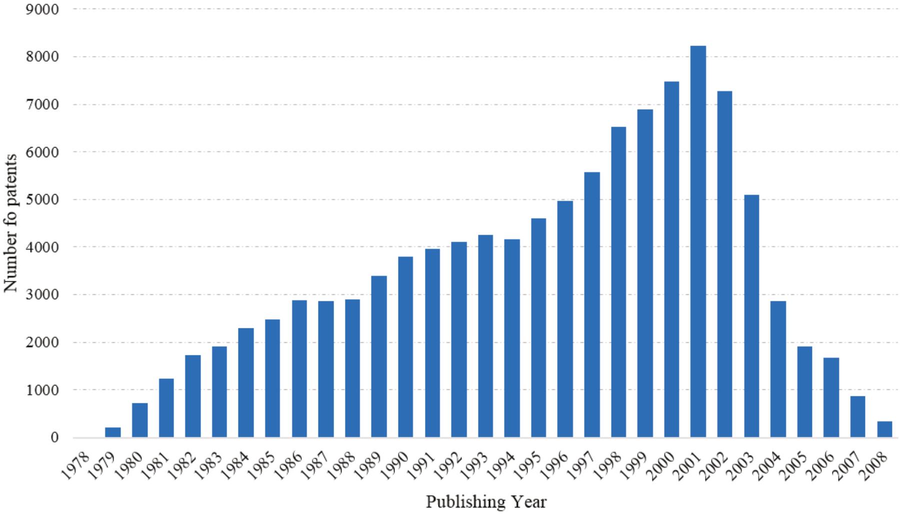


Figure 4. The distribution of the dataset.

Table 1. Brief information of the training, validation and test datasets.

Datasets	# of Documents	Average # Labels per Document	# of Categories
Training dataset	72,532	1.4	96
Validation dataset	18,133	1.4	96
Test dataset	2679	1.5	93

The M-CLEF dataset consists of the four sections of subset title, abstract, claims, and description. In previous studies, researchers found that methods based on only patent title and abstract score lower than those based on the full content of patent documents [34]. Therefore, we used the entire document content as different input channels for our models to determine the effect of various sections in the patent content on the classification performance.

Figure 5 shows the detailed text preprocessing procedure and research framework. First of all, we selected the patents from the CLEF-IP 2011 dataset, which are filled in English and belong to section F. Then, we conducted the followings text preprocessing procedures on the selected patent texts (M-CLEF): (1) Remove all punctuation and convert to lowercase; (2) Replace all contiguous whitespace sequences with a single space; (3) Separate unrelated blocks of text with a newline character. After these text preprocessing procedures, the M-CLEF dataset was converted to the M-CLEF corpus. Then the corpus was fed to the CBOW model to train word embeddings. The remaining descriptions of Figure 5 are presented above.

For each experiment, we used the followings evaluation metrics to evaluate various methods, since we conducted multi-label patent classification experiments. Firstly, we predicted one, five and 10 IPC labels for each patent document respectively. Then we calculated $Precison_{weighted}$, $Recall_{weighted}$, and $F1_{weighted}$ for each prediction.

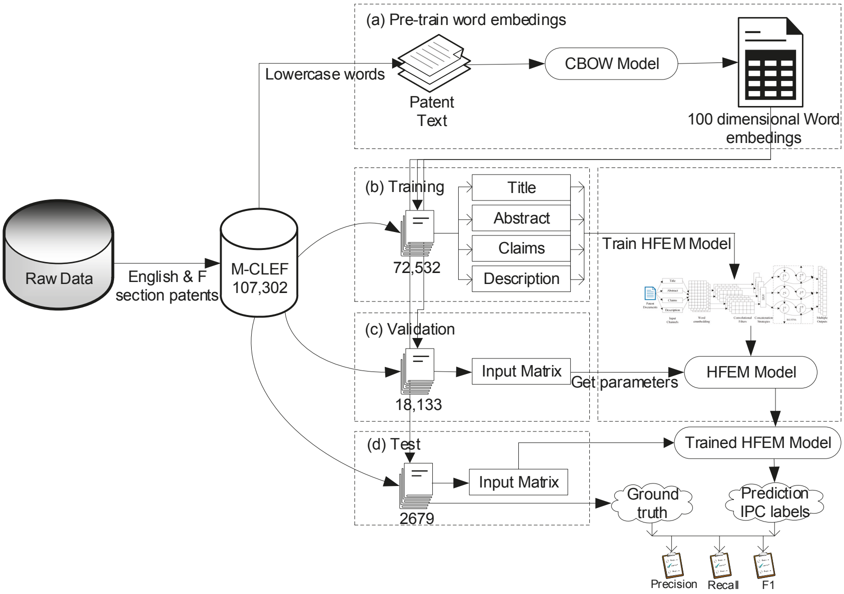


Figure 5. Data preprocessing and data flow diagram.

For each patent document, the number of outcome labels of our approach (prediction labels) that matched the available IPC labels (available labels), without taking the exact order into account, are considered to be true positives (TP). False positives (FP) are the labels predicted by our approach that do not match the available IPC labels. As false negatives (FN), we considered the labels that should have been predicted by our approach, but were not. True negatives (TN) are the labels that, correctly, were not predicted by our approach. Then the precision and recall could be calculated as follows:

$$Precison_{score} = \frac{|\text{available labels} \cap \text{prediction labels}|}{|\text{prediction labels}|} \tag{13}$$

$$Recall_{score} = \frac{|\text{available labels} \cap \text{prediction labels}|}{|\text{available labels}|} \tag{14}$$

After calculating each $Precison_{score}$ and $Recall_{score}$ for each patent, then the weighted average $Precison_{weighted}$, $Recall_{weighted}$ and $F1_{weighted}$ can be calculated as follows:

$$Precison_{weighted} = \frac{1}{Total\ Samples} \sum_{i=1}^{Total\ Samples} Precison_i \tag{15}$$

$$Recall_{weighted} = \frac{1}{Total\ Samples} \sum_{i=1}^{Total\ Samples} Recall\ _i \tag{16}$$

$$F1_{weighted} = 2 \times \frac{Precison_{weighted} \times Recall_{weighted}}{Precison_{weighted} + Recall_{weighted}} \tag{17}$$

The $Precison_{weighted}$, $Recall_{weighted}$ and $F1_{weighted}$ are denoted as P, R, and $F1$ respectively. We use # to denote the number of topmost labels returned by the model, and then we can denote the measures as P@#, R@#, and $F1$@# respectively.

5. Performance Analysis of HFEM with Different Concatenation Strategies

Since the concatenation strategies influence the performance of our algorithm, we implemented the HFEM with different concatenation strategies using Keras [35], a Python deep learning library, which supports efficient symlic dntiation and transparent use of a Graphics Processing Unit (GPU). To benefit from the efficiency of parallel computation of the tensor, we trained the model on a GPU.

First, since the convolutional layers in our model demand fixed-length inputs, we used MAX_LEN to denote the maximum length of text for each input section in the training and test dataset. We employed the CBOW algorithm [36] to pre-train our M-CLEF corpora into word embeddings, with a dimensionality of 100, 200, and 300.

We initialized each section text that had a length less than MAX_LEN with a $(MAX_{LEN} - CUR_{LEN}) \times d$ zero vector at the end of the representation matrix. Nevertheless, for the texts which had a length longer than MAX_LEN, we simply cut extra words at the end of these texts to reach MAX_LEN. Therefore, all text was converted to representation matrixes of the same shape. The shape of the final matrices is $MAX\_LEN \times d$. For the classification labels, the whole number of categories in the M-CLEF dataset was 96. Each patent had at least one classification label. In the case of multi-label classification task, we represented the joint set of labels with a binary indicator matrix. For example, given several patents with labels as follows—$p_1 = \{F03G\}$, $p_2 = \{F16B, F16L\}$, $p_3 = \{F16B\}$—the label set should be $S = \{F03G, F16B, F16L\}$. Each patent could be represented as one row of the label matrix with binary values; the label matrix $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ represents label $F03G$ in the first patent, labels $F16B$ and $F16L$ in the second patent and labels $F16B$ in the third patent. The non-zero elements correspond to the subset of labels.

In order to determine which concatenation strategy offered the best effect, we implemented four versions of HFEM with different concatenation strategies. The detailed configuration parameters are listed in Table 2. Each variant of HFEM takes four input channels and each channel uses 150 word vectors to concatenate into a text matrix of 150 × 100. A total of 128 convolutional filters with size of 3 × 100 were applied in the convolutional layer and a ReLU function was subsequently used as the non-linear activation function. Then the concatenation strategy was employed to concatenate the extracted feature maps. The jointed feature maps were fed to BiLSTM, consisting of 128 forward and backward LSTM neurons. Finally, a fully-connected layer with sigmoid activation was used to calculate the probabilities of 96 IPC labels.

Table 2. The parameters of the HFEM model.

HFEM				
channel names	title	abstract	claims	description
training epochs	40			
input size	150 × 100	150 × 100	150 × 100	150 × 100
# of filters	128			
filter size	3 × 100			
activation layer	ReLU			
concatenation strategies	Concatenation	Maximum	Average	Summation
memory size	128			
activation layer	Sigmoid			
# of target classes	96			

Five experiments were conducted to determine which concatenation strategy can provide the best performance, and the statistical results are shown in Table 3. As we can summarize from Table 3, the single channel HFEM achieved decent results, taking the entire text as the single channel input. At the same time, all the other four multi-channel variants of HFEM improved the performance in nine evaluation metrics. In the scenario of predicting one IPC label for each patent, the concatenation scheme obtained the best performance for three evaluating criteria, namely improving the precision, recall and F1 scores by 1.61%, 0.42% and 1.17%, respectively, compared with the single channel HFEM. The maximum scheme had a slight advantage in predicting five IPC labels for each patent, and achieved the best performance for two out of three criteria. Regarding the prediction of 10 IPC labels, the best performance was achieved by different concatenation schemes. Nevertheless, we found that the concatenation scheme obtained the best performance in five out of nine evaluation metrics. Hence, we choose the HFEM with concatenation scheme as our ultimate version model for all experiments.

Table 3. Performance of HFEM with different strategies using the narrative text as input.

Strategies	Precison Top 1 (%)	Precison Top 5 (%)	Precison Top 10 (%)	Recall Top 1 (%)	Recall Top 5 (%)	Recall Top 10 (%)	F1 Top 1 (%)	F1 Top 5 (%)	F1 Top 10 (%)
Single channel	78.93	31.31	18.21	54.57	89.49	94.6	62.8	45.26	29.71
Concatenation	**80.54**	31.69	19.04	**54.99**	90.28	95.59	**63.97**	**46.55**	30.8
Maximum	79.81	**31.92**	**19.28**	54.77	91.05	96.01	63.52	46.32	30.22
Average	80.26	31.63	19.02	54.89	90.67	95.98	63.58	46.1	30.69
Summation	80.21	31.88	19.04	54.87	**90.88**	**96.31**	63.65	45.95	**30.82**

6. Comparison and Analysis with Other Methods

In this section, we first describe the implementation of three baseline models for the classification task, consisting of CNN, LSTM, and BiLSTM, then give a detailed description of how to conduct the experiments. Five groups of experiments are carried out to validate the feasibility and effectiveness of our HEFM model, and then the results and related analyses are presented.

6.1. Experimental Setup

We compare the HFEM with the following baseline neural network models for mechanical patent classification.

1. CBOW+CNN: We converted patent text using word embeddings pre-trained with the CBOW algorithm into the input matrix and then trained a CNN model with 128 filters for classifying mechanical patent documents.
2. CBOW+LSTM: We converted patent text using word embeddings pre-trained with the CBOW algorithm into the input matrix and then trained an LSTM model with 128 memory LSTM units for classifying mechanical patent documents.
3. CBOW+BiLSTM: We converted patent text using word embeddings pre-trained with the CBOW algorithm into the input matrix and trained a BiLSTM model with 128 forward and backward LSTM units for classifying mechanical patent documents.

The detailed hyper-parameters are listed in Table 4. For each baseline method, the training epoch was fixed to 40, and the number of input words was set to 150 when only taking one section from the entire patent document. The number was set to 600 when the entire text was used by the model, and finally, a fully-connected layer with sigmoid activation function was connected to 96 categories from the IPC label matrix.

Table 4. Hyper-parameters for baseline methods.

Hyper-Parameters	CNN	LSTM	BiLSTM
training epochs	40	40	40
input size	600 × 100	600 × 100	600 × 100
# of filters	128	-	-
memory size	-	128	128
max-pooling size	2	-	-
# of target classes	96	96	96

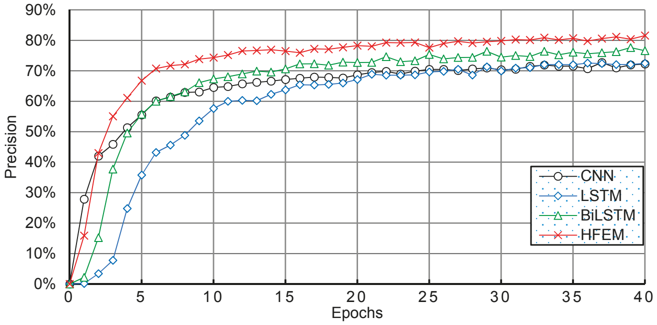
6.2. Experimental Results and Discussion

We conducted a series of experiments on our HFEM and the baseline models, using the entire narrative text from the M-CLEF dataset. The results are shown in Figures 6–11 and Tables 5 and 6.

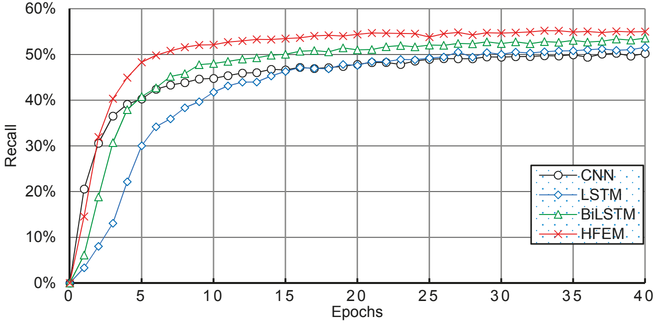
According to Figure 6, it can be seen that HFEM obtained approximate precision, recall, and F1 scores of 81%, 55%, and 64%, respectively, while the best performance of the three baseline models was 78%, 52%, and 64%.This indicates that HFEM improved the performance by 3% in terms of the three evaluation criteria. Figure 6a illustrated the precision scores achieved by four models. HFEM showed its overwhelming advantage in precision when compared with the three baseline methods. Besides, HFEM also demonstrated its superiority in recall, as shown in Figure 6b. From the evaluation criterion of F1, the four approaches show very similar performance in terms of precision and recall, and the results are displayed in Figure 6c. Furthermore, as we can see from Figure 6, the HFEM model converged faster than the three baseline models. The precision, recall, and F1 scores for HFEM tended to converge before 15 epochs, while the other models needed at least 20 epochs to reach steady state.

In addition, we report the performance of these four models for nine evaluation indictors in Table 5. HFEM obtained the best performance in predicting one label for each patent document as well as in predicting five and ten labels. The experimental results demonstrate and verify the feasibility and effectiveness of our HFEM model for mechanical patent classification.

Moreover, we were interested in understanding which section in the patent document has more representative features for classification. We conducted a series of orthogonal experiments by separately using four sections as input and four models as classifiers. More specifically, we used the title, abstract, claims and description separately as input for the CNN, LSTM, BiLSTM, and HFEM models.

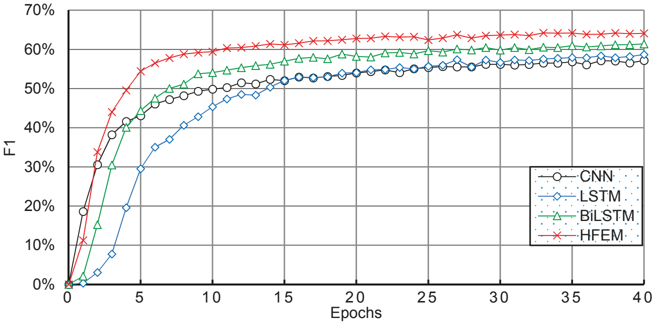


(**a**) Precision scores of four models using the entire text as input.



(**b**) Recall scores of four models using the entire text as input.

Figure 6. *Cont.*



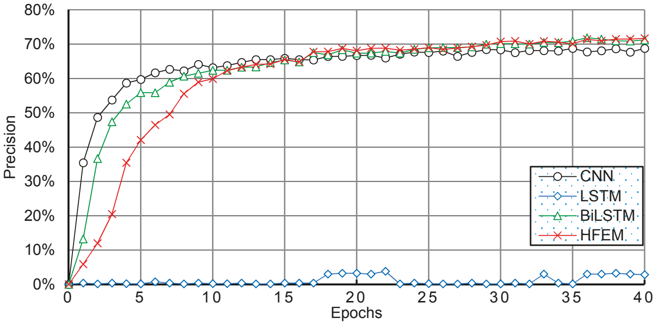
(**c**) F1 scores of four models using the entire text as input.

Figure 6. Performance using the entire text as input.

Table 5. Results of various models using the narrative text as input.

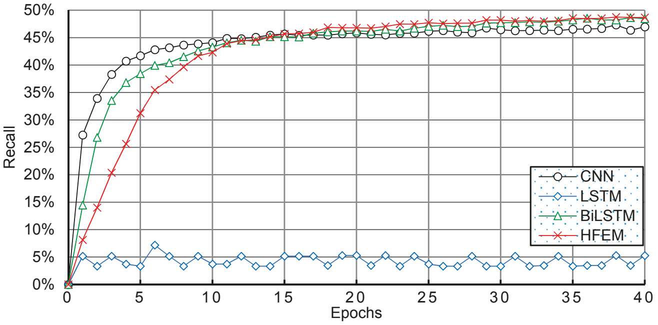
Algorithms	P@1%	P@5%	P@10%	R@1%	R@5%	R@10%	F1@1%	F1@5%	F1@10%
CNN	71.34	29.89	17.43	50.08	86.81	92.93	57.02	43.09	28.35
LSTM	74.44	30.53	18.44	51.96	86.14	92.96	59.26	43.72	29.73
BiLSTM	77.71	30.96	18.83	53.57	88.1	94.67	61.55	44.53	30.24
HFEM	**80.54**	**31.69**	**19.04**	**54.99**	**90.28**	**95.59**	**63.97**	**46.55**	**30.8**

First of all, we used the title section as the input for the models. Figure 7 illustrates the performance of the four models on the title section. From the figure, we can find that the LSTM model achieved quite poor performance whether in terms of precision (see Figure 7a), recall (see Figure 7b), or F1 score (see Figure 7c). The CNN and BiLSTM models obtained decent results, while the convergence of CNN was faster than the other models. Although the convergence of our HFEM model was not as fast as CNN, after 10 epochs, it began to achieve comparable results. After 20 epochs, our model began to outperform the other methods. The best precision of our model was around 72%, while it was only 5% for LSTM. We inspected the M-CLEF dataset and found the title section in the patent documents to be quite short. On average, there are less than ten words in the title section, which explains why the LSTM model achieved poor performance.

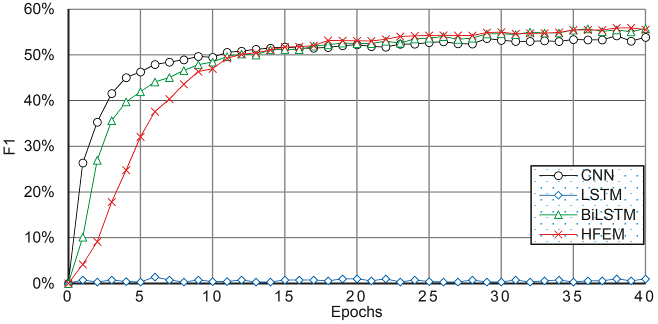


(**a**) Precision scores of four models using the title section as input.

Figure 7. *Cont.*



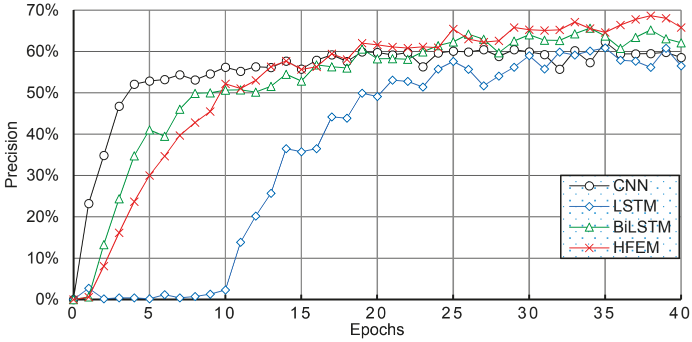
(**b**) Recall scores of four models using the title section as input.



(**c**) F1 scores of four models using the title section as input.

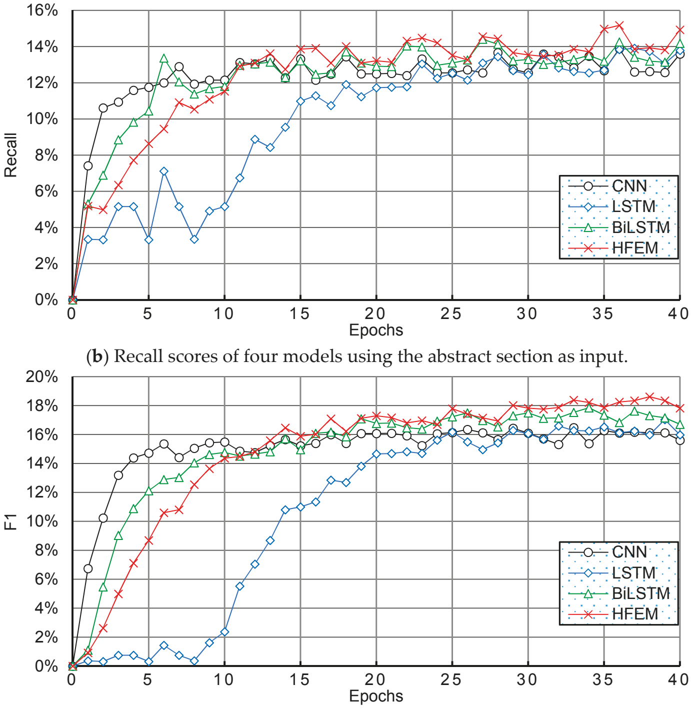
Figure 7. Performance of four models on the title section.

Then, we conducted experiments using the abstract section. Figure 8 shows the performance of the four models on the abstract section. As illustrated in Figure 8a below, the best performance for precision achieved by each model was 60.4%, 63.1%, 67.2%, and 70.3%. At the very beginning, the BiLSTM and CNN models performed better in recall than HFEM, but the HFEM model obtained better recall scores than the others after 15 epochs (see Figure 8b). Furthermore, we found that the curves of Figure 8c are not as smooth as in Figure 7c. After inspecting the M-CLEF dataset, we found that some documents have a missing abstract section, which may have led to the unsmooth curve phenomenon.



(**a**) Precision scores of four models using the abstract section as input.

Figure 8. *Cont.*

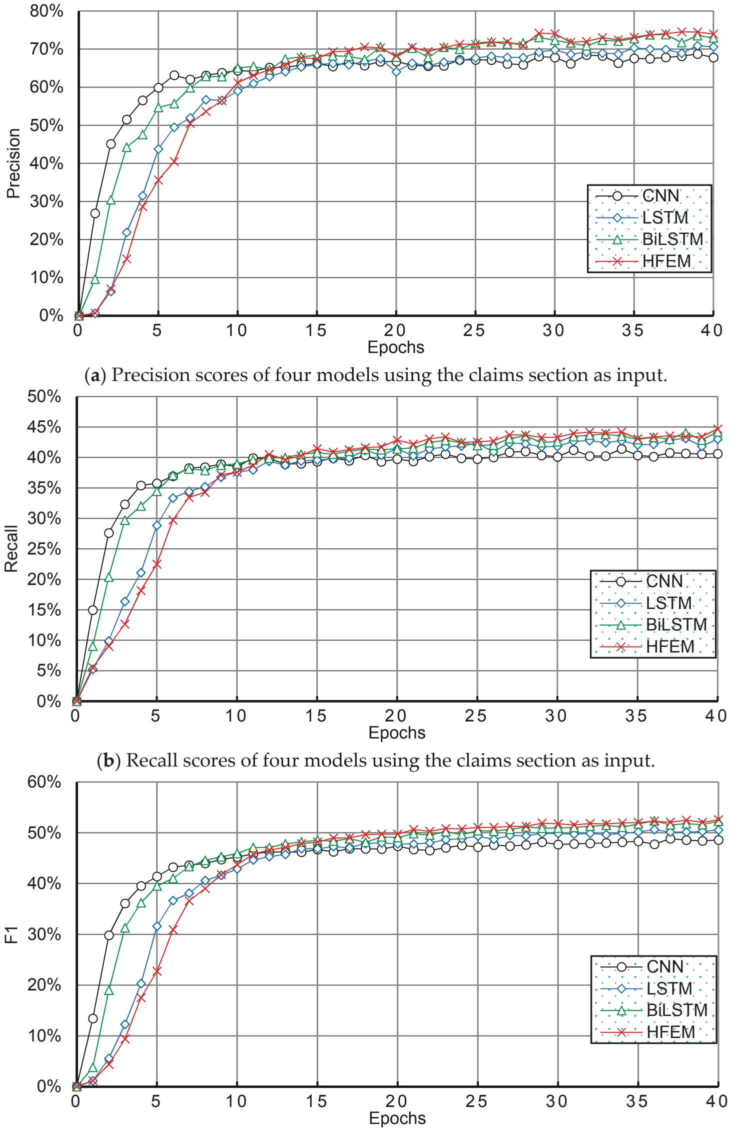


(**b**) Recall scores of four models using the abstract section as input.

(**c**) F1 scores of four models using the abstract section as input.

Figure 8. Performance of four models using the abstract section as input.

After that, the claims section was used in the following experiments. Figure 9 illustrates the performance of the four models using the claims section as input. As shown in Figure 9a, for the first 10 epochs the CNN and BiLSTM models converged rapidly with decent precision scores, while after 15 epochs HFEM began to show its superiority. The LSTM approach obtained a precision score of at least 70%, but the CNN model performed worse than the others. We can find similar phenomena in Figure 9b, where the CNN, LSTM and BiLSTM models show better convergence in the recall scores, but HFEM achieved the best performance at last. Figure 9c shows the F1 scores achieved by the four models, after 25 epochs the four curves reached a relatively steady state. We found that our HFEM model slightly improved the precision, recall, and F1 results. Our results show that the claims section can produce better classification performance as it contains more information than the title or the abstract section.



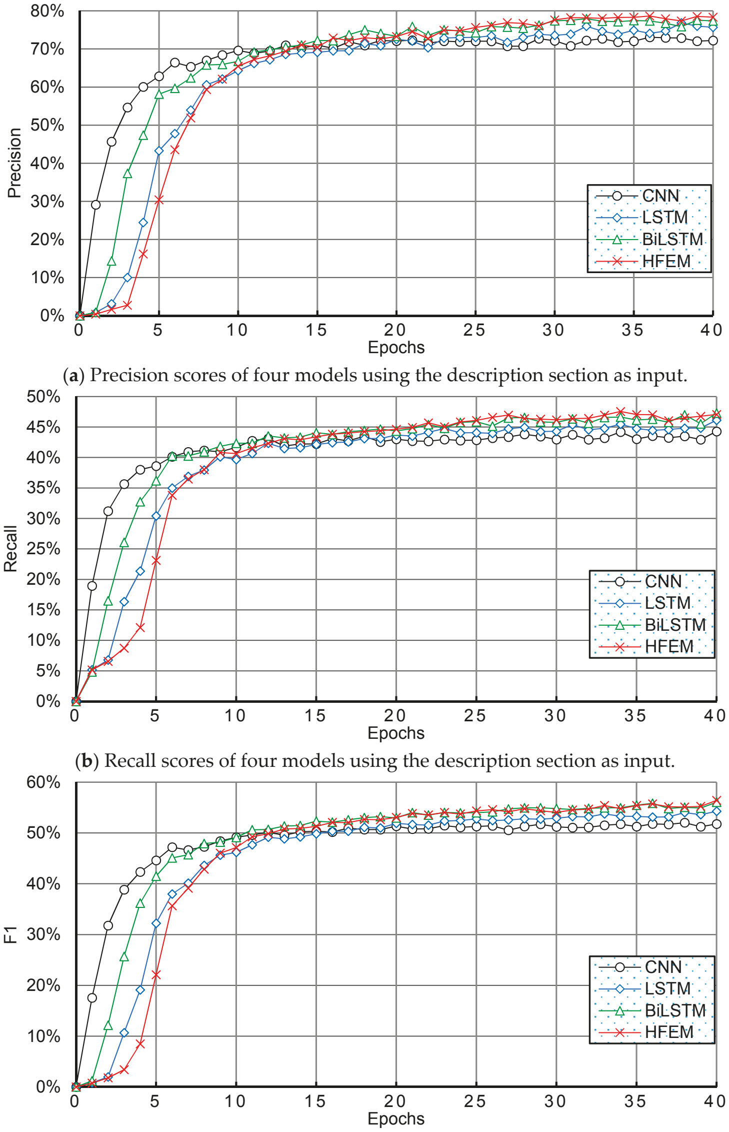
(**a**) Precision scores of four models using the claims section as input.

(**b**) Recall scores of four models using the claims section as input.

(**c**) F1 scores of four models using the claims section as input.

Figure 9. Performance of four models using the claims section as input.

Furthermore, we used the description section as the input for the four models. Figure 10 displays the performance achieved by these four models. We found that using the description section as the input for the models could lead to relatively decent performance. As a result, all the approaches achieved precision scores of more than 70% (see Figure 10a). In addition, our HFEM model achieved the best performance among all the approaches. Compared to previous experiments, using the description section can obtain better performance for all models, because the description section consists of more information than the title, abstract or claims sections. Therefore, it can be concluded that the description section has the most discriminating features for mechanical patent classification.



(**a**) Precision scores of four models using the description section as input.

(**b**) Recall scores of four models using the description section as input.

(**c**) F1 scores of four models using the description section as input.

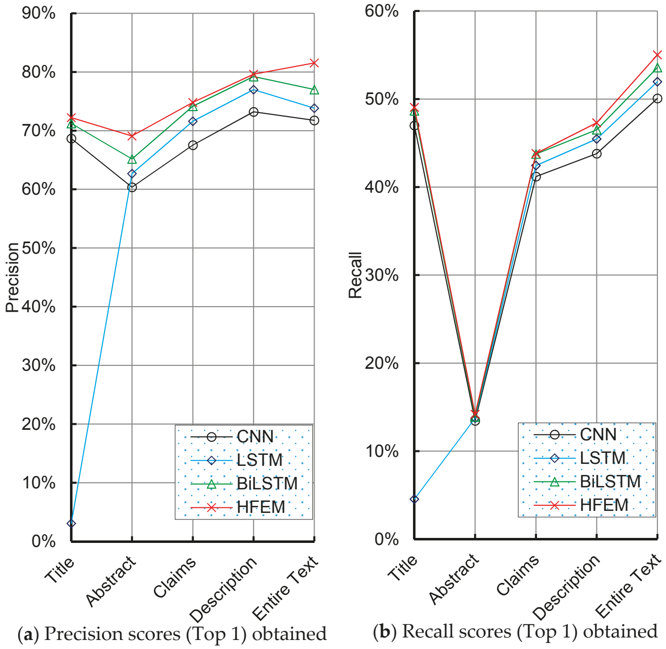
Figure 10. Performance of four models using the description section as input.

According to the five experimental results above, we summarize all detailed experimental results in Table 6. As shown in the table, when predicting one label for each patent document, HFEM achieved better performance in all evaluation metrics regardless of which input scheme was used.

Table 6. Results of various models using the narrative text as input.

Metrics	Algorithms	Input Schemes				
		Title	Abstract	Claims	Description	Entire Text
Precision Top 1	CNN	68.67%	60.36%	67.54%	73.21%	71.75%
	LSTM	3.08%	62.68%	71.60%	76.98%	73.82%
	BiLSTM	71.22%	65.19%	74.16%	79.22%	77.02%
	HFEM	**72.22%**	**69.09**%	**74.81**%	**79.62%**	**81.55%**
Recall Top 1	CNN	46.98%	13.44%	41.19%	43.79%	50.07%
	LSTM	4.54%	13.74%	42.44%	45.46%	51.96%
	BiLSTM	48.66%	13.89%	43.75%	46.50%	53.58%
	HFEM	**49.06%**	**14.18%**	**43.81%**	**47.29%**	**55.02%**
F1 Top 1	CNN	53.85%	16.26%	48.43%	51.91%	57.02%
	LSTM	0.81%	17.16%	50.80%	54.36%	59.27%
	BiLSTM	55.40%	17.54%	52.37%	55.98%	61.55%
	HFEM	**56.18%**	**18.73%**	**52.38%**	**56.55%**	**63.60%**

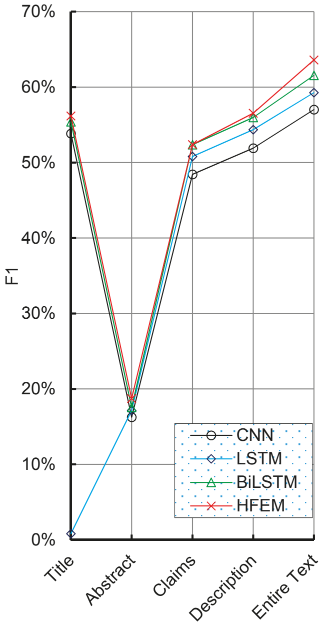
Next, we provide a comprehensive comparison of the performance of the four models under five input schemes, as shown in Figure 11, which shows that our HFEM model outperformed the other models under the same circumstances. From Figure 11a, we can see that HFEM has a slight advantage when using the title, claims, and description section. Moreover, it shows a clear superiority in terms of precision when using the abstract section and entire text. From the view of the recall score (Figure 11b), HFEM still demonstrates its advantages. Additionally, the F1 scores of all these methods show similar trends (Figure 11c). Since HFEM can take benefits from the CNN and BiLSTM model, thus it could maximally leverage the local lexical-level and global sentence-level features from patent texts.



(**a**) Precision scores (Top 1) obtained under various inputs.

(**b**) Recall scores (Top 1) obtained under various inputs.

Figure 11. *Cont.*



(**c**) F1 scores (Top 1) obtained under various inputs.

Figure 11. Classification performance (precision, recall, and F1 scores) under different inputs.

Besides, we found that algorithms using the entire text as input outperformed those that only used a single section of patent documents. We also found that when the claims and description section are used as input, all algorithms can achieve decent performance, no matter which model is used. When only the title and abstract sections are separately used, it leads to relatively low recall and F1 scores, especially for the LSTM model. When we investigate further, we found that the title and abstract sections usually contain less than 100 words, particularly for the title sections. This means that to achieve competitive performance in patent classification, enough information is required, regardless of which model is used.

7. Conclusions

This study proposed HFEM for multi-label mechanical patent classification, which does not rely on sophisticated feature engineering, external language knowledge or complicated pre-processing. The results of extensive experiments on the M-CLEF dataset showed that our approaches can improve the classification performance of three baseline models, all of which are single neural network models.

A number of insights were obtained when we tried to investigate the features of our algorithm that led to its better performance. We found that in the architecture of the HFEM model, the CNN layers are in charge of capturing salient local lexical-level features, while the BiLSTM layer learns long-term dependencies from sequences of higher-level representations in the patent text. Therefore, our proposed model can take full advantage of both CNN and BiLSTM to make significant improvements in precision, recall and F1 scores, respectively.

Secondly, we found that algorithms with the entire text as input always achieved better performance. When the claims and description sections were used, decent performances were obtained. Finally, the title and abstract sections when used as input separately all led to relatively low scores. Simply concatenating four sections as input provided a clear performance improvement for all tested patent classifiers. However, our experiment results with HFEM showed that more significant improvement can be achieved when we separately apply CNN to extract features from four channels and then use various concatenation strategies to jointly concatenate feature maps as an input of BiLSTM. The experiment results indicate that our hybrid model can take full advantage of the entire patent text.

Several additional studies are planned in our future work. One is adding meta-information to improve the classification performance. Beyond that, classifying a patent to the lowest IPC level is a crucial step in building a complete multi-label IPC auto-classification system. We plan to design a hierarchical algorithm, which uses both narrative text and meta-information as input.

Acknowledgments: This work is supported by the National Natural Science Foundation of China under Grant Nos. 51475097, 91746116 and 51741101, and the China Scholarship Council, and Science and Technology Foundation of Guizhou Province under Grant Nos. [2015]4011, R[2015]13, JZ[2014]2004, JZ[2014]2001, ZDZX[2013]6020, [2016]5013 and LH[2016]7433. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

Author Contributions: Jie Hu and Shaobo Li conceived and designed the experiments; Jie Hu and Shaobo Li performed the experiments; Jianjun Hu and Guanci Yang analyzed the data; Guanci Yang implemented the baseline methods; Jie Hu wrote the paper; Jianjun Hu, Guanci Yang, and Jie Hu revised and polished the manuscript.

References

1. Park, Y.; Yoon, J.; Phillips, F. Application technology opportunity discovery from technology portfolios: Use of patent classification and collaborative filtering. *Technol. Forecast. Soc. Chang.* **2017**, *118*, 170–183. [CrossRef]
2. Cong, H.; Tong, L.H. Grouping of TRIZ Inventive Principles to facilitate automatic patent classification. *Expert Syst. Appl.* **2008**, *34*, 788–795. [CrossRef]
3. D'hondt, E.; Verberne, S. Patent classification on subgroup level using Balanced Winnow. In *Current Challenges in Patent Information Retrieval*; Springer: Berlin, Germany, 2017; pp. 299–324.
4. Al Shamsi, F.; Aung, Z. Automatic patent classification by a three-phase model with document frequency matrix and boosted tree. In Proceedings of the IEEE 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, UAE, 6–8 December 2016.
5. Stutzki, J.; Schubert, M. Geodata supported classification of patent applications. In Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data, San Francisco, CA, USA, 26 June 2016.
6. Lim, S.; Kwon, Y. IPC Multi-label Classification Based on the Field Functionality of Patent Documents. In Proceedings of the 12th International Conference on Advanced Data Mining and Applications (ADMA 2016), Gold Coast, QLD, Australia, 12–15 December 2016; Springer: Berlin, Germany, 2016.
7. Wu, J.L.; Chang, P.C.; Tsao, C.C.; Fan, C.Y. A patent quality analysis and classification system using self-organizing maps with support vector machine. *Appl. Soft Comput.* **2016**, *41*, 305–316. [CrossRef]
8. D'hondt, E.; Verberne, S.; Koster, C.; Boves, L. Text Representations for Patent Classification. *Comput. Linguist.* **2013**, *39*, 755–775. [CrossRef]
9. Meng, L.E.; He, Y.; Li, Y. Research of Semantic Role Labeling and Application in Patent Knowledge Extraction. In Proceedings of the IPaMin 2014 Co-Located with Konvens 2014 1st International Workshop on Patent Mining and Its Applications (IPaMin@ KONVENS), Cincinnati, OH, USA, 6–7 October 2014.
10. Noh, H.; Jo, Y.; Lee, S. Keyword selection and processing strategy for applying text mining to patent analysis. *Expert Syst. Appl.* **2015**, *42*, 4348–4360. [CrossRef]
11. Joung, J.; Kim, K. Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. *Technol. Forecast. Soc. Chang.* **2017**, *114*, 281–292. [CrossRef]
12. Taeyeoun, R.; Yujin, J.; Byungun, Y. Developing a Methodology of Structuring and Layering Technological Information in Patent Documents through Natural Language Processing. *Sustainability* **2017**, *9*, 2117.
13. Kim, G.; Lee, J.; Jang, D.; Park, S. Technology Clusters Exploration for Patent Portfolio through Patent Abstract Analysis. *Sustainability* **2016**, *8*, 1252. [CrossRef]

14. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Stateline, NV, USA, 5–10 December 2013.
15. Kuang, S.; Davison, B.D. Learning Word Embeddings with Chi-Square Weights for Healthcare Tweet Classification. *Appl. Sci.* **2017**, *7*, 846. [CrossRef]
16. Zeng, Y.; Yang, H.; Feng, Y. A convolution BiLSTM neural network model for Chinese event extraction. In Proceedings of the International Conference on Computer Processing of Oriental Languages, Kunming, China, 2–6 December 2016; Springer: Berlin, Germany, 2016.
17. Kiperwasser, E.; Goldberg, Y. Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations. *arXiv* **2016**, arXiv:1603.04351v3.
18. Derieux, F.; Bobeica, M. Combining Semantics and Statistics for Patent Classification. In Proceedings of the CLEF 2010 LABs and Workshops, Notebook Papers, Padua, Italy, 22–23 September 2010.
19. Benson, C.L.; Magee, C.L. A hybrid keyword and patent class methodology for selecting relevant sets of patents for a technological field. *Scientometrics* **2013**, *96*, 69–82. [CrossRef]
20. Brants, T.; Franz, A. Web 1T 5-gram Version 1. Linguistic Data Consortium. Available online: https://catalog.ldc.upenn.edu/ldc2006t13 (accessed on 13 January 2018).
21. Lim, J.; Choi, S.; Lim, C.; Kim, K. SAO-Based Semantic Mining of Patents for Semi-Automatic Construction of a Customer Job Map. *Sustainability* **2017**, *9*, 1386. [CrossRef]
22. Zhang, D.; Xu, H. Chinese comments sentiment classification based on word2vec and SVM perf. *Expert Syst. Appl.* **2015**, *42*, 1857–1863. [CrossRef]
23. Xu, H.; Dong, M.; Zhu, D. Text Classification with Topic-based Word Embedding and Convolutional Neural Networks. In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Washington, DC, USA, 2–5 October 2016.
24. Verberne, S.; D'hondt, E. Patent Classification Experiments with the Linguistic Classification System LCS in CLEF-IP 2011. In Proceedings of the CLEF 2011 Notebook Papers/Labs/Workshop, Amsterdam, The Netherlands, 19–22 September 2011.
25. Li, Z.; Tate, D.; Lane, C.; Adams, C. A framework for automatic TRIZ level of invention estimation of patents using natural language processing, knowledge-transfer and patent citation metrics. *Comput. Aided Des.* **2012**, *44*, 987–1010. [CrossRef]
26. Zhang, L.; Suganthan, P.N. A survey of randomized algorithms for training neural networks. *Inf. Sci.* **2016**, *364*, 146–155. [CrossRef]
27. LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks*; MIT Press: Cambridge, MA, USA, 1995; Volume 3361.
28. Llamas, J.; M Lerones, P.; Medina, R.; Zalama, E.; Gómez-García-Bermejo, J. Classification of Architectural Heritage Images Using Deep Learning Techniques. *Appl. Sci.* **2017**, *7*, 992. [CrossRef]
29. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 7–12 December 2015.
30. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
31. Kim, Y. Convolutional neural networks for sentence classification. *arXiv* **2014**, arXiv:1408.5882.
32. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
33. Piroi, F.; Lupu, M.; Hanbury, A.; Zenz, V. CLEF-IP 2011: Retrieval in the intellectual property domain. In Proceedings of the CLEF 2011 Labs and Workshop, Notebook Papers, Amsterdam, The Netherlands, 19–22 September 2011.
34. Han, T.L.; He, C.; Shen, L. Automatic classification of patent documents for TRIZ users. *World Pat. Inf.* **2006**, *28*, 6–13.
35. Chollet, Franois, and others, *Keras*, in *GitHub*. 2015. Available online: https://github.com/keras-team/keras (accessed on 13 January 2018).
36. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.

sustainability

MDPI

Article

Ontology Design for Solving Computationally-Intensive Problems on Heterogeneous Architectures

Hossam M. Faheem [1,†], Birgitta König-Ries [2,†], Muhammad Ahtisham Aslam [3,*,†], Naif Radi Aljohani [3,†] and Iyad Katib [3,†]

[1] Computer Systems Department, Faculty of Computer and Information Sciences, Ain Shams University, Abbassia 11566, Cairo, Egypt; hmfaheem@cis.asu.edu.eg

[2] Heinz-Nixdorf Endowed Chair for Distributed Information Systems, Friedrich-Schiller-Universitat Jena, 07743 Jena, Germany; birgitta.koenig-ries@uni-jena.de

[3] Faculty of Computing and Information Technology, King Abdulaziz University, 21589 Jeddah, Saudi Arabia; nraljohani@kau.edu.sa (N.R.A.); iakatib@kau.edu.sa (I.K.)

* Correspondence: maaslam@kau.edu.sa; Tel.: +966-56-332-1977

† These authors contributed equally to this work.

Received: 13 November 2017; Accepted: 15 December 2017; Published: 8 February 2018

Abstract: Viewing a computationally-intensive problem as a self-contained challenge with its own hardware, software and scheduling strategies is an approach that should be investigated. We might suggest assigning heterogeneous hardware architectures to solve a problem, while parallel computing paradigms may play an important role in writing efficient code to solve the problem; moreover, the scheduling strategies may be examined as a possible solution. Depending on the problem complexity, finding the best possible solution using an integrated infrastructure of hardware, software and scheduling strategy can be a complex job. Developing and using ontologies and reasoning techniques play a significant role in reducing the complexity of identifying the components of such integrated infrastructures. Undertaking reasoning and inferencing regarding the domain concepts can help to find the best possible solution through a combination of hardware, software and scheduling strategies. In this paper, we present an ontology and show how we can use it to solve computationally-intensive problems from various domains. As a potential use for the idea, we present examples from the bioinformatics domain. Validation by using problems from the Elastic Optical Network domain has demonstrated the flexibility of the suggested ontology and its suitability for use with any other computationally-intensive problem domain.

Keywords: ontology design; knowledge management; heterogeneous architectures; Big Data

1. Introduction

Solving computationally-intensive problems is an attractive topic for researchers in the field of parallel processing and high-performance computing. In parallel to this, ontologies can play vital role in solving domain specific problems by making use of knowledge representation and reasoning [1–4]. Due to the nature of existing hardware clusters [5], it is now common to have a cluster of hardware architectures [6] equipped with an NVIVIA GPGPUs and Xeon-Phi coprocessor in addition to traditional Intel CPUs. Task scheduling on such heterogeneous architectures is considered to constitute a major challenge and is one of the most important topics in current scientific research. The major factors that contribute to the intensity of an integrated architecture are the complexity of modern hardware architectures, and the parallel computing paradigms and the memory management techniques associated with them [7]. Task scheduling can be characterized as the assignment of time-constrained jobs to time-constrained resources within a pre-defined time interval, representing the complete timescale of the schedule. The domain of solving computationally-intensive problems comprises several entities, namely the jobs that solve the problem [8,9], the computing devices with the hardware architectures on which the jobs will run, the scheduling strategies used to schedule the tasks of the jobs on the hardware and the algorithms used to solve specific problems in a specific domain.

A mapping scheme from the problem domain to the computer domain should be clearly identified. For this purpose, ontologies can be developed by describing entities from one or multiple domains and relations among these entities [10,11]. Ontologies are recognized as conceptual information models that describe the entities in a specific domain, such as classes, relationships and functions [12]. In this paper, we present an ontology in the domain of High-Performance Computing (HPC) and show how an ontology-based approach may be used to solve computationally-intensive problems on heterogeneous architecture.

Many ontological reasoning-based approaches have been presented so far to address the challenges related to HPC and parallel computing. For example, a framework as an integrated solution of parallel computing and ontological reasoning is presented in [13]. It can be used on scalability issues in parallel computing by using ontological reasoning. An ontology learning-based framework to address scalability issues in parallel computing is presented in [14]. Similarly, an ontology-drive solution for cloud service handling and discovery is presented in [15,16]. In [17], the authors present an ontology editor (i.e., ONTOLIS) that can be used for concept mapping and inferencing for better content coverage in Big Data and HPC courses. In [18], the authors present a parallel ABox reasoning algorithm for increased scalability in HPC and parallel computing environments. computationally-intensive problems may be treated and solved as self-contained problems by identifying hardware, software and scheduling strategies, depending on the complexity and nature of the problem. Ontologies can play a significant role in mapping domain concepts to an inferencing environment and suggesting the best possible solutions at run time. Due to the diversity of domains and involvement of multidirectional software and hardware, very little work has been undertaken on solving computationally-intensive problems on heterogeneous architectures by designing and using ontologies.

In this paper, we present an ontology for solving computationally-intensive problems by performing reasoning on the data for given jobs. The work presented covers multidimensional domain knowledge, such as the versatility of hardware architectures, software, scheduling strategies and memory management techniques, which makes this work appropriate for domain users who belong to either the HPC or parallel computing domains. We also show the potential use of this work by presenting a real case study at the HPC Center of King Abdulaziz University.

The rest of this paper is organized as follows: Section 2 describes some work related to use of ontologies in solving various domain specific problems. Section 3 describes the structure of the ontology, including its conceptual basic building blocks. Section 4 describes the main concepts of the ontology, such as the classes, along with its attributes. Section 5 provides an evaluation to the flexibility of the proposed ontology and shows how we can easily insert computationally-intensive problem domains to it. Finally, in Section 6 we conclude our work.

2. Related Work

Ontological reasoning plays a key role in solving diverse complex problems in various domains. With the help of domain experts, semantic web experts and ontology engineers are developing ontologies that use ontological reasoning to solve problems that otherwise cannot be solved (or might not be addressed efficiently) regarding textual entities or data available in relational databases. In this section, we present work related to ontology development in various domains and the use of these ontologies in data representation, as well as solving complex problems.

An integrated solution of ontological reasoning and parallel computing to address issues of scalability is presented in [13]. Parallel computer architectures, such as home computing environments, multi-core machines, grid and peer-to-peer, lead to great demand for efficient handling of computational resources through making use of software architecture, algorithms and now a new paradigm-making use of ontological reasoning. In [13], the authors propose using ontology modularization and queries. This helps to solve the problem of reasoning in individual modules, ultimately to achieve maximum efficiency in the use of architectures and parallel computing algorithms.

In [15], the authors present an ontology-driven platform for using and integrating the services provided by various cloud environments. The proposed framework may also be used to describe the functionality of given services by making use of ontological concepts, looking for other services from target providers to generate client adapters to use the required services. The inference engine of the proposed framework performs semantic matching to find the best possible matched service from the cloud, facilitating the resulting applications as an integrated resource of diverse services.

An ontology learning-based framework is presented in [14]. The proposed framework aims to improve scalability by making efficient use of processing power, computing resources and processing time. The proposed approach tackles the difficulties in distributed and low-level parallel programming by coupling high-level semantic descriptions with programming models.

Another approach to discover suitable services from an increasing choice on the cloud, according to user requirements such as cost, security and performance, is presented in [16]. The authors propose mapping services attributes to ontologies to minimize the gap between service descriptions, types, features and naming conventions. Using ontological reasoning on such services makes service discovery easier and more accurate on the cloud.

An ontology-based approach and framework for organizing courses covering Big Data and HPC contents is described in [17]. In this work, the authors introduce an ontology editor (i.e., ONTOLIS) that can be used to stimulate the staffing of various domains, as well as IT companies. The framework presented in [17] makes use of ontological reasoning to bridge the learning gap between domain expertise (having IT expertise; medium-level IT users; or introductory IT students). An analysis and an interpretation of archives and collections through historical data are presented in [19]. This study proposes methods for semantic associations of social networks by linking them with external datasets.

An ontology-based knowledge framework for material selection in engineering domain is presented in [20]. In this work, the authors provide a semantic representation of labeled instances, as well as the material products, representing them as RDF instances and then generating a knowledge graph from the RDF triples. The graph is generated from reasoning on the domain knowledge (modeled as classes, sub-classes, properties and individuals of the ontology). The knowledge graph generated by the knowledge framework makes suggestions for material selection.

In [21], the authors use ontological reasoning to find situations from calendar events to support users in fulfillment of their jobs. They present an ontology developed for the event domain that can be used to accommodate various situations from such events as classes and properties of the ontology. The proposed approach infers the situations by using both temporal and semantic aspects of situations. Further, the authors implemented the approach as an application for mobile phones to manage incoming calls based on the inferred situation of the user.

A parallel ABox reasoning algorithm for increasing the scalability in parallel computing environment is presented in [18]. The proposed algorithm can be used to model disjointness and inconsistencies in the ontology model, which ultimately can improve parallelization and reduce resource cost. The separation of ABox reasoning from TBox reasoning in the proposed algorithm makes the derivation simple and paralyses the integration steps, and this ultimately improves efficiency and reduces memory access.

3. Ontology Structure

The ontology's design philosophy depends mainly on its flexibility in dealing with problems from various domains. The key point is to map the algorithms used to solve problems in a given domain to equivalent algorithms in the computer science domain. Figure 1 shows the conceptual basic building blocks of an ontology. The computationally-intensive problem can be mapped to a computer algorithm using the mapping scheme. The problem to be solved needs a computing device that contains at least a hardware architecture. The algorithm can use different parallel paradigms that run on different architectures. The scheduling strategy is responsible for assigning the appropriate architecture to the problem. In this paper, we use the bio problem as a case study of a computationally-intensive problem domain. It is clear that we can consider the proposed ontology as if it were comprised of multiple ontologies, in addition to a mapping scheme.

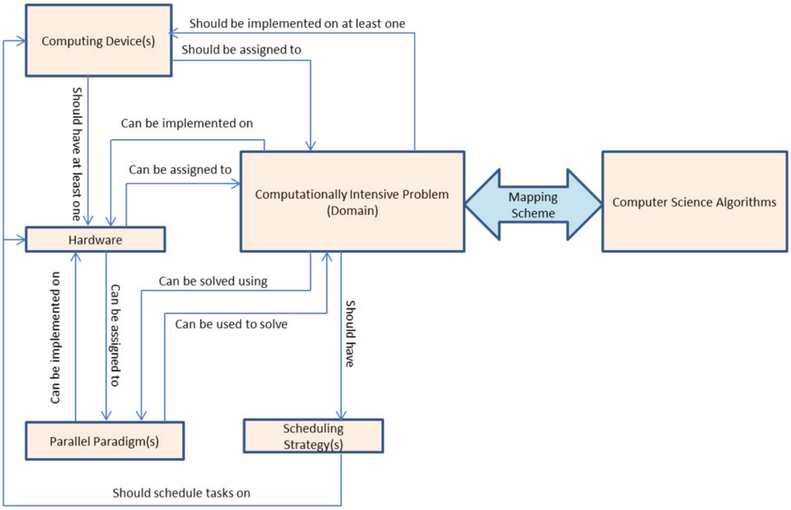


Figure 1. Conceptual Basic Building Blocks of the Ontology.

4. Ontology Basics

In this section, we describe the named classes of the ontology, the object properties and some instances. We will also outline axioms that can be used to define the meaning of several components of the ontology. Moreover, we briefly describe the mapping from the bio problem domain to the computer science domain.

4.1. Named Classes

Classes are interpreted as sets that contain individuals. They are described using formal (mathematical) descriptions that state precisely the requirements for membership of the class. The class tree contains one class called owl:Thing, which is superclass of everything. We have created five disjoint subclasses: “Algorithms”, “BioProblem”, “ComputingDevice”, “Hardware”, “ParallelParadigm”, and “SchedulingStrategy”.

Focusing on the BioProblem subclass can lead us to the following facts:

- BioProblem class has a set of bio problems such as “Comparing Sequences”, “DNA Arrays”, “Finding Signals”, “Genome Rearrangements”, “Identifying Proteins”, “Mapping DNA”, “Molecular Evolutions”, “Predicting Genes”, “Repeat Analysis”, and “Sequencing DNA”. This is shown in Figure 2. Defining these classes can help domain experts in finding the best scheduling strategy as a solution for integrated problem of bio and HPC domains.
- BioProblem class has a relation with SchedulingStrategy, Hardware, ComputingDevice, and ParallelParadigm classes, such that the BioProblem needs ComputingDevice equipped with at least hardware architecture to run the job on it. ParallelParadigm is needed to write code that best fits the selected hardware. SchedulingStrategy is used to schedule the tasks on the hardware of the ComputingDevice.

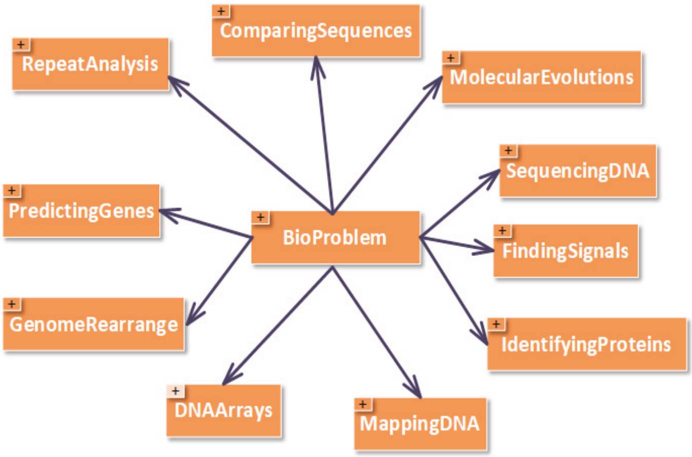


Figure 2. Bio Problem Subclasses.

The *"ParallelParadigm"* class has *MPI, OpenMP, CUDA*, and *MIC* subclasses while *"Hardware"* class has *IntelCPU, NVIDIAGPGPU*, and *Xeon-Phi* subclasses. We can see that *MPI* and *OpenMP* can work on *IntelCPU*. *CUDA* can only work on *NVIVIAGPGPU*, while *MIC* can only work on *Xeon-Phi*. *ComputingDevice* class should have at least one type of *Hardware* while it has a domain of all the hardware available. The relation between *ComputingDevice*, *Hardware*, and *ParallelParadigm* is shown in Figure 3. Ontology is designed in such a way that other parallel computing paradigms and hardware architectures can be added easily to the ontology. For example, we can add any hardware such as *ARM* which can work on *MPI* and *OpenMP*. We can also add *OpenCL* as a new *ParallelComputingParadigm* class to the *ParallelParadigm* class. Similarly, we can add *OpenCL* class that can work on *IntelCPU* and *NVIDIAGPGPU*.

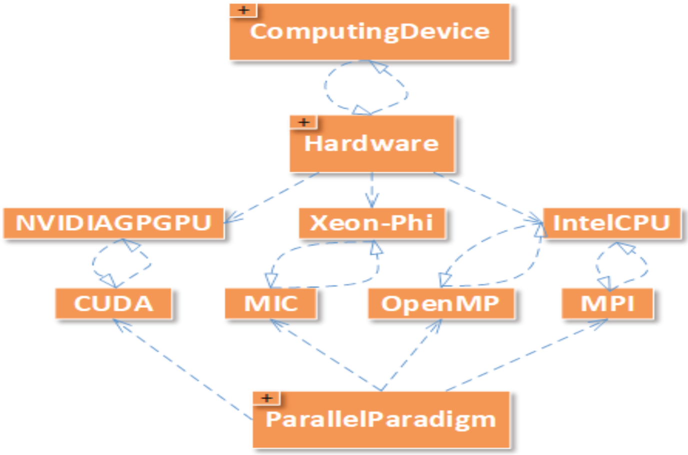


Figure 3. Relation between *ComputingDevice*, *ParallelParadigm*, and Hardware Subclasses.

4.2. Object Properties

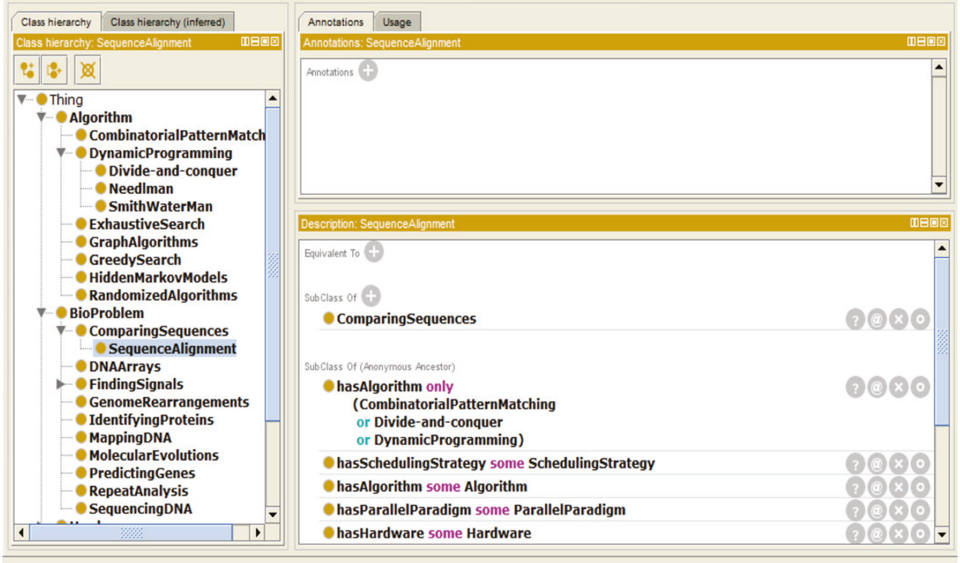
The object feature tree contains 12 features that can be assigned to the bio problem *"hasAlgorithm"*, *"hasHardware"*, *"hasParallelParadigm"*, *"hasSchedulingStrategy"*, *"hasArchitecture"*, *"hasComputingDevice"*, *"hasParameter"*, *"IsAlgorithmOf"*, *"IsHardwareOf"*, *"IsParallelParadigmOf"*, *"IsComputingDeviceOf"*, and *"IsSchedulingStrategyOf"*.

- Property "hasAlgorithm"
  This property assigns an algorithm (from the computer science domain) to solve a given bio problem. *"IsAlgorithmOf"* is an inverse property of it.
- Property "hasHardware"
  This property assigns a hardware architecture on which the algorithm will run to solve a given bio problem. *"IsHardwareOf"* is an inverse property of it.
- Property "hasParallelParadigm"
  This property assigns a parallel paradigm used to write an algorithm to solve a given bio problem. *"IsParallelParadigmOf"* is an inverse property of it.
- Property "hasSchedulingStrategy"
  This property assigns the scheduling strategy used to deploy the hardware used to solve a given bio problem. *"IsSchedulingStrategyOf"* is the inverse of it.
- Property "hasComputingDevice"
  This property assigns at least one computing device to solve a given bio problem. *"IsComputingDevice"* is the inverse of it.
- Property "hasParameters"
  This property assigns the parameters required for each algorithm. For example, a motif-finding problem has L, d, n, and T, where: L is the length of motif, d is the permitted mutation, n is the number of characters in each sequence, and T is the number of sequences. These parameters are all of the type 'integer'.
- Property "hasArchitecture"
  This property assigns at least architecture to a computing device. Each computing device can be equipped with one or more architectures.

4.3. Instances

We initially selected two famous bio problems: "DNA sequence alignment"; and "Motif-finding Problem".

- Instance "SequenceAlignment"
  The DNA sequence alignment is one of the most famous bio problems. It is classified under "ComparingSequences". It can be solved using either combinatorial pattern matching or divide-and-conquer or dynamic programming algorithms (as shown in Figure 4).
- Instance "MotifFindingProblem"
  The motif-finding problem is one of the most famous bio problems. It is classified under "FindingSignals". It can be solved using exhaustive or greedy searches, hidden Markov models or randomized algorithms (as described in Figure 5).


Class hierarchy
Class hierarchy (inferred)
Class hierarchy: SequenceAlignment
Annotations
Usage
Annotations: SequenceAlignment
Annotations
Description: SequenceAlignment
Equivalent To
SubClass Of
SubClass Of (Anonymous Ancestor)

- Thing
  - Algorithm
    - CombinatorialPatternMatch
    - DynamicProgramming
      - Divide-and-conquer
      - Needlman
      - SmithWaterMan
    - ExhaustiveSearch
    - GraphAlgorithms
    - GreedySearch
    - HiddenMarkovModels
    - RandomizedAlgorithms
  - BioProblem
    - ComparingSequences
      - SequenceAlignment
    - DNAArrays
    - FindingSignals
    - GenomeRearrangements
    - IdentifyingProteins
    - MappingDNA
    - MolecularEvolutions
    - PredictingGenes
    - RepeatAnalysis
    - SequencingDNA

SubClass Of

ComparingSequences

SubClass Of (Anonymous Ancestor)

hasAlgorithm only
(CombinatorialPatternMatching
or Divide-and-conquer
or DynamicProgramming)

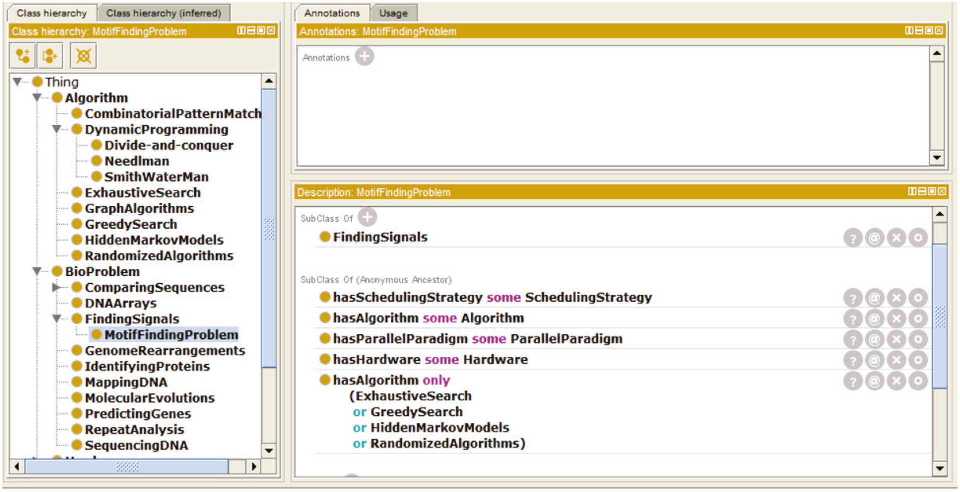
hasSchedulingStrategy some SchedulingStrategy

hasAlgorithm some Algorithm

hasParallelParadigm some ParallelParadigm

hasHardware some Hardware

Figure 4. Instance "Sequence Alignment".


Class hierarchy
Class hierarchy (inferred)
Class hierarchy: MotifFindingProblem
Annotations
Usage
Annotations: MotifFindingProblem
Annotations
Description: MotifFindingProblem
SubClass Of
SubClass Of (Anonymous Ancestor)

- Thing
  - Algorithm
    - CombinatorialPatternMatch
    - DynamicProgramming
      - Divide-and-conquer
      - Needlman
      - SmithWaterMan
    - ExhaustiveSearch
    - GraphAlgorithms
    - GreedySearch
    - HiddenMarkovModels
    - RandomizedAlgorithms
  - BioProblem
    - ComparingSequences
    - DNAArrays
    - FindingSignals
      - MotifFindingProblem
    - GenomeRearrangements
    - IdentifyingProteins
    - MappingDNA
    - MolecularEvolutions
    - PredictingGenes
    - RepeatAnalysis
    - SequencingDNA

SubClass Of

FindingSignals

SubClass Of (Anonymous Ancestor)

hasSchedulingStrategy some SchedulingStrategy

hasAlgorithm some Algorithm

hasParallelParadigm some ParallelParadigm

hasHardware some Hardware

hasAlgorithm only
(ExhaustiveSearch
or GreedySearch
or HiddenMarkovModels
or RandomizedAlgorithms)

Figure 5. Instance "Motif-finding Problem".

4.4. Axioms

This section describes the axioms used to define the meaning of various components of the ontology and relationships. Table 1 lists the axioms for the ontology.

Table 1. Axioms for the ontology.

Concept Name	Axiom Description	Logical Expression
Algorithm	A collection of computer algorithms such that an algorithm is a software procedure or formula used to solve a specific problem in this domain, based on conducting a sequence of specified actions.	$A \sqsubseteq solves.Problem$
Computationally_ Intensive_Problem	A generic term to describe any problem in any domain that needs intensive computations.	$CIP \sqsubseteq Problem$
Computing_Device	A device that should contain at least one physical hardware processor.	$CD \sqsubseteq \exists contains.HardwareProcessor$
Hardware	A generic term to express the processors used in performing the computations.	$HW \sqsubseteq usedFor.Computation$
IntelCPU	A traditional type of central processing unit developed by Intel.	$ICPU \sqsubseteq CPU \sqcap madeBy.Intel$
NVIDIAGPGPU	A general-purpose graphics processor unit developed by NVIDIA.	$NGPU \sqsubseteq GPU \sqcap madeBy.NVidia$
Xeon-Phi	Term for many core processors or coprocessors developed by Intel.	$XPhi \sqsubseteq hasManyCores.ICPU$
ParallelParadigm	A paradigm used to parallelize sequential code.	$PP \sqsubseteq parallelize.SeuqentialCode$
MPI	A Message Passing Interface parallel computing paradigm that supports distributed memory multiprocessing.	$MPI \sqsubseteq PP \sqcap supports.Distributed-Memory$
OpenMP	An Open Multi-Processing parallel computing paradigm that supports shared memory multiprocessing.	$OMP \sqsubseteq PP \sqcap supports.SharedMemory$
CUDA	A Compute Unified Device Architecture parallel computing paradigm used on NVIDIA GPGPUs.	$CUDA \sqsubseteq PP \sqcap uses.NGPU$
MIC	A Many Integrated Core Architecture parallel computing paradigm used on Xeon-Phi coprocessors.	$MIC \sqsubseteq PP \sqcap uses.XPhi$
SchedulingStrategy	A scheduling strategy is used to assign suitable hardware resources to perform the jobs in a way to achieve a certain goal such as speeding up, load balancing, or optimization of power consumption, etc.	$SS \sqsubseteq assigns.HW \sqcap perfoms.Jobs$
FIFO	A First-In-First-Out scheduling strategy.	$FIFO \sqsubseteq SS \sqcap uses.FirstInFirstOut$
HEFT	A Heterogeneous Earliest Finish Time scheduling strategy.	$HEFT \sqsubseteq SS \sqcap uses.EearliestFinishTime$
PEFT	A Predictable Earliest Finish Time scheduling strategy.	$PEFT \sqsubseteq SS \sqcap uses.PredictableEarlies-tFinishTime$
Speed-based	A scheduling strategy that assigns data sets to be processed based on the speed of the processors being used to solve a problem.	$SB \sqsubseteq SS \sqcap uses.ProcessorSpeed$

5. Evaluation of the Flexibility

To evaluate the flexibility of the proposed ontology, we tested it on an additional domain called the "Elastic Optical Network (EON)". The International Telecommunication Union (ITU) divides the optical spectrum range of 1530–1565 nm (the so-called C-band) into fixed 50 GHz spectrum slots.

This fixed spectrum allocation wastes a great deal of the spectrum. EON is implemented to allow better utilization of the C-band. EON introduces plenty of challenges:

- Finding an optical path from source to destination that passes through multiple links, all of which have the same free spectrum range. This problem can be solved using either the exhaustive search algorithm, the heuristic algorithm or linear programming techniques.
- Finding a set of links that constitute an optical path, on condition that all the links have enough free contiguous spectra. This problem can be solved by using the computer science algorithm known as an exhaustive search.
- Load balancing of traffic to minimize spectrum fragmentation. This can be solved by a sorting algorithm or a binary search algorithm.

As we can see, the EON problem domain can replace the Bio Problem domain. It is also possible to create a new class called *"Computationally_Intensive_Problem"* that has several subclasses such as *"BioProblem"*, *"EON"* and so on. This can be shown as in Figure 6. The user can add as many computer algorithms as the user needs. Various scheduling strategies can be added. Eventually, it will be possible to use a ready-made ontology for computationally-intensive problems from any domain for merging with the proposed ontology.

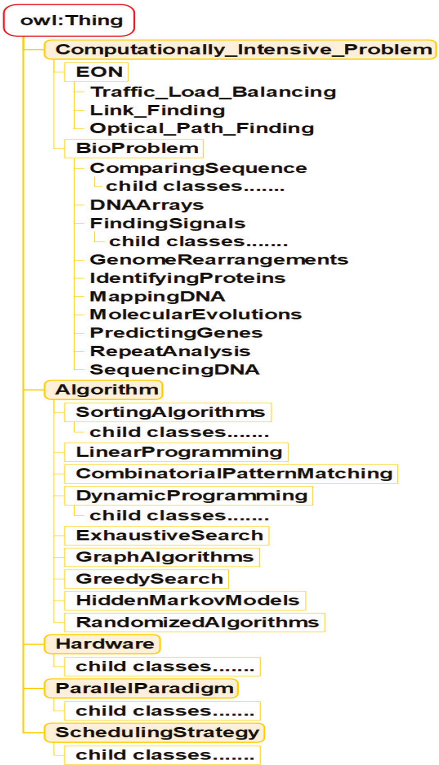


Figure 6. Class Hierarchy using *"Computationally Intensive Problem"* Class.

Now, we will describe two problems; the "motif finding problem" from the bioinformatics domain, and "Optical Path Finding" from the elastic optical network domain. The main purpose is to describe how the ontology is used to simplify and automate the whole process of problem solving.

5.1. Motif Finding Problem

The Motif Finding Problem (MFP) can be simply considered as a string matching problem. Solving the MFP to find a motif of length L with permitted mutation d can be implemented using a brute-force algorithm. All the possible *L-mers* (4^L) are compared with each possible motif of length L. If we have a sequence of size N then we can have ($N - L + 1$) motifs. In this paper, we present a problem in which the motif has a length $L = 16$, allowed mutations $d = 4$, and the number of sequences we are searching in is $T = 20$ each of size $N = 600$. All these parameters are described in the ontology. Inputs of this problem could be text files or specific DNA databases. The proposed ontology was used to describe the following:

- Computing device and its hardware used in solving the problem (*IntelCPU* and *NVIDIAGPGPU* and *Xeon-Phi*)
- Parallel computing paradigm used on each architecture (*MPI* and *OpenMP* on *IntelCPU*, *CUDA* on *NVIDIAGPGPU*, and *MIC* on *Xeon-Phi*)
- Scheduling strategy used to solve the problem (Speed-based)
- Computer algorithm used to solve the problem (exhaustive search)

We used SPARQL queries to extract all the required components from the knowledge base to implement the problem in an efficient way. The ontology was part of a large system used to solve the problem. This system includes computing cluster consisting of heterogeneous architectures, smart scheduling strategy, queuing system (PBS Pro), and algorithms library. Of course, we obtained the same results achieved when we previously deployed customized codes in [22] but in this case, the complete cycle was automated and the PBS scripts are automatically generated. This is obvious because we didn't change the algorithms used to solve the problem on the same hardware. The main concern here was how to use the ontology to simplify and automate the implementation.

5.2. Optical Path Finding

Optical path finding problem can be considered "routing and spectrum assignment problem" (RSA) which depends mainly on both the traffic demand bit-rate and the distance between the source and the destination. Routing and Spectrum Assignment (RSA) problem in a mesh network can be defined as in (1) and (2).

$$G = (v, A) \tag{1}$$

$$T = [T_{sd}] \tag{2}$$

where:

- ***G*** is a directed graph
- ***V*** is the set of nodes in the graph
- ***A*** is the set of unidirectional arcs connecting the graph nodes
- ***T*** is the traffic demand matrix

All these items are included in the ontology part related to the EON domain. Each traffic demand from a source node ***s*** to a destination node ***d*** is assigned an Elastic Optical Path (EOP) which is a physical path and contiguous spectrum based on the RSA algorithm. The assigned EOP is selected to minimize the total required spectrum used on any link. This is according to the following three conditions:

- Each demand is assigned a contiguous spectrum (spectrum contiguity constraint)
- Each demand is assigned the same spectrum across all links of its path (spectrum continuity constraint)
- Demands that share the same link are assigned non-overlapping spectrum parts

The proposed ontology was used to describe:

- Computing device and its hardware used in solving the problem (*IntelCPU* and *NVIDIAGPGPU*)
- Parallel computing paradigm used on each architecture (*OpenMP* on *IntelCPU* and *CUDA* on *NVIDIAGPGPU*)
- Scheduling strategy used to solve the problem (Speed-based)
- Computer algorithm used to solve the problem (exhaustive search)

The same methodology used to deploy the ontology in solving the "motif finding problem" was used to solve the RSA problem. We achieved the same results as in [23]. The same simplicity was achieved when the ontology was deployed in this case.

5.3. A Case Study of Bioinformatics Problem

This section describes the scheme for mapping from the bioinformatics domain to the computer science domain. Bioinformatics problems can be mapped to equivalent computer algorithms. Figure 7 shows how the graph algorithms from the computer science domain can be used to solve three different sets of problems in the bioinformatics domain. These include identifying proteins, sequencing DNA and DNA arrays. Figure 8 shows that "Finding Signals"—a problem from the bioinformatics domain—can be solved using four different sets of computer algorithms. These include the greedy search, randomized algorithms, hidden Markov models and the exhaustive search. We can conclude the cross domain problem mapping as, that one computer algorithm can solve many problems (as shown in Figure 7) and one problem can be solved by using different computer algorithms (as shown in Figure 8).

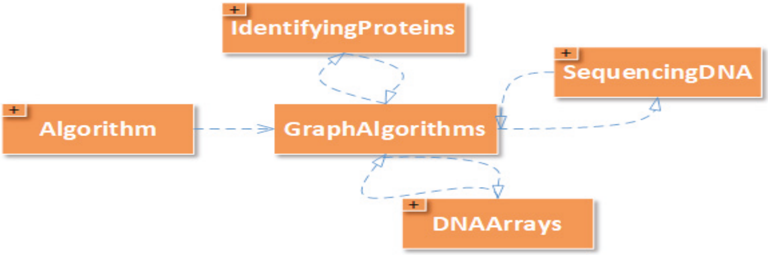


Figure 7. Use of Graph Algorithms (Computer Domain) to Solve Three Different Problems in the Bioinformatics Domain.

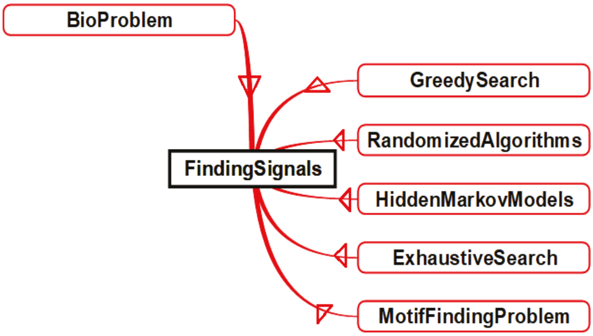


Figure 8. Solving the "Finding Signals" Problem (Bioinformatics Domain) Using Four Different Sets of Computer Algorithm.

6. Conclusions

In this paper, we presented an ontology as a semantically enriched schema for computational intensive problems. We showed how data for a domain-specific problem can be mapped to classes of the ontology and linked with each other by the object and data type properties of the ontology, so that we can perform reasoning on the given data. We also showed that computer science algorithms and hardware architectures can be flexibly appended to respond to various domain requirements. Schemes for mapping from a given domain to computer science domain are presented. To prove the usage of ontologies in solving computationally-intensive problems, we took a real-life problem from the bio domain (at HPC Center of King Abdulaziz University), mapped problem-specific data to the ontology and used the mapped data. In fact, the ability of adding new computationally-intensive problem domains is the main focus of this work to satisfy the important flexibility and reusability concepts. Flexibility is clear in the ability of the proposed ontology to add as many computationally-intensive problem domains as possible. Reusability concept is clear since the use of computer algorithms and hardware resources is common in all domains. Our case study demonstrated validation by investigating problems from two different domains.

Acknowledgments: The work described in this paper was supported by King Abdulaziz University's High Performance Computing Center (Aziz Supercomputer) (http://hpc.kau.edu.sa).

Author Contributions: Hossam M. Faheem, Birgitta König-Ries and Muhammad Ahtisham Aslam coined and worked on the feasibility of the idea. Hossam M. Faheem and Muhammad Ahtisham Aslam designed the ontology. Hossam M. Faheem, Naif Radi Aljohani and Iyad Katib designed the experiments. Hossam M. Faheem and Naif Radi Aljohani conducted the experiments in the HPC environment. Hossam M. Faheem, Muhammad Ahtisham Aslam and Naif Radi Aljohani analyzed the results and flexibility of the proposed ontology. Hossam M. Faheem and Muhammad Ahtisham Aslam wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Compton, M.; Barnaghi, P.; Bermudez, L.; Garcıa-Castro, R.; Corcho, O.; Cox, S.; Graybeal, J.; Hauswirth, M.; Henson, C.; Herzog, A.; et al. The SSN Ontology of the W3C Semantic Sensor Network Incubator Group. *Web Semant. Sci. Serv. Agents World Wide Web* **2012**, *17*, 25–32.
2. Keet, C.M.; Ławrynowicz, A.; d'Amato, C.; Kalousis, A.; Nguyen, P.; Palma, R.; Stevens, R.; Hilario, M.H. The Data Mining OPtimization Ontology. *Web Semant. Sci. Serv. Agents World Wide Web* **2015**, *32*, 43–53.
3. Baker, C.; Shaban-Nejad, A.; Su, X.; Haarslev, V.; Butler, G. Semantic Web Infrastructure for Fungal Enzyme Biotechnologists. *Web Semant. Sci. Serv. Agents World Wide Web* **2006**, *4*, 168–180.
4. Shekarpour, S.; Marx, E.; Ngomo, A.C.N.; Auer, S. SINA: Semantic Interpretation of User Queries for Question Answering on Interlinked Data. *Web Semant. Sci. Serv. Agents World Wide Web* **2015**, *30*, 39–51.
5. Heino, N.; Pan, J.Z. RDFS Reasoning on Massively Parallel Hardware. In Proceedings of the 11th International Semantic Web Conference (ISWC2012), Boston, MA, USA, 11–15 November 2012.
6. Atahary, T.; Taha, T.M.; Douglass, S. Hardware Accelerated Cognitively Enhanced Complex Event Processing Architecture. In Proceedings of the 2013 14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Honolulu, HI, USA, 1–3 July 2013; pp. 283–288.
7. Miksa, T. Using ontologies for verification and validation of workflow-based experiments. *Web Semant. Sci. Serv. Agents World Wide Web* **2017**, *43*, 25–45.
8. Koumenides, C.; Alani, H.; Shadbolt, N.; Salvadores, M. Global Integration of Public Sector Information. In Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, Raleigh, NC, USA, 26–27 April 2010.
9. Höchtl, J.; Reichstädter, P. Linked Open Data—A Means for Public Sector Information Management. In *Electronic Government and the Information Systems Perspective*; Andersen, K.N., Francesconi, E., Groenlund, A., van Engers, T.M., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6866, pp. 330–343.

10. Bechhofer, S.K.; Stevens, R.D.; Lord, P.W. GOHSE: Ontology Driven Linking of Biology Resources. *Web Semant. Sci. Serv. Agents World Wide Web* **2006**, *4*, 155–163.
11. Jonquet, C.; LePendu, P.; Falconer, S.; Coulet, A.; Noy, N.F.; Musen, M.A.; Shah, N.H. NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources, information retrieval, biomedical data and ontologies. *Web Semant. Sci. Serv. Agents World Wide Web* **2011**, *9*, 316–324.
12. Uschold, M. *Building Ontologies: Towards a Unified Methodology*; Technical Report; University of Edinburgh: Edinburgh, UK, 1996.
13. Bock, J. Parallel Computation Techniques for Ontology Reasoning. In *The Semantic Web—ISWC 2008, Proceedings of the 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, 26–30 October 2008*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 901–906.
14. Arguello, M.; Gacitua, R.; Osborne, J.; Peters, S.; Ekin, P.; Sawyer, P. Skeletons and Semantic Web Descriptions to Integrate Parallel Programming into Ontology Learning Frameworks. In Proceedings of the 2009 11th International Conference on Computer Modelling and Simulation, Cambridge, UK, 25–27 March 2009; pp. 640–645.
15. Gonidis, F.; Paraskakis, I.; Simons, A.J.H. On the Role of Ontologies in the Design of Service Based Cloud Applications. In *Euro-Par 2014: Parallel Processing Workshops, Proceedings of the Euro-Par 2014 International Workshops, Porto, Portugal, 25–26 August 2014*; Revised Selected Papers; Springer International Publishing: Cham, Switzerland, 2014; Part II, pp. 1–12.
16. Ali, A.; Shamsuddin, S.; Eassa, F. Ontology-based cloud services representation. *Res. J. Appl. Sci. Eng. Technol.* **2014**, *8*, 83–94.
17. Chuprina, S. Steps towards Bridging the HPC and Computational Science Talent Gap Based on Ontology Engineering Methods. *Procedia Comput. Sci.* **2015**, *51*, 1705–1713.
18. Ren, Y.; Pan, J.Z.; Lee, K. Parallel ABox Reasoning of EL Ontologies. In Proceedings of the First Joint International Conference of Semantic Technology (JIST 2011), Hangzhou, China, 4–7 December 2011.
19. Pattuelli, M.C.; Miller, M. Semantic network edges: A human-machine approach to represent typed relations in social networks. *J. Knowl. Manag.* **2015**, *19*, 71–81.
20. Zhang, Y.; Luo, X.; Zhao, Y.; Zhang, H.C. An ontology-based knowledge framework for engineering material selection. *Adv. Eng. Inform.* **2015**, *29*, 985–1000.
21. Kabir, M.A.; Han, J.; Colman, A.; Aljohani, N.R.; Basheri, M.; Aslam, M.A. Ontological Reasoning about Situations from Calendar Events. In *On the Move to Meaningful Internet Systems: OTM 2016 Conferences, Proceedings of the Confederated International Conferences: CoopIS, C&TC, and ODBASE 2016, Rhodes, Greece, 24–28 October 2016*; Springer: Cham, Switzerland, 2016; pp. 810–826.
22. Faheem, H.M.; Park, S.J.; Shires, D.R. A New Scheduling Strategy for Solving the Motif Finding Problem on Heterogeneous Architectures. *Int. J. Comput. Appl.* **2014**, *101*, 27–31.
23. Fayez, M.; Katib, I.; Rouskas, G.N.; Faheem, H.M. Spectrum Assignment in Mesh Elastic Optical Networks. In Proceedings of the 2015 24th International Conference on Computer Communication and Networks (ICCCN), Las Vegas, NV, USA, 3–6 August 2015; pp. 1–6.

sustainability

Article

Semantic Modeling of Administrative Procedures from a Spanish Regional Public Administration

Francisco José Hidalgo López [1,*], Jose Emilio Labra Gayo [2] and Patricia Ordóñez de Pablos [3]

[1] General Directorate of Information and Communications Technologies, Principality of Asturias, 33005 Oviedo, Spain

[2] Department of Computer Science, University of Oviedo; 33007 Oviedo, Spain; labra@uniovi.es

[3] Department of Business Administration, University of Oviedo, 33007 Oviedo, Spain; patriop@uniovi.es

[*] Correspondence: franciscojose.hidalgolopez@asturias.org

Received: 31 December 2017; Accepted: 16 February 2018; Published: 28 February 2018

Abstract: Over the past few years, Public Administrations have been providing systems for procedures and files electronic processing to ensure compliance with regulations and provide public services to citizens. Although each administration provides similar services to their citizens, these systems usually differ from the internal information management point of view since they usually come from different products and manufacturers. The common framework that regulations demand, and that Public Administrations must respect when processing electronic files, provides a unique opportunity for the development of intelligent agents in the field of administrative processes. However, for this development to be truly effective and applicable to the public sector, it is necessary to have a common representation model for these administrative processes. Although a lot of work has already been done in the development of public information reuse initiatives and common vocabularies standardization, this has not been carried out at the processes level. In this paper, we propose a semantic representation model of both processes models and processes for Public Administrations: the procedures and administrative files. The goal is to improve public administration open data initiatives and help to develop their sustainability policies, such as improving decision-making procedures and administrative management sustainability. As a case study, we modelled public administrative processes and files in collaboration with a Regional Public Administration in Spain, the Principality of Asturias, which enabled access to its information systems, helping the evaluation of our approach.

Keywords: administrative file; administrative procedure; sustainability; open data; linked data; provenance; RDF; PROV-O; P-PLAN

1. Introduction

We cannot deny the revolution that Information and Communication Technologies (ICT) have had in our society in the last twenty years. This revolution has reached all levels of society, and Public Administrations have not been an exception. The progressive implementation of ICT has promoted the development of new rights for citizens and the creation of new communication channels between citizens and Administrations, as well as between the Administrations themselves.

The progressive digitization of work made by civil servants has led to the development of new tools that allow the reception, comprehensive management, and complete traceability of administrative procedures in general, as well as citizen's requests.

The legal system has been advancing in an attempt to try to define and regulate the way that the application of new technologies is developed in Public Administrations. Although it is not the objective of this work to make an exhaustive analysis of the regulations, it is considered necessary to point out the most important regulations in relation to this work. Given that this work will be based on the case study of a regional administration in Spain, the normative to which we will refer will be that of this country:

- Resolution of 19 February 2013, from the Secretariat of State for Public Administrations, which approves the Technical Standard for the Reuse of information resources that establishes the common conditions for the selection, identification, description, format, conditions of use, and availability of documents and information resources prepared or guarded by the public sector [1].
- Law 19/2013, of 9 December, on transparency, access to information, and good governance. This recognizes and guarantees access to information on administrative activity [2].
- Law 39/2015, of 1 October, on the Common Administrative Procedure of Public Administrations [3].

To comply with the regulations in all matters that relate to public information access by citizens, different initiatives have emerged from the Administrations that have led to open data catalogs development.

Taking into account the datasets published in the Spanish open data catalog [4], which federates most of the Spanish Public Administrations open data catalogs, it can be concluded that in spite of the existing initiatives, there is a lack of homogeneity which means that two Administrations at the same level do not to publish the same type of information, as well as a lack of standardized vocabularies which causes data to be represented in different ways even if these Administrations publish the same information.

Another aspect that stands out from this analysis is that information is only being published at the data level and that information about administrative processes is not being published at all. It is considered that representing processes, of whatever kind, can be interesting for several reasons:

- Modeling processes, whether following linked data principles or not, converts them into actionable objects at the same level as data.
- Once these processes can be represented and automatically actionable, it is possible to build intelligent agents that interact with them and perform tasks that go from traceability to auditing.
- Administrative procedures are common to all Public Administrations since they derive from the same regulations. This means that all the work done on them can be reused by the entire public sector.

The result of modeling Public Administrations procedures can help sustainability policies in two ways:

- Improving decision-making procedures. A better processes knowledge, administrative or not, of those that are carried out in the scope of the organizations leads to improved decision-making procedures as much as they are based on a global and integrating vision.
- Administrative management sustainability. Apart from the undoubted improvement in the environmental, social, and economic impact that the implementation of a whole electronic administration implies, it is necessary to highlight the benefits that an adequate knowledge of the internal processes can imply to improve material, economic, and human resources management.

2. Motivation

The goal of the present paper is not only to develop an ontology to represent administrative procedures, but also to facilitate tools which can make these processes actionable objects and facilitate the future development of intelligent agents that analyze and process them favoring organization sustainability through an improvement of policy decision taking.

2.1. Administrative Procedure Visualization

The information of administrative processes represented in a semantic model can be employed to improve the information provided by Public Administrations when a service is applied by a citizen.

Through the information available in administrative procedure catalogues, it is possible to provide high quality information about the tasks that Public Administrations perform when processing citizen applications. Later, through files consultant services, citizens will be able to know the state of a file compared to the tasks established by a procedure.

In this way, citizens can know which part of a procedure has been performed and which part remains to be done, whether they need to get in contact with the administration, and important deadlines, etc.

2.2. Predict Input/Output Registry Activity

The daily registry annotations made through the different channels can be predicted through the information stored in the Public Administrations. In this way, human resources available in the information offices and the capacity of the TI infrastructure can be planned in a more efficient way.

2.3. Administrative Procedures Scheduling

Thanks to historical processing information, the calls for grants and subsidies can be rescheduled to provide a workload of the service that is as linear as possible, avoiding the civil servants work peaks and valleys.

2.4. Analysis and Comparison of Administrative Procedures

From historical administrative processing and given that for the same family of procedures the processing scheme is similar, it will be possible to study the Public Administrations processes and to obtain information about:

- The accuracy between the processing of files and what is defined in the standard.
- The differences in the definition of the same administrative procedure in two different Public Administrations.
- How the processing of records of a specific procedure has evolved over time in the same Service.
- Other processing analyses of most frequent actions, which documentation is most commonly remedied, and the detection of bottlenecks, etc.

3. Background and Related Work

In this section, we present work related to our study in different subjects. First of all, we present different approaches to represent workflows that go from BPMN to provenance, and finally we present cases related to the e-Government domain.

Before reviewing related work, it is necessary to define what we understand by administrative procedures. We define administrative procedures and files as:

- Administrative procedures are the ordered set of procedures and actions normally carried out, according to the channel legally envisioned, to dictate an administrative act or express some Administration requirement [3].
- Administrative files are the materialization of an administrative procedure and consist of an ordered set of documents and actions that serve as a background and form the basis of an administrative resolution, as well as the proceedings directed to its execution.

In this way, we can consider an administrative procedure as an action protocol, a "future" process model, which is the collection of actions that can be carried out in the processing of each individual file. Furthermore, administrative files are the execution traces, instances of each particular procedure that have already happened or are happening, which will require and generate information.

Given the importance for organizations of knowledge about their own processes, it has been necessary to obtain a set of tools that allows the design and modeling of this process management (Business Process Management, BPM). This set of tools is called Business Process Management Software

(BPMS) and uses a common notation called Business Management Modeling Notation (BPMN). There have been several attempts to combine BPM with semantic web technology, as we can see in [5] and [6]. Although BPMN has been used successfully in the industry, only a few electronic file processing systems implement BPMN as a workflow engine. This is because civil servants prefer tools that are more flexible, compared to the rigid ones that BPMN tends to include. This was the main reason why we decided not to follow BPMN semantic technology to represent administrative process, because we were searching for a model that could fit in almost every system.

The World Wide Web Consortium has been working to represent and model provenance data with the development of the Provenance (PROV-O) ontology [7], which was designed to represent the actions carried out in the preparation of any information element or software. Provenance describes the way that information entities are created and handled by activities when different agents are involved [8].

Provenance can only be used as a process representation base once activities have happened, i.e., “in the past”, because it can describe activities that have already happened in some record and it describes who has done those activities. However, provenance does not help to represent administrative procedures “in the future”, which would be the expected processes and all proceedings that will have to be performed in each of the records.

This problem has already been addressed in other domains that have previously applied provenance, like the definition of scientific processes [9,10], where they need to establish the relationship between a process execution and its theoretical execution plan. In that way, a Provenance specialization for process modeling called P-PLAN was proposed [11].

In the Public Administration domain, there has been a lot of work undertaken to achieve the goal of representing e-Government information. At the level of data representation, there are many references like:

- Core vocabularies [12]. The core vocabularies are simplified, re-usable, and extensible data models that capture the fundamental characteristics of an entity in a context-neutral fashion. Nowadays, the current core vocabularies are: Business, location, person, public service, criterion & evidence, and public organization vocabulary.
- Common Directory of Organization Units and Offices (DIR3) [13]. The Common Directory is conceived as an information repository about the organizational structure of a Public Administration and its customer offices. It is a catalogue of administrative units and bodies, administrative registry offices, and public administration citizen services offices.
- Contsem project [14]. The PPROC ontology defines the necessary concepts to describe public procurement processes and the contracts of the public sector (public e-procurement). The ontology has been designed with the main purpose of publishing data about public contracts. This ontology extends the Public Contracts Ontology, an ontology developed by the Czech Open Data initiative.

In order to have a view that is as complete as possible about the administrative domain, it is necessary to join both perspectives: the data and the process level representation. The challenge to support administrative processes in public administrations was described in [15], where the need to model process patterns, as well as workflow and record management, was identified. The use of ontologies to model public administration processes was also proposed in [16], where a specific domain ontology was proposed. In this paper, we present a different approach, based on the reuse of existing and more general ontologies like PROV-O and P-PLAN. We propose an extension that reuses these ontologies to increase reusability and to represent not only the processes themselves, but the processes that have been carried out for the generation of some information.

The objective of this paper is to present a representation framework for administrative processes according to the latest standards of the semantic web that have been applied in other areas, such as scientific processes using PROV-O and P-PLAN.

4. Modelling an Extending the Ontologies

In order to use the provenance ontology (PROV-O) and P-PLAN, it is necessary to define several extensions to leverage and reuse their concepts in a new context. We have used the mechanisms provided both in the Provenance Data Model (PROV-DM) and PROV-O to extend and adapt those ontologies to the administrative procedure domain.

Before describing the extensions developed, we will define the most important entities that we want to represent:

- Service Catalogue. This is an inventory of the services provided by public administrations and is available to citizens in the scope of their competences.
- Procedures Catalogue. This is the internal view of the Service Catalogue which contains the set of administrative acts and phases that are part of the procedure as an answer to a public service.
- Administrative phase. This represents an activity that must be done in the context of a procedure.
- Administrative act. Each of the administrative actions made in the context of a file and performed inside a phase.

Because of this extension, we have developed a new vocabulary for administrative procedures called A-PROC, which is available at [17].

In addition to the above, there are several reasons to choose an ontology to model this representation. Ontologies can be considered as a set of representational primitives with which it is possible to model some domain of knowledge. Those representational primitives describe the entities, properties, and relationships that are possible in some specific domain [16]. As we are describing the entities and relationships that take part in administrative procedures, it is suitable to use an ontology to describe that specific domain. Ontology reuse allows us to improve the ontology development process, saving time and money, and promoting the application of good practices [18]. The objective of the ontology proposed is not only to describe the entities of administrative procedures, but also to describe information about the entities, activities, and people involved in producing them. In this way, we reuse PROV-O and P-PLAN ontologies that have been successfully applied in other domains.

A key advantage of using ontologies is to enable knowledge sharing. In Section 5, we show how the ontology proposed can be applied to the information systems of the Principality of Asturias. As most procedures from Public Administrations come from a common legal regulation, having a specific domain ontology for them can improve knowledge sharing between different administrations and improve the transparency of this process for the citizens.

To improve government open data initiatives, Public Administrations will preferably use formats that offer semantic representation of the information, enabling a better understanding of the information represented and its automated treatment [1].

4.1. Extending PROV-O to Represent Administrative Procedures

In the same way that Provenance is employed to declare the origin of a given information resource, it is very useful to know the steps that have been followed in the context of a Public Administration to manage and generate the information resources that are employed. In the context of Public Administrations, one possibility is to declare the tasks that have been carried out, who has undertaken them, and what new information has been generated from them, with the goal of being able to analyze, track, and share that information or even to warrant the quality and integrity of the process that has been used.

Figure 1 depicts the way in which an administrative process is created. It shows the PROV-O extension to represent administrative files. We have created subclasses of prov:Entity for the entities: AdministrativeFile, FilePhase, FileAction, and Document. In the same way, we have defined activities which are subclasses of prov:Activity to describe the actions that manipulate previous entities. Finally, we defined the class Employee as a subclass of prov:Agent to represent public administration employees.

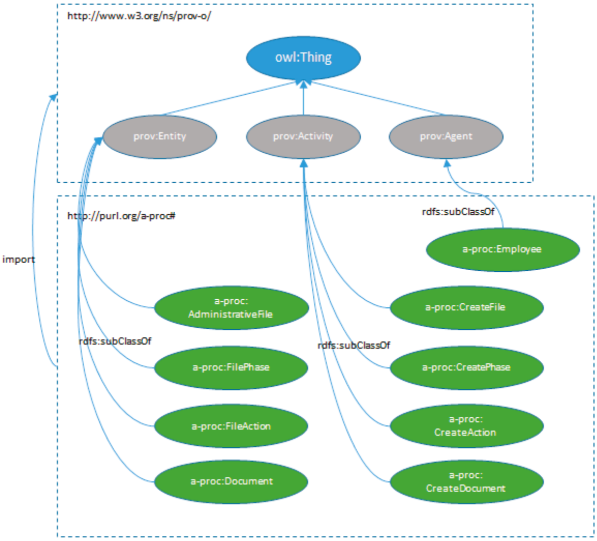


Figure 1. PROV-O extension for administrative procedures.

As an example, Figure 2 shows an example of an administrative file identified as :2016_208584. From a citizen request by citizen: 646176 and its corresponding input registry annotation (:annotation201701) generated by employee: Mathew, an administrative file is created which relates the citizen request and the provided documentation. Figure 3 shows the same information depicted in Figure 2 but in Turtle format.

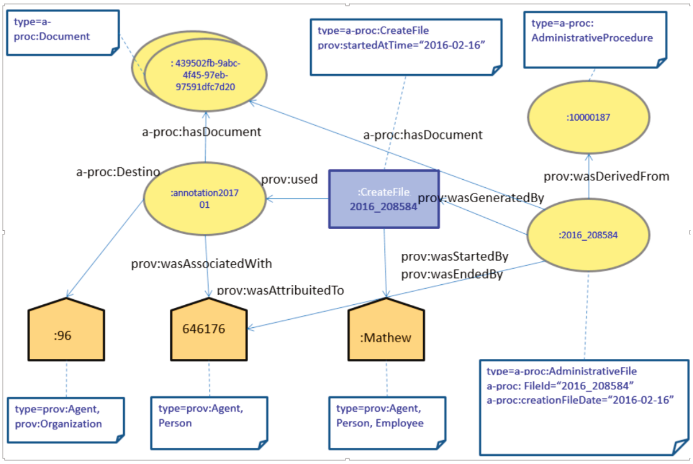
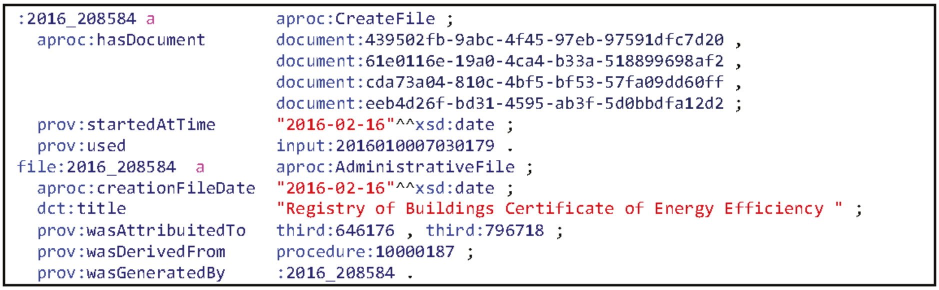


Figure 2. Example of administrative record.



```
:2016_208584 a           aproc:CreateFile ;
  aproc:hasDocument      document:439502fb-9abc-4f45-97eb-97591dfc7d20 ,
                         document:61e0116e-19a0-4ca4-b33a-518899698af2 ,
                         document:cda73a04-810c-4bf5-bf53-57fa09dd60ff ,
                         document:eeb4d26f-bd31-4595-ab3f-5d0bbdfa12d2 ;
  prov:startedAtTime     "2016-02-16"^^xsd:date ;
  prov:used              input:2016010007030179 .
file:2016_208584  a      aproc:AdministrativeFile ;
  aproc:creationFileDate "2016-02-16"^^xsd:date ;
  dct:title              "Registry of Buildings Certificate of Energy Efficiency " ;
  prov:wasAttribuitedTo  third:646176 , third:796718 ;
  prov:wasDerivedFrom    procedure:10000187 ;
  prov:wasGeneratedBy    :2016_208584 .
```

Figure 3. Turtle representation of information depicted in Figure 2.

4.2. Extending PROV-O to Represent Registry Annotations

All documentation entering into or exiting from a Public Administration must have its corresponding annotation in a registry. In the case of input information, the documentation that is managed includes requests or writings, made by citizens, companies, institutions, or other administrations, which are presented to some organization department. The path which these requests undergoes, from the point in which they are presented until they arrive to their destination, varies depending on which channel they have been presented to, and above all, depending on the nature of that documentation, but it is always mandatory to keep a record about the point at which a request is at. In the case of the output record, it is usually different, as the only mandatory information that is usually represented is the instant at which the documentation has exited the organization, its origin, and its destination.

All this information is referred to as an annotation registry, and can be used to keep evidence about documentation traceability, which can be especially important for organizations in order to improve their internal processes and the response and delay time for the citizens.

Registry annotations do not have legal regulation as administrative procedures, and it is not so easy to establish a relationship between the registry annotation and its theoretical execution plan. This is the reason why we do not use P-PLAN to represent registry annotations.

Figure 4 depicts the representation of annotation records and their relationship with PROV-O concepts. Annotation records are defined as subclasses of prov:Entity to represent each annotation. Given that the information recorded differs depending on an input or output annotation, we created two classes :InputRegistryAnnotation and :OutputRegistryAnnotation. In the same way as we did for administrative records, we created the corresponding activities to manage these activities. The Employee class is also defined as a subclass of prov:Agent to represent public administration employees.

Figure 5 shows an example of how an annotation record is defined from the documentation presented by a citizen in some registry office. The annotation record has been created and has a given department from the organization as the destination. Figure 6 shows the same information in Turtle.

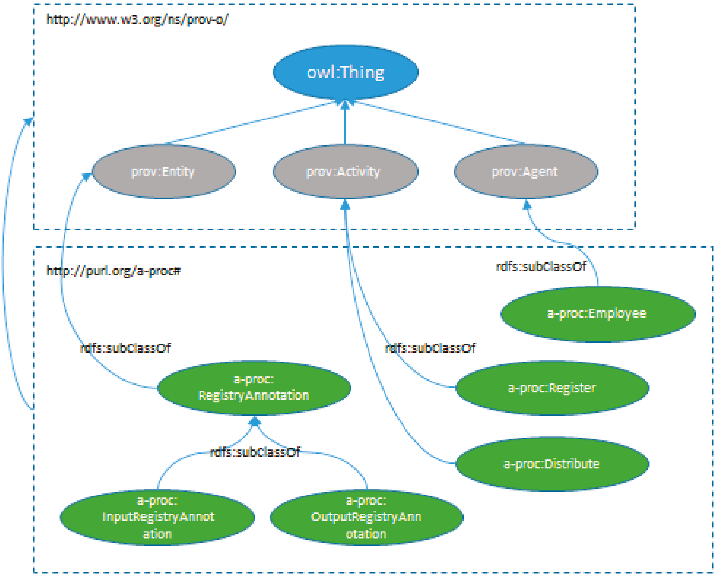


Figure 4. Input-Output registry annotation representation related to PROV-O.

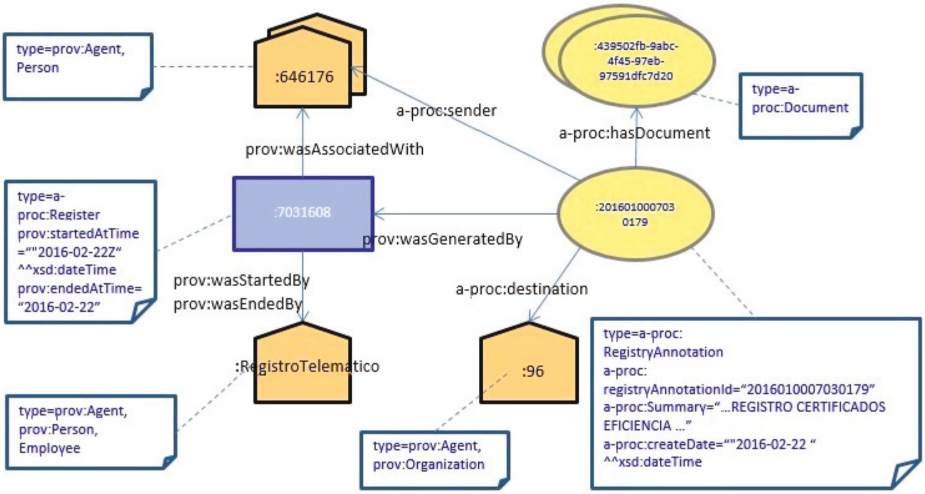
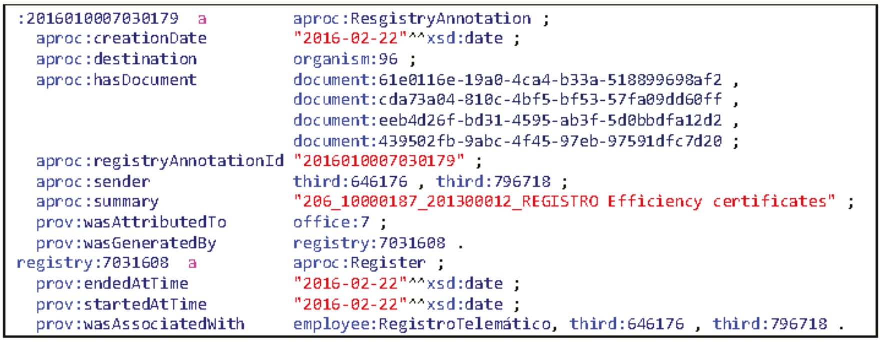

type=prov:Agent, Person
:646176
:439502fb-9abc-4f45-97eb-97591dfc7d20
type=a-proc:Document
a-proc:sender
prov:wasAssociatedWith
a-proc:hasDocument
type=a-proc:Register
prov:startedAtTime =""2016-02-22Z"
^^xsd:dateTime
prov:endedAtTime= "2016-02-22"
:7031608
:2016010007030179
prov:wasGeneratedBy
prov:wasStartedBy
prov:wasEndedBy
a-proc:destination
:RegistroTelematico
:96
type=prov:Agent, prov:Person, Employee
type=prov:Agent, prov:Organization
type=a-proc:RegistryAnnotation
a-proc:registryAnnotationId="2016010007030179"
a-proc:Summary="...REGISTRO CERTIFICADOS EFICIENCIA ..."
a-proc:createDate=""2016-02-22 "
^^xsd:dateTime

Figure 5. Example of annotation record.



```
:201601000703O179  a          aproc:ResgistryAnnotation ;
  aproc:creationDate          "2016-02-22"^^xsd:date ;
  aproc:destination           organism:96 ;
  aproc:hasDocument           document:61e0116e-19a0-4ca4-b33a-518899698af2 ,
                              document:cda73a04-810c-4bf5-bf53-57fa09dd60ff ,
                              document:eeb4d26f-bd31-4595-ab3f-5d0bbdfa12d2 ,
                              document:439502fb-9abc-4f45-97eb-97591dfc7d20 ;
  aproc:registryAnnotationId "201601000703O179" ;
  aproc:sender                third:646176 , third:796718 ;
  aproc:summary               "206_10000187_201300012_REGISTRO Efficiency certificates" ;
  prov:wasAttributedTo        office:7 ;
  prov:wasGeneratedBy         registry:7031608 .
registry:7031608  a           aproc:Register ;
  prov:endedAtTime            "2016-02-22"^^xsd:date ;
  prov:startedAtTime          "2016-02-22"^^xsd:date ;
  prov:wasAssociatedWith      employee:RegistroTelemático, third:646176 , third:796718 .
```

Figure 6. Turtle representation of annotation registry.

4.3. Extending P-PLAN to Represent Administrative Procedures

We have seen how we can define administrative procedures as process execution models that will later be carried out in the context of a public administration: the administrative records. We have also seen that it is possible to represent them as an extension of P-PLAN using the ontology mechanisms.

In Figure 7 depicts the representation of administrative procedures and their relationship with P-PLAN concepts. Figure 8 shows the representation of an individual administrative procedure in turtle format.

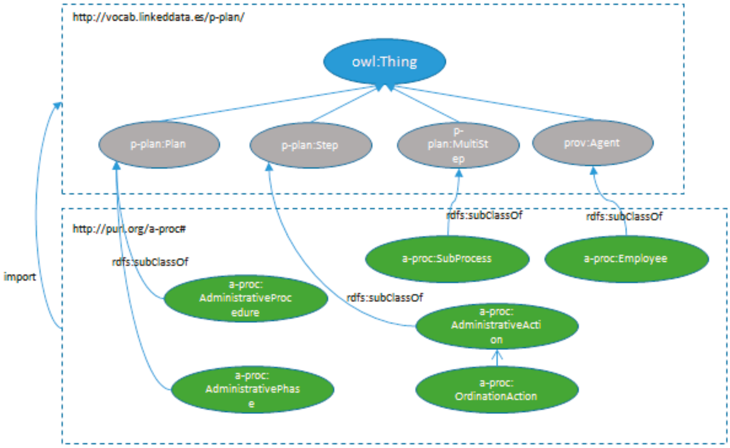
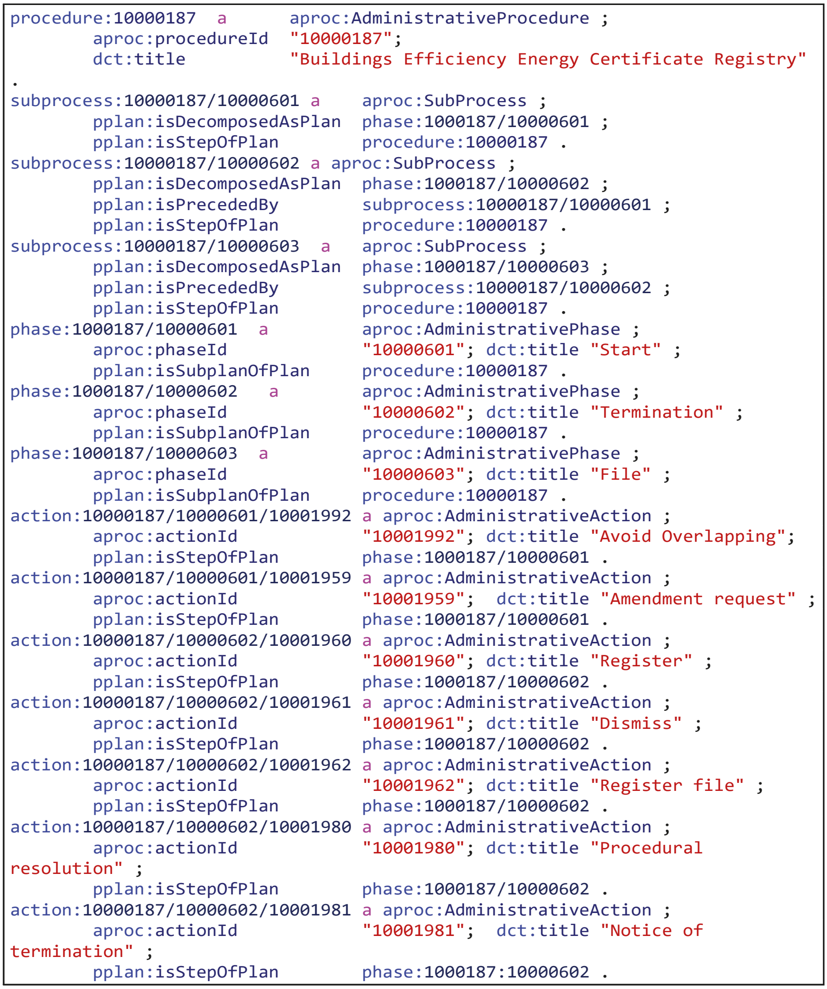


Figure 7. Relation between P-Plan concepts and Administrative Procedure, Administrative Phase, and Administrative Action.



```
procedure:10000187  a      aproc:AdministrativeProcedure ;
        aproc:procedureId  "10000187";
        dct:title          "Buildings Efficiency Energy Certificate Registry"
.
subprocess:10000187/10000601 a    aproc:SubProcess ;
        pplan:isDecomposedAsPlan  phase:1000187/10000601 ;
        pplan:isStepOfPlan        procedure:10000187 .
subprocess:10000187/10000602 a aproc:SubProcess ;
        pplan:isDecomposedAsPlan  phase:1000187/10000602 ;
        pplan:isPrecededBy        subprocess:10000187/10000601 ;
        pplan:isStepOfPlan        procedure:10000187 .
subprocess:10000187/10000603  a   aproc:SubProcess ;
        pplan:isDecomposedAsPlan  phase:1000187/10000603 ;
        pplan:isPrecededBy        subprocess:10000187/10000602 ;
        pplan:isStepOfPlan        procedure:10000187 .
phase:1000187/10000601   a        aproc:AdministrativePhase ;
        aproc:phaseId             "10000601"; dct:title "Start" ;
        pplan:isSubplanOfPlan     procedure:10000187 .
phase:1000187/10000602    a       aproc:AdministrativePhase ;
        aproc:phaseId             "10000602"; dct:title "Termination" ;
        pplan:isSubplanOfPlan     procedure:10000187 .
phase:1000187/10000603   a        aproc:AdministrativePhase ;
        aproc:phaseId             "10000603"; dct:title "File" ;
        pplan:isSubplanOfPlan     procedure:10000187 .
action:10000187/10000601/10001992 a aproc:AdministrativeAction ;
        aproc:actionId            "10001992"; dct:title "Avoid Overlapping";
        pplan:isStepOfPlan        phase:1000187/10000601 .
action:10000187/10000601/10001959 a aproc:AdministrativeAction ;
        aproc:actionId            "10001959";  dct:title "Amendment request" ;
        pplan:isStepOfPlan        phase:1000187/10000601 .
action:10000187/10000602/10001960 a aproc:AdministrativeAction ;
        aproc:actionId            "10001960"; dct:title "Register" ;
        pplan:isStepOfPlan        phase:1000187/10000602 .
action:10000187/10000602/10001961 a aproc:AdministrativeAction ;
        aproc:actionId            "10001961"; dct:title "Dismiss" ;
        pplan:isStepOfPlan        phase:1000187/10000602 .
action:10000187/10000602/10001962 a aproc:AdministrativeAction ;
        aproc:actionId            "10001962"; dct:title "Register file" ;
        pplan:isStepOfPlan        phase:1000187/10000602 .
action:10000187/10000602/10001980 a aproc:AdministrativeAction ;
        aproc:actionId            "10001980"; dct:title "Procedural
resolution" ;
        pplan:isStepOfPlan        phase:1000187/10000602 .
action:10000187/10000602/10001981 a aproc:AdministrativeAction ;
        aproc:actionId            "10001981";  dct:title "Notice of
termination" ;
        pplan:isStepOfPlan        phase:1000187:10000602 .
```

Figure 8. Turtle representation of an administrative procedure.

5. Evaluation

With the goal of evaluating the expressiveness of the ontology developed and to foster future research, we have created a knowledge base from data of the Principality of Asturias Public Administration. The data has been collected from the administrative procedures recorded from 2012 to 2016.

The systems selected as information sources have been:

- Administrative Procedures Management Systems. We employed custom information systems which are being used by the Principality of Asturias public administration: EUG (Unified Management Desktop or *Escritorio Unificado del Gestor,* in Spanish*)* and SPIGA (Administrative Management and Production Support Integration or *Soporte Producción Integración y Gestión Administrativa,* in Spanish). SPIGA is a clinet-server application used since 2002 with file processing functionalities and a document management system. EUG was a project developed following the principles of openFWPA [19] and a Service Oriented Architecture to provide file processing functionalities, digital signature, interoperability, and a mechanism to allow integration with other applications through web services.

- Input/Output registry (*Registro E/S,* in Spanish). The input/output document registry system is the system where registry annotations are completed and oversees the distribution of records from their creation until they arrive at their destination. It is also the system where output annotations are recorded for output records.

The period of 2012–2016 was chosen in order to have closed files which will no longer be modified by civil servants. The three systems identified as the origin of the information use relational databases to store the information, so we used R2RML [20] scripts to transform the data of the administrative procedures and store them in an RDF triple-store, as is shown in Figure 9.

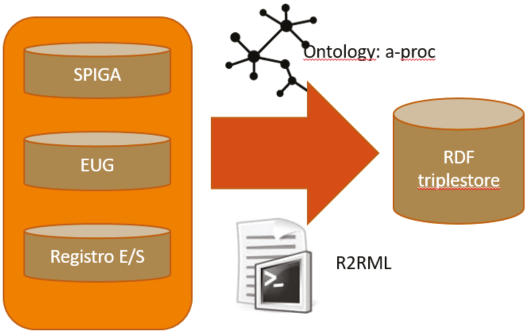


Figure 9. Transformation data schema.

The current system metrics are:

- 6.8 million annotation records.
- 428 administrative procedures.
- 310.000 administrative files.

According to the legal regulation, the administrative procedures are classified as follows:

- Grants and subsidies: 225.
- Authorizations, licenses and registrations: 127.
- Complaints and sanctions: 39.
- Human resources: 6.
- Others: 31.

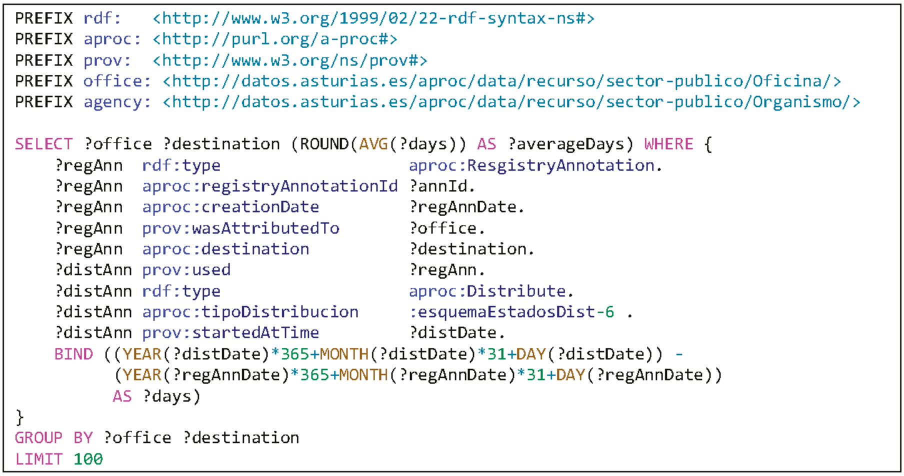
There is a disproportion between the number of annotation records and the number of administrative files and procedures. This is because the annotation registry system covers the whole public administration, and the Administrative Procedures Management System only covers certain sectors in the Principality of Asturias administration. These systems cover sectors like tourism, retail, or industry, but do not cover other main sectors like healthcare, education, or agriculture, which have their own specific systems.

We defined several SPARQL queries that can be applied to the knowledge base to obtain some information that could help decision taking, as an example.

5.1. Compute Average Time of Annotation Record Distribution

A good way to know how well the mechanisms of document distribution work for information provided by a citizen is to calculate the average time it takes between a citizen presenting some documentation in a public administration office and the point at which that information arrives at its destination. Figure 10 shows a SPARQL query that computes, from annotation records, those distribution actions with a state value of 6 (which means “accepted”), and from those two elements, it computes the number of days as the difference between the creation and acceptance date and returns the average of all those values for the same office and the same destination code.

Table 1 shows the results of that SPARQL query. As can be seen, the results vary between 0 to more than 85 days. From this information, an administration policy manager can decide that there are some cases where the document distribution process must be reviewed.



```
PREFIX rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX aproc: <http://purl.org/a-proc#>
PREFIX prov:  <http://www.w3.org/ns/prov#>
PREFIX office: <http://datos.asturias.es/aproc/data/recurso/sector-publico/Oficina/>
PREFIX agency: <http://datos.asturias.es/aproc/data/recurso/sector-publico/Organismo/>

SELECT ?office ?destination (ROUND(AVG(?days)) AS ?averageDays) WHERE {
    ?regAnn  rdf:type                     aproc:ResgistryAnnotation.
    ?regAnn  aproc:registryAnnotationId ?annId.
    ?regAnn  aproc:creationDate          ?regAnnDate.
    ?regAnn  prov:wasAttributedTo        ?office.
    ?regAnn  aproc:destination           ?destination.
    ?distAnn prov:used                    ?regAnn.
    ?distAnn rdf:type                     aproc:Distribute.
    ?distAnn aproc:tipoDistribucion      :esquemaEstadosDist-6 .
    ?distAnn prov:startedAtTime          ?distDate.
    BIND ((YEAR(?distDate)*365+MONTH(?distDate)*31+DAY(?distDate)) -
          (YEAR(?regAnnDate)*365+MONTH(?regAnnDate)*31+DAY(?regAnnDate))
          AS ?days)
}
GROUP BY ?office ?destination
LIMIT 100
```

Figure 10. SPARQL query to calculate average annotation record distribution time.

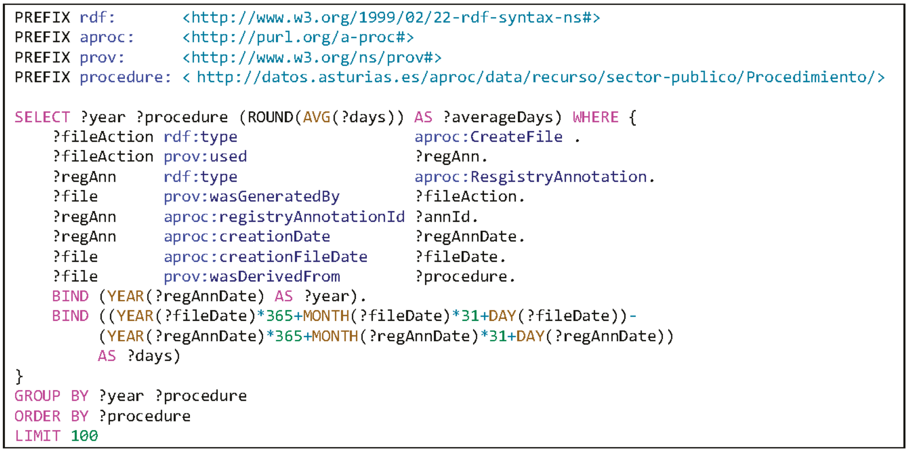
Table 1. Query results.

	Office	Destination	AverageDays
1	office:0	agency:97	“9”^^xsd:decimal
2	office:611	agency:830055	“2”^^xsd:decimal
3	office:611	agency:830057	“0”^^xsd:decimal
4	office:611	agency:831061	“0”^^xsd:decimal
5	office:614	agency:510	“15”^^xsd:decimal
6	office:622	agency:96	“0”^^xsd:decimal
7	office:629	agency:431	“85”^^xsd:decimal

5.2. Compute Average Record Creation Time

Another possible query is to measure the impact that some policy decisions can have on the organization performance in order to speed up the citizen response time. For example, it is possible to measure the time between requests are presented by a citizen and the point at which a record has been created by a management body. In this example, we have taken the creation date as a reference, but it could be extended to other indicators like the record resolution date or the payment date.

Figure 11 shows how to obtain the records which are related to some annotation registry and computes the number of days as the difference between the record creation and the record annotation date. The results are represented in Table 2, where it is possible to see that some values are abnormally high, like 22 or 12 days, which should be reviewed by a policy maker.



```
PREFIX rdf:       <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX aproc:     <http://purl.org/a-proc#>
PREFIX prov:      <http://www.w3.org/ns/prov#>
PREFIX procedure: < http://datos.asturias.es/aproc/data/recurso/sector-publico/Procedimiento/>

SELECT ?year ?procedure (ROUND(AVG(?days)) AS ?averageDays) WHERE {
    ?fileAction rdf:type                    aproc:CreateFile .
    ?fileAction prov:used                   ?regAnn.
    ?regAnn     rdf:type                    aproc:ResgistryAnnotation.
    ?file       prov:wasGeneratedBy         ?fileAction.
    ?regAnn     aproc:registryAnnotationId ?annId.
    ?regAnn     aproc:creationDate          ?regAnnDate.
    ?file       aproc:creationFileDate      ?fileDate.
    ?file       prov:wasDerivedFrom         ?procedure.
    BIND (YEAR(?regAnnDate) AS ?year).
    BIND ((YEAR(?fileDate)*365+MONTH(?fileDate)*31+DAY(?fileDate))-
         (YEAR(?regAnnDate)*365+MONTH(?regAnnDate)*31+DAY(?regAnnDate))
         AS ?days)
}
GROUP BY ?year ?procedure
ORDER BY ?procedure
LIMIT 100
```

Figure 11. SPARQL query to compute average record creation time.

Table 2. Results of SPARQL query.

	Year	Procedure	Average Days
1	“2016”^^xsd:integer	procedure:10000189	“0”^^xsd:decimal
2	“2016”^^xsd:integer	procedure:10000208	“7”^^xsd:decimal
3	“2016”^^xsd:integer	procedure:20000388	“22”^^xsd:decimal
4	“2016”^^xsd:integer	procedure:40500	“3”^^xsd:decimal
5	“2016”^^xsd:integer	procedure:40600	“12”^^xsd:decimal
6	“2016”^^xsd:integer	procedure:40800	“2”^^xsd:decimal
7	“2016”^^xsd:integer	procedure:63200	“11”^^xsd:decimal
8	“2016”^^xsd:integer	procedure:64000	“2”^^xsd:decimal

6. Conclusions

The work presented in this paper can be used to open a new research line about how to represent knowledge in the Public Administration Procedures domain, leveraging general propose vocabularies that have already been applied in other domains.

We did not try to develop a new ontology vocabulary from scratch, but we tried to reuse and extend existing ontology concepts from PROV-O and P-PLAN in order to obtain the necessary expressiveness. Nevertheless, these extensions are not intended to be a final work, but a basis on which to develop future research or standardization works.

We have detected several circumstances in records and administrative procedures that are not completely covered with our proposed extensions. Defining an administrative procedure with P-PLAN represents an ideal situation where records are always forwardly processed; however, this is not always the case. In practice, administrative files processing can become more complex, affected by several decisions, some of which are derived from some concrete data from the file, which can affect the procedure linearity producing forward leaps or even backward leaps. One of the intervening factors in this distortion is the human factor, i.e., the people who control the record processing. It is possible that, depending on the information or situation of some file, some person decides to take an action which is different from those actions that were originally planned or to omit some of them. We consider that being able to detect these situations can also help policy makers to take some decisions to detect or avoid these exceptions, or even to mitigate possible corrupt behavior.

The extensions proposed in this paper have been implemented in accordance with the Spanish rules for Public Administration as a case study; it is possible that during some future standardization process, it will be necessary to adapt and generalize some concepts so that they can be applied in other domains.

Author Contributions: Francisco José Hidalgo López analyzed, designed, developed the A-Proc ontology and made the evaluation of the work. Francisco José Hidalgo López, Jose Emilio Labra Gayo and Patricia Ordóñez de Pablos conceptualized and wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Secretariat of State for Public Administrations. Technical norm about Reusability Interoperability of Information Resources. 2013. Available online: https://administracionelectronica.gob.es/pae_Home/dam/jcr:5842379d-8d7f-4542-87c2-041056cc1e24/2017-06-nota-tecnica-RISP.pdf (accessed on 18 February 2018).
2. Head of State. Law 19/2013, of December 9, on transparency, access to information and good governance. 2013. Available online: https://www.boe.es/boe/dias/2013/12/10/pdfs/BOE-A-2013-12887.pdf (accessed on 18 February 2018).
3. Law 39/2015, of October 1, on the Common Administrative Procedure of Public Administrations. 2015. Available online: https://www.boe.es/boe/dias/2015/10/02/pdfs/BOE-A-2015-10565.pdf (accessed on 18 February 2018).
4. Spanish Open Data Catalog. 2017. Available online: http://datos.gob.es/en (accessed on 18 February 2018).
5. Kalogeraki, E.-M.; Apostolou, D.; Panayiotopoulos, T.; Tsihrintzis, G.; Theocharis, S. A Semantic Approach for Representing and Querying Business Processes. Available online: https://link.springer.com/chapter/10.1007/978-3-662-49179-9_4 (accessed on 14 February 2018).
6. Hepp, M.; Leymann, F.; Domingue, J.; Wahler, A.; Fensel, D. Semantic Business Process Management: A Vision towards Using Semantic Web Services for Business Process Management. Available online: http://ieeexplore.ieee.org/abstract/document/1552942/ (accessed on 14 February 2018).
7. Lebo, T.; Sahoo, S.; McGuinness, D. PROV-O: The PROV Ontology. Available online: https://www.w3.org/TR/prov-o/ (accessed on 18 February 2018).
8. Moreau, L.; Groth, P. Provenance: An Introduction to PROV. Available online: http://www.morganclaypool.com/doi/abs/10.2200/S00528ED1V01Y201308WBE007 (accessed on 14 February 2018).
9. Gil, Y. From Data to Knowledge to Discoveries: Scientific Workflows and Artificial Intelligence. Available online: http://trellis.semanticweb.org/~gil/papers/gil-sp08.pdf (accessed on 14 February 2018).
10. GIl, Y. Intelligent Workflow Systems and Provenance-Aware Software. In Proceedings of the 7th International Congress on Environmental Modelling and Software, San Diego, CA, USA, 16 June 2014.
11. Garijo, D.; Gil, Y. Augmenting PROV with Plans in P-PLAN: Scientific Processes as Linked Data. Available online: http://linkedscience.org/wp-content/uploads/2012/05/lisc2012_submission_12.pdf (accessed on 18 February 2018).
12. European Commission. Semantic Interoperability Community. e-Government Core Vocabularies. 2017. Available online: https://joinup.ec.europa.eu/page/core-vocabularies (accessed on 18 February 2018).
13. Ministerio de Hacienda y Función Pública Secretaría General de Administración Digital. Directorio Común de Unidades Orgánicas y Oficinas (DIR3). Available online: https://administracionelectronica.gob.es/ctt/dir3 (accessed on 18 February 2018).
14. Muñoz, J.; Esteban, G.; Bernal, M.; Serón, F. Public Procurement Ontology (PPROC). Available online: http://contsem.unizar.es/ (accessed on 14 February 2018).
15. Savvas, I.; Bassiliades, N. A process-oriented ontology-based knowledge management system for facilitating operational procedures in public administration. *Expert Syst. Appl.* **2009**, *36*, 4467–4478. [CrossRef]
16. Chandrasekaran, B.; Josephson, J.; Benjamins, V. What Are Ontologies, and Why Do We Need Them? Available online: http://ieeexplore.ieee.org/abstract/document/747902/ (accessed on 14 February 2018).
17. Hidalgo López, F.J. Administrative Procedures. Available online: http://www.purl.org/a-proc (accessed on 18 February 2018).
18. Poveda-Villalón, M.; Suárez-Figueroa, M.; Gómez-Pérez, A. Reusing ontology Desing Patterns in a Context Ontology Network. In Proceedings of the Second Workshop on Ontology Patterns (WOP 2010), Shanghai, China, 8 November 2010.

19. Government of the Principality of Asturias. openFWPA: Open Framework for the Public Administration. Available online: https://www.asturias.es/openFWPA (accessed on 18 February 2018).
20. Das, S.; Sundara, S.; Cyganiak, R. R2RML: RDB to RDF Mapping Language. Available online: https://www.w3.org/TR/r2rml/ (accessed on 14 February 2018).

sustainability

MDPI

Article

An Empirical Study on Visualizing the Intellectual Structure and Hotspots of Big Data Research from a Sustainable Perspective

Feng Hu [1,2,3], Wei Liu [4,*], Sang-Bing Tsai [5,6,*], Junbin Gao [3], Ning Bin [7] and Quan Chen [5,*]

[1] School of Management, Guangdong University of Technology, Guangzhou 510520, China; fenghu@gdut.edu.cn
[2] Institute of Big Data Strategic Research, Guangdong University of Technology, Guangzhou 510006, China
[3] Discipline of Business Analytics, The University of Sydney Business School, The University of Sydney, Camperdown, NSW 2006, Australia; junbin.gao@sydney.edu.au
[4] School of Business, Wuyi University, Nanping 354300, China
[5] Zhongshan Institute, University of Electronic Science and Technology of China, Guangzhou 528400, China
[6] Economics and Management College, Civil Aviation University of China, Tianjin 300300, China
[7] School of Management, Guangdong University of Technology, Guangzhou 510520, China; bbb8087@gdut.edu.cn
* Correspondence: wei.liu2@sydney.edu.au (W.L.); sangbing@hotmail.com (S.-B.T.); zschenquan@gmail.com (Q.C.)

Received: 6 February 2018; Accepted: 26 February 2018; Published: 1 March 2018

Abstract: Big data has been extensively applied to many fields and wanted for sustainable development. However, increasingly growing publications and the dynamic nature of research fronts pose challenges to understand the current research situation and sustainable development directions of big data. In this paper, we visually conducted a bibliometric study of big data literatures from the Web of Science (WoS) between 2002 and 2016, involving 4927 effective journal articles in 1729 journals contributed by 16,404 authors from 4137 institutions. The bibliometric results reveal the current annual publications distribution, journals distribution and co-citation network, institutions distribution and collaboration network, authors distribution, collaboration network and co-citation network, and research hotspots. The results can help researchers worldwide to understand the panorama of current big data research, to find the potential research gaps, and to focus on the future sustainable development directions.

Keywords: big data; visualizing; intellectual structure; big data environment; co-citation network; collaboration network; sustainability

1. Introduction

With the growing popularity of mobile terminals, Internet of Things (IoT), social networks, cloud computing and mobile commerce, myriad data are generated, and the era of big data is coming. The advent of big data has promoted the revolution of data-driven thinking and decision making. Governments, industry and academia have paid great attention to big data strategy, technologies and applications. More and more people worldwide have made tremendous efforts in large-scale heterogeneous data collection, organization, storage, analysis, mining and applications under the big data environment. Big data has become a hot topic of discussion. For example, Nature and Science published special issues "Big Data" in 2008 and "Dealing with Data" in 2011 respectively. In May 2011, the McKinsey global institute (MGI) released the research report "Big Data: The Next Frontier for Innovation, Competition, and Productivity" [1]. In March 2012, U.S. President Office of Science and Technology Policy declared in public that the United States government would invest

$200 million to launch "The Big Data Research and Development Initiative" [2]. At the same time, big data had been extensively applied into many fields, such as IoT, social networks, health care, intellisense, environment and sustainable development, and so on [3]. For example, according to the UN Sustainable Development Goals (SDG) [4], big data such as from satellite imagery and sensor networks make environment and development indicators increasingly measurable. Worldwide research institutions and scholars had devoted themselves to big data science research and wanted big data for sustainable development. However, more and more research outcomes have been emerging and growing rapidly [5–7]. Moreover, the dynamic nature of a research front poses challenges for scientists, research policy makers, and many others to keep up with the rapid advances of the state of the art in science [8]. It is still difficult for scholars to understand the current research situation and sustainable development trends of big data. Therefore, how to identify intellectual structure, to detect emerging trends and sudden changes of big data research is increasingly essential.

In recent years, with the rapidly increasing publications related to big data, some scholars have begun to aggregate relevant existing literatures, performed the bibliometric analysis, and visualized the intellectual structure, hotspots and evolution paths to provide knowledge support for other researchers in different fields based on bibliometrics [9–11]. Bibliometrics comprehensively utilizes multi-disciplinary knowledge and methods, such as mathematics, statistics, philology, etc., to analyse the distribution regularities, the developments and research trends of a certain scientific field, and finally visualizes the research results. However, so far few quantitative depictions have been given of the intellectual structure and hotspots of big data research. A few existing surveys mainly focus on specific big data subfields and themes, such as big data and IoT applications on circular economy [10], social networks, health care [11], and supply chain [12]. However, these surveys were absent in the panorama of the big data field and were not conducted on the sustainability of big data research. It is still difficult for readers to deeply understand the current intellectual structure and sustainable development directions of big data research.

In this paper, we performed a bibliometric analysis distinct from the above existing surveys in several aspects. Firstly, this study retrieved all journal articles of big data between 2002 and 2016 in the WoS database, which include Science Citation Index Expanded (SCI-EXPANDED) journals, Social Sciences Citation Index (SSCI) journals, and Arts & Humanities Citation Index (A&HCI) journals. Secondly, this study did not simply describe the traditional concentrated distribution regularities. More importantly, visualization techniques and co-word analysis were used to demonstrate visually the intellectual structures, collaboration networks, and research hotspots of big data between 2002 and 2016 from the following perspectives: publications distribution, core journals, core institutions and collaboration network, core authors and collaboration network, as well as high-frequency keywords network. It provided a vivid overall picture of big data research. This study will help would-be big data researchers know the current research situation, research gaps, what journals they should follow, what authors they should focus, how to seek co-researchers, and work out the details in big data research activities. Moreover, it will also be helpful to improve and upgrade the sustainable research and development, applications, and policy making of big data at different levels in the future. Thirdly, this study went beyond traditional citation counts. Journal co-citation analysis (JCA) and author co-citation analysis (ACA), provided by CiteSpace, were used to detect some special pioneers and journals in the big data field from the following perspectives: the most co-cited frequency, intellectual turning points, and highest citation bursts. These pioneers and journals had contributed to the sustainable development of big data research from different perspectives.

The paper is organized as follows: In Section 2, we describe the methodology, including original data sources and research methods. In Section 3, we demonstrate the bibliometric analysis results, and visualize the intellectual structure and hotspots of big data research. In particular, we detect the distribution characteristics, intellectual turning points, strongest citation bursts, and research hotspots. In Section 4, we finally present the discussion and conclusions.

2. Methodology

2.1. Data Source

The first step consisted in collecting bibliographic data from robust and reliable data sources. Previous bibliographic data were extracted from different data sources. Some collected data from a single journal [13] or multiple journals [14]. Others did not discriminate the journals sources but regarded the citation databases [9,10,15]. Commonly used citation databases include the Web of Science (WoS), Scopus, Google Scholar (GS), and PubMed [15]. Each database has its own advantages and drawbacks. A certain database could be stronger in types, quantities, and countries of publications, while the others focus more on literature evaluation methods and indicators. For example, compared to WoS, Scopus significantly alters the relative ranking of those scholars who appear in the middle of the rankings and GS stands out in its coverage of conference proceedings as well as international, non-English language journals [16]. In this paper, we select the WoS as our data source. The WoS is an ideal single research destination to explore the citation universe across subjects and around the world, and provides everyone access to the most reliable, integrated, multidisciplinary research connected through linked content citation metrics from multiple sources within a single interface. Furthermore, the WoS adheres to a strict evaluation process, and only the most influential, relevant, and credible information is included.

In addition, with the growing emergence of social media, there are a variety of important ways for scientists to spread their academic ideas, such as monographs, conference proceedings, and personal blogs or web pages. Nevertheless, compared with the books, reports, and other equal ways, academic journals tend to be more direct, consistent and important channels for scientists to publish, spread, accumulate, comment on and assume the lead in a specific scientific research fields [17,18]. Furthermore, most key studies are usually published in core international journals [10]. We therefore target the journal articles in the Web of Science.

On 20 May 2017, the WoS database was searched using the following basic terms: topic = “big data”, literature type = “article”, and publication years were restricted to “2002–2016”. We eventually obtained 4927 effective journal articles. The bibliographic records, including titles, authors, institutions, keywords, references, etc., were downloaded. These journal articles were distributed across 1729 journals, and contributed by 16,404 authors from 4137 institutions.

2.2. Research Methods

The methods of bibliometrics have been widely applied in quantitative analyses in many knowledge fields [8–10,12–15,17,19]. It comprehensively uses the professional knowledge and methods of mathematics, statistics, information science, philology, and other disciplines to analyze the distribution regularities, intellectual base, research front, and evolution paths. Commonly used bibliometrics methods include co-word analysis [20], document co-citation analysis (DCA) [11,21], author co-citation analysis (ACA) [22,23] and many other variations [15]. In addition, information visualization, raised by Robertson in 1989, focuses on interactive visual representations of abstract data to reinforce human cognition. Visualization techniques include visualizations of hierarchies or trees [24], graph or network structures [25], temporal structures [26], geospatial visualizations, and coordinated views of multiple types of visualizations [15].

With the development of information technologies, many representative software tools were exploited to facilitate the information visualization and science mapping of knowledge domains. Frequently used information visualization and science mapping software tools include some nonspecific science mapping software (e.g., Pajek, Gephi, or UCINET) and specific science mapping software tools, such as IN-SPIRE [27], VantagePoint [28], CiteSpace II [11,29,30], CoPalRed [31,32], Leydesdorff’s Software [33], Bibexcel [34], Sci2 Tool [35] (Sci2 Team, 2009), VOSViewer [36], Network Workbench Tool [37], SciMAT [38], and so on. Each one presents different features, advantages, and drawbacks due to its own different analysis techniques and algorithms. As a result, there was no single software tool effective and flexible enough to fulfill overall science mapping analysis [39]. Therefore, in this paper, we adopt using more than one software tool to perform deep science mapping analyses. For example, we use an ad hoc software tool SATI3.2 [40] to clean the data in the preprocessing stage, and apply UCINET6 and CiteSpace V to build networks and visualize scientific mapping. SATI3.2, developed by Qiyuan Liu at Zhejiang University (China), is also applied to field data extraction, item frequency statistics, co-occurrence matrix construction, and visual analysis based on NetDraw. It is freely downloadable at http://sati.liuqiyuan.com. UCINET6 for Windows, developed by Lin Freeman, Martin Everett and Steve Borgatti, is a software package for the analysis of social network data. It comes with the NetDraw network visualization tool and can be downloaded at https://sites.google.com/site/ucinetsoftware/home. CiteSpace V, developed by professor Chaomei Chen at Drexel University (USA), is used to focus on visual analysis and scientific mapping. It is a Java-based information visualization and scientific mapping software package and can be freely available at http://cluster.cis.drexel.edu/~cchen/citespace/. The main functions include co-word networks analysis and co-citation networks analyses of authors, documents, institutions and journals. More importantly, CiteSpace V facilitates the identification of the chronologic patterns of a specific knowledge domain, including research hotspots, intellectual turning points, and citation burst.

The aim of this article is to demonstrate visually the intellectual structure and hotspots in big data research from 2002 to 2016. Particularly, the distribution characteristics, intellectual turning points, and emerging trends are examined from the following perspectives: publications, journals, institutions and authors, as well as keywords analysis.

3. Results

3.1. Publications Distribution

To evaluate the outcomes of big data research between 2002 and 2016, we collected 4927 journal articles from WoS databases and tracked the annual publications distribution of big data research (shown in Figure 1). There were few journal articles on big data research before 2009. However, a growth spurt was generated from 2010 to 2016, when dozens, and eventually thousands, of journal articles emerged.

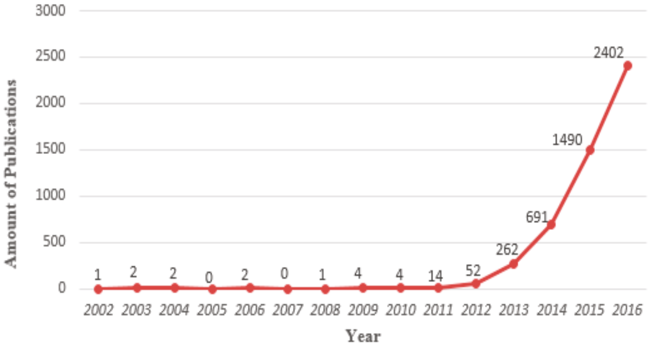


Figure 1. Annual publications distribution of big data research.

As shown in Figure 1, we roughly divided the development of big data research into two stages. Stage I (2002–2009) is an embryonic stage with few annual articles, which indicate that big data exploration just starts. The topics of big data research mostly are the introductions of theories, techniques, and methods related to big data, such as data-mining application architecture [41], SINFONI [42], MapReduce [43], Hive [44], the pathologies of big data [45], large-scale electrophysiology of big data [46], and so on. Stage II (2010–2016) has a rapid growth spurt in annual research outcomes.

In this stage, there were four articles in 2010; by 2016, the number of annual articles sharply increased to 2402, which represented that the number of annual articles had increased 600 times over the past six years. Such a significant change is attributable, to a great extent, to the growing research enthusiasm of governments, scholars and enterprises, such as the research report "Big Data: The Next Frontier for Innovation, Competition, and Productivity" [1], the declaration "The Big Data Research and Development Initiative" [2], the book "Big Data: A Revolution That Will Transform How We Live, Work, and Think" [47], and the worldwide opening of the big data subject. All of these promoted effectively the rapid development of scientific research works related to big data. The studies of big data gradually matured.

To further verify the rapid growth trend of research literatures related to big data in Stage II, we develop a curve-fitting, and find that the curve conforms to the exponential distribution: $y = 5.076e^{1.175t}$, where y is the amount of annual publications, and t is a time sequence between 2010 and 2016. Moreover, according to goodness of fit test, the closer R^2 (R Square, coefficient of determination) is to 1, the better fitting degree of the regression line. The quantitative result shows that $R^2 = 0.974$; R^2 is very close to 1. This result indicates the fitting regression curve has a good reliability of forecast and goodness of fit. Therefore, the annual publications of big data research between 2010 and 2016 grow exponentially and big data has become a hot topic. It is worth worldwide scholars to pay more attention.

Figure 2 shows the annual number of authors who published articles from 2002 to 2016. The line in Figure 2 is similar to the annual publications distribution in Figure 1. There were four authors in 2002, and 14 authors in 2010. However, the number of authors sharply increased to 9558 in 2016, which shows that the number of annual authors has increased hundreds of times over the past several years.

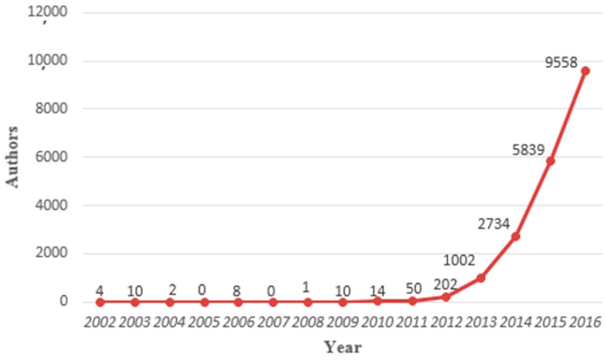


Figure 2. Annual authors distribution.

To further evaluate the annual collaboration ratio of researchers in the big data research field, we depicted average participants per article from 2002 to 2016 (shown in Figure 3). However, we excluded 2005 and 2007 because of mathematics. Figure 3 reveals a trend of collaboration among authors in the big data research field. In 2003, the average number of participants per article reached a maximum of five. However, the value hits rock bottom twice at 2004 and 2008 because of having an independent author in each article. After 2008, this number continued to rise, and reached 3.98 in 2016. Moreover, there was only a slight fluctuation from 2012 to 2016, which indicated that the average number of participants per article in the big data field were between three and four authors. The research collaboration, to some extent, ensured the quality of the publications.

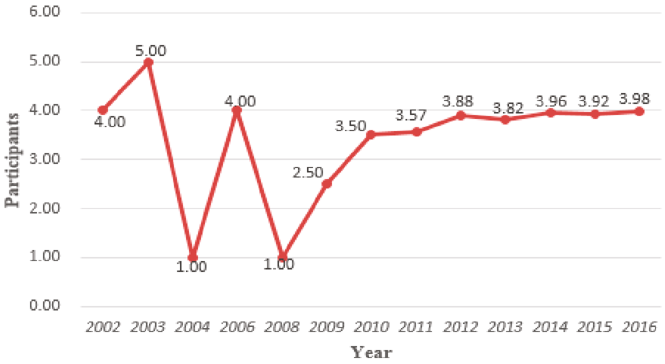


Figure 3. Average participants distribution per article.

3.2. *Journals Distribution and Co-Citation Network*

3.2.1. Core Journals Identification

In this section, we examined 1729 different academic journals. According to Price law, core journals must be the journals which published more than N (note: N = 0.749 × square (69) ≈ 7) articles. According to the statistical analysis, there are 154 core journals. Table 1 lists the top 10 academic journals in descending order of publications. The core academic journal with the most publications in big data research is PLoS One (69), followed by IEEE Access (63), and Big Data (52). There is a narrow gap of less than five publications among Cluster Computing the Journal of Networks, Software Tools and Applications (45), Neurocomputing (45), Journal of Supercomputing (43), and Concurrency and Computation: Practice and Experience (41). IEEE Network, Information Sciences, and International Journal of Distributed Sensor Networks have equal publications (31).

In addition, according to the Journal Citation Reports in the WoS, *IEEE Network* simultaneously has the highest impact factor (IF, 7.230) and immediacy index (1.638) in these top 10 most publications core academic journals of big data research. Moreover, the top 10 academic journals published 450 articles, which account for 9.1% of overall published articles from 2002 to 2016. Simply, it indicates that 0.6% of academic journals in the big data research field published 9.1% of overall articles from 2002 to 2016. It conforms to what is known as a “Matthew effect” in academic journals distribution.

Table 1. Top 10 most publications core academic journals.

Journal	The Number of Publications	Impact Factor in 2016	Five Year Impact Factor	Immediacy Index
PLoS One	69	2.806	3.394	0.396
IEEE Access	63	3.244	3.870	0.607
Big Data	52	1.239	2.292	0.286
Cluster Computing—The Journal of Networks, Software Tools and Applications	45	2.040	2.076	0.339
Neurocomputing	45	3.317	3.211	0.819
Journal of Supercomputing	43	1.326	1.349	0.282
Concurrency and Computation: Practice and Experience	40	1.133	1.219	1.065
IEEE Network	31	7.230	6.410	1.638
Information Sciences	31	4.832	4.732	1.041
International Journal of Distributed Sensor Networks	31	1.239	1.315	0.238

3.2.2. Journals Co-Citation Network

Journals co-citation analyses usually are employed to discover the journals that formed the intellectual base of a knowledge domain. Figure 4 shows the highly cited journals co-citation network from 2002 to 2016. This network is constructed by the top 50 most cited references in each given time slices based on 337 iterations. It contains 195 journals and 489 links among them. Table 2 lists the top 10 highest co-cited journals from 2002 to 2016. The journals with frequencies more than 1000 include Nature (1899), Science (1844), Lecture Notes in Computer Science (1436), PLoS One (1210), Communications of the ACM (1197), and Proceedings of the National Academy of Sciences of the United States of America (1128). These six journals are the primary publishing outlets and the dominant citing sources for big data scholars, and contribute to the sustainable intellectual base formation of big data.

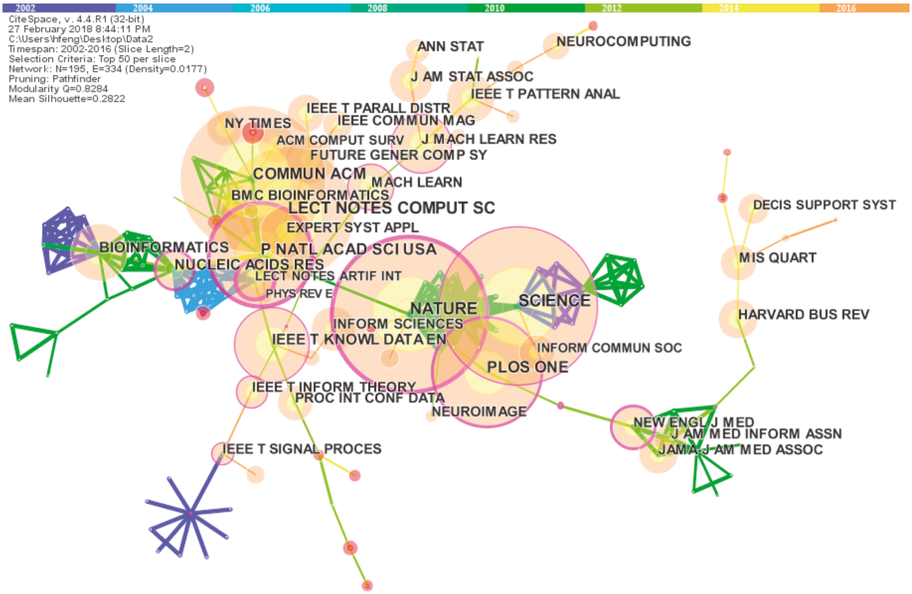


Figure 4. Journals co-citation network.

More interestingly, the nodes with purple tree rings around the outer rim indicate that some highly cited journals have high betweenness centrality (betweenness centrality ≥ 0.23), such as Nature (0.54), Proceedings of the National Academy of Sciences of the United States of America (0.56), Nucleic Acids Research (0.31), and PLoS One (0.23). These pivotal journals make connections to others in the journal co-citation network (see Figure 4). Some big nodes with thinner purple rings indicate that high co-citation scores do not necessarily have a high betweenness centrality. For example, Lecture Notes in Computer Science has a high co-citation frequency node (1436) and a lower betweenness centrality (0.04). Moreover, the journals in multidisciplinary sciences and Computer Science received more citations. It means that knowledge from multidisciplinary sciences and computer science is therefore a major intellectual resource for big data scholars. In addition, a significant co-citation burst journal is visualized by the node with red inner tree rings. The size of the red inner tree rings node represents the strength of its burst property. As shown in Figure 4, Big Data Revolution is a journal with red inner tree rings, suggesting that its citations have rapidly increased between 2014 and 2016.

Table 2. Frequency distribution and between centrality of the highest co-cited Journals.

Journal	Frequency	Centrality	IF	Categories
Nature	1899	0.54	40.137	Multidisciplinary Sciences
Science	1844	0.16	37.205	Multidisciplinary Sciences
Lecture Notes in Computer Science	1436	0.04	0.402	Computer Science (Theory and Methods)
PloS One	1210	0.23	2.806	Multidisciplinary Sciences
Communications of the ACM	1197	0.06	4.027	Computer Science (Hardware and Architecture; Software Engineering; Theory and Methods)
Proceedings of the National Academy of Sciences of the United States of America	1128	0.56	9.661	Multidisciplinary Sciences
Bioinformatics	908	0.1	7.307	Biochemical Research Methods; Biotechnology and Applied Microbiology; Mathematical and Computational Biology
Nucleic Acids Research	773	0.31	10.162	Biochemistry and Molecular Biology
IEEE Transactions on Knowledge and Data Engineering	720	0.11	3.438	Computer Science (Artificial Intelligence; Information Systems); Engineering, Electrical and Electronic
Journal of Machine Learning Research	579	0.12	5.000	Automation and Control Systems; Computer Science, Artificial Intelligence

Source: the Web of Science and Journal Citation Reports 2016; IF, impact factor in 2016.

3.3. Institutions Distribution and Collaboration

3.3.1. Core Institutions Identification

It is significant to study the institutions distribution in a research field. Commonly the number of publications is an important index to measure academic level, scientific research ability, and status of the authors and their institutions in a specific field. Core institutions are important leaders in a research field. However, the names of academic institutions might change over time. Therefore, to avoid inconsistent signatures, we firstly need to standardize the names of academic institutions. In this section, we reserved the top level names, and constructed uniform names of academic institutions. Eventually we achieved 4137 different institutions.

According to Price law, core institutions must be the institutions who published more than N (note: N = 0.749 $\times$ square (153) $\approx$ 10) articles. According to the statistical analysis, there are 265 core institutions in development history of big data research from 2002 to 2016. Table 3 lists the top 10 most prolific academic institutions in descending order of publications. The top 10 most prolific academic institutions are almost all colleges and universities from USA and China. Among them, Chinese Academy of Sciences is the most prolific institution (153), followed by Tsinghua University (126) and University of California, Los Angeles (91). Stanford University and MIT just have a narrow gap less than two articles. The top 10 academic institutions published 895 articles, which account for 18.2% of overall published articles from 2002 to 2016. Simply, it reveals that 0.2% of institutions in the big data research field published 18.2% of overall articles published from 2002 to 2016. It conforms to what is known as a "Matthew effect" in institutions distribution.

Table 3. Top 10 most prolific academic institutions.

Institution	Country	The Number of Publications	Centrality	Year
Chinese Academy of Sciences	China	153	0.15	2012
Tsinghua University	China	126	0.04	2013
University of California, Los Angeles	USA	91	0.33	2009
Stanford University	USA	84	0.29	2013
MIT	USA	82	0.06	2011
University of Washington	USA	75	0.08	2012
University of Michigan	USA	73	0.06	2013
Harvard University	USA	72	0.17	2012
University of California, San Diego	USA	70	0.18	2010
University of Minnesota	USA	67	0.16	2011

3.3.2. Institutions Collaboration Network

To enhance overall research strength in a scientific field, scientific research collaboration usually is an important means, which allows researchers to play their own academic advantages and share information [48]. Moreover, the level of scientific research collaboration is one of important indexes to evaluate the academic level, scientific research ability, and status of institutions in a specific field. To discuss the scientific research collaboration in the big data research field, we constructed a scientific research collaboration network (shown in Figure 5).

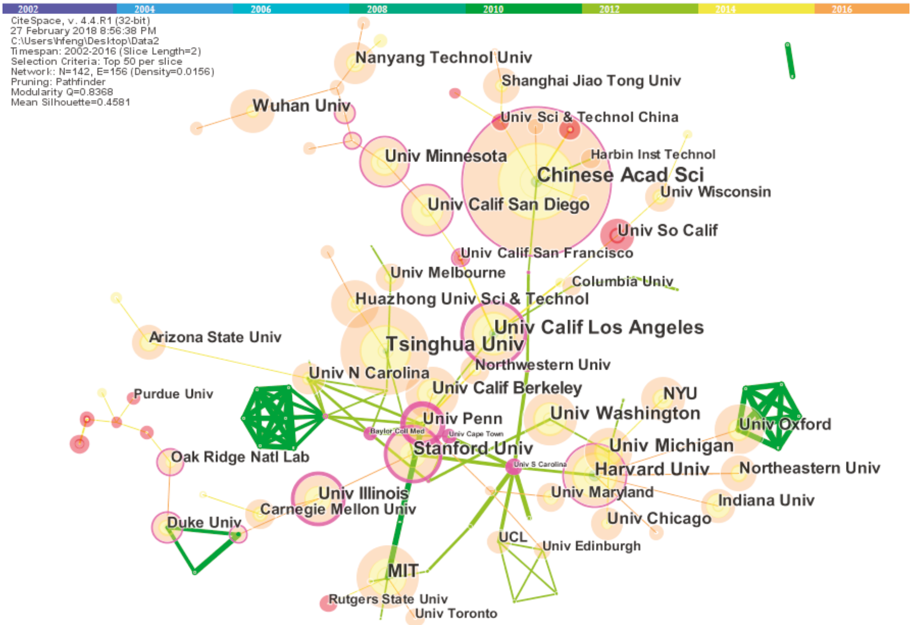


Figure 5. Institutions collaboration network.

This scientific research collaboration network consists of 142 nodes and 342 links. Each node represents an institution, and is depicted with a series of tree rings across multiple time slices. The size of each node is proportional to the total number of publications in each institution [8]. Each link between two nodes represents a scientific research collaboration relationship, and the thickness of a link represents the scientific research collaboration strength [49]. As shown in Figure 5, there are a wider scientific research collaboration among different institutions. For example, Chinese Academy of Sciences is a red tree ring node, which has the most publications 153 and cross-connects with University of Sydney, Harbin Institute of Technology, University of Science and Technology China, Peking University, Beijing Normal University, and Otto Von Guericke University. The gold-colored link between University of Sydney represents that the first scientific research collaboration year is between 2014 and 2015. However, the nodes with more publications do not certainly have stronger betweenness centrality scores. As listed in Table 3, compared with Stanford University (0.29), Chinese Academy of Sciences has a weaker betweenness centrality score (0.15). This means that Chinese Academy of Sciences plays a weaker intellectual pivotal role among the institutions collaboration network. Furthermore, University of South Carolina with the highest betweenness centrality score (0.63) has a very low co-occurrence frequency. These results reveal that the current research relationship is rather weak and diffuse. In addition, three thicker lines, which are linked with Otto Von Guericke University (link strength: 0.3), University of Sydney (link strength: 0.23), and Harbin Institute of Technology (link strength: 0.21) respectively, indicate the stronger collaboration relationships. Moreover, two green lines, which are linked with Otto Von Guericke University and Harbin Institute of Technology, indicate that the first collaboration among them is in the 2012–2013 time slice.

3.4. Authors Distribution and Co-Citation Network

3.4.1. Core Authors Identification

It is interesting to study the core authors distribution in the big data research field. Usually the amount of publications is an important index to evaluate the academic level, advancement, and position of an author in a specific research field. In addition, core authors also are particular important leaders in a research field. However, the names of authors may be full and abbreviated names downloaded from the WoS. The same abbreviated name might stand for different full names. For example, Y ZHANG represent Yin ZHANG, Yi ZHANG, or Yong ZHANG. Similarly, Y WANG may represent Yige WANG, Yi WANG, or Yuhang WANG, et al. Moreover, a same full name may be different authors. For example, Yin ZHANG can be Yin ZHANG who comes from the School of Computer Science or Information Technology at Huazhong University of Science and Technology (HUST), or even Yin ZHANG who comes from the School of Economics and Law at Zhongnan University. They are different persons. To avoid inconsistent signatures, we therefore need to examine seriously the unique full names and affiliated institutions of the authors, count the amount of the articles, and order the different authors in descending articles. Eventually, we got 16,404 different authors who published 4927 articles from 2002 to 2016. It indicates that the average number of collaborator per article is between three and four in the big data research field. This result coincides with the publications distribution (see "publications distribution" section).

According to Price law, core authors must be the authors who published more than M (note: M = 0.749 × square (18) ≈ 3) articles. According to the statistical analysis, there are 229 core authors. Table 4 lists the top 10 most prolific authors by the amount of articles from 2002 to 2016. Among them, Ranjan, Rajiv ranks first with 18 articles, and Zomaya, Albert Y ranks second with 17 articles. If we do not exclude the collaborative articles, the top 10 authors published 138 articles, which account for 2.8% of overall articles published from 2002 to 2016. This means approximately 0.6% of overall authors published 2.8% of overall articles between 2002 and 2016. It conforms to what is known as a "Matthew effect" in core authors distribution. However, the number of all core authors is only 229, which accounts for 1.4% of overall authors. This means that 98.6% authors are not core authors. This result shows that research strengths are still comparatively weak and fragmented. Moreover, from the geographical perspective, the core authors from Australia account for 50% of the top 10 core authors, which means that Australia currently has a stronger research strength in the big data field compared with other countries.

Table 4. Top 10 most prolific authors.

Author	Institution	Country	Publications
Rajiv Ranjan	Computational Informatics, CSIRO, Australian National University	Australia	18
Albert Y. Omiya	School of Information Technologies, University of Sydney	Australia	17
Lizhe Wang	Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences	China	16
Xuyun Zhang	Engineering and Information Technology, University of Technology Sydney	Australia	14
Jinjun Chen	Engineering and Information Technology, University of Technology Sydney	Australia	14
Laurence T Yang.	School of Computer Science and Technology, Huazhong University of Science and Technology	China	13
Chang Liu	Engineering and Information Technology, University of Technology Sydney	Australia	12
Keqin Li	Department of Computer Science State University of New York New Paltz	USA	12
Francisco Herrera	Dept. of Computer Science and Artificial Intelligence, CITIC-UGR (Research Center on Information and Communications Technology), University of Granada	Spain	11
Samee U. Khan	electrical and computer engineering at North Dakota State University	USA	11

3.4.2. Core Authors Collaboration Network

To deeply understand the current research collaboration of core authors, we also developed the social network analysis based on UCINET (shown in Figure 6). Because the original network has lower density (0.0240), we deleted some isolates and pendants (nodes with degree one) to increase the identifiability of the network. Eventually, the core authors collaboration network consists of 44 nodes and four small networks. The two bigger networks have 25 nodes and 12 nodes separately. This means that more core authors nodes tend to be the isolates or the pendants. In general, the overall core authors collaboration network is relatively decentralized. This result reveals that the research collaboration among core authors is not enough close in the big data field.

As shown in Figure 6, the size of each node represents the between centrality score. According to the between centrality measure, Rajiv Ranjan is the central node with highest between centrality, as it form the densest bridges with other nodes. In addition, Laurence T. Yang, Albert Y. Zomaya, and Kim-Kwang Raymond Choo also have a higher between centrality. More interesting, nine authors (Rajiv Ranjan, Albert Y. Zomaya, Lizhe Wang, Xuyun Zhang, Jinjun Chen, Laurence T. Yang, Chang Liu, Keqin Li, and Samee U. Khan) listed in Table 4 have close bonds with each other in the biggest network.

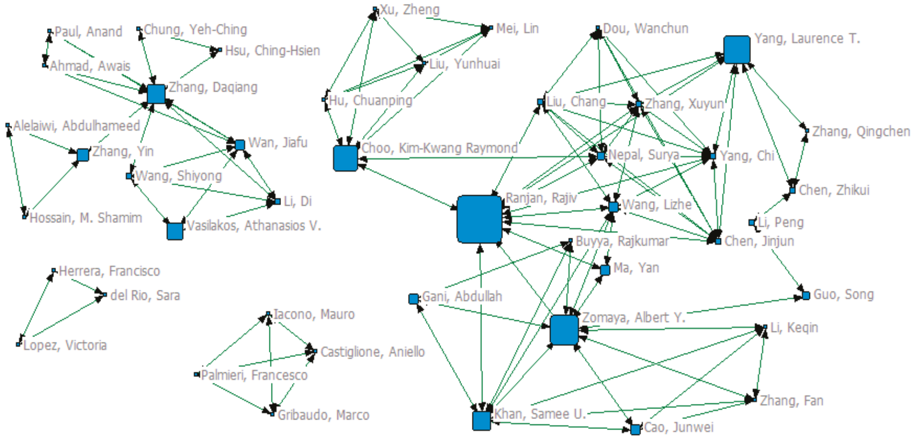


Figure 6. Core authors collaboration network.

3.4.3. Authors Co-Citation Network

Unlike the core authors collaboration analysis, authors co-citation analysis focuses on the co-cited authors who published the co-cited articles. Authors co-citation relationship is critical to understand the academic communication and knowledge base diffusion in a specific research field [11]. The more two authors are co-cited, the closer the intellectual relationship is. Figure 7 shows the overall landscape view of authors co-citation network in the big data research field. The top 50 most cited authors in each slice are used to construct the authors co-citation network based on 137,929 valid distinct references. This network consists of 262 nodes and 593 links. Moreover, this network has a very high modularity (0.9102), which can be considered that the specialties in science mapping are clearly defined in terms of co-citation clusters. The mean silhouette score (0.4179) is relatively lower mainly because of the numerous small clusters [15]. Therefore, we just need to focus on the major clusters.

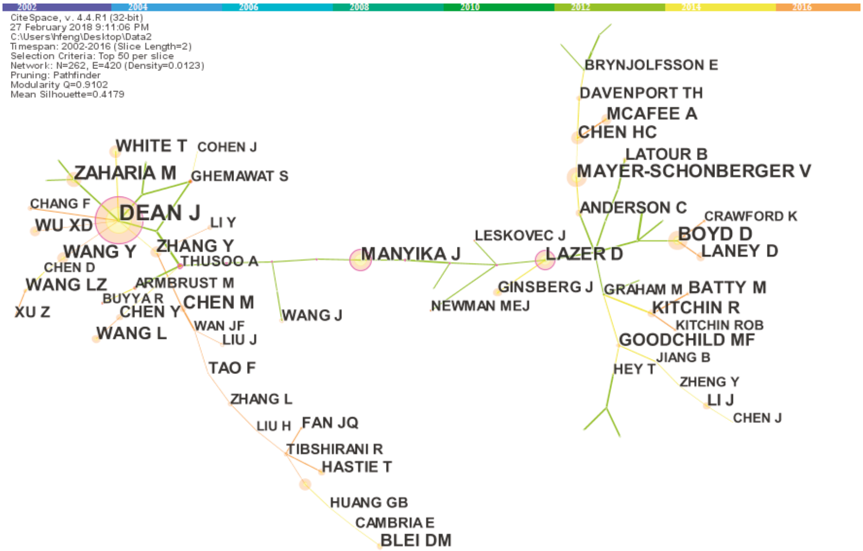


Figure 7. Authors co-citation network.

As shown in Figure 7, each node with a series of tree rings across multiple time slices represents an author. The size of each node is proportional to the total authors co-citation frequency. Each link between two nodes represents a co-citation relationship, and the thickness of a link shows the co-citation link strengths [49]. For example, Dean J is the biggest tree rings node, which has the most co-citation articles (493) and cross-connects with White T, Wu XD, Wang Y, Zaharia M, Isard M, and Condie T. The green-colored link with Zaharia M represents that the first co-citation year is 2012. In addition, three thicker lines, which are linked with Zaharia M (link strength: 0.57), Isard M (link strength: 0.53), and Condie T (link strength: 0.53) respectively, indicate some stronger co-citation relationships. Table 5 lists the most co-citation authors in the big data research field. Most of them come from USA. The highest co-citation author is Jeffrey Dean (493) at Google Inc., followed by Danah Boyd (224) at Microsoft Research, Matei Zaharia (212), James Manyika (202), Viktor Mayer-schönberger (192), Lazer David (192), Breiman Leo (170), LiZhe Wang (136), Hsinchun Chen (135), and Andrew Mcafee (128).

Table 5. Top 10 most co-citation authors.

Author	Country	First Co-Citation Year	Frequency	Centrality
Jeffrey Dean	USA	2012	493	0.1
Danah Boyd	USA	2012	224	0.01
Matei Zaharia	USA	2012	212	0.01
James Manyika	USA	2012	202	0.11
Viktor Mayer-schönberger	USA	2014	192	0.04
Lazer David	USA	2012	192	0.1
Breiman Leo	USA	2014	170	0.01
LiZhe Wang	China	2014	136	0
Hsinchun Chen	China	2014	135	0.03
Andrew Mcafee	USA	2014	128	0

The node with purple tree rings around the outer rim indicates this co-cited author has a high betweenness centrality, and this author tends to be a pivotal scholar whose work linked different disciplines, research topics, or stages in the big data field. Table 6 lists all authors with high betweenness centrality (betweenness centrality $\geq$ 0.1). For example, Savage M (0.12) proposed "The Coming Crisis of Empirical Sociology" (2007) and "Contemporary Sociology and the Challenge of Descriptive Assemblage" (2009) to argue the challenges of "social" transactional data and descriptive assemblage. Savage M is a milestone author who argues how to develop sociology within the big data environment. Other authors with a strong betweenness centrality include Manyika J (0.11), Thusoo A (0.11), Schadt EE (0.11), Barabasi AL (0.11), and Chaudhuri S (0.11). Thusoo A (2009; 2010) presented the well-known Hive—a petabyte scale data warehouse using Hadoop. Schadt EE (2010) proposed the computational solutions to large-scale data management and analysis. Barabasi AL (2010) discussed the emergence of scaling in random networks and the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems. However, it is not the case that a highly co-cited author positively has a high betweenness centrality. These authors are visualized by the small nodes with thicker purple tree rings, such as Savage M, Thusoo A, Schadt EE, Barabasi AL, and Chaudhuri S. Only a node simultaneously with a high co-citation frequency and a betweenness centrality is the milestone author. For example, as listed in Table 5, Manyika J (0.11) at McKinsey global institute (MGI, San Francisco, CA, USA) firstly released the research report "Big Data: The Next Frontier for Innovation, Competition, and Productivity" in May 2011. This report is a milestone publication, which triggered the research enthusiasm of scholars worldwide.

Table 6. High betweenness centrality authors.

Author	First Co-Citation Year	Frequency	Centrality
Savage M	2012	15	0.12
Manyika J	2012	202	0.11
Thusoo A	2012	57	0.11
Schadt EE	2012	11	0.11
Barabasi AL	2012	10	0.11
Chaudhuri S	2012	9	0.11
Dean J	2012	493	0.1
Lazer D	2012	192	0.1
Tien JM	2012	15	0.1
Stonebraker M	2012	11	0.1
Isard M	2012	9	0.1
Zhang D	2013	7	0.1

In addition, the node with red inner rings in Figure 7 means a significant co-citation burst. It reveals that the co-citation frequency of authors increased rapidly within a given time period. The size of the red inner tree rings node represents the strength of its burst property. As shown in Figure 7, there are 25 nodes with red inner tree rings. It means that there are 25 authors with co-citation bursts in big data research from 2002 to 2016. These authors may have profound impacts on the big data research, and their work should be paid more attention because they may impact the sustainable development directions of big data research. Table 7 lists the top 25 cited authors with strongest citation bursts. Among them, Ghemawat S with the strongest citation burst (11.6177) demonstrated the Google file system, a scalable distributed file system for large distributed data-intensive applications, which guided the big data storage research. Thusoo A, with the second strongest citation burst (9.8427), presented the well-known Hive based on Hadoop. In addition, Hey T, Armbrust M, Wang C, Cohen J, and Buyya R, etc. also made important contributions to the sustainable development of big data research from different perspectives.

Table 7. Top 25 Cited Authors with Strongest Citation Bursts.

Cited Authors	Year	Strength	Begin	End	2002–2016
Ghemawat S	2002	11.6177	2013	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟥
Thusoo A	2002	9.8427	2012	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟥🟥
Hey T	2002	9.437	2012	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟥🟥
Armbrust M	2002	9.0315	2012	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟥🟥
Wang C	2002	8.9019	2014	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥
Cohen J	2002	8.6382	2014	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥
Buyya R	2002	8.3746	2014	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥
Newman MEJ	2002	7.8172	2012	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟥🟥
Chen J	2002	7.5844	2014	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥
Leskovec J	2002	7.2113	2012	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟥🟥
Schadt EE	2002	6.9707	2012	2013	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟦🟦🟦
Savage M	2002	6.9707	2012	2013	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟦🟦🟦
Brynjplfsson E	2002	6.4045	2012	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟥🟥
Chen D	2002	5.2202	2014	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥
Stonebraker M	2002	5.066	2012	2013	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟦🟦🟦
Isard M	2002	5.066	2012	2013	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟦🟦🟦
Barabasi AL	2002	4.4317	2012	2013	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟦🟦🟦
Lotan G	2002	4.4317	2012	2013	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟦🟦🟦
Yang H	2002	4.4317	2012	2013	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟦🟦🟦
Christakis NA	2002	3.7977	2012	2013	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟦🟦🟦
Chaudhuri S	2002	3.7977	2012	2013	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟦🟦🟦
Callon M	2002	3.164	2012	2013	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟦🟦🟦
Von Ahn L	2002	3.164	2012	2013	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟦🟦🟦

3.5. Keywords Co-Word Network

Keywords usually provide the core content and principal research methods of each article. Keyword co-word analysis can be applied to identify research topics and monitor research frontiers of a knowledge domain [50]. To construct a reasonable keywords co-word network, SATI3.2 was used to extract the high frequency keywords and form keywords co-occurrence matrix. Moreover, commonly keywords must be integrated and unified because of synonymy and polysemy. We removed some broad words (such as algorithm, model, design, analysis, research, etc.), and eventually got the top 80 keywords. Table 8 lists the top 80 high frequency keywords.

Table 8. Top 80 high frequency keywords.

	Keywords	Frequency		Keywords	Frequency		Keywords	Frequency
1	Big Data	1834	28	bioinformatics	35	55	Reliability	19
2	cloud computing	213	29	technology	35	56	deep learning	19
3	machine learning	207	30	Database	35	57	Education	19
4	MapReduce	164	31	Security	33	58	parallel processing	19
5	data mining	144	32	data science	31	59	Sentiment analysis	19
6	big data analysis	128	33	text mining	30	60	computational social science	19
7	Hadoop	106	34	crowdsourcing	29	61	NoSQL	19
8	social media	101	35	Internet	29	62	Design	19
9	Internet of Things	80	36	ethics	29	63	innovation	18
10	privacy	77	37	scalability	29	64	knowledge	18
11	data analysis	71	38	Business Intelligence	28	65	informatics	18
12	Prediction	69	39	Data quality	26	66	software	17
13	computing	59	40	surveillance	26	67	epidemiology	17
14	Algorithm	58	41	open data	26	68	Spark	17
15	Twitter	54	42	Genomics	24	69	natural language processing	17
16	Classification	52	43	systems	24	70	precision medicine	17
17	networks	52	44	GIS	23	71	time series	17
18	optimization	51	45	Distributed computing	23	72	methodology	17
19	Cloud	49	46	Distributed	22	73	data management	16
20	model	49	47	Feature selection	21	74	Decision making	16
21	visualization	47	48	Measurement	20	75	sampling	16
22	Social network	46	49	statistics	20	76	Scheduling	16
23	performance	44	50	Healthcare	20	77	social	16

Table 8. *Cont.*

	Keywords	Frequency		Keywords	Frequency		Keywords	Frequency
24	clustering	44	51	Ontology	20	78	Parallel	16
25	smart city	39	52	Data protection	20	79	Storage	16
26	Information	37	53	Parallel computing	20	80	energy	15
27	management	35	54	energy efficiency	20			

As listed in Table 8, the highest frequency keyword absolutely is Big Data (1834), followed by cloud computing (213), machine learning (207), MapReduce (164), data mining (144), big data analysis (128), Hadoop (106), and social media (101). Other keywords with less than 100 frequency include Internet of Things (IoT, 80), privacy (77), data analysis (71), Prediction (69), Computing (59), Algorithm (58), Twitter (54), Classification (52), networks (52), and optimization (51). These keywords reflect the current research hotspots. Besides these high frequency keywords, the identification of some other relevant keywords indicate the current emerging research areas such as social network (46), smart city (39), bioinformatics (35), crowdsourcing (29), ethics (29), Genomics (24), GIS (23), Healthcare (20), Education (19), epidemiology (17), precision medicine (17), and energy (15).

To understand the relationship among these keywords, we construct the keywords co-word network (shown in Figure 8). Each node represents a keyword. The size of each node is proportional to the betweenness centrality of keywords. It is not surprising that some well-known words, such as well-known topics including data mining, cloud computing, machine learning, MapReduce, Hadoop, social media, and visualization, have higher co-occurrence frequencies and betweenness centrality scores. Besides these topics, we find that data science, including data privacy, data management, data protection, and data quality, etc., also gradually enter the researchers' considerations. In addition, deep learning, algorithm, model, performance, optimization are some interesting findings in big data research. These keywords reveal the popular research hotspots and will have profound impacts on future sustainable development research directions of big data.

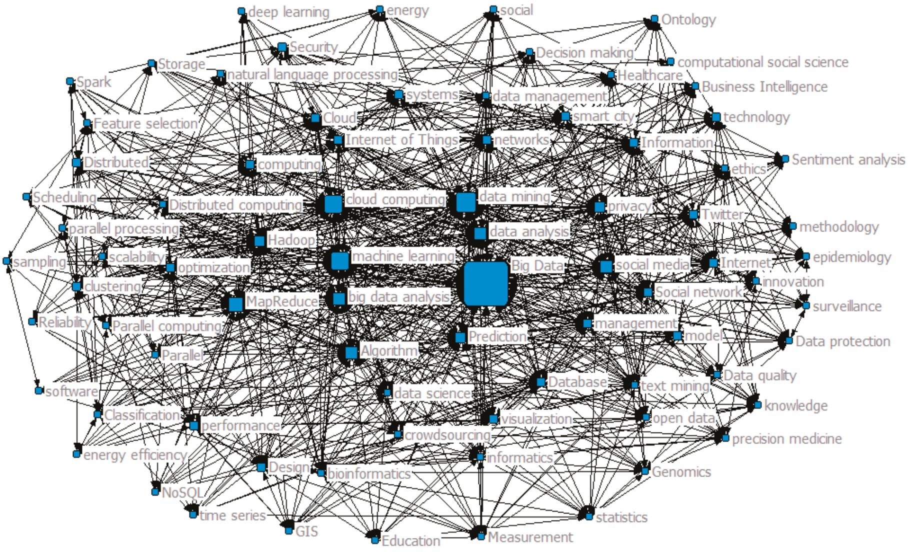


Figure 8. Keywords co-word network.

4. Discussion and Conclusions

In this study, we extracted the bibliometric data of 4927 effective journal articles listed in the WoS between 2002 and 2016, visualized the intellectual structure and hotspots of big data research from the bibliometric perspective, and presented the results in terms of publications distribution, journals distribution and co-citation network, institutions distribution and collaboration network, authors distribution, collaboration network and co-citation network, and keywords co-word network. The main findings of this study are as follows:

According to publications distribution, we found the annual growth trend of big data research outcomes and authors, as well as the changes of co-author numbers in each article. The research outcomes in the embryonic stage (2002–2009) were very few, but an exponential growth spurt was generated from 2010 to 2016. In addition, the growth trend of annual authors is similar to the annual publications distribution. Moreover, we found that the average number of participants per article in the big data field were between three and four authors.

The current core journal with the most publications was PLoS One, followed by IEEE Access and Big Data. However, the top five co-citation journals, which contributed to the sustainable intellectual base formation of big data, were Nature, Science, Lecture Notes in Computer Science, PLoS One, and Communications of the ACM. Among them, Nature had the highest betweenness centrality. Moreover, the most categories of top 10 co-citation journals were multidisciplinary sciences and computer science, which is closely related to the nature of big data science.

There was a wider scientific research collaboration among institutions in big data research. The top three core institutions in terms of publications were Chinese Academy of Science, Tsinghua University, and University of California, Los Angeles. However, the institutions with most publications had lower betweenness centrality scores, signifying that these institutions still were scattered and did not get general consent. Hence, the current research relationships among the institutions were rather weak and diffuse in the big data research field. With sustainable development and prosperity of big data research, the research collaboration relationships will be strengthened and increasingly firm.

According to the core authors identification, compared with USA and China, Australia had a current greater research strength in big data research. However, according to authors co-citation analysis, the top 10 most co-cited authors mainly came from USA and China. Moreover, some special authors with most co-citation frequency, high betweenness centrality and strong citation bursts were also identified, such as the most co-citation authors, pivotal scholars or intellectual turning pointers, and the strongest citation burst authors. These authors had contributed to the sustainable development of big data from different perspectives, and have a profound impact on the big data field. More attention should be paid to their work.

Keywords co-word analysis detected the current research hotspots and emerging topics, including not only the well-known research hotspots like data mining, cloud computing, machine learning, MapReduce, Hadoop, social media, and visualization, but also some emerging research topics, such as data science (data privacy, data management, data protection, and data quality, etc.), deep learning, and so on. Moreover, algorithm, model, performance and optimization are also gradually entering researchers' considerations. In addition, keywords co-word analysis also detected the current emerging and sustainable development applications areas of big data, such as social network, smart city, bioinformatics, crowdsourcing, ethics, Genomics, GIS, Healthcare, Education, epidemiology, precision medicine, and energy.

As an emerging hot topic, big data has changed the lives of human beings, and driven some changes in thinking, decision making, and research paradigms. Moreover, big data itself contains important strategic resources for social trends, market changes, scientific and technological development and national security. Many colleges and universities have opened big data disciplines and courses. However, as a new emerging cross-discipline, the sustainable development of big data still faces many very complicated and difficult challenges, such as the heterogeneity and incompleteness of data, the efficiency of big data processing, big data security and privacy protection, high energy consumption, and so on. On the one hand, these challenges indicate some sustainable development directions of future big data research. On the other hand, these challenges are also unprecedented opportunities of big data sustainable development. With the increasing improvement of physical infrastructure constructions and policy making at national and institutional levels, and the further breakthroughs of information technologies (computer networks, distributed systems, cloud computing, data storage, machine learning, and so on), these above issues will be gradually solved. A bright future of big data science is coming.

Acknowledgments: This work was financially supported by Project 14YJC870008 (Humanity and Social Science Youth Foundation of Ministry of Education, China), Project GD13YTS01 (Philosophy and Social Science Foundation of Guangdong Province during the "12th Five-Year Plan"), Project 2016A070705052 (Public Welfare and Ability Construction Special Foundation of Science and Technology Plan, Guangdong Province), Project 2017GZYB98 (Philosophy and Social Science Foundation of Guangzhou City during the "13th Five-Year Plan"), Provincial Nature Science Foundation of Guangdong (2015A030310271 and 2015A030313679)and Project 2017B1015 (Zhongshan City Science and Technology Bureau).

Author Contributions: Writing: Feng Hu; Providing case and idea: Feng Hu, Wei Liu, Sang-Bing Tsai; Providing revised advice: Sang-Bing Tsai, Junbin Gao, Ning Bin, Quan Chen.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McKinsey Global Institute. Big Data: The Next Frontier for Innovation, Competition, and Productivity. 2011. Available online: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (accessed on 10 February 2017).
2. Office of Science and Technology Policy Executive Office of the President. Big Data Research and Development Initiative. 2012. Available online: https://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf (accessed on 10 February 2017).
3. Feng, Z.Y.; Guo, X.H.; Zeng, D.J.; Chen, Y.B.; Chen, G.Q. On the research frontiers of business management in the context of Big Data. *J. Manag. Sci. China* **2013**, *16*, 1–9. (In Chinese)
4. Sustainable Development Goals. Available online: http://www.un.org/sustainabledevelopment/sustainable-development-goals/ (accessed on 30 January 2018).
5. McAfee, A.; Brynjolfsson, E. Big data: The management revolution. *Harv. Bus. Rev.* **2012**, *90*, 60–68. [PubMed]
6. Naimi, A.I.; Westreich, D.J. Big Data: A revolution that will transform how we live, work, and think. *Am. J. Epidemiol.* **2014**, *179*, 1143–1144. [CrossRef]
7. Jeon, S.; Hong, B. Monte Carlo simulation-based traffic speed forecasting using historical big data. *Future Gener. Comput. Syst.* **2016**, *65*, 182–195. [CrossRef]
8. Chen, C. CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. *J. Am. Soc. Inf. Sci. Technol.* **2006**, *57*, 359–377. [CrossRef]
9. Zhao, R.; Wang, Q. The Mining and Analysis of Big Data Research Hotspots in the Field of Humanities and Social Science from Perspective of Information Measurement in China. *J. Intell.* **2016**, *35*, 93–98. (In Chinese)
10. Nobre, G.C.; Tavares, E. Scientific literature analysis on big data and internet of things applications on circular economy: A bibliometric study. *Scientometrics* **2017**, *111*, 463–492. [CrossRef]
11. Gu, D.; Li, J.; Li, X.; Liang, C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. *Int. J. Med. Inform.* **2017**, *98*, 22–32. [CrossRef] [PubMed]
12. Isasi, N.K.G.; Frazzon, E.M.; Uriona, M. Big Data and Business Analytics in the Supply China: A Review of the Literature. *IEEE Lat. Am. Trans.* **2015**, *13*, 3382–3391. [CrossRef]
13. Tsay, M.Y. Journal Bibliometric Analysis: A Case Study on the JASIST. *Malays. J. Libr. Inf. Sci.* **2008**, *13*, 121–139.
14. Chen, C.; Ibekwe-SanJuan, F.; Hou, J. The Structure and Dynamics of Co-Citation Clusters: A Multiple-Perspective Co-Citation Analysis. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 1386–1409. [CrossRef]
15. Chen, C. Science Mapping: A Systematic Review of the Literature. *J. Data Inf. Sci.* **2017**, *2*, 1–40. [CrossRef]
16. Meho, L.I.; Yang, K. Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science versus Scopus and Google Scholar. *J. Am. Soc. Inf. Sci. Technol.* **2007**, *58*, 2105–2125. [CrossRef]

17. Pan, L.; Wang, S. A bibliometrics analysis on Chinese education research hotspots based on literature keywords co-occurrence knowledge map. *Educ. Res. Exp.* **2011**, *6*, 20–24.
18. No, H.J.; An, Y.; Park, Y. A structured approach to explore knowledge flows through technology-based business methods by integrating patent citation analysis and text mining. *Technol. Forecast. Soc. Chang.* **2015**, *97*, 181–192. [CrossRef]
19. Gautam, P. An overview of the Web of Science record of scientific publications (2004–2013) from Nepal: Focus on disciplinary diversity and international collaboration. *Scientometrics* **2017**, *113*, 1245–1267. [CrossRef]
20. Callon, M.; Courtial, J.P.; Turner, W.A.; Bauin, S. From translations to problematic networks—An introduction to co-word analysis. *Soc. Sci. Inf. Sci. Soc.* **1983**, *22*, 191–235. [CrossRef]
21. Small, H. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Am. Soc. Inf. Sci.* **1973**, *24*, 265–269. [CrossRef]
22. Chen, C. Visualising semantic spaces and author co-citation networks in digital libraries. *Inf. Process. Manag.* **1999**, *35*, 401–420. [CrossRef]
23. White, H.D.; McCain, K.W. Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *J. Am. Soc. Inf. Sci. Technol.* **1998**, *49*, 327–355.
24. Johnson, B.; Shneiderman, B. Tree-maps: A space filling approach to the visualization of hierarchical information structures. In Proceedings of the 2nd Conference on Visualization '91, San Diego, CA, USA, 22–25 October 1991; pp. 284–291.
25. Herman, I.; Melançon, G.; Marshall, M.S. Graph visualization and navigation in information visualization: A survey. *IEEE Trans. Vis. Comput. Graph.* **2000**, *6*, 24–44. [CrossRef]
26. Morris, S.A.; Yen, G.; Wu, Z.; Asnake, B. Timeline visualization of research fronts. *J. Am. Soc. Inf. Sci. Technol.* **2003**, *55*, 413–422. [CrossRef]
27. Wise, J.A. The ecological approach to text visualization. *J. Am. Soc. Inf. Sci.* **1999**, *50*, 1224–1233. [CrossRef]
28. Porter, A.L.; Cunningham, S.W. *Tech Mining: Exploiting New Technologies for Competitive Advantage*; John Wiley and Sons, Inc.: Hoboken, NJ, USA, 2004; ISBN 9780471475675.
29. Chen, C. *Information Visualization: Beyond the Horizon*, 2nd ed.; Springer: New York, NY, USA, 2004.
30. Chen, C. Searching for intellectual turning points: Progressive knowledge domain visualization. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5303–5310. [CrossRef] [PubMed]
31. Bailón-Moreno, R.; Jurado-Alameda, E.; Ruíz-Baños, R. The scientific network of surfactants: Structural analysis. *J. Am. Soc. Inf. Sci. Technol.* **2006**, *57*, 949–960. [CrossRef]
32. Bailón-Moreno, R.; Jurado-Alameda, E.; Ruíz-Baños, R.; Courtial, J.P. Analysis of the scientific field of physical chemistry of surfactants with the unified scienctometric model. Fit of relational and activity indicators. *Scientometrics* **2005**, *63*, 259–276. [CrossRef]
33. Leydesdorff, L.; Schank, T. Dynamic animations of journal maps: Indicators of structural changes and interdisciplinary developments. *J. Am. Soc. Inf. Sci. Technol.* **2008**, *59*, 1810–1818. [CrossRef]
34. Persson, O.; Danell, R.; Wiborg Schneider, J. How to use Bibexcel for various types of bibliometric analysis. In *Celebrating Scholarly Communication Studies: A Festschrift for Olle Persson at His 60th Birthday*; Åström, F., Danell, R., Larsen, B., Schneider, J.W., Eds.; International Society for Scientometrics and Informetrics: Leuven, Belgium, 2009; Volume 5, pp. 9–24.
35. Sci2 Team. Science of Science (Sci2) Tool. Indiana University and SciTech Strategies. 2009. Available online: https://sci2.cns.iu.edu (accessed on 10 August 2017).
36. Van Eck, N.J.; Waltman, L. Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics* **2010**, *84*, 523–538. [CrossRef] [PubMed]
37. Börner, K.; Huang, W.; Linnemeier, M.; Duhon, R.; Phillips, P.; Ma, N.; Price, M. Rete-netzwerk-red: Analyzing and visualizing scholarly networks using the network workbench tool. *Scientometrics* **2010**, *83*, 863–876. [CrossRef]
38. Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. SciMAT: A new Science Mapping Analysis Software Tool. *J. Am. Soc. Inf. Sci. Technol.* **2012**, *63*, 1609–1630. [CrossRef]
39. Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. Science Mapping Software Tools: Review, Analysis, and Cooperative Study among Tools. *J. Assoc. Inf. Sci. Technol.* **2011**, *62*, 1382–1402. [CrossRef]
40. Liu, Q.; Ye, Y. A Study on Mining Bibliographic Records by Designed Software SATI: Case Study on Library and Information Science. *J. Inf. Sources Manag.* **2012**, *1*, 50–58. (In Chinese)
41. Petersohn, H. Data-mining application architecture. *Wirtschaftsinformatik* **2004**, *46*, 15–21. [CrossRef]

42. Abuter, R.; Schreiber, J.; Eisenhauer, F.; Ott, T.; Horrobin, M.; Gillesen, S. SINFONI data reduction software. *New Astron. Rev.* **2006**, *50*, 398–400. [CrossRef]
43. Dean, J.; Ghemawat, S. MapReduce: Simplified data processing on large clusters. *Commun. ACM* **2008**, *51*, 107–113. [CrossRef]
44. Thusoo, A.; Sarma, J.S.; Jain, N.; Shao, Z.; Chakka, P.; Zhang, N.; Antony, S.; Liu, H.; Murthy, R. Hive—A Petabyte Scale Data Warehouse Using Hadoop. In Proceedings of the 26th International Conference on Data Engineering (ICDE 2010), Long Beach, CA, USA, 1–6 March 2010. [CrossRef]
45. Adam, J. The Pathologies of Big Data. *Commun. ACM* **2009**, *52*, 36–44.
46. Brinkmann, B.H.; Bower, M.R.; Stengel, K.A.; Worrell, G.A.; Stead, M. Large-scale Electrophysiology: Acquisition, Compression, Encryption, and Storage of Big Data. *J. Neurosci. Methods* **2009**, *180*, 185–192. [CrossRef] [PubMed]
47. Mayer-Schönberger, V.; Cukier, K. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*; Houghton Mifflin Harcourt: Boston, MA, USA, 2013; ISBN-10: 0544227751; ISBN-13/EAN: 9780544227750.
48. Ebadi, A.; Schiffauerova, A. How to become an important player in scientific collaboration networks? *J. Informetr.* **2015**, *9*, 809–825. [CrossRef]
49. Navonil, M.; Nik, B.; Simon, J.E.T.; Stelios, S. Exploring the e-science knowledgebase through co-citation analysis. *Procedia Comput. Sci.* **2013**, *19*, 586–593.
50. Callon, M.; Courtial, J.; Laville, F. Co-word analysis as a tool for describing the network of interactions between basic and technological research—The case of polymer chemistry. *Scientometrics* **1991**, *22*, 155–205. [CrossRef]

sustainability

Article

Big Social Network Data and Sustainable Economic Development

Umit Can [1],* and Bilal Alatas [2]

[1] Computer Engineering Department, Munzur University, 62000 Tunceli, Turkey
[2] Software Engineering Department, Firat University, 23119 Elazig, Turkey; balatas@firat.edu.tr
* Correspondence: ucan@munzur.edu.tr; Tel.: +90-555-844-5537

Received: 9 October 2017; Accepted: 4 November 2017; Published: 7 November 2017

Abstract: New information technologies have led to the rapid and effective growth of social networks. The amount of data produced by social networks has increased the value of the big data concept, which is one of the popular current phenomena. The immediate or unpredictable effects of a wide array of economic activities on large masses and the reactions to them can be measured by using social media platforms and big data methods. Thus, it would be extremely beneficial to analyze the harmful environmental and social impacts that are caused by unsustainable business applications. As social networks and big data are popular realms currently, their efficient use would be an important factor in sustainable economic development. Accurate analysis of people's consumption habits and economic tendencies would provide significant advantages to companies. Moreover, unknown consumption factors that affect the economic preferences of individuals can be discovered and economic efficiency can be increased. This study shows that the numerous solution opportunities that are provided by social networks and big data have become significant tools in dynamic policy creation by companies and states, in solving problems related to women's rights, the environment, and health.

Keywords: social networks; big data; big data analysis; sustainable development

1. Introduction

New information technologies have led to the rapid and effective growth of social networks. Initially aimed at establishing communication between individuals or groups, these platforms are currently used by millions of people to express their political views, personal emotional states, or economic preferences. The data produced by these platforms have reached to an unbelievable magnitude. This large amount of data has increased the value of the big data concept, which is one of the popular and rapidly developing phenomena currently. Big data has started presenting new solutions to the challenges in many areas by helping in the development of new methods. One of these areas is sustainable economic development, which contains solutions to various issues in the world. Even though there are multiple definitions of sustainable economic development, the most frequently used definition is as follows: "a development model that meets the requirements of the present without compromising the ability of future generations to meet their requirements" [1]. In other words, it signifies the planning of present and future life and development, such that it will allow for the requirements of future generations to be met by maintaining a balance between human beings and nature without exploiting natural resources until they are exhausted. Hence, sustainable development constitutes an extremely critical issue for a sustainable economy, for the environment and for society, and to provide a livable world that can meet the requirements of future generations. Several policies have been developed to achieve sustainable economic development. Big data, big data analysis tools, and social network analysis, which are products of technology in the 21st century, have acquired new and significant roles as tools for achieving sustainable economic development.

The emergence of big data analysis capabilities has further increased the value of big data that is generated by social networks and has led to the concept of big social networking. The fact that billions of people can instantly share content including sports, economy, politics, and environment is extremely crucial in terms of the comprehensiveness of information. These data groups, which indicate the trends of large groups of people, can be analyzed through big data facilities. The results of this analysis can be used as a new tool to develop effective sustainable investment policies and policies and strategies in areas such as environment, health, and women's rights. The world economy has been growing since the Industrial Revolution. Its influence on environment and nature became more visible by the end of the 20th century. The inability of the world's resources to respond to this growth is the reason for the increase in poverty, inequalities, and conflicts. Hence, sustainable development has become critical in this context. It is significant in the economic sense and in terms of social norms to provide a fair and controllable approach, and model to meet the requirements of future generations. Widespread communication in social networks and all of these dimensions with big data analysis tools makes understanding billions of people's ideas and reactions easier. Thus, it is possible to create policies to meet these requirements, and this will help consumption to be more sustainable. In particular, this study can significantly contribute to sustainable development being discussed in the context of the 2030 sustainable development objectives [2], which the United Nations (UN) has set as global objectives.

Big data is a popular phenomenon aimed at providing alternatives to conventional solutions based on databases and data analysis. In addition to providing data storage and data access, big data aims to perform analysis to produce required solutions, comprehend data, and realize its value. Big data refers to data sets with sizes of the order of terabytes, petabytes, and even exabytes. As these data sets are so large, efficiently storing, managing, and analyzing them is beyond the capabilities of average database software tools [3]. New technologies have resulted in a high-level increase in current data volume and types, and have created unprecedented possibilities for informing and transforming society, and for protecting the environment. States, corporations, researchers, and citizens are attempting to become familiar with the new data world via experiences and innovations. This change is referred to as the data revolution [4]. Prior to the great data revolution, companies could not store all of their archives for long periods of time and effectively manage large data sets. Conventional technologies have limited storage capacities and inflexible management tools, which are considerably expensive. Their lack of scalability, flexibility, and performance has been resolved through new technologies in the context of big data [5].

The effect of social networks on sustainable development has been demonstrated by a few studies. However, the concepts of big data and facilities have not been discussed in these studies.

Willard Terri presented the impact of social networks on managing sustainable development. In general, the discussion included how individuals interested in sustainable development would use social networks, how communities and groups would benefit from social networks during their decision-making processes, and how the capabilities of social networks would be developed in sustainable development processes [6].

Herry and Vollan [7] conducted a study regarding sustainability science literature and discussed social networks. They proved that shortcomings still exist with regard to social networks and sustainability. The conclusions of their study are as follows:

1. Social relationships (and therefore social networks) are important in numerous sustainability issues, such as transfer of knowledge, cooperation and management of shared resources, and the creation of policies aimed at influencing behavior.
2. Concepts of networks emerge in relation to the following three difficulties in sustainability: associating knowledge with action, developing collective action, and encouraging social learning. Practical strategies can be developed to better understand the structure and dynamics of networks and to solve or manage these problems.

3. Networks can be conceptualized as structural limitations that make interactions (such as participating in shared spaces) or social processes (such as cooperating in or coordinating shared activities, sharing resources, or influencing the comprehension or behaviors of network members) possible.

The methods employed to solve various social network analysis issues can be used to handle extremely big social network data to realize the objective of sustainable development. Thus, the purpose of this study is to analyze big social network data using popular big data analysis tools and to demonstrate the effects of data on sustainable development.

The rest of the article is organized as follows: In Section 2, sustainable economic development is described with general dimensions. In the Section 3, general information about social networks is provided. In addition, online social networks, which are the focus of this study, are discussed in detail. The Section 4 discusses big data. Section 5 discusses the potential effects of big social network data on sustainable development. Section 6 presents the conclusions obtained from the study and the direction of future work.

2. Sustainable Economic Development

The concept of sustainable development has been proposed as an alternative to economic process, which began with the Industrial Revolution and continues today, that prevents long-term environmental and social development, and only based on the economic benefit obtained while processing raw materials into finished products. The concept began to be accepted when it was discussed in a number of international meetings, since the 1970s and 1980s. The concept was first introduced by the Limits of Growth report. This report was the first to demonstrate the contradiction between unlimited and uncontrolled growth and the world's limited resources, and it highlighted the options for society to create a sustainable development process that would be consistent with environmental constraints [8]. The second step that led to the development of the concepts of sustainability and sustainable development was the United Nations Conference on the Human Environment held in Stockholm, United Nations, in June 1972. The concerns about the protection and improvement of the ecological environment inhabited by humans and the transferability of the environment to the next generations were on the agenda. Thus, they constituted a basis for the development of the concepts of sustainability and sustainable development [9]. The concepts of "sustainability of the world system in a balanced manner" and "eco-development", which were introduced to the agenda by Dennis and Donella Meadows, were included for the first time in an official conceptual framework in the "Our Common Future" report, which was published in 1987 by the World Commission on Environment and Development (also known as the Brundtland Commission) [1]. Thus, all of the international organizations that manage the political and economic processes of the world have begun using the concept of sustainable development. The United Nations Conference on Environment and Development (also known as the Earth Summit), which was held in Rio de Janeiro in June 1992, used the conceptual framework of Brundtland Commission's report for sustainability and sustainable development. The Earth Summit is significant because an agreement consisting of five points with international consensus on the concepts of sustainability and sustainable development was declared [10]. Since the Rio summit [11], where sustainable development gained international importance, sustainable development has become a significant global phenomenon. Today, the 2030 sustainable development goals [2] constitute the most updated targets of sustainable development. Even though these goals include various topics, they include the following three dimensions that are generally accepted since the world started discussing and using sustainable development [12]:

- Economical: An economically sustainable system should be able to produce goods and services based on ongoing principles; the economy should manage the government and foreign debts and prevent imbalances between the sectors that would damage agricultural and industrial production.

- Environmental: An environmentally sustainable system should consume resources that are sufficiently stable so that the base of the resources remains, prevent exploitation of renewable resource systems or environmental investment functions, and use only nonrenewable resources if the resources are renewed through investments. In addition, this process should include the conservation of biodiversity, atmospheric balance, and other ecosystem functions that are not classified as economic resources.
- Social: A socially sustainable system should ensure an equal distribution of health, education, gender equality, political responsibility, participation, and the sufficient delivery of social services.

These three dimensions are related to each other and constantly affect each other. Hence, the concept of sustainable economic development does not require only economic success. It is necessary to simultaneously consider the "economic", "social", and "environmental" components for sustainable development. It was at the United Nations meeting in 2015 that the goals outlined in these dimensions of sustainable development were set as new and updated targets for the world. At the most recent UN Sustainable Development Summit held on 25–27 September 2015, the goals of sustainable development for 2030 were accepted with the signatures of 193 countries [2]. Sustainable development should be evaluated in the context of the following goals, and policies should be produced in the direction of these goals:

2030 Sustainable Development Goals

- Goal 1. Ending all kinds of poverty, no matter where they are.
- Goal 2. Ending hunger, providing food safety, developing nutritional resources, and supporting sustainable agriculture.
- Goal 3. Ensuring a healthy life for all people and prosperity for everybody, at every age.
- Goal 4. Providing everybody equally quality education and the possibility of life-long education.
- Goal 5. Providing gender equality in society and strengthening the social status of women and girls.
- Goal 6. Ensuring the availability and sustainable management of water and sanitation for all.
- Goal 7. Providing accessible, reliable, sustainable, and modern energy to everyone.
- Goal 8. Ensuring sustainable and inclusive economic development and ensuring full and productive employment and decent jobs for human dignity.
- Goal 9. Constructing durable infrastructure and encouraging sustainable and inclusive industrialization and new inventions.
- Goal 10. Reducing inequalities in and between countries.
- Goal 11. Making cities and human settlements strong, secure, and sustainable.
- Goal 12. Providing sustainable consumption and production.
- Goal 13. Taking emergent steps to address climate change and its impacts.
- Goal 14. Protecting oceans, seas, and marine resources for sustainable development and using them in a sustainable manner.
- Goal 15. Preserving and restoring terrestrial ecosystems, ensuring their sustainable use, and addressing desertification.
- Goal 16. Encouraging peaceful and embracing communities for sustainable development, ensuring access to justice for everyone, and establishing effective, accountable, and embracing institutions at every level.
- Goal 17. Strengthening the application tools of global partnership for sustainable development and reviving global partnership.

3. Social Networks

Towards the end of the 20th century, approximately between 1997 and 2000, the internet became highly widespread and revolutionized a considerable part of our economic and social life [13]. This change had a significant impact on online social networks. This was because the internet provides important communication means. This communication opportunity created social networks, which are a part of the internet revolution, and made them extremely influential. In a general sense, a social network is a social structure composed of individuals and organizations [14]. Any community or social interaction unit can be defined as a social network in a particular manner. Theoretically, a social network is defined as an (*N*, *R*) graph, where *N* is the node cluster that represents individuals (persons, organizations, countries, etc.) and *R* is the cluster of edges (or connections) that constitutes relations [15]. In other words, a social network is a social structure that is composed of nodes (generally individuals or institutions) that are connected to each other with one or more than one type of dependency, such as values, visions, ideas, financial change, friendship, kinship, being unloved, conflict, or trade [6]. Figure 1 shows the diagram of a simple social network.

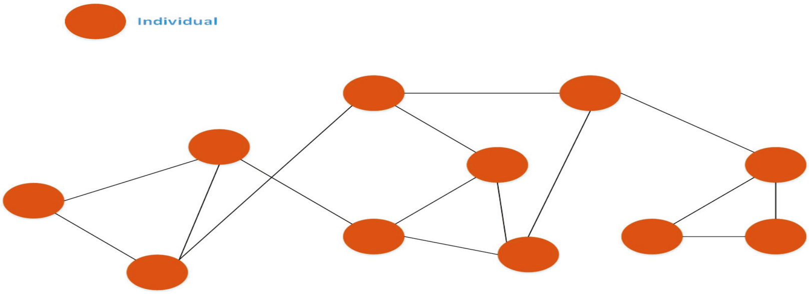


Figure 1. Social network diagram [16].

In this study, we use online social sites and a new concept of social networks, which have emerged through technology, instead of a definition. These social networks are a type of social media and involve social networking sites. They enable internet users to connect with people and create knowledge [17]. The collective adoption of online social network sites has changed the manner in which people convey and share knowledge, how establishments activate and compete, and how politicians make policy [18]. By virtue of these social networks, it is currently possible to analyze revealing confidential information, such as who would be the new crime focus, observing the spread of any disease, and what are the important political views [15]. Various applications have emerged along with the widespread utilization of social network data. Moreover, online social networks produce a vast amount of text and multimedia content that express a large number of views [18].

Social media analysis refers to the analysis of structural and nonstructural data obtained via various social media channels, and obtaining various results. Social media can be classified as follows: social networks (Facebook, LinkedIn), blogs (BlogSpot, WordPress), microblogs (Twitter, Tumblr), social news (Digg, Reddit), social bookmarks (Delicious, StumbleUpon), media sharing (Instagram, YouTube), Wiki (Wikipedia, Wikihow), question & answer sites (Yahoo! Answers, Ask.com), and comment sites (Yelp, TripAdvisor) [19,20]. Several of these sites either directly or indirectly refer to a social network.

Social networking sites have reached incredible amounts of users. For example, according to Twitter statistics, Twitter has 310 million users and 40% of these users are technical people [21]. In addition, according to 2017 data, YouTube has more than one billion users globally, and it has more viewers than any TV network in the United States of America in the age group of 18–34 years. YouTube is localized in 88 countries, and it is accessible in 76 different languages. Videos are being watched for 1 billion hours per day [22].

Figure 2 shows the number of active users of popular social networks throughout the world, according to a statistics report produced in August 2017. Market leader Facebook became the first social network to exceed 1 billion registered accounts, and it currently has 2.047 billion active users per month. The seventh-ranked photo sharing application, Instagram, has 700 million active accounts per month, and blogging service Tumblr has more than 357 million active blog users.

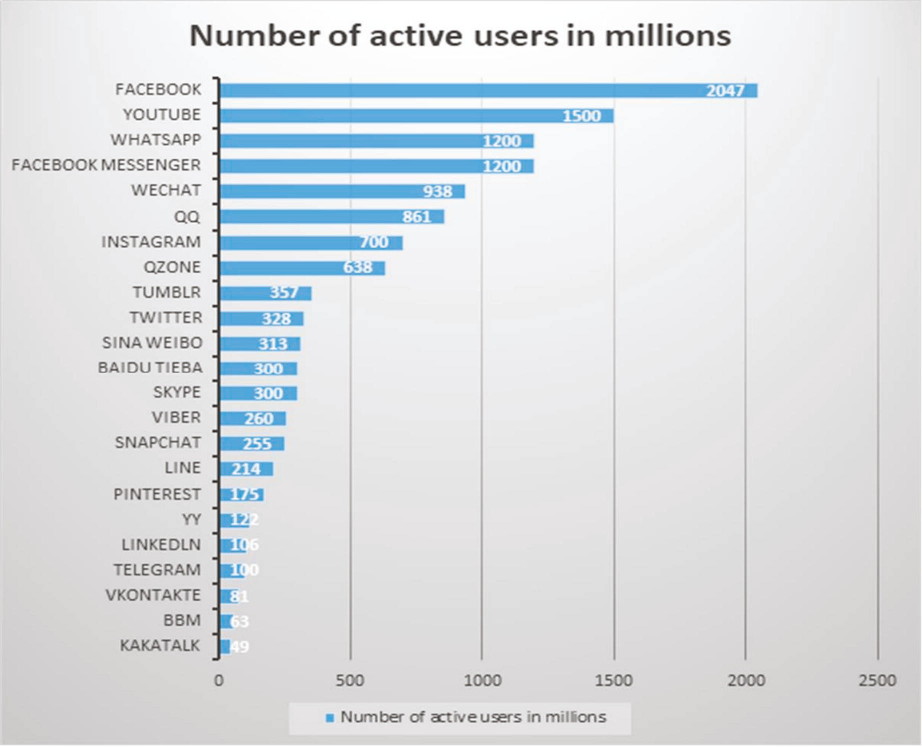


Figure 2. Number of active users of social network sites (August 2017) [23].

The figures stated in above graph are important as they show the dimensions of social networks established by social network platforms worldwide. The number of individuals interacting via these platforms is extremely high, and this number is increasing daily.

4. What Is Big Data?

Data collection and benefiting from this data has been one of the important fields throughout the history of humanity. In the beginning, the scarcity of data relatively facilitated obtaining knowledge from data. The slow data collection process underwent a significant change in the 21st century, owing to the digital revolution, and almost had a revolution within itself. Computers have allowed for data production in extremely short periods and for the storage of large amounts of data. Such data would normally be produced in tens of years. The size of collected data considerably increased again in the 21st century as a result of technological development. This initiated the big data era referred to as the petabyte era.

The concept of big data was first introduced by Francis X. Diabold in 2000 [24]. Subsequently, the developments in the internet made big data extremely important. The rapid development of sensors, wireless communication and network communication technology, cloud data processing, and smart mobile devices and the affordability of these devices enabled them to become more widespread. Thus, a large amount of data is accumulated almost in all fields of our lives. Moreover, data volume rapidly continues to increase via more complex structures and forms [3,25]. Large amounts of data are being produced by all devices that can connect to a network, such as social networks, financial markets, and connections with web servers, smart phones, tablets, credit cards, trains, and planes. While 100 GB of data were generated per day in 1992, this increased to 100 GB per hour in 1997 and 28.875 GB per second in 2013. It is presumed that this amount will increase to 50.000 GB per second in 2018 [26].

In parallel with the increasing amount of data, establishments all around the world began recognizing that the skills required to analyze and use big and complex datasets would be the most important source of competitive advantage in the 21st century. Big data has the potential of offering a better customer experience, increasing internal productivity, and consequently, improving the profitability and competitive advantage of establishments in all industries. Organizations can use big data to become smarter and more innovative through methods that were not possible prior to the "zettabyte era" [27].

According to data obtained from IBM, more than 50% of the population of the US owns smartphones, and the total number is expected to be 10 billion in 2020. Two hundred and ninety four billion emails are sent per day, and more than 1 billion searches are performed on Google every day. The size of the data produced by users on Facebook is more than 30 petabytes per day. There are more than 230 million tweets per day on Twitter. However, only 23% organizations use big data strategies [28].

According to another study, the amount of data produced in 2020 will be exactly 44 times of the amount that was produced in 2009 [29]. Every hour, the operations of Wal-Mart customers provide approximately 2.5 petabytes of data to the company, and Wal-Mart is mentioned in social networks approximately 300,000 times per week [30].

Owing to the interesting information obtained from big data, numerous actors in different countries have initiated important projects. The USA was one of the first leading countries to realize the benefits of big data. In March 2012, the Obama administration commenced big data research and development, with a budget of 200 million dollars [31]. In Japan, big data development became an important element of the technological strategy in July 2012 [32]. Furthermore, the United Nations published a paper titled "Big Data for Development: Challenges and Opportunities". This paper aims at outlining the main concerns regarding big data and improving the discussion on how big data would serve international development [33].

4.1. Big Data and 4V's

Big data was initially explained by 3V's, i.e., volume, velocity, and variety. Subsequently, the "value" dimension was added to the concept of big data. Thus, big data is currently explained by 4V's. Figure 3 shows the volume, velocity, variety, and value dimensions. These dimensions are discussed below.

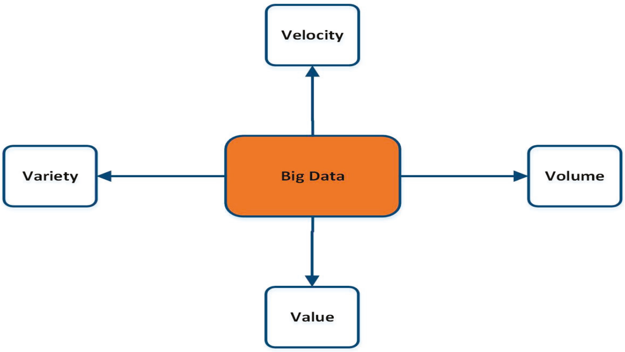


Figure 3. Big data and 4V's.

4.1.1. Volume

Volume refers to the data collected and/or produced by an organization or individual. Big data requires the processing of social network data, clicks on a web page, web traffic, and the high-volume and low-intensity data obtained by sensors that capture the data at the speed of light, the value of which is unknown [34]. While the current threshold for big data is 1 terabyte, the minimum size that would be qualified as big data can be defined as a function of technological development. One terabyte of data refers to an area where approximately 16 million Facebook pictures can be stored, and this refers to the data that would fit into 1500 CDs or 220 DVDs [35].

4.1.2. Velocity

Data are obtained at a certain velocity. High-velocity data typically flow in memory instead of being saved to a disk. The Internet of Things applications consist of health and security sub-applications that require real-time evaluation and processing. In addition, other smart products that can access the internet are managed in real time or close to real time. For example, e-commerce consumer applications merge the locations of mobile devices with consumer preferences and aim at providing offers when required. Operationally, mobile application experiences have broad user masses, increasing network traffic, and instant response expectation [34]. For example, Wal-Mart (an international discount retail chain) produces more than 2.5 petabytes of data per hour via the transactions of its customers. YouTube is a good example that shows the velocity of big data [5].

4.1.3. Variety

Variety refers to new unstructured data types. Unstructured or semi-structured data types, such as text, voice, and video necessitate additional processing to obtain meaning and supportive metadata. If unstructured data are comprehended once, they satisfy most of the requirements of structural data, such as summarization, ancestry, controllability, and confidentiality. The data obtained from a known source become more complex when altered without prior notification. Frequent or real-time schema changes create a huge burden for processing and analytical environments [34]. Big data is composed of more than one form of various sources (e.g., videos, documents, comments, and journals). Big data sets are composed of public or private, local or distant, shared or confidential, completed or missing, and structured or unstructured data [5].

4.1.4. Value

Data have value; however, this value must be discovered. Several research techniques are used to obtain value from data. Technological developments have led to an incremental decrease in data storage and computing expenses, and thus provided abundant data related to statistical sampling and other techniques to derive meaning. However, finding value requires new discovery processes that involve smart analysts, users from the business world, and administrators. The real big data challenge implies asking the right questions, recognizing patterns, making conscious assumptions, and predicting behaviors [34].

4.2. Big Data Analysis Tools

Big data comprises large amounts of multidimensional unstructured, semistructured, and structured data. Big data has become a production factor, and this amount of data cannot be processed using classic methods anymore. The most popular technology used to analyze and process big data today is Hadoop [36], an open source software. The biggest advantages of Hadoop are its ability to process big data rapidly and the fact that it is free. Many distributors, such as Cloudera [37] and Hortonworks [38] provide big data platform services by using the Hadoop framework. These distributors provide a simplified Hadoop experience to users by combining various Hadoop components together.

The Hadoop project contains many separate subprojects. Big data can be processed easily under these different projects, and various conclusions can be drawn.

The Hadoop project involves the following main modules:

- Hadoop Common: Common tools that support other Hadoop modules [36].
- Hadoop Distributed File System (HDFS): A distributed file system providing high-efficiency access to application data [36,39].
- Hadoop YARN: A framework for business planning and cluster source management [36].
- Hadoop MapReduce: A YARN-based parallel processing and programming model used for processing big data sets and producing information from them [36,40].

Other Hadoop projects:

- Ambari: A web-based tool containing support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig, and Sqoop projects and for enabling one to monitor and manage Apache Hadoop clusters. Ambari also provides a dashboard to monitor heat maps and applications like MapReduce, Pig, and Hive and to display the cluster efficiency [36].
- Avro: A data serialization system [36].
- Cassandra: A distributed, open source, unrelated, column-oriented database that was developed by Facebook to store a vast amount of structured data [41].
- Chukwa: A data collection system used for managing big distributed systems [36].
- HBase: A scalable and distributed database supporting structural data storage for big tables. Google's BigTable is an important Apache-Hadoop-based project that was developed recently and modelled using HBase. Hbase adds a fault-tolerant scalable database that is installed and distributed on an HDFS file system and that has random real-time read/write access [36,42].
- Hive: A data warehouse built on Hadoop that enables summarizing, querying, and analyzing big data clusters stored in Hadoop files. It is not designed to offer real-time queries; however, it may support text files and queue files [43].
- Mahout: A scalable machine learning and data mining library. Mahout currently focuses on algorithms for clustering, classification, data mining (frequent item set), and evolutionary programming [36,44].
- Pig: Provides a high-level parallel mechanism to execute MapReduce works on Hadoop clusters. It uses a command file language called Pig Latin; the data streaming language is collaterally directed to data processing [43].
- Spark: A fast and general calculation engine for Hadoop data. It provides a simple and effective programming model that supports a wide range of applications such as Spark, ETL, machine learning, stream processing, and graph calculation [36].
- TEZ: A generalized data stream programming framework installed on Hadoop YARN that provides a powerful and flexible engine to execute the directed acyclic gap (DAG) of tasks to process data for collective and interactive usage [36,45].
- ZooKeeper: Provides a high-performance coordination service for distributed applications [36].

Figure 4 [34] shows general details about the information management platforms.

Big data comprises unstructured, semistructured, and structured data. Big data mainly consists of unstructured data of types such as text, image, video, and audio. Semistructured data are produced by machines. Structured data mainly include transaction data.

As shown in Figure 4, when various types of data are obtained from many sources, they can be directly (in real time) written to memory processes or to disk as messages, files, or database processes. After obtaining the data, more than one option is available to keep them. They can be written to a file system or to distributed cluster systems, such as traditional RDBMS, NoSQL, and HDFS. The basic techniques used to rapidly evaluate unstructured data include Running Map

Reduce (Hadoop) and Memory Map Reduce (Spark). Additional evaluation options are available for real-time data streams. The integration layer in the middle is (while being organized) comprehensive and provides an open import, data store, data warehouse, and analytical architecture. The data can be analyzed using various tools. The business intelligence layer (under decision) contains interactive, real-time, and data modelling tools. These tools can leave a vast amount of data in their place and query, report, and model data. Along with traditional components, such as reports, dashboards, warnings, and queries, these tools also include advanced analytics, statistical analysis of a database or reservoir, and advanced visualization [34].

Big data management tools and analysis methods are greatly important today because they enable the analysis of data that is otherwise impossible to handle owing to its various dimensions. These tools and methods can be used to obtain new results from such data.

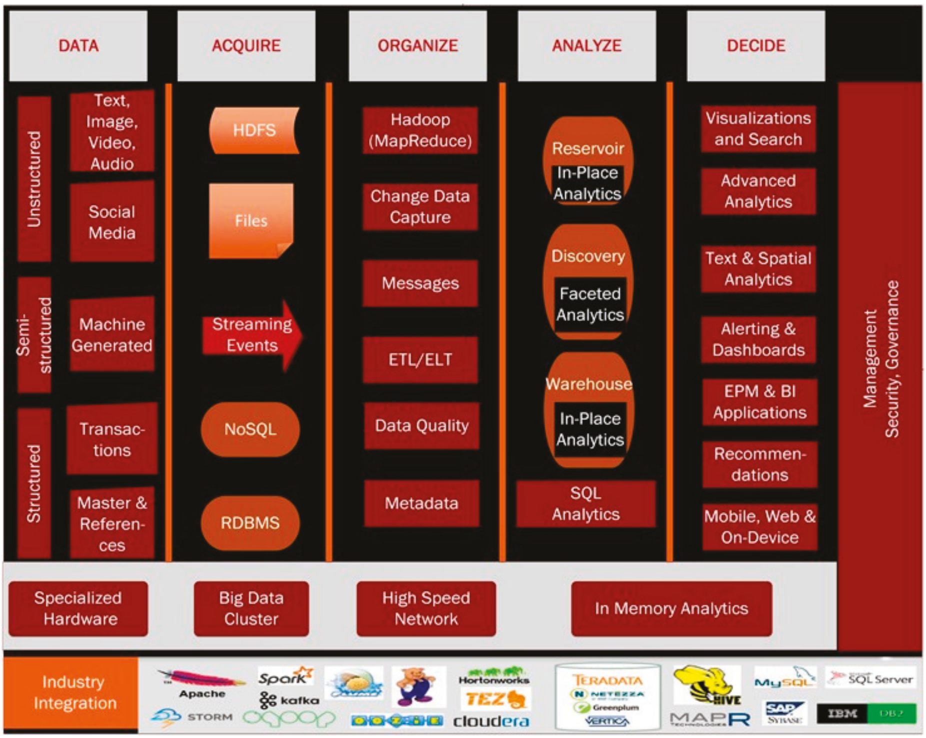


Figure 4. Big data and consolidated information management [34].

5. Big Social Network Data and Sustainable Development

Sustainable development has emerged as an important issue worldwide with industrialization and technological developments. Therefore, it is imperative to discuss this concept from different viewpoints and to develop it in parallel with new technological developments. One of the most important new technological developments was the widespread use of the internet at the turn of the century. Later, online social networks were developed; these can greatly influence sustainable development and are also responsible for producing big data, which is an important economic parameter. This study focuses on the large amounts of data (i.e., big data) produced on social networks, how new technological resources can be used to gauge the vast majority's opinions, and how this would affect sustainable development. Big data can provide very valuable information once it is analyzed. In fact, in 2012, attendees of the World Economic Forum held in Davos, Switzerland, declared big data to be a strategic economic resource, which is as important as money and gold [25,46].

In the most recent report prepared by the Independent Expert Advisory Group (IEAG) [4] under the UN Secretary General, the data revolution for sustainable development was defined as producing more detailed and high-quality information to encourage and monitor sustainable development by combining the data that was obtained from new technologies with traditional data. It clearly reemphasized the importance of big data by declaring that access to data should be improved to produce better policies and decisions to obtain better results for humans and our planet and to provide more participation, accountability, and transparency.

5.1. Effect of Big Social Network Data on Developmental Goals

The concept of big social data is very important because it affords possibilities, such as involving billions of people worldwide, providing a communication network between them, rapidly revealing real-time solutions by evaluating immediate data streams with big data analysis tools, and making suggestions and predictions by using these results to determine sustainable development policies, generate new policies, and dynamically shape them when needed. Big social network data may play an important role in generating a clearer and more up-to-date picture of the world, planning required policies and programs together, monitoring and assessing these programs, and evaluating the processes of sharing resources that could affect people's lives and influence political decision-making [4]. Social networks are quite important in many sustainability issues, such as information transfer, cooperation on management of shared resources, and the formulation of policies aimed at influencing various behaviors [7].

Many social network analysis methods can be used for this purpose. These methods, although still very new, are already very popular. Methods used for network analyses, such as anomaly detection [47], discrimination discovery [48], opinion leaders detection [49], event detection [50], role mining [51], rumor propagation detection system [52], conflict detection [53], and topic detection [54], can also greatly contribute in the field of sustainable development.

5.1.1. Helping Companies Invest in Accordance with Sustainable Development Policies

Big social networking data can contribute to the promotion of investments because they may enable companies to invest or encourage investing in accordance with sustainable economic objectives by using social network analysis facilities. In addition, these data may contribute to the prevention or rehabilitation of investments that are unsuitable for this objective. This means that investments that can permanently damage natural resources can be prevented; therefore, this will influence the efficient use of economic resources and the sustainability of existing investments. Nowadays, large numbers of people can directly influence companies and their image and investments through social media. In other words, social media has become a preferred and inexpensive marketing tool that encourages communication among companies and consumers, thus liberating communication to a hitherto unseen extent. Customers have made social media a marketing tool by expressing their views about the community and their reactions. Social media has become a channel that allows for companies to address its customer's concerns and questions and to interact with its customers. When it is used effectively, it can contribute to brand building [55]. Brand building on social media may be oriented toward sustainable development. Therefore, companies can enjoy great opportunities for customer interaction when formulating sustainable policies.

Today, communicating with customers via social media has become an integral part of functions, such as marketing, public relations, and customer services [56]. Companies consider social sharing networks as the perfect tool for spreading marketing information, attracting new customers, or merely obtaining valuable feedback from dissatisfied customers [57].

Technological developments in big data infrastructures, analytics, and services have allowed for companies to transform themselves into data-driven entities. Owing to its great potential, big data has been portrayed as a factor that has changed the rules of the game, and therefore, companies have improved their capacity to use big data to compete in the market [58]. All social networking and big data facilities help companies to be updated in terms of sustainable policies and to generate dynamic economic policies.

Similarly, states, related institutions, and nongovernmental organizations that are striving to produce sustainable development policies can also contribute to the formulation of policies and sustainable investments by using big social network data and analysis opportunities. Big data analysis allows for large quantities of data to be analyzed both retrospectively and instantly, and to reveal undiscovered information and even unknown parameters. For example, thousands of tweets from various accounts over a certain period can be analyzed rather than 50,000 tweets on a topic per day. In addition, big social network analysis is related to many different social network challenges that are unique to them. For example, the rumor propagation detection system [52] usually addresses the problem of finding an expansion of false information in social networks. As the number of social media tools increases, the amount of information and its spread is increasing, and therefore, information is increasingly exposed to modification and deterioration [59]. It is possible to quickly find information that supports investments that are unsuitable for sustainable development policies or that aim to produce such policies. After defining wrong policies, we can produce campaigns and policies against them. This approach may provide serious opportunities for the adoption of sustainable development policies by society.

Many companies actively use social network platforms. Figure 5 shows the importance of and usage ratio of these platforms by marketers.

The statistical graph in Figure 5 presents worldwide data about social media platforms used by marketers as of January 2017.

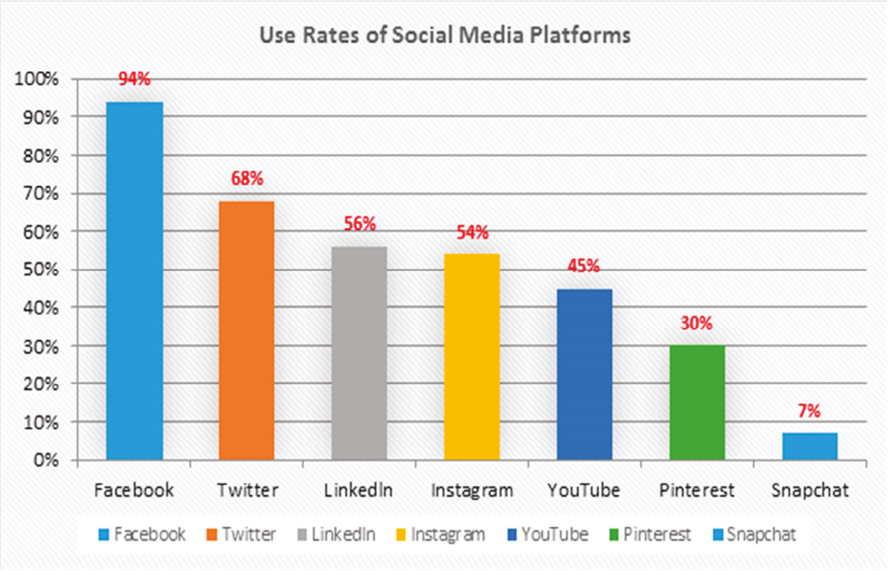


Figure 5. January 2017, usage ratio of social platforms by marketers worldwide.

As of January 2017, 94%, 56%, and 30% of respondents stated that Facebook, LinkedIn, and Pinterest, respectively, were the most important social platform for their establishments [60]. These data show the importance and comprehensiveness of the social platforms and the big data they produce as they enable establishments to rapidly reach more than 2 billion people.

The fact that companies and large numbers of people can reach each other through social networks may lead to the use of more efficient technologies that are suitable for sustainable development in many sectors. When considering the investment size and economic share of the private sector worldwide, private companies clearly have significant influence. For example, the construction sector is on its way to becoming the world's largest energy consumer; it uses 40% of global energy and causes one-third of global greenhouse gas emissions. Thus, campaigns that encourage a sustainable construction industry can be created through social networks, because energy efficiency has increasingly become one of the biggest concerns for a sustainable society [3]. The impact of social networks is huge and can lead the sector in this direction.

For example, Yazdanifard and Yee studied the effects of social network sites on the hospitality and tourism industries. They found that the establishment of a broad communication network through social networks is an important way to rapidly spread knowledge worldwide and to build a brand image. It also becomes easier to reach potential customers who are members of social networks. Finally, they stated that social networks enable people to express their views on companies. All of the above factors affect the hospitality and tourism sectors, making social networks an appropriate tool. Social networks can help make industries, such as tourism, more productive [17], as well as more efficient in terms of sustainable development.

The effective use of social network data would increase productivity in many sectors and contribute to sustainable development. This can have a great impact on economic productivity. This means that the effective use of big social network data might affect and support developmental goals 7–10 specified in the sustainable goals for 2030.

5.1.2. Helping States Formulate Sustainable Policies

Social networks and big data have led to the addition of digital data to the paradigm of the state organization. Big social network data have led to a great revolution and currently have great potential for state structures, which remain awkward, slow to react, and slow in formulating real-time policies that can meet society's needs to renew itself and stay up-to-date. Policy-making is now being actualized in an ever-prospering environment that brings new promises and creates difficulties for policy-makers. Social network data offers the chance to make and implement policies by considering the needs, preferences, and real public service experiences of citizens, and to be citizen-oriented. When citizens express their political views on social network sites like Twitter and Facebook, by rating the service they receive at state institutions or by debating issues on the sites of various social institutions and NGOs, they generate a series of data that state institutions can use to better themselves. Policy-makers can access a wide range of information about citizens' actual behaviors when they interact with government institutions or undertake a number of citizenship commitments, such as signing a petition [61].

These data obtained from social media or via administrative operations also provide new data that can enable government institutions to monitor and improve their own performances by way of daily usage data of their electronic existence or recorded internal transactions. Governments can learn what people say about the government; understand which policies, services, or contractors are subject to negative views and complaints; identify unsuccessful institutions such as schools, hospitals, or contractors; and, use social media data to make automatic improvements. They can obtain such data by using the data on their own sites or that on social networks. They can find out what people worry about or what people are searching for through Google Search APIs that save most users' search terms [62].

These new opportunities that have emerged with technological developments can transform the relationship between the state and the society into a new paradigm. In light of this transformation, if big social network data and analysis opportunities are used for achieving sustainable development objectives, they may lead states to the creation of economically sustainable policies, more equitable education policies, gender equality in government institutions, new policies for protecting the ecosystem, etc. This role of big social networks has begun to be discussed in international institutions, such as the UN. In turn, this has led to the production of related policies. For example, on 23 September 2016, Twitter and the UN Global Pulse announced a partnership that would provide access to Twitter data tools to support the efforts of the UN to reach the Sustainable Development Goals that were adopted by world leaders last year. Robert Kirkpatrick, the Director of UN Global Pulse, said that "The Sustainable Development Goals are first and foremost about people, and Twitter's unique data stream can help us truly take a real-time pulse on priorities and concerns—particularly in regions where social media use is common—to strengthen decision-making. Strong public-private partnerships like this show the vast potential of big data to serve the public good" [63].

Public institutions can partner with social networks and exploit their capabilities to become more effective and to create more sustainable policies in any field in which government institutions are active. These dynamic policies signify a direct positive contribution to many sustainable development objectives that governments are influencing. These emerging new opportunities carry great hope in terms of government policies.

5.1.3. Contributing to the Protection of Women's Rights

One of the objectives of sustainable development is to strengthen the role of women and girls in social life by ensuring gender equality. This objective includes goals, such as ensuring equal right to education, access to health services, and a safe work environment [64]. Big social network data and its analysis can contribute greatly toward achieving this objective. Current and past communications about women can be collected from social media platforms to help create an image of women's social and economic situation, education and working conditions, and harassment. This information can help in achieving social equality by offering valuable benefits to many institutions that work to strengthen the position of women in society.

For example, the ILO (International Labour Organization) used social media as a tool to monitor workplace discrimination faced by women in the Asia-Pacific region, where gender-based discrimination remains widespread. Female labor participation in this region has increased greatly over the last two decades. This increase has brought along not only sexual and ethnic discrimination, but also sexual harassment. For example, Indonesian women have limited access to employment, face a 35% wage difference, and enjoy unequal employment and education possibilities, as well as professional responsibilities. In the last decade, women's labor participation rate was 50–53%, whereas that of men was 80–83% [65].

Big data provides new and important opportunities to extract real-time information from community behaviors. In Indonesia, social media data mining, especially of tweets, is a good alternative to expensive traditional data collection methods in long-term studies for obtaining new information about workplace discrimination. Together with the Indonesian Government and the ILO, the UN Global Pulse Laboratory in Jakarta tested whether social media supervision would provide indications regarding real-time workplace discrimination against women. They analyzed tweets in the Bahasa language between 2010 and 2013 and found tens of thousands of tweets about work permits, job suitability, workload of working women, and employment discrimination [65].

Because women fear highlighting the injustices at workplaces or institutions where they are working, they may use social media platforms instead, making these platforms the place where accurate information is available. Such projects can provide governments with accurate information on women's working conditions, as well as gender equality.

5.1.4. Contributing to Sustainable Environment Policies

Sustainable environment policies are one of the main elements of sustainable development. It is necessary to leave a healthy environment to live in for future generations. It is very important for environmental policies to be implemented and adopted by the community today as healthy environmental conditions are gradually disappearing. In particular, global warming and climate change are alarming for human life. Big social network platforms have a key role in the implementation and expansion of these policies, and to measure the community's reaction to them in terms of the number of people reached.

For example, the Asia-Pacific region is one of the most important areas from the viewpoint of climate change. Even though climate change is not a regional priority, to deal with it in regions where it has greater effect, the awareness of people in these regions should be increased. However, sufficient data cannot be obtained using traditional data collection tools to increase the awareness and participation of people to solve the problem of climate change. Global Pulse and the Secretary General's Climate Change Support Team created a tool to monitor real-time social media interactions before and after the 2014 Climate Summit. Daily tweets on climate change and related issues in English, French, and Spanish were examined. These tweets helped to measure the pulse of the community and to produce new policies [66].

Nwagbara's work considered the position of the Niger Delta, which has strategic importance in the socioeconomic and political development of Nigeria in the supply of global energy. He stated that environmental sustainability in this region should be urgently monitored, social media tools should be used effectively for this purpose, and sufficient efforts should be made in the struggle against climate change for a more sustainable future. This study mentions criticisms against sustainability commitments by the shareholders of multinational companies in shaping corporate social responsibility (CSR) policies and lays the emphasis on the role of communication in the formation of immaterial assets, such as corporate reputations. Owing to corporate pressures along with climate change and the pressure of social media technology, this article suggests that communication on social media should be used to advance the CSR obligations of multinational companies toward a more sustainable future in the Niger Delta in Nigeria [67].

5.1.5. Contributing toward Sustainable Development Goals in Healthcare

Another goal of sustainable economic development is to ensure that people live a healthy life and that people of all ages prosper [68]. In 1992, the World Health Organization defined development as "the process to improve the quality of life for humans"; furthermore, attainable standard of health was defined as "one of the fundamental rights of every human being without distinction of race, religion, political belief, economic or social condition" [69,70]. The 2030 targets for sustainable development objectives include, reducing child mortality, protecting maternal health, and combating diseases such as malaria, tuberculosis, polio, and HIV/AIDS. In this context, big social networking platforms allow for people to access platforms on health without making any distinction on economic or social condition. Thus, big social data may contribute toward achieving sustainable development objectives in health and other areas. Because people share their experiences about health problems through social networks, these networks can play a significant role in informing the masses. In addition, by increasing awareness about protective health measures, which is an important factor in the prevention of diseases, the possibility of infection may be reduced. The analysis of social media communications about diseases, such as malaria, tuberculosis, polio, and HIV/AIDS, may help to take effective decisions regarding sustainable objectives. In addition, it is conduct a social-media-based campaign against these diseases. These measures will contribute greatly to the health policies stated in the sustainable development objectives.

Google used big data analysis to measure the spread of flu and claimed that current public services could warn people about flu epidemics two weeks in advance. They detected a direct proportion between the searches that were made on relevant search engines by millions of people, and the occurrence of illness. If the public and the authorities can be warned in advance, it would provide a great advantage in preventing the potential increase of flu incidents [71]. In today's highly connected world, social media provides a data point on people searching about a flu epidemic by monitoring how people search about symptoms. In fact, social media, mobile phones, and other means of communication have provided a portal for providing information to the public by opening a two-way road in healthcare investigations and also enabled people to express their concerns, locations, and physical movements [72]. A study conducted by ICF Macro consulting company on behalf of the Toxic Substances and Illnesses Organization Agency in China regarding two environmental health issues, perchlorates found in baby food and mold on dry walls, determined that the obtained results matched blogs, Facebook posts, and published official reports on the same topics [73]. In another study, Twitter data were used to try to detect flu trends in real time. Early diagnosis is very important to fight against epidemics. The traditional approach used by the Centers for Disease Control and Prevention (CDC) involves collecting influenza-like illnesses' (ILI) activity data from medical practices. This introduces a delay of 1–2 weeks between a patient's diagnosis and the entry of this information into reports. In this study, a framework to monitor the flu epidemic and make relevant guesses is recommended by examining tweets under the name of Social Network Enabled Flu Trends (SNEFT). The comparisons showed that tweets and illness ratios in the reports agreed with each other [74].

If early disease detection is possible using social network analysis, early protective measures can be taken to save the millions of dollars that are otherwise spent on these diseases, and these resources can instead be spent on improving human health. In addition, policies that are consistent with sustainable development goals can be organized by analyzing big social network data.

6. Conclusions

In this study, by discussing the concept of social networks and big data together, the contribution of new data analysis opportunities that have emerged with recent technological developments to the 2030 sustainable development objectives is discussed. Unlike previous studies, in addition to the topics of sustainable development and social networks, the issue of big data, which has become significant today, is also discussed. The concept of big social data is evaluated as a parameter that may influence sustainable development. The prominence of big social networking data in sustainable development is revealed by explaining companies' investment policies and the dynamism of state policies through examples such as women's rights, sustainable environment, and health.

Moreover, it is proved that the use of solutions that have been used for resolving various social network analysis problems for achieving sustainable development objectives and handling big social data will contribute directly or indirectly to every sustainable development objective. It is now essential to analyze social network platforms using big data analysis tools to obtain effective results. However, few studies and analyses have focused on big social network data for sustainable development objectives. In the future, many disciplines have to analyze social network data using big data tools to better understand the discipline and make correct predictions in a given field. Such analyses can also provide real-time results that measure the community's pulse as never before. This advantage of big social networks will be a significant factor in the establishment of a sustainable world economy.

Author Contributions: The research was conceived and designed by Umit Can and Bilal Alatas. Bilal Alatas studied social networks and big data and contributed to the related sections. Umit Can contributed to the introduction and the sections on big social network data and sustainable development.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. UN Documents, Report of the World Commission on Environment and Development: Our Common Future. Available online: http://www.un-documents.net/wced-ocf.htm (accessed on 25 September 2017).
2. Sustainable Development Knowledge Platform, Transforming Our World: The 2030 Agenda for Sustainable Development. Available online: https://sustainabledevelopment.un.org/post2015/transformingourworld (accessed on 5 September 2017).
3. Koseleva, N.; Ropaite, G. Big Data in Building Energy Efficiency: Understanding of Big Data and Main Challenges. *Procedia Eng.* **2017**, *172*, 544–549. [CrossRef]
4. A World that Counts: Mobilizing a Data Revolution for Sustainable Development. Available online: http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf (accessed on 20 September 2017).

5. Oussous, A.; Benjelloun, F.Z.; Lahcen, A.A.; Belfkih, S. Big Data Technologies: A Survey. *J. King Saud Univ. Comput. Inf. Sci.* **2017**, in press. [CrossRef]
6. Willard, T. *Social Networking and Governance for Sustainable Development*; International Institute for Sustainable Development: Winnipeg, MB, Canada, 2009; pp. 1–34.
7. Henry, A.D.; Vollan, B. Networks and the challenge of sustainable development. *Annu. Rev. Environ. Resour.* **2014**, *39*, 583–610. [CrossRef]
8. Meadows, D.H.; Meadows, D.L.; Randers, J.; Behrens, W.W. *The Limits to Growth*, 5th ed.; Universe Books: New York, NY, USA, 1972, ISBN 0-87663-165-0.
9. Declaration of the United Conference on the Human Environment. Available online: http://www.un-documents.net/aconf48-14r1.pdf (accessed on 25 September 2017).
10. The Rio Declaration on Environment and Development. Available online: http://www.unesco.org/education/pdf/RIO_W.PDF (accessed on 26 September 2017).
11. Wheeler, S.M.; Beatley, T. *The Sustainable Urban Development Reader*, 3rd ed.; Routledge: New York, NY, USA, 2014; pp. 79–87, ISBN 978-0-415-70775-6.
12. Harris, J.M. Basic principles of sustainable development. In *Dimensions of Sustainable Development*; Seidler, R., Bawa, K.S., Eds.; Eolss Publishers Co. Ltd.: Oxford, UK, 2009; Volume 1, pp. 21–41, ISBN 978-1-84826-207-2.
13. Milano, R.; Baggio, R.; Piattelli, R. The effects of online social media on tourism websites. In Proceedings of the 18th International Conference on Information Technology and Travel & Tourism, Innsbruck, Austria, 26–28 January 2011.
14. Ozkan-Canbolat, E.; Beraha, A. Evolutionary knowledge games in social networks. *J. Bus. Res.* **2016**, *69*, 1807–1811. [CrossRef]
15. Brandão, M.A.; Moro, M.M. Social professional networks: A survey and taxonomy. *Comput. Commun.* **2016**, *100*, 20–31. [CrossRef]
16. File:Social-network.svg. Available online: http://en.wikipedia.org/wiki/Image:Social-network.svg (accessed on 14 September 2017).
17. Yazdanifard, R.; Yee, L.T. Impact of social networking sites on hospitality and tourism industries. *Glob. J. Hum. Soc. Sci. Econ.* **2014**, *14*. Available online: https://globaljournals.org/GJHSS_Volume14/1-Impact-of-Social-Networking.pdf (accessed on 5 November 2017).
18. Sapountzi, A.; Psannis, K.E. Social networking data analysis tools & challenges. *Future Gener. Comput. Syst.* **2016**, in press.
19. Özköse, H.; Arı, E.S.; Gencer, C. Yesterday, today and tomorrow of big data. *Procedia Soc. Behav. Sci.* **2015**, *195*, 1042–1050. [CrossRef]
20. Li, H.; Lu, K.; Meng, S. Big Provision: A Provisioning Framework for Big Data Analytics. *IEEE Netw.* **2015**, *29*, 50–56. [CrossRef]
21. About. Available online: https://about.twitter.com/company (accessed on 17 September 2017).
22. YouTube for Press. Available online: https://www.youtube.com/yt/about/press/ (accessed on 3 September 2017).
23. The Statistics Portal, Most Famous Social Network Sites Worldwide as of August 2017, Ranked by Number of Active Users (in Millions). Available online: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed on 28 September 2017).
24. Diebold, F.X. *"Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting, Advances in Economics and Econometrics, Eighth World Congress of the Econometric Society*; Cambridge University Press: Cambridge, UK, 2003; pp. 115–122, ISBN-10: 052152413X.
25. Alharthi, A.; Krotov, V.; Bowman, M. Addressing barriers to big data. *Bus. Horiz.* **2017**, *60*, 285–292. [CrossRef]
26. Every Day Big Data Statistics—2.5 Quintillion Bytes of Data Created Daily. Available online: http://www.vcloudnews.com/every-day-big-data-statistics-2--5-quintillion-bytes-of-data-created-daily/ (accessed on 5 April 2015).
27. LaValle, S.; Lesser, E.; Shockley, R.; Hopkins, M.S.; Kruschwitz, N. Big data, analytics and the path from insights to value. *MIT Sloan Manag. Rev.* **2011**, *52*, 21.
28. Quick Facts and Stats on Big Data. Available online: http://www.ibmbigdatahub.com/gallery/quick-facts-and-stats-big-data (accessed on 16 September 2017).
29. Big Data Statistics & Facts for 2017. Available online: https://www.waterfordtechnologies.com/big-data-interesting-facts/ (accessed on 16 September 2017).

30. How Big Data Analysis helped increase Walmarts Sales Turnover? Available online: https://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109 (accessed on 9 September 2017).
31. The White House. Obama Administration Unveils Big Data Initiative: Announces 200 Million in New R&D Investments. Available online: https://obamawhitehouse.archives.gov/the-press-office/2015/11/19/release-obama-administration-unveils-big-data-initiative-announces-200 (accessed on 18 September 2017).
32. Chen, M.; Mao, S.; Zhang, Y.; Leung, V.C.M. *Big Data: Related Technologies, Challenges and Future Prospects*; Springer: New York, NY, USA, 2014; pp. 2–9, ISBN 978-3-319-06245-7. (online).
33. United Nations Global Pulse. White Paper: Big Data For Development: Opportunities & Challenges. 2012. Available online: https://www.unglobalpulse.org/projects/BigDataforDevelopment (accessed on 2 September 2017).
34. An Enterprise Architect's Guide to Big Data, Oracle. Available online: http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf (accessed on 15 September 2017).
35. Gandomi, A.; Haider, M. Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manag.* **2015**, *35*, 137–144. [CrossRef]
36. Apache Hadoop. What Is Apache Hadoop? Available online: http://hadoop.apache.org/ (accessed on 13 September 2017).
37. Cloudera. Available online: https://www.cloudera.com/ (accessed on 14 September 2017).
38. Hortonworks. Available online: https://hortonworks.com/ (accessed on 14 September 2017).
39. Shvachko, K.; Kuang, H.; Radia, S.; Chansler, R. The hadoop distributed file system. In Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Incline Village, NV, USA, 3–7 May 2010.
40. Dean, J.; Ghemawat, S. MapReduce: A flexible data processing tool. *Commun. ACM* **2010**, *53*, 72–77. [CrossRef]
41. Lakshman, A.; Malik, P. Cassandra: Structured storage system on a p2p network. In Proceedings of the 28th ACM symposium on Principles of distributed computing, Calgary, AB, Canada, 10–12 August 2009.
42. Chang, F.; Dean, J.; Ghemawat, S.; Hsieh, W.C.; Wallach, D.A.; Burrows, M.; Chandra, T.; Fikes, A.; Gruber, R.E. Bigtable: A distributed storage system for structured data. *ACM Trans. Comput. Syst. (TOCS)* **2008**, *26*, 4. [CrossRef]
43. Kulkarni, A.P.; Khandewal, M. Survey on Hadoop and Introduction to YARN. *Int. J. Emerg. Technol. Adv. Eng.* **2014**, *4*, 82–87.
44. Taylor, R.C. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinform.* **2010**, *11*, S1. [CrossRef] [PubMed]
45. Saha, B.; Shah, H.; Seth, S.; Vijayaraghavan, G.; Murthy, A.; Curino, C. Apache tez: A unifying framework for modeling and building data processing applications. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Australia, 31 May–4 June 2015.
46. Johnson, J.E. Big data + big analytics = big opportunity: Big data is dominating the strategy discussion for many financial executives. As these market dynamics continue to evolve, expectations will continue to shift about what should be disclosed, when and to whom. *Financ. Executive* **2012**, *28*, 50–54.
47. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly Detection: A Survey. *ACM Comput. Surv.* **2009**, *41*, 1–72. [CrossRef]
48. Ruggieri, S.; Pedreschi, D.; Turini, F. Data mining for discrimination discovery. *ACM Trans. Knowl. Discov. Data (TKDD)* **2010**, *4*, 9. [CrossRef]
49. Cho, Y.; Hwang, J.; Lee, D. Identification of effective opinion leaders in the diffusion of technological innovation: A social network approach. *Technol. Forecast. Soc. Chang.* **2012**, *79*, 97–106. [CrossRef]
50. Zhou, X.; Chen, L. Event detection over twitter social media streams. *VLDB J.* **2014**, *23*, 381–400. [CrossRef]
51. Rossi, R.A.; Ahmed, N.K. Role discovery in networks. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 1112–1131. [CrossRef]
52. Zhang, Q.; Zhang, S.; Dong, J.; Xiong, J.; Cheng, X. Automatic detection of rumor on social network. In *Natural Language Processing and Chinese Computing*; Springer: Cham, Germany, 2015; Volume 9362, pp. 113–122.
53. Gürses, S.; Berendt, B. The social web and privacy: Practices, reciprocity and conflict detection in social networks. In *Privacy-Aware Knowledge Discover: Novel Applications and New Techniques*, 1st ed.; Bonchi, F., Ferrari, E., Eds.; Chapman and Hall/CRC: New York, NY, USA, 2010; pp. 395–432, ISBN 978-1-4398-0366-0.

54. Cataldi, M.; Di Caro, L.; Schifanella, C. Emerging topic detection on twitter based on temporal and social terms evaluation. In Proceedings of the Tenth International Workshop on Multimedia Data Mining, New York, NY, USA, 25 July 2010.
55. Carraher, S.M.; Buchanan, J.K.; Puia, G. Entrepreneurial need for achievement in China, Latvia, and the USA. *Balt. J. Manag.* **2010**, *5*, 378–396. [CrossRef]
56. Berkowitch, A. What does success look like for your company: Social media starting points with measurable returns. *People Strateg.* **2010**, *33*, 10.
57. Ioanid, A.; Scarlat, C. Factors Influencing Social Networks Use for Business: Twitter and YouTube Analysis. *Procedia Eng.* **2017**, *181*, 977–983. [CrossRef]
58. Lee, I. Big data: Dimensions, evolution, impacts, and challenges. *Bus. Horiz.* **2017**, *60*, 293–303. [CrossRef]
59. Nel, F.; Lesot, M.J.; Capet, P.; Delavallade, T. Rumour detection and monitoring in open source intelligence: Understanding publishing behaviours as a prerequisite. In Proceedings of the Terrorism and New Media Conference, Dublin, UK, 8–9 September 2010.
60. Which Social Media Platform(s) Do You Use to Market Your Business? Available online: https://www.statista.com/statistics/259379/social-media-platforms-used-by-marketers-worldwide (accessed on 25 September 2017).
61. Marsh, D.; Stoker, G. *Theory and Methods in Political Science*, 3rd ed.; Palgrave Macmillan: New York, NY, USA, 2010; pp. 1–34, ISBN 978-0-230-57626-1.
62. Margetts, H. The Promises and Threats of Big Data for Public Policy-Making. Available online: http://blogs.oii.ox.ac.uk/policy/promises-threats-big-data-for-public-policy-making/ (accessed on 17 September 2017).
63. Twitter and UN Global Pulse Announce Data Partnership. Available online: http://www.unglobalpulse.org/news/twitter-and-un-global-pulse-announce-data-partnership (accessed on 17 September 2017).
64. Goal 5: Achieve Gender Equality and Empower All Women and Girls. Available online: www.un.org/sustainabledevelopment/gender-equality (accessed on 24 September 2017).
65. United Nations Global Pulse. Feasibility Study: Identifying Trends in Discrimination against Women in the Workplace in Social Media (2014). Available online: https://www.unglobalpulse.org/projects/indonesia-women-employment (accessed on 16 September 2017).
66. United Nations Global Pulse. Using Twitter to Measure Global Engagement on Climate Change. Available online: https://www.unglobalpulse.org/projects/Twitter-Climate-Change (accessed on 24 September 2017).
67. Nwagbara, U. The effects of social media on environmental sustainability activities of oil and gas multinationals in Nigeria. *Thunderbird Int. Bus. Rev.* **2013**, *55*, 689–697. [CrossRef]
68. Goal 3: Ensure Healthy Lives and Promote Well-Being for All at All Ages. Available online: https://www.un.org/sustainabledevelopment/health/ (accessed on 24 September 2017).
69. Our Planet, Our Health. Report of the WHO Comission on Health and Environment. Available online: http://apps.who.int/iris/handle/10665/37933 (accessed on 25 September 2017).
70. WHO. *Basic Documents*, 48th ed.; WHO: Geneva, Switzerland, 2014; Available online: http://apps.who.int/gb/bd/PDF/bd48/basic-documents-48th-edition-en.pdf (accessed on 26 September 2017).
71. Google Predicts Spread of Flu Using Huge Search Data. Available online: https://www.theguardian.com/technology/2008/nov/13/google-internet (accessed on 10 September 2017).
72. Schmidt, C.W. Trending now: Using social media to predict and track disease outbreaks. *Environ. Health Perspect.* **2012**, *120*, a30. [CrossRef] [PubMed]
73. Vincent, N. Social Media and Environmental Health Crises: An Examination of Public Response to Imported Drywall and Perchlorate Health Risks. Presented at American Public Health Association Annual Meeting and Exposition, Washington, DC, USA, 31 October 2011.
74. Achrekar, H.; Gandhe, A.; Lazarus, R.; Yu, S.H.; Liu, B. Predicting flu trends using twitter data. In Proceedings of the 2011 IEEE Conference, Computer Communications Workshops (INFOCOM WKSHPS), Shanghai, China, 10–15 April 2011.

sustainability

MDPI

Article

Exploring the Technological Collaboration Characteristics of the Global Integrated Circuit Manufacturing Industry

Yun Liu [1,2], Zhe Yan [1], Yijie Cheng [1] and Xuanting Ye [1,*]

[1] School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China; liuyun@ucas.ac.cn (Y.L.); yanzhe987456@163.com (Z.Y.); cheng.yijie2008@163.com (Y.C.)
[2] School of Public Policy and Management, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: yexuant@bit.edu.cn; Tel.: +86-10-6891-8823

Received: 31 October 2017; Accepted: 11 January 2018; Published: 22 January 2018

Abstract: With the intensification of international competition, there are many international technological collaborations in the integrated circuit manufacturing (ICM) industry. The importance of improving the level of international technological collaboration is becoming more and more prominent. Therefore, it is vital for a country, a region, or an institution to understand the international technological collaboration characteristics of the ICM industry and, thus, to know how to enhance its own international technological collaboration. This paper depicts the international technological collaboration characteristics of the ICM industry based on patent analysis. Four aspects, which include collaboration patterns, collaboration networks, collaboration institutions, and collaboration impacts, are analyzed by utilizing patent association analysis and social network analysis. The findings include the following: first, in regard to international technological collaboration, the USA has the highest level, while Germany has great potential for future development; second, Asia and Europe have already formed clusters, respectively, in the cooperative network; last, but not least, research institutions, colleges, and universities should also actively participate in international collaboration. In general, this study provides an objective reference for policy making, competitiveness, and sustainability in the ICM industry. The framework presented in this paper could be applied to examine other industrial international technological collaborations.

Keywords: international technological collaboration; IC manufacturing; patent association analysis; social network analysis; collaboration network

1. Introduction

The Integrated Circuit (IC) industry is the foundation and power source of high-speed development in the information technology industry [1] and it has infiltrated every area of national economic and social development. The technological level and development scale of a country's IC industry has become the key for achieving competitive advantage and sustainable development for companies, economies, and societies [2,3]. For example, as one of the research directions of the IC industry, Radio Frequency Identification (RFID) technology is a catalyst which can promote big data analysis and application [4]. RFID can revolutionize big data analytics, benefitting various sustainability themes: the use of big data analytics based on RFID technology in areas such as smart city, smart power grid, smart building, smart industry, supply chain analytics, and better logistics planning can effectively reduce the environmental stress from broader social and economic activities [5,6]. In addition, countries all over the world have accelerated the strategic adjustment for their economic structures after the international financial crisis, which makes the strategic adjustment fundamental, and the leading position of the IC industry more obvious. For instance, the USA regards the IC industry as one of the four major technical areas that can radically transform the manufacturing industry over the next 20 years [7]. The European Union implemented its Micro- and Nanoelectronic Technology Industry Strategy Roadmap [8]. Also, China issued the policy "Outline of Development of the National Integrated Circuit Industry" and a series of policy documents [9–11] that promoted the status of the IC industry from inside the industry to the national level. As a result of the crucial technological and economic importance of its innovations, it is no surprise that this industry is the focus of many innovation researchers [12]. Located in the middle of the IC industrial value chain [13], the Integrated Circuit Manufacturing (ICM) industry is both technology- and capital-intensive, and the level of technological development also matters a great deal in this sector [14,15].

In fact, the ICM industry is experiencing a rapid technological development [16,17]. The number of transistors on a single chip is up to 10^9 orders in magnitude from 10 of thirty years ago [18]. As ICM technologies become more complicated, it is impossible for any single institution to endure merely with its own patents and technologies [15]. Technological cooperation is becoming increasingly popular [19]. Also, because of the intensification of the international competition and the existence of a technological gap among countries, international technological collaboration is increasing gradually [20]. International technological collaboration plays a pivotal role in promoting the emergence of a knowledge flow among countries [21]. The benefit from cross-border collaboration, particularly pronounced for a small country, is that the pool of knowledge a firm usually pulls from is no longer limited. Firms may thus benefit from the larger pool of knowledge provided by international collaboration partners that facilitate spillovers from a richer pool of other Research and Development (R&D)-active firms [22]. Therefore, collaborating with international partners has become an important strategy for firms in the ICM industry to develop next-generation technology by sharing expenses and risks [23].

More importantly, if a country, a region or an institution has a high level of international technological collaboration, it can further promote the transmission of technological knowledge to other organizations [24], thus contributing to achieving and utilizing technological knowledge resources worldwide with higher efficiency [25]. This will improve its innovation efficiency and technological level in the ICM industry because an organization's accumulated knowledge is key to its continued ability to innovate. Specifically, having a diverse technological knowledge-base within the organization can facilitate innovation through novel combinations of readily accessible pieces of knowledge [26]. On the other hand, this efficient technological knowledge collaboration, which is based on knowledge management, is essential to bring about sustainable innovation [27]. The necessary condition to generate a shift towards sustainable innovation is that the various actors and stakeholders involved can share knowledge and learn from experiments, practices, and other kinds of R&D activities [27]. It should be emphasized that the air pollution and water pollution generated in the process of IC manufacturing are receiving more and more attention [28,29]. The sustainable innovation which is promoted by the efficient international collaboration in the context of knowledge management can lead to sustainable consumption and production that are ecologically sound, are accepted and adopted by society, and effectively relieve the environmental pollution from the ICM industry. Therefore, a high level of international technological collaboration will enable a country, a region or an institution to enhance its innovation efficiency, technological level, and competencies to sustainably develop in the global ICM industry.

In summary, with the development of technology and the intensification of international competition, there are many international technological collaborations in the ICM industry [15,23], and the important role of improving the level of international technological collaboration is becoming more and more prominent. Therefore, it is vital for a country, a region or an institution in the ICM industry to understand the international technological collaboration characteristics of the ICM industry and thus to know its own advantages and disadvantages for international collaboration, which technology partners have a strong technical force and collaboration potential, and, finally, how to improve its own level of international technology collaboration. Nevertheless, although various computer chips are being investigated, few articles can be obtained about these research topics in the ICM industry.

As one of the objects of knowledge management, patents provide a reliable quantification basis for technology or industry development studies because they are the manifestation of the latest and most valuable technological knowledge for inventions [30,31]. In consequence, the majority of studies were done using patent bibliometric methods [32]. The ICM industry is also ideal for information on innovation, quality, and technological knowledge that can be obtained through the analyses of pertinent patent data [33–35]. This is because this industry places a strong emphasis on the importance of patents and patent rights [35]. The ICM industry is somewhat unique in that it is characterized by a highly cumulative process of innovation and has patented all landmark inventions dating back to the early 1980s [36,37]. IC firms often require access to a "thicket" of patent rights in order to advance the technology, or to legally produce or sell their products [16]; therefore, many of them have drawn considerable revenues from royalties and the licensing of patented technologies [38]. For example, in an earlier study, patents issued to ICM enterprises from 1975 to 1994 were utilized to examine the effects of the value of a firm's technological advancement and of the newness of its technology on the probability of its failure [12]. In 2010, another study showed a bibliometric analysis of the patent publications of 73 Taiwanese IC design firms covering the period from 1995 to 2007 [39]. Using the patents from Taiwan and U.S. IC design and manufacturing firms, Tsai [40] shed light on the patent and Research and Development (R&D) spillover effects on productivity. Later, patent citation data was analyzed to demonstrate that Taiwan's Hsinchu Science Park is a healthy and knowledge-based cluster surrounded by the IC sector [41]. Recently, researchers investigated the different developmental paths in the IC industry in China and India using qualitative and quantitative data, including U.S. utility patent holdings [3,42]. Patent data can also be used in the study of international technological cooperation [43]. Scholars have studied international technological collaboration by patent data from various angles, including the collaboration network [21,44,45], collaboration pattern [20], driving factors [46], knowledge flow [24], industrial degree [25], national degree [47,48], etc. Although the academia has actively promoted the research of international technological collaboration and ICM industrial innovation by patent analysis, ICM international technological collaboration studies using patent data are not yet fully developed.

Therefore, this paper attempts to explore the international technological collaboration characteristics of the ICM industry by discussing the issue from the following three points of view based on patent analysis:

(1) What are the differences among each country and region in the distribution of technical topics? What is the key cooperative direction for international collaboration of each country and region? By answering these questions, we can know for a specific country and region which countries and regions are easy to collaborate with and how to promote a technological collaboration with a country or region by focusing on some specific collaboration directions, and we can thus learn the potential for future development of each country and region in the ICM international technological collaboration.

(2) What position does each country and region occupy in the ICM international collaboration network? Which institutions from each country and region participate in international collaboration? Is there any difference in the distribution of institution types? These questions are mainly about the current situation of the ICM international technological collaboration. By answering these questions, we can know the performance of each country and region in the collaboration network or at the institutional level and the performance of each institution in the international technological collaboration, and thus we can put forward some suggestions specifically to improve the performance.

(3) What are the impact characteristics of each country and region in the ICM international collaboration network? What are the differences among the collaboration impact characteristics from typical countries and regions? By answering these questions, we can know which countries and regions have a high influence in the collaboration network, understand the relative strengths and

weaknesses of each country and region for their collaboration impact, and thus we can know the consequences of the international collaboration of each country and region.

The research results can help the policymakers of all countries and regions and the decision makers of different institutions understand the technological collaboration characteristics of the global ICM industry, recognize their own advantages and disadvantages, look for appropriate technology partners, and improve their performance in the collaboration network or at the institutional level, respectively, based on the specific suggestions. We try to explore the international technological collaboration characteristics to provide a reference for both the countries and regions and the institutions to improve their level of international technological collaboration.

The paper is organized as follows. First, we describe the data and methodology. For the data, we construct a technical classification system and patent search strategy for each subfield of the ICM industry. A sample of 11,621 patents in the ICM industry from the European Patent Office (EPO) was collected for analysis. In addition, the methods of patent bibliometric analysis and social network analysis utilized in this paper are also described. Second, the characteristics of the collaboration pattern, collaboration network, collaboration institution, and collaboration impact are described, based on patent analysis. Finally, conclusions are given and specific questions are discussed. Overall, this study explores the international technological collaboration characteristics in the ICM industry and provides an objective reference for both the countries and regions and institutions to facilitate their decisions for future policy making, to improve the level of international technological collaboration, and to sustainably develop in the global ICM industry.

2. Data and Methodology

2.1. Research Framework

We analyzed the characteristics of the international technology cooperation in the ICM industry from four aspects: collaboration pattern, collaboration network, collaboration institution, and collaboration impact, as shown in Figure 1. Specifically, through the analysis of the collaboration patterns, we could find out the differences among countries or regions in the main technological research directions, as well as their main technological collaboration subfields, and understand the potential for future technological collaborations for these countries and regions. Through the research on the collaboration networks and collaboration institutions, we could understand the developmental situation for each country and region. And through the research on the collaboration impact, we could understand to some extent the consequences of the international technological collaboration for each country and region.

2.2. Technical Classification System and Data Collection

Because the ICM includes various types of manufacturing processes and technologies, it is difficult to retrieve all patents by a single International Patent Classification (IPC) code or IPC combination. Thus, we first established an ICM technology classification system. Next, based on the improved vocabulary query methods, a patent retrieval strategy was formulated by selecting keywords in every subfield. As shown in Figure 2, the determination of the ICM technology classification system and patent retrieval strategy refer to three kinds of documents: (1) research literature related to the ICM, whose research contents can be divided into patent bibliometrics [12,39–41], developing situations [11–13,15], technology fields [14,18], etc.; (2) industrial information related to the ICM, which includes industrial research reports [49,50], popular science books [51], etc.; and (3) policy documents related to the ICM.

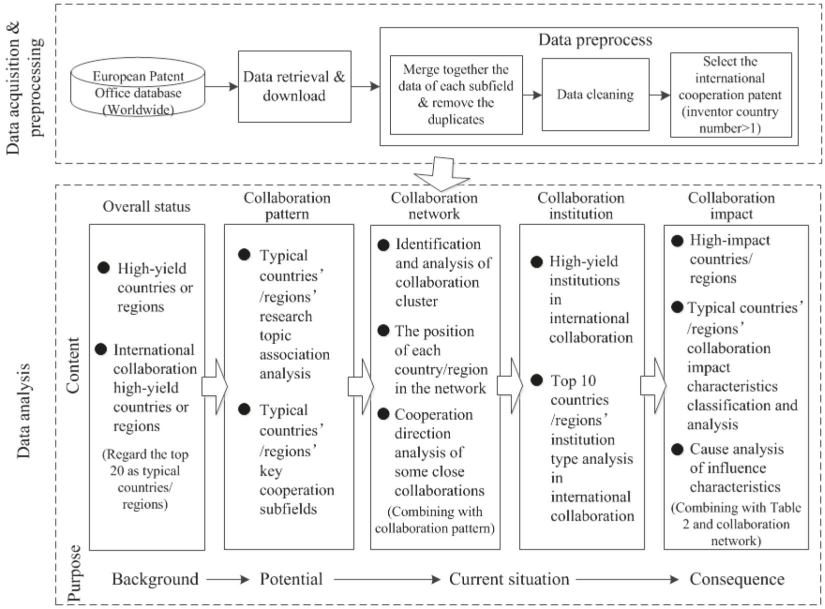

Data acquisition & preprocessing
European Patent Office database (Worldwide)
Data retrieval & download
Data preprocess
Merge together the data of each subfield & remove the duplicates
Data cleaning
Select the international cooperation patent (inventor country number>1)
Data analysis
Content
Overall status
High-yield countries or regions
International collaboration high-yield countries or regions
(Regard the top 20 as typical countries/ regions)
Collaboration pattern
Typical countries'/regions' research topic association analysis
Typical countries'/regions' key cooperation subfields
Collaboration network
Identification and analysis of collaboration cluster
The position of each country/region in the network
Cooperation direction analysis of some close collaborations
(Combining with collaboration pattern)
Collaboration institution
High-yield institutions in international collaboration
Top 10 countries/regions' institution type analysis in international collaboration
Collaboration impact
High-impact countries/ regions
Typical countries'/regions' collaboration impact characteristics classification and analysis
Cause analysis of influence characteristics
(Combining with Table 2 and collaboration network)
Purpose
Background
Potential
Current situation
Consequence

Figure 1. Research framework.

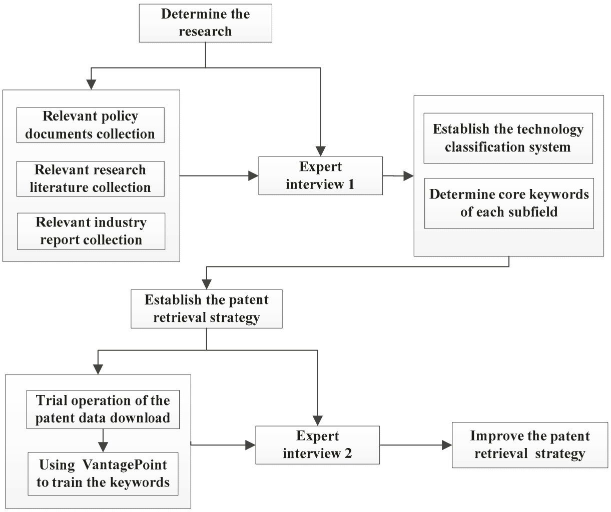

Determine the research
Relevant policy documents collection
Relevant research literature collection
Relevant industry report collection
Expert interview 1
Establish the technology classification system
Determine core keywords of each subfield
Establish the patent retrieval strategy
Trial operation of the patent data download
Using VantagePoint to train the keywords
Expert interview 2
Improve the patent retrieval strategy

Figure 2. The process of the establishment of the technical classification system and search strategy for the ICM.

As shown in Figure 2, in addition to the three document types for reference, interviews with technical experts and keyword training were also used for the optimization of the technology classification system and patent retrieval strategy. Among them, the expert interview mainly refers to communication with experts in the form of a semistructured interview. The interview is divided into two rounds, as shown in the figure. In the first round, the experts worked in the Chinese Academy of Science and Technology for

Development, and their research area is technology assessment and foresight. We consulted three experts for advice, mainly focusing on the establishment and improvement of the ICM industry's technology classification system and core keyword selection in each subfield. In the second round, we invited four experts from the Institute of Microelectronics of the Chinese Academy of Sciences, whose research area is the ICM. These experts mainly provided suggestions on the processing of the keyword training results and on the further improvement of the retrieval strategy. It is worth noting that every round of expert interviews involved multiple interviews. We expected to get guidance and recommendations from the experts based on their professional background knowledge through a number of exchanges about the interview results and a feedback between the two sides involved in the interview.

Keyword training is a method that searches and downloads patent data roughly on the basis of core keywords for each subfield. It also includes extracting vocabularies from the download patent data on the basis of the VantagePoint software and selecting the appropriate vocabularies as new keywords to improve the retrieval strategy by replacing the previous keywords or adding them directly. Patent data is then downloaded on the basis of the improved strategy, and a new round of keyword updates is conducted over and over, until no suitable new keyword appears [52,53]. Furthermore, the VantagePoint is a software application for structured text mining and analysis [52]. In this paper, the VantagePoint software was used to recognize and extract vocabularies from the title and summary information of the patent data.

Eventually, as shown in Table 1, the ICM technology was divided into a total of eight subfields: cleaning technology, lithography technology, etching technology, thin film technology, doping technology, annealing technology, planarization technology, and packaging technology. It is worth noting that the 8 subfields have more than 40 key technical topics and thus derived more than 60 keywords. The specific retrieval strategies are shown in Table A1.

Table 1. The Technical Classification System and Key Technical Topics of the ICM Industry.

No.	Technical Subfields	Key Technical Topics	Number of Patents Retrieved
1	Cleaning Technology (CT)	Plasma cleaning, Megasonic cleaning, Laser cleaning	566
2	Lithography Technology (LT)	X-ray Lithography (XRL), Focused Icon Beam Lithography (FIBL), Extreme Ultraviolet Lithography (EUV), Nanoimprint Lithography (NIL), Electron Projection Lithography (EPL)	539
3	Etching Technology (ET)	Wet Chemical Etching, Plasma Etching (PE), Reactive Ion Etching (RIE), Inductively Coupled Plasma Etching (ICP)	2495
4	Thin Film Technology (TFT)	Thermal Oxidation, Physical Vapor Deposition (PVD), Chemical Vapor Deposition (CVD), Electroplate, Vapor Phase Epitaxy (VPE), Molecular Beam Epitaxy (MBE), Complementary Metal Oxide Semiconductor	1021
5	Doping Technology (DT)	Ion Implantation, diffusion	855
6	Annealing Technology (AT)	Rapid Thermal Annealing (RTA), Laser Annealing	409
7	Planarization Technology (PLT)	Stress-Free Polishing (SFP)	736
8	Packaging Technology (PAT)	Copper Interconnect, Low K Dielectric, Optical Interconnect, Carbon Nanotubes (CNT), Through Silicon Via (TSV), Ball Grid Array Package (BGA), Chip Size Package (CSP), Multi Chip Package (MCP), Wafer Level Package (WLP), Flip Chip Package, Product In Package, System On Package (SOP), System In Package (SIP), 3D Packaging, Pin Gird Array Package (PGA)	5000

Note: In the third column, brackets contain the special abbreviation of the technology mentioned before the brackets. The acronym of each technical topic, if any, is also added to the patent search strategy to make the retrieved patent data more complete.

According to this established patent retrieval strategy, a total of 11,621 patents between 2004 and 2013 were downloaded from the world patent database from the EPO [32]. After the process of data cleaning and removal of duplicates, the remaining 7077 patents were built into a database. Data cleaning included the unification of the names of the countries and regions of the inventors and the deletion of patent data that lacked the inventor's nationality information. Next, 533 patents in which the number of the inventor's countries or regions was greater than or equal to 2 were selected to establish the ICM international cooperation patent database. It is of note that both the international co-inventor information and the international co-assignee information in the patents could be used for the international cooperation research. However, if inventors from two or more different nations worked together (i.e., international co-inventors), it implied that the inventive human resources from different nations were combined. If institutions from two or more different nations shared the ownership of the patent (i.e., international co-assignee), it would typically imply that these institutions engaged in collaboration in terms of finance, human resources, etc. [46]. In this paper, we believe that a direct contact among persons would be critical in identifying and designing an international technological collaborative research. Thus, we analyzed the ICM international technical cooperation based on the international co-inventions. This practice is not only the choice of many scholars for the study of the international technical cooperation [20,21,24,25,46,48,54], but it has also been discussed and proven effective [43,55].

2.3. Patent Association Analysis

The association analysis measures the degree of association among some related factors, such as the assignee, country or region, and application year. This analytic process transfers the degree of association into a matrix and maps it into a two-dimensional figure using a multidimensional scaling method [20]. Of the various measures used to produce association matrices, this study employed the similarity-based association analysis [56].

To be specific, we could obtain the degree of association by cosine similarity. Consider the patent inventor's country or region as an example, and assume there are m topic words in the patent having vectors $A = (a_1, a_2, \dots, a_m)$ and $B = (b_1, b_2, \dots, b_m)$, representing the eigenvectors of two different countries (or regions). Using cosine similarity, we define the degree of association between vectors A and B as follows:

$$cos(A,\ B) = \frac{\sum_{i=1}^{m} a_i b_i}{\left(\sum_{i=1}^{m} a_i^2 \times \sum_{i=1}^{m} b_i^2\right)^{\frac{1}{2}}},\ (i = 1,\ 2, \ldots, m), \tag{1}$$

where a_i is the number of times topic word i appears in the patent of an inventor belonging to country or region A; similarly, b_i is the number of times topic word i appears in the patent of an inventor belonging to country or region B. With this formula, we can calculate the degree of association in each country or region and further enhance the research on the collaboration pattern.

In this paper, the topic words' thesaurus was established according to the title and summary information from the ICM patents. Then, the degrees of technological correlation among countries or regions were calculated by the patent association analysis method based on all the ICM patents. Finally, a visual technology map was generated from the results. In addition, a stop words' thesaurus, synonym thesaurus, normal vocabulary of patent documents, thematic vocabulary were all constructed on the basis of specific technology fields, which could improve the effectiveness of the association analysis results.

2.4. Social Network Analysis

2.4.1. Cohesive Subgroups Analysis

The cohesive subgroups in the social network analysis mainly refer to subgroups with certain cohesion. Wasserman and Faust have defined it as that there are strong, direct, close, or positive relationships among nodes inside the cohesive subgroups [57]. At present, the scholars analyze the cohesive subgroups mainly from four perspectives: the reciprocity of the relationship, the proximity or accessibility between the nodes inside the subgroup, the frequency of the relationships among the nodes inside the subgroup, and the difference degree of the density of relationship among the nodes inside the subgroup relative to the density of the relationship between internal nodes and external nodes. This paper identifies the collaboration clusters in the international collaboration network based on the principle that the relationship among nodes inside the cluster is close, and the nodes inside the cluster have less collaboration with the nodes outside the cluster. This is the same as the fourth perspective of the cohesive subgroups research, and the CONCOR (Convergence of Iterated Correlations) algorithm is a common method to analyze cohesive subgroups from this perspective [58,59]. CONCOR is a kind of iterative correlation convergence method, clustering nodes by multiple iterations [60]. Specifically, the CONCOR process computes the Pearson product-moment correlation coefficients among the rows and columns of the input matrices by comparing the value of a given cell to the mean value of both the row and the column in which it occurs at first. It then uses the correlation matrices as input for a new round of correlation computations. The output from this calculation is used as input for yet another round of correlations, and the process continues in this fashion. After several iterations of this procedure, the values of all correlations in the matrix are equal to either +1 or −1. The final correlation matrices are dichotomized to allow all nodes to be grouped into different subgroups [61,62]. The CONCOR algorithm is frequently used for cluster analysis of multivariate relational data and multivalued relational matrices [58,63]. In this paper, we identified collaboration clusters in the ICM international collaboration network by the CONCOR algorithm.

2.4.2. Centrality Analysis

This paper assumes that there are close technological knowledge exchanges among inventors who have co-invented patents; thus, the technological information can be transmitted among different individuals. It is worth noting that this assumption has been supported by several empirical studies [24,64,65]. In this paper, actors in the collaboration network are taken as nodes, and the manifested relationships among them are seen as the links among the nodes [66]. A good position occupied by a node presents advantages in that it lets a central node enjoy profits in terms of information collection, processing, and transfer, and, therefore, it lets it have a strong impact in the network [54,67]. In other words, the position of a node in the network determines its collaboration impact. Thus, as a common evaluation criterion for the locations of nodes in a network, the centrality of a node can be used to analyze its collaboration impact within the network. Of the various indexes used to measure centrality, this study selected degree centrality, closeness centrality, and betweenness centrality [68]. It should be emphasized, for convenient comparisons, that the following formulas were used to calculate the relative centralities:

(1) Degree Centrality (DC)

The DC measures the number of nodes in the network that are directly connected to node *i* [69]. The higher the DC of a node, the higher the number of nodes it directly cooperates with, so the greater the direct collaboration impact range it has in the ICM collaboration network.

We define the DC as follows:

$$C_{RD}(i) = \frac{C_{AD}(i)}{C_{ADmax}} = \frac{b_i}{n-1}, \tag{2}$$

where $C_{RD}(i)$ is the relative DC of node *i*; $C_{AD}(i)$ is the absolute DC of node *i*; C_{ADmax} is the maximum possible absolute DC; b_i is the number of nodes which are directly connected to node *i*; *n* is the number of nodes in the network.

(2) Closeness Centrality (CC)

The CC measures to what degree the node *i* is not controlled by other nodes in the network [68]. Here, we introduce the concept of *distance*, which means the number of lines contained in the geodesics linking two nodes in the network. In the ICM collaboration network, a country or region node with a high CC value is at a short distance from all other countries and regions. This implies that it can easily both receive technological information from all countries and regions in the network and spread information to them. That is to say, its cooperation impact spreads faster in the network.

We define the CC as follows:

$$C_{RC}(i) = \frac{C_{AC}(i)}{C_{ACmax}} = \frac{\frac{1}{C_{AC}^{-1}(i)}}{\frac{1}{n-1}} = \frac{\frac{1}{\sum_{j=1}^{n} d_{ij}}}{\frac{1}{n-1}} = \frac{n-1}{\sum_{j=1}^{n} d_{ij}}, \tag{3}$$

where $C_{RC}(i)$ is the relative CC of node i; $C_{AC}(i)$ is the absolute CC of node i; C_{ACmax} is the maximum possible absolute CC; n is the number of nodes in the network; d_{ij} is the minimum distance between node i and node j.

(3) Betweenness Centrality (BC)

The BC measures the degree of node *i* in the shortest geodesics linking two other nodes in the network. That is to say, a high BC value indicates that node *i* plays the role of *intermediary* in the ICM technology cooperation network and can control the transmission of technological information in the network to a greater extent [69]. In short, the strength of the collaboration impact of *i* is greater.

We define the BC as follows:

$$C_{RB}(i) = \frac{C_{AB}(i)}{C_{ABmax}} = \frac{\sum_{j}^{n}\sum_{k}^{n}\frac{g_{jk}(i)}{g_{jk}}}{\frac{n^2-3n+2}{2}} = \frac{2\left[\sum_{j}^{n}\sum_{k}^{n}\frac{g_{jk}(i)}{g_{jk}}\right]}{n^2-3n+2},\ j \neq k \neq i,\ j < k, \tag{4}$$

where $C_{RB}(i)$ is the relative BC of node i; $C_{AB}(i)$ is the absolute BC of node i; C_{ABmax} is the maximum possible absolute BC; g_{jk} is the number of the shortest geodesics linking node j and node k; $g_{jk}(i)$ is the number of shortest geodesics linking node j and node k that contain node i; n is the number of nodes in the network.

(4) Analysis model of the collaboration impact characteristics and location characteristics

In this paper, three kinds of centrality of typical nodes in the ICM international cooperation network were calculated and sorted according to the formulas above. Based on the above description about DC, CC, and BC, we used the three kinds of centrality of a node to symbolize the three dimensions of the node's collaboration impact in the network: the range of direct collaboration impact, the speed of collaboration impact transmission and reception, and the strength of the collaboration impact. Afterwards, the average value of the rankings for three centralities (ARC) of one node was calculated and employed in this paper to analyze the collaboration impact of countries and regions in the collaboration network. To a certain extent, the ARC measure is a combination of one node's results for three kinds of centrality which can reflect the three dimensions of the collaboration impact. Therefore, we used the ARC value of one node to represent its comprehensive collaboration impact in the cooperative network.

We define the ARC as follows:

$$ARC(i) = \frac{R_{DC}(i) + R_{CC}(i) + R_{BC}(i)}{3} \tag{5}$$

where $ARC(i)$ is the ARC value of node i; $R_{DC}(i)$ is the ranking for the DC value of node i in all nodes within the network; $R_{CC}(i)$ is the ranking for CC value of node i; $R_{BC}(i)$ is the ranking for the BC value of node i.

In order to analyze the relative strengths and weaknesses of each country and region regarding the collaboration impact, the rankings for three kinds of centrality of each node were compared, separately, with the node's ranking of ARC. As we mentioned above, the three kinds of centrality of a node can be used to represent the three dimensions of the node's collaboration impact. Thus, the results of the comparisons between the rankings for three kinds of centrality and the ARC ranking of typical nodes can be used to analyze their corresponding collaboration impact characteristics and location characteristics in the network. The location characteristics' analysis model is shown in Table 2.

Table 2. Location Characteristics' Analysis Model.

	Low Degree Centrality (DC)	Low Closeness Centrality (CC)	Low Betweenness Centrality (BC)
High DC		The cooperative nodes of this kind of node are all in the same cooperative cluster, which is far away from the other nodes in the network.	The cooperative relation of this kind of node can be regarded as a redundant relation by its cooperative nodes. That is to say, there is usually a direct cooperation among the partners of the node.
High CC	A node with this ranking is the key node that has a direct cooperation with the core node.		There may be many ways of technological information flow in the network, which means that the node is close to a lot of nodes but other nodes are closer to others.
High BC	Few cooperative relationships of this kind of node are very important for the flow of technological information in the network.	This kind of node is relatively rare because it monopolizes the flow of information among some nodes and the other nodes.	

Note: Whether the value of each centrality is high or low is determined by the comparison results among the rankings of the three individual centralities of each node and its ARC ranking. If the ranking of the centrality for an individual node is higher than that of the sum of the three kinds of centrality, then it is considered high.

3. Results

3.1. Overall Status

There is a total of 45 countries and regions that have ICM international cooperation patents in the ICM industry from 2004 to 2013. The top 20 countries and regions with the most international cooperation patents are shown in Table 3. This paper defines them as typical countries and regions for international cooperation in the ICM industry and focuses on them in the following analysis. It can be seen that the USA is ranked first in the sorting. Its cooperation patent quantity is nearly four times that of Taiwan, which is ranked second; Taiwan is followed by Singapore and China. In terms of geographical distribution, the top 20 countries and regions are divided as follows; 10 in Asia, 7 in Europe, 2 in North America, and 1 (Australia) in Oceania. On the whole, Asian and European countries appeared to have more international technological cooperation output in the ICM industry.

Table 3. Top 20 High-Yield Countries and Regions in ICM International Collaboration Patents.

Country (Region)	Integrated Circuit Manufacturing (ICM) International Collaboration Patents	ICM Patents
	Rank (Number)	Rank (Number)
USA	1 (322)	1 (2951)
Taiwan	2 (73)	4 (651)
Singapore	3 (69)	8 (232)
China	4 (61)	3 (801)
Germany	5 (58)	6 (325)
Japan	6 (47)	2 (1207)
Korea	6 (47)	5 (513)
United Kingdom	8 (38)	9 (98)
India	9 (34)	13 (45)
France	10 (30)	7 (249)
The Netherlands	11 (26)	10 (89)

Table 3. *Cont.*

Country (Region)	Integrated Circuit Manufacturing (ICM) International Collaboration Patents	ICM Patents
	Rank (Number)	Rank (Number)
Malaysia	12 (22)	15 (43)
Belgium	13 (19)	16 (30)
Canada	13 (19)	11 (52)
Philippines	15 (13)	19 (19)
Austria	16 (12)	18 (23)
Hong Kong	17 (11)	19 (19)
Australia	18 (10)	12 (46)
Israel	19 (9)	16 (30)
Russia	20 (8)	14 (44)

3.2. Collaboration Pattern

Based on all the ICM patents, the Topic Association Map was generated by the association analysis method, as shown in Figure 3. The larger the circles, the more patents there are; the thicker the lines, the stronger the relevance among them. The different meaning of various formats of connection lines are indicated in the legend on the upper left corner of Figure 3. To make the map as clear as possible, only the typical countries and regions and the top 40 connections with the strongest correlations are shown. In order to study combine with the current situation of cooperation, the top three subfields with the most collaboration patents and the exact number for each country and region are displayed around the circle.

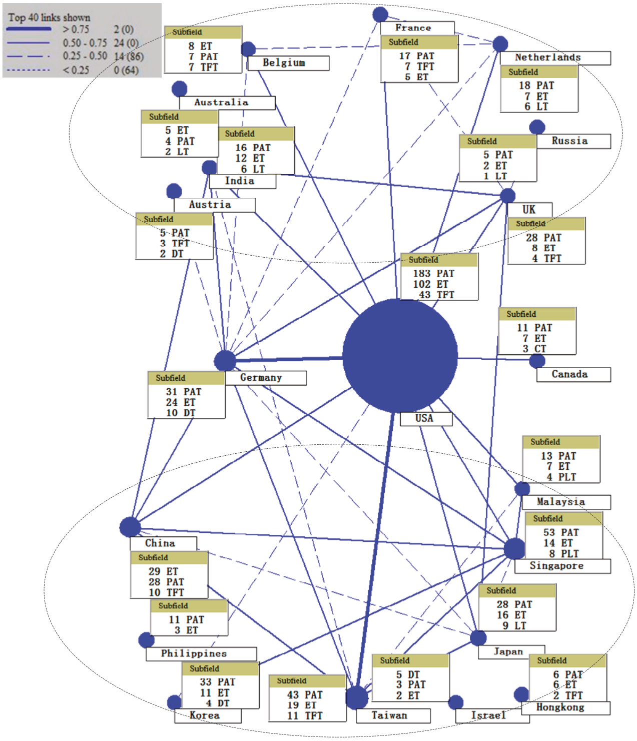

Top 40 links shown
> 0.75 2 (0)
0.50 - 0.75 24 (0)
0.25 - 0.50 14 (86)
< 0.25 0 (64)
Belgium
Subfield 8 ET 7 PAT 7 TFT
France
Subfield 17 PAT 7 TFT 5 ET
Netherlands
Subfield 18 PAT 7 ET 6 LT
Australia
Subfield 5 ET 4 PAT 2 LT
India
Subfield 16 PAT 12 ET 6 LT
Russia
Subfield 5 PAT 2 ET 1 LT
Austria
Subfield 5 PAT 3 TFT 2 DT
UK
Subfield 28 PAT 8 ET 4 TFT
USA
Subfield 183 PAT 102 ET 43 TFT
Canada
Subfield 11 PAT 7 ET 3 CT
Germany
Subfield 31 PAT 24 ET 10 DT
Malaysia
Subfield 13 PAT 7 ET 4 PLT
Singapore
Subfield 53 PAT 14 ET 8 PLT
China
Subfield 29 ET 28 PAT 10 TFT
Philippines
Subfield 11 PAT 3 ET
Korea
Subfield 33 PAT 11 ET 4 DT
Subfield 43 PAT 19 ET 11 TFT
Taiwan
Subfield 5 DT 3 PAT 2 ET
Japan
Subfield 28 PAT 16 ET 9 LT
Israel
Hongkong
Subfield 6 PAT 6 ET 2 TFT

Figure 3. Topic Association Map of the ICM international collaboration patents for typical countries and regions.

It was observed that the USA has strong associations with almost every one of the typical countries and regions. Among them, Germany and Taiwan are the most similar to the USA. It is easier for them to have a technological cooperation with the USA in the ICM industry. In addition, it is easy to see that the map can be divided into two clusters. From the perspective of the geographical position, the top circles are mostly European countries and the remaining circles are Asian countries, and, thus, they can be called the European cluster and the Asian cluster, respectively. As can be seen in Figure 3, the two clusters show strong clustering characteristics: the associations among the internal circles of each cluster are stronger; the associations among the internal circles and the external circles are relatively weak. Germany is a notable country outside the two clusters. Although it is located in Europe, it is very similar to the Asian cluster of countries and regions. Therefore, like the USA, Germany is also a country with diversified research topics in the ICM industry that can cooperate with a number of countries and regions easily.

From the point of view of cooperative subfields, almost all the typical countries and regions have more international cooperation patents about packaging technology and etching technology. This illustrates that the internationalization level of scientists in these two subfields is higher, producing more cooperative innovation achievements. Of course, this has also a certain relationship with the rapid development of technological innovation in these two subfields. Except for these two high-yield subfields, it appeared that the European cluster tends to cooperate in subfields of lithography technology and thin film technology, while the Asian cluster tends to cooperate in subfields of planarization technology and thin film technology.

3.3. Collaboration Network

In this section, we identified the clusters in the collaboration network by the international cooperation matrix derived from the ICM international collaboration patents based on the CONCOR algorithm; the analysis results are shown in Figure 4. It can be seen from the dotted line box in the figure that the countries and regions are divided into four subgroups at the second level, and are divided into two subgroups at the first level. From the first level, the countries and regions in the left cluster are mostly in Asia, while the countries in the right cluster are all in Europe, with the exception of Mexico.

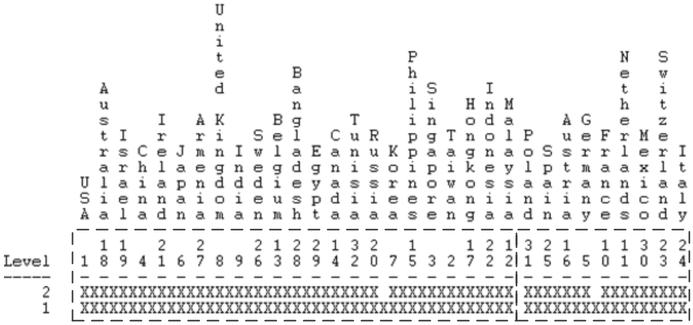


Figure 4. CONCOR Partition Diagram of ICM international cooperation.

We then created the ICM international cooperation network map by using the international cooperation matrix. In order to see the position of all countries and regions in the ICM cooperation network more clearly, Figure 5 was created with the Ucinet software using Gower Metric Scaling, with the main principle that linked nodes plotted adjacent to each other and non-linked nodes pushed apart [70]. As shown in Figure 5, the larger the nodes, the higher the number of international collaboration patents; the thicker the line, the more the patents that inventors from two different countries or regions co-invented. It can be seen that the cooperation network map can be divided into two subgroups, as shown by the dotted-line circle. The nodes in the upper circle are mostly located in Asia, while the nodes in the lower circle are mostly located in Europe. By comparing the outcomes of CONCOR clustering in Figure 4 and the cooperative network map in Figure 5, we found that, although two different methods were adopted, the results were quite consistent: Asia and Europe both showed an obvious characteristic of clustering in international technological cooperation.

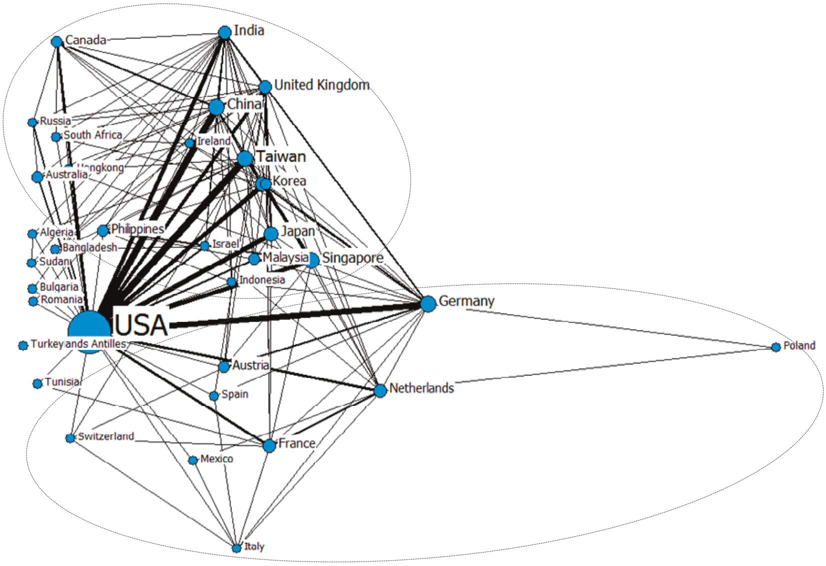


Figure 5. The ICM collaboration network of all countries and regions from 2004 to 2013 based on the Gower Metric Scaling layout.

From the point of view of the location of nodes in the network map, the USA occupies the core position in the network. The inventors from the USA have co-invented patents with inventors from 43 of the 44 countries and regions in the network. Also, almost all of the thickest connecting lines are related to the USA in the network map; relationships with China, Taiwan, Germany, and Japan are the closest. The technological cooperation of USA with China and Germany is mainly focused on the subfields of etching technology and packaging technology, whereas the cooperation with Japan and Taiwan is mainly focused on the packaging technical subfield. In addition, India, China, Taiwan, Korea, Singapore, and Germany all have cooperative links with multiple nodes in their respective cluster, therefore they are active nodes.

3.4. Collaboration Institution

The patent outputs of institutions are the basis for a country or region to own patents [54]. This section explains the current situation of the ICM international technological cooperation through an analysis at the institutional level. Table 4 shows the top 20 high-yield institutions in ICM international collaboration patents. Among them, the USA has the most of the top 20 high-yield institutions. Furthermore, it can be seen that 19 of them are enterprises, so enterprises are the major force driving the ICM international technological cooperation.

Subsequently, the type distribution of institutions owning ICM international collaboration patents in the top 10 countries and regions with the most international cooperation patents was analyzed, as shown in Figure 6. It can be seen that the distribution characteristics of the inner and outer circular rings of these countries and regions are basically consistent, apart from Korea, Singapore, and China. In Singapore and Korea, the average number of patents owned by research institutions is smaller than the average number of patents owned by companies, which means that the technological output of international cooperation for their research institutions is not as significant as that of their enterprises. Relatively speaking, the average number of patents owned by China's enterprises is smaller than the average number of patents owned by China's colleges and universities. Compared to enterprises, some universities in China have more international collaboration patents.

Table 4. Top 20 High-Yield Institutions in ICM International Collaboration Patents.

Rank	Institutions	Country (Region)	Type
1	IBM	USA	Enterprise
2	Taiwan Semiconductor Mfg.	Taiwan	Enterprise
3	Infineon Technologies AG	Germany	Enterprise
4	Stats Chippac Ltd.	Singapore	Enterprise
5	Texas Instruments Inc.	USA	Enterprise
6	Samsung Electronics Co. Ltd.	Korea	Enterprise
7	Chartered Semiconductor Mfg.	Singapore	Enterprise
8	Applied Materials Inc.	USA	Enterprise
9	Globalfoundries Sg Pte Ltd.	Singapore	Enterprise
10	NXP Bv	The Netherlands	Enterprise
11	Agere Systems Inc.	USA	Enterprise
12	D2S Inc.	USA	Enterprise
13	Koninkl Philips Electronics NV	The Netherlands	Enterprise
14	St. Microelectronics Sa	France	Enterprise
15	Univ Fudan	China	College and university
16	ASAT Ltd.	Hong Kong	Enterprise
17	Broadcom Corp.	USA	Enterprise
18	Intel Corp.	USA	Enterprise
19	Pulsic Ltd.	United Kingdom	Enterprise
20	Synopsys Inc.	USA	Enterprise

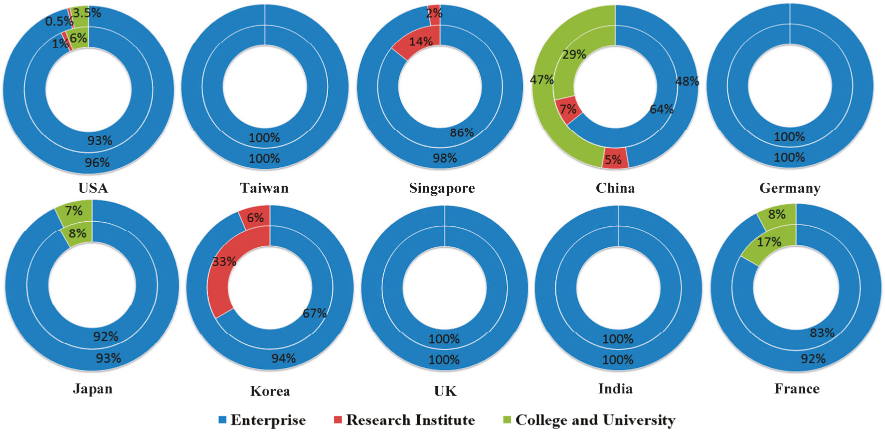


Figure 6. Type distribution of institutions that owned ICM international collaboration patents in the top 10 countries and regions. Note: The outer circular ring represents the quantity distribution of international cooperation patents owned by various types of institutions; the inner ring represents the quantity distribution of various types of institutions that have international cooperation patents.

The top 10 countries and regions can be divided into four groups according to Figure 6. The first group includes Taiwan, Germany, the United Kingdom, and India. Their international cooperation patents are all owned by enterprises, which means that enterprises dominate the ICM international technological cooperation in these countries and regions. The second group includes Singapore and Korea. Most of their patents belong to enterprises, and the rest belong to research institutions. The third group includes the USA, Japan, and France. The majority of the international cooperation patents in these countries are owned by enterprises, but colleges and universities have also achieved some results. The last group includes only China, where China's enterprises, colleges, and universities are the dominant force in international technological cooperation. Beyond that, research institutions also share part of the international cooperation output.

3.5. Collaboration Impact

According to the formula discussed in Section 2.4.2, the three kinds of centrality and ARC values of the typical countries and regions in the ICM international cooperation network were calculated separately, and their rankings were also given, as shown in Table 5. It can be seen that the ARC of the USA ranked first for all countries, which also confirms the core position of the USA in the network. In terms of geographical distribution, the ARC-rank for the top 20 countries and regions includes 10 countries or regions in Asia, 8 in Europe, and 2 in North America. Overall, the countries and regions in Asia and Europe have not only many technological cooperation outputs, but also a great impact on the international cooperation network. Therefore, Asia and Europe are very successful in ICM international technological cooperation.

Table 5. The Centralities and the ARC of the Typical Countries and Regions in the ICM International Collaboration Network.

Country (Region)	Degree Centrality (DC)	Closeness Centrality (CC)	Betweenness Centrality (BC)	Average Value of the Rankings for Three Centralities (ARC)
	Rank (Value)	Rank (Value)	Rank (Value)	Rank (Value)
USA	1 (97.73)	1 (97.78)	1 (64.70)	1 (1)
Taiwan	6 (34.09)	6 (60.27)	7 (1.90)	6 (6.33)
Singapore	7 (31.82)	7 (59.46)	8 (1.11)	7 (7.33)
China	4 (38.64)	4 (61.97)	5 (3.14)	5 (4.33)
Germany	4 (38.64)	4 (61.97)	3 (4.48)	4 (3.67)
Japan	10 (22.73)	10 (56.41)	13 (0.33)	11 (11)
Korea	13 (20.46)	12 (55.70)	18 (0.11)	14 (14.33)
United Kingdom	2 (45.46)	2 (64.71)	4 (4.00)	3 (2.67)
India	2 (45.46)	2 (64.71)	2 (4.68)	2 (2)
France	10 (22.73)	10 (56.41)	9 (0.83)	10 (9.67)
The Netherlands	8 (27.27)	8 (57.90)	6 (2.61)	7 (7.33)
Malaysia	14 (18.18)	14 (55.00)	19 (0.10)	16 (15.67)
Belgium	9 (25.00)	9 (57.14)	10 (0.43)	9 (9.33)
Canada	10 (22.73)	12 (55.70)	11 (0.39)	11 (11)
Philippines	16 (15.91)	18 (53.66)	21 (0.07)	18 (18.33)
Austria	19 (13.64)	18 (53.66)	15 (0.17)	17 (17.33)
Hong Kong	19 (13.64)	21 (53.01)	23 (0.05)	22 (21)
Australia	24 (11.36)	24 (52.38)	24 (0.02)	24 (24)
Israel	19 (13.64)	18 (53.66)	20 (0.09)	20 (19)
Russia	19 (13.64)	21 (53.01)	22 (0.05)	21 (20.67)
Average	- (28.64)	- (59.22)	- (4.46)	- (-)

Note: In order to analyze the collaboration impacts of the typical countries and regions in the whole ICM international cooperation network more accurately, all the rankings in this table are in the scope of all 45 countries and regions in the network.

Next, the rankings of each node for the three centralities were compared to the node's ARC ranking. If the ranking of a centrality was higher than the ARC ranking, we used H to denote the centrality of the node (represented in Table 6 by a solid circle); in contrast, L was used to denote a centrality lower than the ARC ranking (represented by an open circle). Thus, according to the different characteristics of the collaboration impact in three dimensions, the countries and regions could be divided into eight categories. Based on the international cooperation network and location characteristics analysis model, three of these categories were analyzed in detail in the following subsections.

(1) HHH

This category of node has more direct cooperation objects and a close indirect relationship with most of the other nodes in the network. In addition, because of the scarce relationship with some conservative collaborative nodes, HHH nodes can control the transmission of technology information to a certain extent. As seen in Table 6, the USA, India, and Germany are categorized as this type of node; they are ranked in the forefront of the three centralities and ARC. In the network, the USA is the core node; India and Germany are active nodes in the cluster of Asian and European clusters and have cooperative relationships with many nodes outside the cluster. Moreover, some of their connections have monopolized the transmission of technological information among the "conservative" nodes and others; this regards, for example, the connections between the USA and Egypt, the USA and Norway, India and Bulgaria, India and Sultan, Germany and Turkey, Germany and Poland, etc. The common features of an HHH node include a wide range of direct impact, faster impact spread speed, and greater impact strength.

(2) HHL

According to Table 2, there is usually a direct cooperation among the partners of an HHL node, which makes the cooperative relation usually regarded as redundant by its connected nodes. In addition, an HHL node is close to a lot of nodes, but other nodes are close to others. Taiwan, Singapore, China, Japan, Korea, the United Kingdom, Malaysia, Hong Kong, and Israel belong to the HHL category. They are both active nodes in the Asian cluster according to the analysis of the collaboration network section, and have direct connections with many nodes in the Asian cluster (because of the characteristics of the cluster cooperation, there is usually a connection among its connected nodes); they also have a direct cooperation with active nodes of the European cluster, so they are very close to most of the nodes in the network, but cannot control the transmission of technological information effectively. Among these countries and regions, the characteristics of China are especially evident. China connects most of the nodes in the Asian cluster. In addition, it also connects with the USA and Germany. Overall, an HHL node can be called a cross-cluster node. This category of nodes has a large direct impact range and a faster information propagation speed, but because its cooperation relationship is regarded as redundant, the impact strength is weak.

(3) HLL

On the basis of the analysis model mentioned in Table 2, an HLL node is directly connected to multiple nodes that are all in the same cooperative cluster, and the cluster is far away from the other nodes in the network. This leads to the average distance among it and the other nodes in the network as being very large. In addition, because of the characteristics of the cluster cooperation, its cooperative relations are also regarded as redundant by its connected nodes. An HLL node can be termed a within-cluster node. The Philippines, Russia, and Canada belong to the within-cluster node category, and the characteristics of the Philippines are particularly prominent. Although the Philippines has a lot of connected nodes, all these connected nodes are located in the Asia cluster except for the node representing the USA, which makes the Philippines far from the European cluster nodes. In general, a within-cluster node has a wide range of direct impact, but has a low information propagation speed and a weak impact strength.

Table 6. Comparison of Rankings among the Three Centralities and the ARC of the Typical Countries and Regions.

Country (Region)	DC vs. ARC	CC vs. ARC	BC vs. ARC	Category
USA	●*	●*	●*	HHH
Taiwan	●*	●*	○	HHL
Singapore	●*	●*	○	HHL
China	●	●	○ *	HHL
Germany	●*	●*	●	HHH
Japan	●	●	○	HHL
Korea	●	●	○	HHL
United Kingdom	●	●	○	HHL
India	●*	●*	●*	HHH
France	○ *	○ *	●	LLH
The Netherlands	○	○	●	LLH
Malaysia	●	●	○	HHL
Belgium	○ *	○ *	○	LLL
Canada	●	○	○ *	HLL
Philippines	●	○ *	○	HLL
Austria	○	○ *	●	LLH
Hong Kong	●	●	○	HHL
Australia	○ *	○ *	○ *	LLL
Israel	●	●	○ *	HHL
Russia	●	○ *	○	HLL

Note: * represents a situation where the centrality ranking of a node is consistent with its ARC ranking, then the concrete value of this centrality is compared to the average value of the centrality among the typical countries and regions. If the concrete value is higher, then it is a solid circle; conversely, it is an open circle.

4. Conclusions and Discussion

(1) In regard to international technological collaboration in the ICM industry, the USA has the highest level in the world; Germany has great potential for future development.

First, the USA has the largest number of patents, international collaboration patents, and top 20 high-yield institutions for international collaboration in the ICM industry worldwide. It is also located at the core position of the topic association map and collaboration network. In addition, the USA has the greatest comprehensive impact and possesses the characteristics of a wide range of direct impact, fast impact spread speed, and great impact strength. Thus, it can be said that the USA has the highest level in the world.

Second, Germany is the country most similar to the USA in the distribution of research topics, and also has similarities with numerous countries and regions in the research topics. In addition, Germany is a HHH node. In the future, Germany should not only strengthen the technological collaboration with the USA to further enhance its technological level by taking advantage of the similarities with the USA in research themes, but also seek technological collaborations with many countries and regions that are similar to it in the technology direction, so as to utilize many external technological knowledge resources on the basis of a wide range of direct impact, fast impact spread speed, and great impact strength. Therefore, Germany has great potential for future development.

(2) There are obvious clustering patterns in Asia and Europe, respectively, in the collaboration network. The nodes within these two clusters should further play the role of the innovation cluster.

The countries and regions in Asia and Europe show, separately, an obvious trend of clustering in the international cooperation network map. Asia and Europe are also successful in the ICM international technical cooperation. Furthermore, on the basis of the analysis of the collaboration pattern, the topic association map can also be divided into two clusters in Asia and Europe. The consistency of the research direction could promote the technical cooperation among countries and regions within a cluster and accelerate the process of agglomeration in the international cooperation network [20].

Therefore, the node within the Asian or European cluster should take advantage of the convergence in research themes inside the collaboration cluster to further play the role of the innovation cluster [71]. For example, in addition to the two high-yield subfields of packaging technology and etching technology, European nodes should focus on the further development of collaboration in the subfields of lithography technology and thin film technology, while Asian nodes should focus on the collaboration in the subfields of planarization technology and thin film technology.

(3) At the same time that the enterprises operate as a major force, research institutions, colleges, and universities of each country and region should also actively participate in international collaboration to improve the level of ICM international technological collaboration.

Although the enterprises have always been the dominant power in promoting the international collaboration in the ICM industry, the types of institutions that owned international collaboration patents in typical countries and regions are different. This means that there are differences in the type of institutions that are actually engaged in international collaboration among countries and regions [72]. For example, only enterprises that participated in international collaboration in Taiwan, Germany, the United Kingdom, and India, while universities in the USA, Japan, and France, and research institutions in Korea and Singapore also carried out some international collaboration activities.

In general, China has the most complete structure for its institution type. China's enterprises, research institutions, colleges, and universities have all participated in international collaborations. Some studies found that the industry–academy collaboration can not only promote the increase in the number of collaboration patents [73,74], but it can also improve the patent value and quality [75]. Although China's international collaboration output in the ICM industry is still lagging behind that of the USA, we think the gap is most likely due to the difference in ICM patent numbers between the USA and China. This is because if a country has more patents, other countries will be more willing to cooperate with it [54], and the number of its own patents also affects its international collaboration output positively in each collaborative relationship [21]. Therefore, we recommend that the research institutions, colleges, and universities of each country and region (especially for Taiwan, Germany, the United Kingdom, and India) should participate in international collaboration actively, while the enterprises operate as the major force. And the industry–academy collaboration will play a facilitating role in improving the level of ICM international technological collaboration.

Despite our careful analysis and resulting outcomes, this work still suffers from several limitations and weaknesses. First, this paper analyzes ICM international technical cooperation based on the international co-inventor of patents, but the inventor cooperative information in patents cannot reflect the international technological collaboration accurately. Above all, the patent has limitations when it is used as a measure of innovation. This is because in the ICM industry, where the pace of technology is rapid, and firms advance quickly utilizing the innovations made by others, firms may patent for strategic reasons [16,36]. For example, firms in this industry are likely to build larger portfolios for their own "legal rights to exclude", to reduce the holdup problem posed by external patent owners and enable firms to negotiate access to external technologies on more favorable terms [76]. In general, the innovation level of patent rights may vary widely among firms over time, even within one industry. Therefore, patents cannot measure innovation effectively in the case that the patent quality is not considered. Furthermore, the inventor cooperative information in patents can also only reflect part of the international technical cooperation. For example, technology transformation, expert visits, R&D organization co-building, academic exchange meetings, etc., also contribute to international technical cooperation.

Second, based on the concrete calculation method of the three kinds of centrality, we regarded the centralities as three separate dimensions of the collaboration impact: direct impact range, impact propagation speed, and impact strength; however, this may not be accurate. This is because a high CC does not necessarily indicate that the country or region can receive and send the technical information in the network more quickly. With the acceleration of globalization, there is a variety of ways to exchange technical information. In addition, a high BC is usually due to the cooperation with conservative collaborative nodes, but the actual correlation among this cooperation and the country or region's impact strength is not as strong as often supposed. Moreover, the three kinds of centrality do not take into account the weight of the collaboration relationships—they only measure the two-valued collaboration matrices. Therefore, they cannot fully reflect the collaboration impact characteristics of nodes in the three dimensions. As Freeman [68] said, it remains to be seen how well each of them will stand up in the light of further empirical work in this area.

Third, the classification of the collaboration impact of a typical node was derived from the comparative analysis of its rankings for three kinds of centrality based on its ARC ranking. In a sense, this is a comparison between a typical node and itself. Because the nodes both had similar rankings for the three kinds of centralities (as shown in Table 5), this comparison could reasonably categorize the country or region more effectively compared to other methods (e.g., the comparison based on the average value of centrality among the typical countries and regions). Although this may cause some countries which have a similar value for three kinds of centralities to be categorized into two different groups (e.g., the United Kingdom is HHL and India is HHH), we could see the relative strengths and weaknesses of each country and region more clearly in the three dimensions of the collaboration impact through this analysis, and thus draw the collaboration impact characteristics more accurately.

There are various avenues for future research. An important research question to be answered is how to find more accurate international technical cooperation information. How do we take patent quality into account, and how do we analyze other types of technological collaboration information? We believe that big data analytics can be used for collecting, transforming, and analyzing different types of raw data about international technological collaboration. For instance, more technological information could be available from patent data by big data analytics [77]. This is our key research direction for the future. In addition, it would be worthwhile to investigate how to improve the measurement method of the three kinds of centrality. How do we make the characteristics reflected by each centrality more obvious, and how do we highlight the weight of the collaboration relationship? Of course, the most important thing is how to combine this research with the actual situation, find the reasons behind the characteristics, and further play the role of knowledge management on the basis of patent analysis.

In summary, this paper studies the international technological collaboration characteristics of the ICM industry from four aspects: collaboration pattern, collaboration network, collaboration institution, and collaboration impact. Through research on these four aspects, we attempted to understand the characteristics from three points of view: development potential, current situation, and consequence. Based on the analysis from these three angles, this paper reveals the advantages and disadvantages of countries and regions for international technological collaboration in the ICM industry, puts forward some suggestions, and provides an objective reference for policy making, competitiveness, and sustainability. Specifically, through this paper, the policymakers of all countries and regions and the decision makers of different institutions can better understand the technological collaboration characteristics of the global ICM industry; they can recognize and implement their own advantages; they can identify technology partners who have a strong technical force and collaboration potential; they can refer to the specific recommendations given. Based on the above measures, they will further enhance their international technological collaborations. This will enable a country, a region, or an institution to enhance its technological level, innovation efficiency, and competencies to sustainably develop in the global ICM industry according to the elaboration in Section 1. It is also worth noting that, so far, no scholars have studied the international technological collaboration based on these three perspectives. The research framework of this paper provides a new research approach for the future research of industrial international collaboration based on patent analysis. It could also be applied to analyze the international technological collaboration of other industries and to provide assistance for the purpose of sustainable development in other industries.

Acknowledgments: The paper was funded by the National Natural Science Foundation of China (No. 71573017, No. 71603019 and No. 71273030). We are thankful for the suggestions and efforts from the experts and editors.

Author Contributions: The manuscript was approved by all authors for publication. Yun Liu, Zhe Yan, and Xuanting Ye conceived and designed the study; Zhe Yan and Xuanting Ye collected the data; Yun Liu, Zhe Yan, and Xuanting Ye analyzed the data; Yun Liu and Zhe Yan wrote the paper. Zhe Yan, Xuanting Ye, and Yijie Cheng reviewed and edited the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Retrieval Strategy for ICM Patents.

Technical Subfields	Patent Retrieval Strategy
Cleaning Technology	TS = ((IC OR "integrated circuit*" OR semiconductor) AND (cleaning* OR "plasma cleaning" OR "plasma-cleaned" OR (megasonic* AND cleaning) OR "laser cleaning" OR "laser-clean*"))
Lithography Technology	TS = ((IC OR "integrated circuit*" OR semiconductor) AND ("lithograph*" or (X-ray AND lithograph*) OR XRL OR ((FIB OR "focused ion beam" OR "focused-ion-beam") AND (Lithograph* OR FIBL)) OR ((EUV OR "extreme ultraviolet lithography" OR "extreme-ultraviolet-lithography") AND (lithograph* OR EUVL)) OR NIL OR "nanoimprint lithography" OR "nanoimprint-lithography" OR EPL OR "electron projection lithograph*" OR "electron-projection-lithograph*"))
Etching Technology	TS = ((IC OR "integrated circuit*" OR semiconductor) AND ("etch*" OR "wet chemical etching" OR "plasma etching" OR PE OR "reactive ion etching" OR RIE OR "inductively coupled plasma etching" OR ICP))
Thin Film Technology	TS = ((IC OR "integrated circuit*" OR semiconductor) AND ("thermal-oxidated" OR "thermal oxidation" OR "physical vapor deposition" OR "physical-vapor-deposition" OR PVD OR "chemical vapor deposition" OR "chemical-vapor-deposition" OR CVD OR electroplate OR "epitax*" OR "vapor phase epitax*" OR "vapor-phase-epitax*" OR VPE OR "molecular beam epitax*" OR "molecular-beam-epitax*" OR MBE OR "complementary metal oxide semiconductor"))
Doping Technology	TS = ((IC OR "integrated circuit*" OR semiconductor) AND ("ion-implant*" OR "ion implant*" OR diffusion))
Annealing Technology	TS = ((IC OR "integrated circuit*" OR semiconductor) AND ("anneal*" OR "rapid thermal anneal*" OR "rapid-thermal-anneal*" OR RTA OR "laser anneal*" OR "laser-anneal*"))
Planarization Technology	TS = ((IC OR "integrated circuit*" OR semiconductor) AND (polish* OR "chemical mechanical polish*" OR "chemical-mechanical-polish*" OR CMP OR "stress-free polish*" OR "stress free polish*" OR SFP))
Packaging Technology	TS = ((IC OR "integrated circuit*" OR semiconductor) AND ("interconnect*" OR "Cu interconnect*" OR "copper interconnect*" OR "low k dielectric*" OR "optical interconnect*" OR "carbon nanotube*" OR CNT OR "through silicon via" OR TSV OR "packag*" OR "ball grid array" OR "ball-grid-array" OR BGA OR "chip size package" OR "chip-size package" OR CSP OR "multi chip package" OR "multi-chip package" OR MCP OR "wafer level package" OR "wafer-level package" OR WLP OR "flip chip packag*" OR "flip-chip packag*" OR "product in packag*" OR "product-in packag*" OR PIP OR "system on packag*" OR "system in packag*" OR SIP OR SOP OR "3D packag*" OR "pin gird array package*" OR "pin-gird-array packag*" OR PGA))

References

1. Wu, X.B.; Dou, W.; Wang, Y.Q. China's ICT Industry: Catch-Up Trends, Challenges and Policy Implications. *China Int. J.* **2013**, *11*, 117–139.
2. Lee, T.L.; Tunzelmann, N. A dynamic analytic approach to national innovation systems: The IC industry in Taiwan. *Res. Policy* **2005**, *34*, 425–440. [CrossRef]
3. Fuller, D.B. China's national system of innovation and uneven technological trajectory The case of China's integrated circuit design industry. *Chin. Manag. Stud.* **2009**, *3*, 58–74. [CrossRef]
4. Zhong, R.Y.; Huang, G.Q.; Lan, S.; Dai, Q.; Chen, X.; Zhang, T. A big data approach for logistics trajectory discovery from RFID-enabled production data. *Int. J. Prod. Econ.* **2015**, *165*, 260–272. [CrossRef]
5. Hashem, I.A.T.; Chang, V.; Anuar, N.B.; Adewole, K.; Yaqoob, I.; Gani, A.; Ahmed, E.; Chiroma, H. The role of big data in smart city. *Int. J. Inf. Manag.* **2016**, *36*, 748–758. [CrossRef]

6. Wang, G.; Gunasekaran, A.; Ngai, E.W.; Papadopoulos, T. Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *Int. J. Prod. Econ.* **2016**, *176*, 98–110. [CrossRef]
7. Cheyre, C.; Kowalski, J.; Veloso, F.M. Spinoffs and the ascension of Silicon Valley. *Ind. Corp. Chang.* **2015**, *24*, 837–858. [CrossRef]
8. Group, E.L. *A European Industrial Strategic Roadmap for Micro- and Nano-Electronic Components and Systems: Implementation Plan*; European Commission: Brussels, Belgium, 2014.
9. Rasiah, R.; Kong, X.X.; Lin, Y. Innovation and learning in the integrated circuits industry in Taiwan and China. *J. Asia Pac. Econ.* **2010**, *15*, 225–246. [CrossRef]
10. China's State Council. *Outline of Development of National Integrated Circuit Industry*; China's State Council: Beijing, China, 2014.
11. Kong, X.X.; Zhang, M.; Ramu, S.C. China's semiconductor industry in global value chains. *Asia Pac. Bus. Rev.* **2016**, *22*, 150–164. [CrossRef]
12. Levitas, E.F.; McFadyen, M.A.; Loree, D. Survival and the introduction of new technology: A patent analysis in the integrated circuit industry. *J. Eng. Technol. Manag.* **2006**, *23*, 182–201. [CrossRef]
13. Tsai, B.H.; Li, Y.M. Cluster evolution of IC industry from Taiwan to China. *Technol. Forecast. Soc.* **2009**, *76*, 1092–1104. [CrossRef]
14. Corsino, M.; Passarelli, M. The competitive advantage of business units: Evidence from the integrated circuit industry. *Eur. Manag. Rev.* **2009**, *6*, 182–194. [CrossRef]
15. Chen, L.; Xue, L. Global Production Network and the Upgrading of China's Integrated Circuit Industry. *China World Econ.* **2010**, *18*, 109–126. [CrossRef]
16. Hall, B.H.; Ziedonis, R.H. The patent paradox revisited: An empirical study of patenting in the US semiconductor industry, 1979–1995. *RAND J. Econ.* **2001**, *32*, 101–128. [CrossRef]
17. Kapoor, R.; Mcgrath, P.J. Unmasking the interplay between technology evolution and R&D collaboration: Evidence from the global semiconductor manufacturing industry, 1990–2010. *Res. Policy* **2014**, *43*, 555–569.
18. Liu, X.P.; Xu, Q.; Ning, F.; Wang, H. A summary of the current development of developing technology in the field of integrated circuit manufacturing. *ACSR Adv. Comput.* **2015**, *15*, 1333–1340.
19. Ernst, D. Global production networks and the changing geography of innovation systems. Implications for developing countries. *Econ. Innov. New Technol.* **2002**, *11*, 497–523. [CrossRef]
20. Wang, X.F.; Ren, J.; Zhang, Y.; Zhu, D.H.; Qiu, P.J.; Huang, M. China's patterns of international technological collaboration 1976–2010: A patent analysis study. *Technol. Anal. Strateg.* **2014**, *26*, 531–546. [CrossRef]
21. Prato, G.D.; Nepelski, D. Global technological collaboration network: Network analysis of international co-inventions. *J. Technol. Transf.* **2014**, *39*, 358–375. [CrossRef]
22. Hottenrott, H.; Lopesbento, C. (International) R&D collaboration and SMEs: The effectiveness of targeted public R&D support schemes. *Res. Policy* **2014**, *43*, 1055–1066.
23. Chang, Y.C. Benefits of co-operation on innovative performance: Evidence from integrated circuits and biotechnology firms in the UK and Taiwan. *R&D Manag.* **2003**, *33*, 425–437.
24. Guan, J.C.; Chen, Z.F. Patent collaboration and international knowledge flow. *Inf. Process. Manag.* **2012**, *48*, 170–181. [CrossRef]
25. Zheng, J.; Zhao, Z.Y.; Zhang, X.; Chen, D.Z.; Huang, M.H. International collaboration development in nanotechnology: A perspective of patent network analysis. *Scientometrics* **2014**, *98*, 683–702. [CrossRef]
26. Singh, J. Distributed R&D, cross-regional knowledge integration and quality of innovative output. *Res. Policy* **2008**, *37*, 77–96.
27. Quist, J.; Tukker, A. Knowledge collaboration and learning for sustainable innovation and consumption: Introduction to the ERSCP portion of this special volume. *J. Clean. Prod.* **2013**, *48*, 167–175. [CrossRef]
28. Chen, H.-W. Gallium, indium, and arsenic pollution of groundwater from a semiconductor manufacturing area of Taiwan. *Bull. Environ. Contam. Toxicol.* **2006**, *77*, 289–296. [CrossRef] [PubMed]
29. Lin, A.Y.-C.; Panchangam, S.C.; Lo, C.-C. The impact of semiconductor, electronics and optoelectronic industries on downstream perfluorinated chemical contamination in Taiwanese rivers. *Environ. Pollut.* **2009**, *157*, 1365–1372. [CrossRef] [PubMed]
30. Narin, F.; Hamilton, K.S. Bibliometric performance measures. *Scientometrics* **1996**, *36*, 293–310. [CrossRef]
31. Meyer, M.; Persson, O. Nanotechnology—Interdisciplinarity, patterns of collaboration and differences in application. *Scientometrics* **1998**, *42*, 195–205. [CrossRef]

32. Chen, N.; Liu, Y.; Cheng, Y.J.; Liu, L.; Yan, Z.; Tao, L.X.; Guo, X.H.; Luo, Y.X.; Yan, A.S. Technology Resource, Distribution, and Development Characteristics of Global Influenza Virus Vaccine: A Patent Bibliometric Analysis. *PLoS ONE* **2015**, *10*, e0136953. [CrossRef] [PubMed]
33. Albert, M.B.; Avery, D.; Narin, F.; Mcallister, P. Direct validation of citation counts as indicators of industrially important patents. *Res. Policy* **1991**, *20*, 251–259. [CrossRef]
34. Podolny, J.M.; Stuart, T.E. A Role-Based Ecology of Technological Change. *Am. J. Sociol.* **1995**, *100*, 1224–1260. [CrossRef]
35. Bapuji, H.; Loree, D.; Crossan, M. Connecting external knowledge usage and firm performance: An empirical analysis. *J. Eng. Technol. Manag.* **2011**, *28*, 215–231. [CrossRef]
36. Ziedonis, R.H. Don't fence me in: Fragmented markets for technology and the patent acquisition strategies of firms. *Manag. Sci.* **2004**, *50*, 804–820. [CrossRef]
37. Tsai, B.H. Does Litigation over the Infringement of Intellectual Property Rights Hinder Enterprise Innovation? An Empirical Analysis of the Taiwan IC Industry. *Issues Stud.* **2010**, *46*, 173–203.
38. Podolny, J.M.; Hannan, M.T. Networks, Knowledge, and Niches: Competition in the Worldwide Semiconductor Industry, 1984–1991. *Am. J. Sociol.* **1996**, *102*, 659–689. [CrossRef]
39. Chen, J.H.; Jang, S.L.; Wen, S.H. Measuring technological diversification: Identifying the effects of patent scale and patent scope. *Scientometrics* **2010**, *84*, 265–275. [CrossRef]
40. Tsai, B.H. Innovation spillover effect in semiconductor industry. In Proceedings of the Technology Management for Global Economic Growth (PICMET), Phuket, Thailand, 18–22 July 2010; pp. 1–4.
41. Hu, M.C. Evolution of knowledge creation and diffusion: The revisit of Taiwan's Hsinchu Science Park. *Scientometrics* **2011**, *88*, 949–977. [CrossRef]
42. Fuller, D.B. Chip design in China and India: Multinationals, industry structure and development outcomes in the integrated circuit industry. *Technol. Forecast. Soc.* **2014**, *81*, 1–10. [CrossRef]
43. Bergek, A.; Bruzelius, M. Are patents with multiple inventors from different countries a good indicator of international R&D collaboration? The case of ABB. *Res. Policy* **2010**, *39*, 1321–1334.
44. Liu, F.; Zhang, N.; Cao, C. An evolutionary process of global nanotechnology collaboration: A social network analysis of patents at USPTO. *Scientometrics* **2017**, *111*, 1449–1465. [CrossRef]
45. Su, H.-N. Global Interdependence of Collaborative R&D-Typology and Association of International Co-Patenting. *Sustainability* **2017**, *9*, 541.
46. Tsukada, N.; Nagaoka, S. Determinants of International Research Collaboration: Evidence from International Co-Inventions in Asia and Major OECD Countries. *Asian Econ. Policy Rev.* **2015**, *10*, 96–119. [CrossRef]
47. Ma, Z.Z.; Lee, Y.; Chen, C.F.P. Booming or emerging? China's technological capability and international collaboration in patent activities. *Technol. Forecast. Soc.* **2009**, *76*, 787–796. [CrossRef]
48. Zheng, J.; Zhao, Z.Y.; Zhang, X.; Chen, D.Z.; Huang, M.H.; Lei, X.P.; Zhang, Z.Y.; Zhao, Y.H. International scientific and technological collaboration of China from 2004 to 2008: A perspective from paper and patent analysis. *Scientometrics* **2012**, *91*, 65–80. [CrossRef]
49. Semiconductor Industry Association. *Global Sales Report*; Semiconductor Industry Association: Washington, DC, USA, 2015.
50. IC Insights. *Global Wafer Capacity 2016–2020*; IC Insights: Scottsdale, AZ, USA, 2015.
51. Van Zant, P.; Chapman, P. *Microchip Fabrication: A Practical Guide to Semiconductor Processing*; McGraw-Hill: New York, NY, USA, 2000; Volume 5.
52. Arora, S.K.; Porter, A.L.; Youtie, J.; Shapira, P. Capturing new developments in an emerging technology: An updated search strategy for identifying nanotechnology research outputs. *Scientometrics* **2013**, *95*, 351–370. [CrossRef]
53. Huang, Y.; Schuehle, J.; Porter, A.L.; Youtie, J. A systematic method to create search strategies for emerging technologies based on the Web of Science: Illustrated for Big Data. *Scientometrics* **2015**, *105*, 2005–2022. [CrossRef]
54. Guan, J.C.; Zhang, J.J.; Yan, Y. The impact of multilevel networks on innovation. *Res. Policy* **2015**, *44*, 545–559. [CrossRef]
55. Ma, Z.; Lee, Y. Patent application and technological collaboration in inventive activities: 1980–2005. *Technovation* **2008**, *28*, 379–390. [CrossRef]
56. Zhu, D.H.; Porter, A.L. Automated extraction and visualization of information for technological intelligence and forecasting. *Technol. Forecast. Soc.* **2002**, *69*, 495–506. [CrossRef]

57. Wasserman, S.; Faust, K. *Social Network Analysis: Methods and Applications*; Cambridge University Press: Cambridge, UK, 1994; Volume 8.
58. Steiber, S.R. Building Better Blockmodels: A Non-Hierarchical Extension of CONCOR With Applications to Regression Analysis. *Mid-Am. Rev. Sociol.* **1981**, *6*, 17–40. [CrossRef]
59. Munene, E.; Mottice, S.; Reid, J. Evaluating a Social Network Analytic Tool to Support Outbreak Management and Contact Tracing in an Outbreak of Pertussis. *Online J. Public Health Inform.* **2013**, *5*, e72. [CrossRef]
60. Radil, S.M.; Flint, C.; Tita, G.E. Spatializing social networks: Using social network analysis to investigate geographies of gang rivalry, territoriality, and violence in Los Angeles. *Ann. Assoc. Am. Geogr.* **2010**, *100*, 307–326. [CrossRef]
61. Breiger, R.L.; Boorman, S.A.; Arabie, P. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *J. Math. Psychol.* **1975**, *12*, 328–383. [CrossRef]
62. Ju, Y.; Sohn, S.Y. Patent-based QFD framework development for identification of emerging technologies and related business models: A case of robot technology in Korea. *Technol. Forecast. Soc. Chang.* **2015**, *94*, 44–64. [CrossRef]
63. Zhang, J.; Zhai, S.; Liu, H.; Stevenson, J.A. Social network analysis on a topic-based navigation guidance system in a public health portal. *J. Assoc. Inf. Sci. Technol.* **2016**, *67*, 1068–1088. [CrossRef]
64. Singh, J. Collaborative networks as determinants of knowledge diffusion patterns. *Manag. Sci.* **2005**, *51*, 756–770. [CrossRef]
65. Agrawal, A.; Kapur, D.; McHale, J. How do spatial and social proximity influence knowledge flows? Evidence from patent data. *J. Urban Econ.* **2008**, *64*, 258–269. [CrossRef]
66. Wang, X.F.; Li, R.R.; Ren, S.M.; Zhu, D.H.; Huang, M.; Qiu, P.J. Collaboration network and pattern analysis: Case study of dye-sensitized solar cells. *Scientometrics* **2014**, *98*, 1745–1762. [CrossRef]
67. Schilling, M.A.; Phelps, C.C. Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Manag. Sci.* **2007**, *53*, 1113–1126. [CrossRef]
68. Freeman, L.C. Centrality in social networks conceptual clarification. *Soc. Netw.* **1978**, *1*, 215–239. [CrossRef]
69. Schiffauerova, A.; Beaudry, C. Collaboration spaces in Canadian biotechnology: A search for gatekeepers. *J. Eng. Technol. Manag.* **2012**, *29*, 281–306. [CrossRef]
70. Beyers, J.; Donas, T. Inter-regional networks in Brussels: Analyzing the information exchanges among regional offices. *Eur. Union Politics* **2014**, *15*, 547–571. [CrossRef]
71. Engel, J.S.; Del-Palacio, I. Global Clusters of Innovation: The Case of Israel and Silicon Valley. *Calif. Manag. Rev.* **2011**, *53*, 27–49. [CrossRef]
72. Ma, J.; Wang, X.; Zhu, D.; Zhou, X. Analysis on patent collaborative patterns for emerging technologies: A case study of nano-enabled drug delivery. *Int. J. Technol. Manag.* **2015**, *69*, 210–228. [CrossRef]
73. George, G.; Zahra, S.A.; Wood, D.R. The effects of business–university alliances on innovative output and financial performance: A study of publicly traded biotechnology companies. *J. Bus. Ventur.* **2002**, *17*, 577–609. [CrossRef]
74. Eom, B.-Y.; Lee, K. Determinants of industry–academy linkages and, their impact on firm performance: The case of Korea as a latecomer in knowledge industrialization. *Res. Policy* **2010**, *39*, 625–639. [CrossRef]
75. Grönqvist, C. The private value of patents by patent characteristics: Evidence from Finland. *J. Technol. Transf.* **2009**, *34*, 159–168. [CrossRef]
76. Blind, K.; Edler, J.; Frietsch, R.; Schmoch, U. Motives to patent: Empirical evidence from Germany. *Res. Policy* **2006**, *35*, 655–672. [CrossRef]
77. Park, S.; Lee, S.J.; Jun, S. Patent Big Data Analysis using Fuzzy Learning. *Int. J. Fuzzy Syst.* **2017**, *19*, 1158–1167. [CrossRef]

sustainability

MDPI

Article

Business Intelligence Issues for Sustainability Projects

Mihaela Muntean

Business Information Systems Department, Faculty of Economics and Business Administration, West University of Timisoara, 300223 Timisoara, Romania; mihaela.muntean@e-uvt.ro

Received: 30 December 2017; Accepted: 23 January 2018; Published: 29 January 2018

Abstract: Business intelligence (BI) is an umbrella term for strategies, technologies, and information systems used by the companies to extract from large and various data, according to the value chain, relevant knowledge to support a wide range of operational, tactical, and strategic business decisions. Sustainability, as an integrated part of the corporate business, implies the integration of the new approach at all levels: business model, performance management system, business intelligence project, and data model. Both business intelligence issues presented in this paper represent the contribution of the author in modeling data for supporting further BI approaches in corporate sustainability initiatives. Multi-dimensional modeling has been used to ground the proposals and to introduce the key performance indicators. The démarche is strengthened with implementation aspects and reporting examples. More than ever, in the Big Data era, bringing together business intelligence methods and tools with corporate sustainability is recommended.

Keywords: corporate sustainability; business intelligence; multi-dimensional data model; key performance indicators

1. Introduction

Corporate sustainability (CS) is "a business approach that creates long-term shareholder value by embracing opportunities and managing risks deriving from economic, environmental and social developments" [1]. Sustainable organizations are capable of maximizing their market value in the short, medium, and long term by managing knowledge and all kind of economic, social, and environmental issues [2,3]. When analyzing the sustainability of a company, the following three dimensions are taken into account: social, environmental, and economic aspects. Business decisions should imply a balance among these dimensions. According to Boyer et al. [4] and Farver [5], business intelligence (BI) frameworks are sustaining the decision-making processes in various business contexts. Driving business performance is possible through business analysis, enterprise reporting, and performance management—all three together are the pillars of the BI framework [6].

Nowadays, performance management systems also cover sustainability. Balanced scorecards, dashboards, and metrics include sustainability measurement items [7,8]. Therefore, sustainability performance can be measured with the support of BI [9]. Key performance indicators (KPIs) for measuring sustainability projects performance have been proposed, with a relevant example introduced in a paper by Lee and Saen [10].

BI frameworks that allow complex analytical processing and reporting are nowadays identified as business analytics (BA) frameworks. Skills, technologies, and practices for continuous iterative exploration and investigation of past business performances to gain insight and drive business planning are used [11]. Exploration of data from many source systems implies the formulation of an integrated multi-dimensional data model that can embrace various implementations. New technologies reduce the complexity of the data layer moving to in-memory approaches. Practically, the BA tool is both an analysis and reporting tool that operates on an in-memory multi-dimensional data model (usually named in-memory database).

In order to support performance management, any BI/BA approach can be enriched with a sustainability dimension for measuring the effort made in this direction. The integration of the sustainability view into a classical BI/BA schema is based on measures and dimensions reengineering. A démarche for defining a BI/BA schema with enriched sustainability measures will be subject to debate.

Sustainability projects are challenging ones. Monitoring them during their execution can also be performed with a BI approach. Monitoring is based on a set of key performance indicators, which have been modeled into a multi-dimensional schema, and further viewed and analyzed by a BA tool [2]. Sustainability implementation, monitoring, and reporting will suffer from transformations due to the changes of technology and information systems. Realizing the potential of big data, companies are interested in exploring this multitude of data and extracting valuable knowledge from it. As a result, firms are investing in business intelligence approaches and are using big data analytics to generate 'big picture' reporting.

2. Integrating the Sustainability View into a BI Data Model

Any BI model used by a company can be enriched with a sustainability dimension. KPIs for measuring sustainability performance are added to the classical schema. The new data model can be obtained using the following approach framework (Figure 1):

Step 1. The initial fact table will be transformed into a dimension—*Old_Fact_Now_Dimension*; as a result, the measures of the classical BI model will become dimensional attributes—*AM_1*, *AM_2* , . . . , *AM_m*; based on this transformation the sustainability perspective of the whole approach can be modeled as an add-on facility of the classical schema;

Step 2. Three new dimensions will be introduced: Economic profitability and transparency, Social responsibility, and Environmental sustainability;

Step 3. The measures of the sustainability data model will be established; they will be aggregated with respect to the new dimensions—*Old_Fact_Now_Dimension*, *Time_sustainability*, *Economic profitability and transparency* (*EPT*), *Social responsibility* (*SR*), and *Environmental sustainability* (*ES*).

Good sustainability performance can be achieved only if sustainability initiatives and projects are part of the corporate business strategy. Therefore, the sustainability performance management system should be handled within the framework of the corporate performance management system, taking into account the environmental, social, and economic aspects. The KPIs approach implies the inclusion of the sustainability KPIs into the corporate performance management system. The sustainability KPIs should be modeled as measures, if possible, or, should be deduced from the defined measures.

The proposed data model is developed by extending the corporate data model with sustainability specific data. The sustainability measures are stored in three fact tables that have two common dimensions—*Old_Fact_Now_Dimension* and *Time_sustainability*. *Old_Fact_Now_Dimension* is a degenerative dimension that was derived from a previous fact table of an initial, traditional BI approach. It provides a direct reference back to the transactional system and to the traditional BI data model. *Time_sustainability* dimension brings time intelligence into the model. Sustainability KPIs are stored directly into the fact tables; the overall business KPIs are deduced from the actual dimensional attributes of the degenerative dimension similar to the classical BI schema.

Based on Lee and Saen [10], we have considered economic profitability and transparency (EPT), social responsibility (SR), and environmental sustainability (ES) for covering the sustainability performance landscape.

In terms of economic profitability and transparency (EPT), for covering corporate governance (CG) performance and corporate transparency and accountability (CTA), four KPIs were introduced [10]: number of board/stakeholders meetings (CG1); personnel costs of communication/meetings (CG2); material costs like design/printing costs of communication (CTA1); and finally, personnel/administrative costs (CTA2).

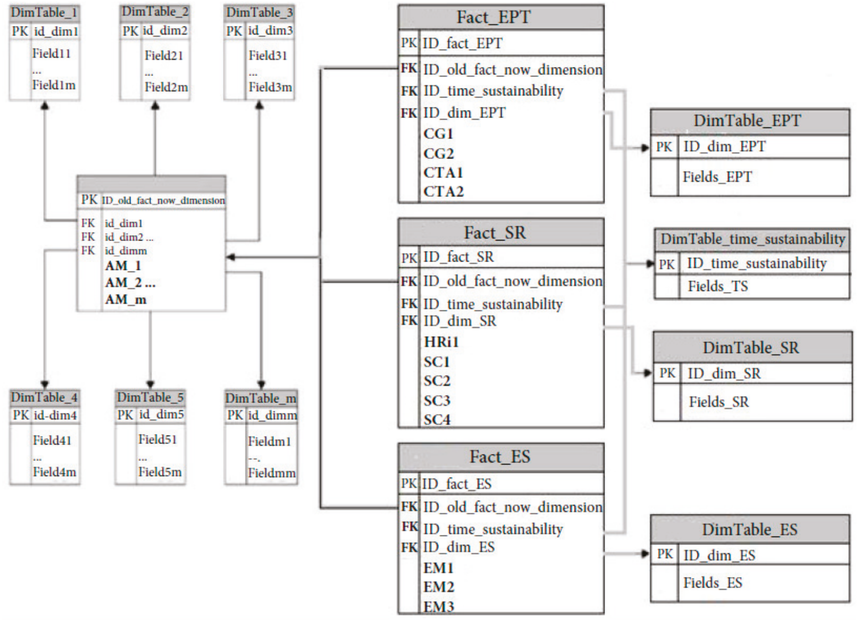


Figure 1. Data model proposal. Integrating sustainability into a business intelligence (BI) approach.

The Social responsibility dimension implies human rights (HRi) aspects and social contribution (SC) related activities. To measure performance, we proposed a mini-set of KPIs: number of employee training hours for corporate social responsibility (HRi1); expenses to train and promote social responsibility (SC1); number of social events with local communities (SC2); amounts of donations (SC3); and volunteering hours/personnel costs (SC4).

In addition, the third dimension, environmental sustainability, is related to environmental management and innovation (EMI) performance based on the following KPIs: number of green technology development projects (EMI1); costs of environmental management (EMI2); and costs of environmental product innovation (EMI3).

The above introduced sustainability KPIs have been chosen to support the data model proposal, although other indicator alternatives would also be suitable; their integration should respect the schema in Figure 1.

In conclusion, sustainability performance management approaches are implying, from a business and data management point of view, the following:

Step 1: analyzing the existing performance management system (BI project);

Step 2: identifying the current KPIs and their dependencies;

Step 3: analyzing the conceptual data model: identifying existing measures and dimensions of the multidimensional view; understanding the implementation details: fact tables and dimensional tables;

Step 4: designing the sustainability performance schema: defining sustainability dimensions—*Economic profitability and transparency* (*EPT*), *Social responsibility* (*SR*), and *Environmental sustainability*;

Step 5: defining sustainability KPIs for the considered sustainability aspects;

Step 6: designing the new data model schema similar to the proposal in Figure 1;

Step 7: establishing all reengineering aspects for transforming a fact table of the initial data model into a dimension—*Old_Fact_Now_Dimension*;

Step 8: establishing all information technology aspects and software procedures for making the BI data model enriched with the sustainability view viable.

The proposed démarche can be applied to any other set of sustainability KPIs.

3. Monitoring Sustainability Projects

Project execution monitoring can be designed into a business intelligence paradigm [12]. Sustainability projects, including sustainability performance management projects such as the approach of integrating the sustainability view into a BI data model, are challenging projects that employ different activities and various resources in order to achieve the desired objectives.

The initiative described in Section 2 has as its main objective the design of a data model that integrates sustainability performance measures into the corporate performance data model. The project itself implies a business intelligence approach based on eight main activities, indicated as Steps 1–8. The obtained output at the end of each step, representing the objective of that activity, will be used as the input for the next step. According to Kerzner [13], project execution monitoring involves monitoring scope, the schedule progress, and the project budget. In Figure 2, we propose a multi-dimensional data model for grounding any initiative of project execution monitoring.

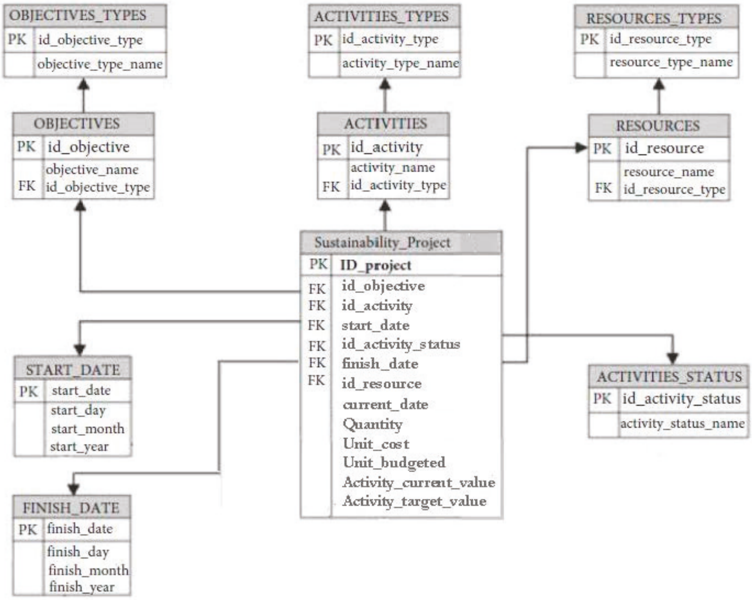


Figure 2. Data model for monitoring sustainability projects.

Key performance indicators can be easily attached to the data model. The following KPIs have been considered [14]: 1—schedule progress: Activity normal average (ANA), Activity normal value (ANV), Activity current average (ACA), Activity average progress (AAvP), Activity absolute progress (AAbP); 2—monitoring the budget: Activity total cost (ATC), Activity total budgeted (ATB), Activity remaining budgeted (ARB); 3—monitoring the scope: Project activities on scope (PAS), Project activities out of scope (PAoS), and Project activities number (PAN). They are calculated directly from the measures (*Quantity*, *Unit_cost*, *Unit_budgeted*, *Activity_target_value*, *Activity_current_value*) and have been subject of the author's previous work [14].

The data model in Figure 2 can be integrated into the global corporate data model by applying a reengineering approach to transform a fact table of the former global model into a degenerative dimension (*Old_Fact_Now_Dimension*) of the monitoring data schema.

4. Data Model Implementation Aspects. Support for BI

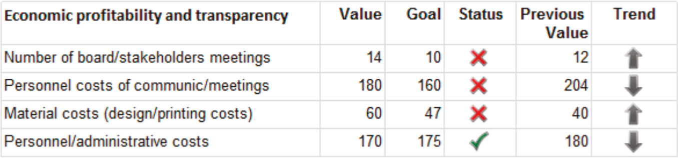
Data modeling is the process of creating a data model by applying a data model theory to create a data model instance. The goal of the data model is to make sure that the all data objects required by the database are completely and accurately represented. Data models describe data for further storage in a data management system. According to Bill Inmon's paradigm [15]: "A multi-dimensional database is one part of the overall BI system. An enterprise has one multi-dimensional database (data warehouse—DM), and optional, further data marts extract their information from it".

Both proposals have been introduced as extensions of an existing BI data model. This implies warehousing procedures for deploying the data models. The démarche is based on Inmon's approach—the top-down design [15]. The goal consists of modeling a subject-oriented, time-variant, integrated DM by passing through Steps 1–8 introduced in Section 2.

Reengineering aspects regarding the implementation of the KPIs and the realization of the degenerative dimension depend on the implementation technology. The degenerative dimension *Old_Fact_Now_Dimension* is derived from a fact table of the initial corporate data model. For example, MS SQL Server and its services offer full assistance in defining the degenerative dimension [16]. Referring to KPIs, technically they are collections of calculations (a combination of multidimensional expressions (MDX) and calculated members) associated with measures groups in the multi-dimensional data model.

The process of defining a KPI implies establishing a measure group, a value expression, a goal expression, a status indicator and expression, and a trend indicator and expression. Finally, all KPIs are included into a balanced scorecard. Results obtained after implementing the sustainability view are presented in Figure 3, where the first group of KPIs is listed. For KPIs with a "negative connotation" [17], the goal represents a maximum accepted value for considering the existence of a good performance status. Thereby, the status expression, for example, for CG2, has been formulated as the following

$$\begin{array}{l} \text{case} \\ \quad \text{when KpiGoal(“CG2”)} - \text{KpiValue(“CG2”))} < 0 \text{ then } -1 \\ \quad \text{when (KpiGoal(“CG2”)} - \text{KpiValue(“CG2”))} = 0 \text{ then } 0 \\ \text{else } 1 \\ \text{end} \end{array} \tag{1}$$



Economic profitability and transparency	Value	Goal	Status	Previous Value	Trend
Number of board/stakeholders meetings	14	10	✗	12	⬆
Personnel costs of communic/meetings	180	160	✗	204	⬇
Material costs (design/printing costs)	60	47	✗	40	⬆
Personnel/administrative costs	170	175	✓	180	⬇

Figure 3. BI approach. Output example of the first data model implementation.

The value −1 stands for the worst performance, 0 for neutral, and 1 for best, with the performance reflected by the status visual indicator.

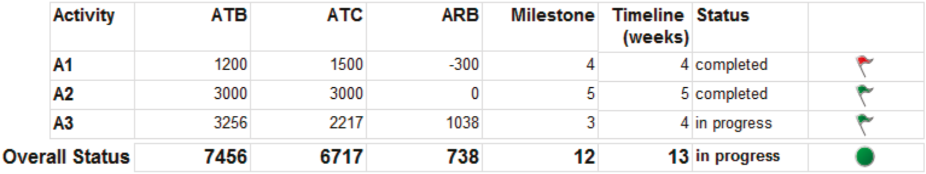
According to the data model proposed for monitoring project execution, the KPIs for monitoring the budget can be calculated with the following formulas

$$\begin{array}{l} \text{ATB} = \text{Unit\_budgeted} \times \text{Quantity}, \\ \text{ATC} = \text{Unit\_cost} \times \text{Quantity}, \\ \text{ARB} = \text{ATB} - \text{ATC}, \end{array} \tag{2}$$

transposed in

$$\begin{array}{ll}[\text{Measures}]\cdot[\text{Unit\_budgeted}] \times [\text{Measures}]\cdot[\text{Quantity}], & \\ [\text{Measures}]\cdot[\text{Unit\_cost}] \times [\text{Measures}]\cdot[\text{Quantity}], & (3)\\ ([\text{Measures}]\cdot[\text{Unit\_budgeted}] - [\text{Measures}]\cdot[\text{Unit\_cost}]) \times [\text{Measures}]\cdot[\text{Quantity}]. & \end{array}$$

For monitoring the budget during the project execution, milestones have been established. In Figure 4, twelve weeks after the project has started, the overall status has been analyzed; also, for each activity, the above KPIs have been determined and analyzed. Analyzing ARB, the budgeting will be analyzed and the corresponding visual status indicator displayed. Although for activity A1, the actual costs are greater than the estimated budged for this activity, the overall estimation is satisfactory. For the last week, we have an ARB of 738, which should cover the cost of the unfinished tasks of activity A3. According to the timeline, the status can be 'started', 'in progress', or 'completed'.



Activity	ATB	ATC	ARB	Milestone	Timeline (weeks)	Status	
A1	1200	1500	-300	4	4	completed	🚩
A2	3000	3000	0	5	5	completed	⚑
A3	3256	2217	1038	3	4	in progress	⚑
Overall Status	**7456**	**6717**	**738**	**12**	**13**	**in progress**	●

Figure 4. BI approach. Output example of the second data model implementation. ATB: Activity total budgeted; ATC: Activity total cost; ARB: Activity remaining budgeted.

Theoretical extensions of the data models together with implementation aspects in different business environments will be subjects of further research.

5. Discussion

Both business intelligence [18–20] and corporate sustainability [21–23] are two themes that have been highly studied in scientific literature. Business intelligence is an umbrella term for various business managing processes based on well informed decisions, which lead to high performance levels within organizations. Considered the art of gaining business advantages from data, BI covers verticals like 'Business Analysis', 'Enterprise Reporting', and 'Performance Management'. Traditionally, BI applications allow users to acquire knowledge from a company's internal data through various technologies. Companies now have access to different kinds of data that are collected from various sources including websites, business applications, social media pages, mobile devices, documents, and archives, all of them together identified as big data. The 'explosion of data', referring to the volume, variety, and velocity of the existing data, involves new forms of BI [18,19].

According to Getz [24], "sustainability is the ability to keep an organization running indefinitely without depleting natural resources or impacting the environment, maintaining economic viability, and conducting fair business practices with human capital". Furthermore, corporate sustainability is a business approach that creates long-term shareholder value by embracing opportunities and managing risks deriving from economic, environmental, and social developments. The sustainability triangle, a conceptual framework for corporate sustainability, aims at economic, eco-, and socio-effectiveness by integrating and linking the economic, environmental, and social dimensions through the concepts of eco-efficiency, socio-efficiency, and eco-justice [9,22]. Theoretical approaches and best practices in measurement, management, and reporting are indispensable for sustainability performance [21,23].

Despite this, only a few studies have approached these two themes in an interdisciplinary démarche. Referring to relevant initiatives [2–4,7,8], we have identified arguments on how BI methods and tools can improve the gathering, analysis, and dissemination of business data among employees, clients, suppliers, and partners, including corporate sustainability information. With respect to performance management principles, sustainability KPIs have been proposed [9,10] and included in balanced scorecards, with the theoretical fundamentals being part of a business intelligence framework [11,12]. Methodologies for measuring sustainability performance and models for measuring corporate sustainability have been proposed and tested, with the debate being developed with respect to the "From-Data-to-Performance" value chain, similar to business intelligence [6]. In summary, the efforts have been made preponderant on business issues, on integrating sustainability management into the corporate performance management system, bringing together business intelligence methods, and corporate sustainability.

However, business models without the support of technology and information systems are not viable. Therefore, the conjunction of business intelligence and sustainability implies beyond the integrated business model, the support of information technology. An integrated multi-dimensional data model would solve most of the integration challenges [25].

We reinforce our opinion that the management of corporate sustainability implies the use of business intelligence methods and tools for analyzing the financial, environmental, and social dimensions of the business [24]. Previous research by the author [6,20] has been focused on some theory and practice issues in business intelligence, with the BI value chain being introduced in terms of a value proposition [26]. Raw data is extracted from different data sources, further integrated into a multi-dimensional data model, which is nowadays usually stored in an "in-memory database". Through analytical processing, knowledge is retrieved from the database in order to support various decision-making processes. Data analytic tools (data analytics/business analytics) are capable, based on the multi-dimensional data model, of performing the necessary processing and reporting to support smart decision making, innovative services, new business models, innovation, and sustainability projects.

Both business intelligence issues presented in this paper represent the contribution of the author in modeling data for supporting further BI approaches in corporate sustainability initiatives. In developing any BI démarche, the design of the data model is crucial for the deployment of the system.

In Romania, most companies have implemented business intelligence systems that are capable of supporting the entire BI value chain: extracting data from data sources, integrating data into the multi-dimensional database, performing business analysis, and generating reports for managers and supporting performance management programs. The last three or four years have marked a change from the traditional BI approaches to the business analytics (BA) based ones [27]. At the same time, first steps have been made into integrating sustainability in business strategies [28,29]. In this context, the demand for reengineering the in-memory database of the existing BI/BA systems is strongly necessary.

Not least, best practices in project management reinforce the importance of monitoring project execution. It turns out that project management is 20% planning and 80% monitoring and control [30]. Remaining in the field of data model proposal, a multi-dimensional schema for monitoring sustainability projects is being introduced.

The two data models are not isolated data structures; thanks to the degenerative dimension *Old_Fact_Now_Dimension*, they are part of the corporate data model, which is, in fact, a big data model.

Despite the effervescence of unstructured data, for example, data gathered from websites, social media pages, and documents, the relational and multi-dimensional data model is still en vogue in corporate and business environments for its capability to model large data sets, commonly referred to as big data. The "in-memory database" technology extends the capability of the database management systems to store huge volumes of data. As mentioned in a recent article [31], dimensional models (multi-dimensional models) are present in the Big Data era. Data marts (departmental data warehouses) are still stored on relational or multi-dimensional platforms, and some companies have also chosen to move into the cloud. Thereby, multi-dimensional modeling approaches are still relevant, and the innovative approach presented in the paper fits in the big data landscape.

Both data model proposals can have a finality in their implementations; BI/BA have highly integrating capabilities and get the most value from big data. Theoretical extensions of the data models together with the implementation of aspects from a particular business environment will be the subject of further research.

6. Conclusions

Companies have implemented various business intelligence approaches in order to sustain performance management. All these initiatives are capable of extracting data from different sources, internal or external, to integrate it into a unitary model for further advanced business analysis and reporting. Relevant information is extracted and delivered to decision makers as valuable knowledge. In big data environments, BI embraces new forms, and BI tools, now evolved into big data analytics, are capable of gaining insight into this large volume of data. According to Forbes, in 2017, 53% of companies adopted big data analytics [32]. Both structured and unstructured data is processed, and descriptive, predictive, and even prescriptive analysis is performed. Despite the effervescence of the of unstructured data, coming from documents, the web, social networks and communities, the structured component of big data is and will remain an important data source for the BI/BA systems. Citing Columbus [32], "the dimensional (multi-dimensional) model becomes the nexus of a holistic approach managing BI, analytics, and governance programs".

Existing BI projects in enterprises need to be extended in order to integrate corporate sustainability approaches. A first step of the démarche consists in adapting the multi-dimensional data model of the BI database. United under *Business Intelligence Issues for Sustainability Projects*, the proposed data models can substantiate any initiative in this domain.

In a top-down approach, from business strategy to underlying data, the following statements represent essential pillars: sustainability is part of the business strategy; sustainability performance management is integrated into the corporate performance management system; the corporate BI system also includes the sustainability dimension; the sustainability data model is integrated into the corporate data model, and sustainability data is part of the corporate data.

Author Contributions: Mihaela Muntean conceived, designed, and wrote this paper.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Dyllick, T.; Muff, K. Clarifying the Meaning of Sustainable Business: Introducing a Typology from Business-as-Usual to True Business Sustainability. *Organ. Environ.* **2015**, *4*, 1–19. Available online: https://www.bsl-lausanne.ch/wp-content/uploads/2015/04/Dyllick-Muff-Clarifying-Publ-Online.full_.pdf (accessed on 21 January 2018). [CrossRef]
2. Petrini, M.; Pozzebon, M. Integrating Sustainability into Business Practices. *Braz. Adm. Rev.* **2010**, *7*, 362–378. Available online: http://www.scielo.br/pdf/bar/v7n4/04.pdf (accessed on 21 January 2018). [CrossRef]
3. Dočekalová, M.P.; Kocmanová, A. Composite indicator for measuring sustainability. *Ecol. Indic.* **2016**, *61*, 612–623. [CrossRef]
4. Boyer, J.; Frank, B.; Green, B.; Harris, T.; Van de Vanter, K. *Business Intelligence Strategy: A Practical Guide for Achieving BI Excellence*, 1st ed.; MC Press Online, LLC: Ketchum, ID, USA, 2010; pp. 35–72. ISBN 978-158347-362-7.
5. Farver, S. *Mainstreaming Corporate Sustainability: Using Proven Tools to Promote Business Success*, 1st ed.; GreenFix, LLC: Aspen, CO, USA, 2013; pp. 133–142. ISBN 978-1484135327.
6. Muntean, M. Business Intelligence Approaches. *Math. Models Methods Appl. Sci.* **2012**, *1*, 192–196. Available online: https://mpra.ub.uni-muenchen.de/41139/ (accessed on 21 January 2018).
7. Ahmad, A. Business Intelligence for Sustainable Competitive Advantage. In *Sustaining Competitive Advantage via Business Intelligence, Knowledge Management, and System Dynamics*; Quaddus, M., Ed.; Emerald Group Publishing Limited: Bingley, UK, 2015; Vol. 22A.

8. Scholtz, B.; Calitz, A. Using Business Intelligence to Support Strategic Sustainability Information Management. In Proceedings of the 2015 Annual Research Conference on South African Institute of Computer Scientists and Information Technologists, Stellenbosch, South Africa, 28–30 September 2015. [CrossRef]
9. Thomas, M.; McElroy, M.W. A Better Scorecard for Your Company's Sustainability Efforts. *Hav. Bus. Rev.* **2015**. Available online: https://hbr.org/2015/12/a-better-scorecard-for-your-companys-sustainability-efforts (accessed on 21 January 2018).
10. Lee, K.H.; Saen, R.F. Measuring corporate sustainability management: A data envelopment analysis approach. *Int. J. Prod. Econ.* **2012**, *140*, 219–226. [CrossRef]
11. Laursen, G.H.N.; Thorlund, J.; DeWees, B. *Business Analytics for Managers: Taking Business Intelligence beyond Reporting*; John Willey & Sons: Hoboken, NJ, USA, 2010; pp. 17–42. ISBN 978-0-470-89061-5.
12. Pondel, J.; Pondel, M. BI and Big Data Solutions in Project Management. *Bus. Inform.* **2015**, *4*, 55–63. Available online: http://www.dbc.wroc.pl/Content/34362/Pondel_BI_And_Big_Data_Solutions_In_Project_Management_2015.pdf (accessed on 21 January 2018). [CrossRef]
13. Kerzner, H. *Project Management: A System Approach of Planning, Scheduling and Controlling*, 11th ed.; John Willey & Sons: Hoboken, NJ, USA, 2013; pp. 549–573. ISBN 978-1-118-02227-6.
14. Muntean, M.; Cabău, L.G. Business Intelligence Support for Project Management. In Proceedings of the 14th International Conference on Informatics in Economy; 2014; pp. 428–432, WOS: 000362796900069. Available online: https://mpra.ub.uni-muenchen.de/51905/ (accessed on 21 January 2018).
15. Inmon, W.H. *Building the Datawarehouse*, 4th ed.; Wiley Publishing Inc.: Hoboken, NJ, USA, 2005; pp. 71–138. ISBN 978-0471081302.
16. Webb, C.; Russo, M.; Ferrari, A. *Expert Cube Development with SSAS Multidimensional Models*, 2nd ed.; Packt Publishing Ltd.: Birmingham, UK, 2014. ISBN 978-1-84968-990-8.
17. Muntean, M.; Cabau, L. Business Intelligence Approach in a Business Performance Context. *Austrian Comput. Soc.* **2012**. Available online: https://mpra.ub.uni-muenchen.de/29914/1/MPRA_paper_29914.pdf (accessed on 21 January 2018).
18. Marz, N.; Warren, J. *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*; Manning Publishing Co.: Shelter Island, NY, USA, 2015; pp. 27–53. ISBN 978-1-617-290343.
19. Provost, F.; Fawcett, T. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*, 1st ed.; O'Reilly Media Inc.: Newton, MA, USA, 2013; pp. 81–110. ISBN 978-1-449-36132-7.
20. Muntean, M.; Muntean, C. Evaluating a Business Intelligence Solution. Feasibility Analysis Based on Monte Carlo Method. *J. Econ. Comput. Econ. Cybern. Stud. Res.* **2013**, *47*, 85–102.
21. Brockett, A.; Razaee, Z. *Corporate Sustainability: Integrating Performance and Reporting*; John Willey & Sons: Hoboken, NJ, USA, 2012; pp. 67–98. ISBN 978-1-118-12236-5.
22. Thiele, L.P. *Sustainability (Key Concepts)*, 2nd ed.; Polity Press: Cambridge, UK, 2016; pp. 90–171. ISBN 978-1509511075.
23. Benn, S.; Edwards, M. *Organizational Change for Corporate Sustainability (Understanding Organizational Change)*, 3rd ed.; Routledge: New York, NY, USA, 2014; pp. 135–180. ISBN 978-0415695497.
24. Getz, A. Using Business Intelligence to Achieve Sustainable Performance. 2014. Available online: http://bi-insider.com/wp-content/uploads/2014/10/BI-to-Enhance-Sustainability-2014-10-10.pdf (accessed on 28 July 2017).
25. Džmuráň, M. Introduction to Data Integration Driven by a Common Data Model. Available online: http://www.galeos.eu/uploads/Soubory/ClankySOI/Introduction_to_Data_Integration_EN.pdf (accessed on 10 November 2017).
26. Muntean, M. Theory and Practice in Business Intelligence. 2012. Available online: https://ssrn.com/abstract=2144440 (accessed on 21 January 2018).
27. Iovan, S. Business Intelligence and the Transition to Business Analytics. Annals of "Constantin Brâncuşi" University of Târgu-Jiu, Engineering Series. 2014, pp. 150–156. Available online: http://www.utgjiu.ro/revista/ing/pdf/2014-4/25_Stefan%20Iovan.pdf (accessed on 21 January 2018).
28. Fistis, G. Demand for Sustainability Expertise on the Rise as Romanian Companies Tune in to European Business Standards. 2015. Available online: http://www.business-review.eu/featured/demand-for-sustainability-expertise-on-the-rise-as-romanian-companies-tune-in-to-european-business-standards-93203 (accessed on 18 September 2017).

29. Stancu, M. Romanian Companies Need to Make Swift and Sustained Steps to catch up on Corporate Responsibility Reporting. 2017. Available online: https://home.kpmg.com/ro/en/home/media/press-releases/2017/10/cr-reporting-2017-survey.html (accessed on 3 November 2017).
30. Berkun, S. *Making Things Happen: Mastering Project Management*, revised ed.; O'Reilly Media, Inc.: Newton, MA, USA, 2008; pp. 241–301. ISBN 978-0-596-51771-7.
31. Adamson, C. Dimensional Models in the Big Data Era. 12 April 2017. Available online: https://tdwi.org/Articles/2017/04/12/Dimensional-Models-in-the-Big-Data-Era.aspx?Page=2 (accessed on 26 November 2017).
32. Columbus, L. 53% of Companies Are Adopting Big Data Analytics. December 2017. Available online: https://www.forbes.com/sites/louiscolumbus/2017/12/24/53-of-companies-are-adopting-big-data-analytics/#4f19636c39a1 (accessed on 8 January 2018).

sustainability

MDPI

Article

Modeling and Quantifying User Acceptance of Personalized Business Modes Based on TAM, Trust and Attitude

Jie Zhao [1], Suping Fang [1] and Peiquan Jin [2],*

[1] School of Business, Anhui University, Hefei 230601, China; 97040@ahu.edu.cn (J.Z.); m16201039@stu.ahu.edu.cn (S.F.)

[2] School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China

* Correspondence: jpq@ustc.edu.cn

Received: 31 December 2017; Accepted: 29 January 2018; Published: 30 January 2018

Abstract: With the rapid development of economics and social businesses, users' business demand has changed a lot. More and more people want to personalize their business modes so that they can get better experiences in business and learning activities. The key factor of personalized business mode is to consider users' individual needs on business activities, so that users can receive differentiated services. Users' satisfaction on personalized services will effectively improve the consuming experience of users, which is helpful for business organizations to strengthen their competitive power in business environments. However, will users wish to participate in personalized businesses? This is a crucial issue for developing personalized businesses. Aiming to solve this problem, this paper analyzes the major factors influencing user acceptance of personalized business modes. Then, we propose a research model that enhances the TAM (Technology Acceptance Model) model with trust and attitude to depict the influence from several variables to user acceptance of personalized business modes. Further, we use the structural equation method to conduct an empirical analysis on questionnaire data from the Internet. The results in terms of many kinds of data analysis show that trust and the TAM factors (perceived usefulness and perceived ease of use) have significant influence on user acceptance of personalized business modes. In addition, there are partial intermediate relationships existing among the factors of the research model.

Keywords: personalized business mode; technology acceptance model; user acceptance; data analysis

1. Introduction

The program of "Made in China 2025", developed based on Germany's "Industrial 4.0", provides new opportunities for transforming and upgrading the manufacturing industry in China. It also becomes the material foundation for customized services. China's realization of the "Internet+" strategy aims to personalize businesses as "Internet+" and to make the coordinated development of manufacturing sectors as an important platform, including manufacturing clothing, home appliances, furniture, and other industries. Consequently, the development of customized services has been a dominant trend in today's business world. In addition, by December 2016, the number of Internet users in China has reached 731 million, and the majority of Internet users in China are 10–39 years old [1]. These people are familiar with and dependent on the Internet world, and gradually become the main consumers in the whole society. Thus, it is important to provide customized or personalized services for them. Customization is also called personalized service, through which enterprises provide customized services according to the needs of different users [2,3]. Nowadays, personalized service has been a key indicator in the development of enterprises, because it enables companies to offer consumers unique products and services by considering users' personal properties such as user profiles, personal preferences, style characteristics, and shopping behavior. The basis for developing personalized businesses is to understand:

(1) Are consumers willing to participate in personalized businesses?
(2) Which factors will impact the participation of users in customized services?

Making these issues clear is helpful for enterprises to improve users' participation of customization.

This paper aims to study the willingness of users in participating in personalized business modes. At present, academia has not formed a unified conclusion about this issue. The research of personalized business modes is still at the beginning stage and has not formed an authoritative theoretical system. Particularly, previous studies in personalized business modes mainly focused on customization technologies and web-based customization systems, and few were towards personalized business modes. Specifically, the intention of users to participate in personalized businesses has not yet been studied systematically.

In this paper, we first study the user acceptance of personalized business modes. In particular, we analyze the major factors that influence user acceptance of personalized business modes. Then, we augment the Technology Acceptance Model (TAM) [4] to construct a theoretical research model to describe the influence from several variables to user acceptance of personalized business modes. The research model integrates TAM with trust, and uses attitude as the mediating factor. After that, we use the structural-equation method to conduct an empirical analysis on questionnaire data from the Internet.

In summary, we make the following contributions in this paper:

(1) We integrate the Technology Acceptance Model (TAM) with trust and attitude to analyze the users' willingness to participate in personalized business modes, and provide some research ideas for the relevant research in the field of personalized business modes. To the best of our knowledge, this is the first study of integrating the TAM model with trust and attitude to analyze user acceptance of personalized business modes.
(2) We conduct questionnaire on the Internet, and perform systematical data analysis over the questionnaire data in terms of various metrics including reliability and validity.
(3) We present empirical data analysis using commercial software and obtain several results. The data analysis consists of many aspects, including factor analysis, correlation analysis, regression analysis, mediating effect analysis, control variable analysis, and hypothesis evaluation. The results show that trust and the TAM factors (perceived usefulness and perceived ease of use) have significant influence on user acceptance of personalized business modes. In addition, there are partial intermediate relationships existing among the factors of the research model.

The study in this paper is helpful for enterprises to realize the importance of developing personalized business modes. It also reveals the major factors that influence personalized business modes, and quantifies the impacts of these factors on user acceptance of personalized business modes. Thus, this study can provide some management ideas for enterprises to develop personalized businesses. In addition, the study in this paper can further advance the relevant theories of personalized customization.

The remainder of this paper is structured as follows. Section 2 describes the related work and the differences between previous studies and this paper. Section 3 presents the research model as well as the hypothesis proposition. Section 4 describes the details of data collection. In Section 5, we discuss the results of data analysis. In Section 6, we conclude the entire paper.

2. Related Work

2.1. Personalized Business Modes

Traditional business modes are typically based on product competition. Enterprises are concerned about the improvement of their own operational efficiency. However, as competition becomes increasingly fierce, enterprises may gradually lose their product-centric competitive advantages in the market. Thus, it is necessary for companies to adapt to the timely transformation from product-centric business modes to customer-centric business modes [5]. On the one hand, personalized customization is helpful to meet the personal needs of customers and to stimulate customer consumption and attract potential consumers. On the other hand, it also benefits the enterprise, and promotes the business development of the enterprise [6]. Choi and Lee [2] found that consumers generally preferred personalized products over standardized ones. This study argued that the consumer preference for personalized products depended on purchasing context and reversibility of choice.

With the development of economy and culture, it is difficult for the mass products to meet the diversified demands of consumers. Some traditional production models gradually lose their competitive advantages in the market, and enterprises need to further segment the market and operate professionally. For consumers, they pursue customized products or services with high quality, low cost and personalized features [3]. Consumers can choose their own style according to their preferences, such as color, size, and location. Businesses do not need to design mass products [7]. In addition, competition among enterprises is becoming more and more rigorous. How to personalize and form diversified and personalized requirements has been the focus of increasing competitive power for enterprises [8]. With the development of personalized business modes as well as the changes in various fields of society, many companies started to reduce costs and improve efficiency by optimizing personalized businesses [9]. Fogliatto et al. [10] reviewed the literature on mass customization over the last decade and provided a conceptual framework to support future research. They reviewed the concept, economics, success factors, and enablers of mass customization. To overcome the gap between customization and personalization, Wang and Ma [11] presented a framework for personalized production based on the concepts of Industry 4.0.

Previous research on personalized business modes focused on customization technologies. The research on product customization technology can be divided into three groups [12,13]. The first group is a management model based on customer demands, the second one is a management model based on product configuration, and the third one is a management mode based on customer online customization. Customization technologies were mainly used to build human–computer interaction platforms. On such platforms, users can select their own product components and participate in product research and development. In addition, web-based technologies offer to enterprises great support to provide personalized online services for their customers. As one of the hottest web-based technologies, recommender systems aim to automatically generate personalized suggestions of products/services to customers (businesses or individuals). Wu et al. [14] pointed out that, although recommender systems have been well studied, there are still two challenges in the development of a recommender system, particularly in real-world B2B e-services. Accordingly, they proposed a method for modeling fuzzy tree-structured user preferences. A recommendation approach to recommending tree-structured items was then developed. This study also applied the proposed recommendation approach to the development of a web-based business partner recommender system. Kim et al. [15] claimed that a Business Activity Monitoring (BAM) system should provide personalized monitoring capabilities. Therefore, they developed a personalized BAM system. Many clothing enterprises adapt custom online customization management mode to develop personalized business. Customers chose the elements of garment design to obtain customized products, such as the United States Mysuit system, VANCL system and other systems. Researchers at Nanyang Technology University set up a product customization platform based on network, the platform will be between all departments of the enterprise network services focused on network platform, proposed a GPF (generic product family) model, using XML document as the middleware for communication should be between the application and the GPF model [16]. Yang et al. added a custom system function module in the enterprise CIM system, established a kind of information resources can be shared e-commerce sales mode [17]. Recently, several mass customizers connected their sales configurators with social-network based software. This is not surprising because social-network software enables an interactive and socially rich shopping experience, which makes shopping with a mass-customization toolkit more similar to retail shopping. However, research on the use of social-network software by mass customizers are very limited, i.e., almost all previous studies on mass-customization toolkits were focused on the dynamic interaction between sales configurators and potential customers. Based on a survey on 277 real online sales configurators, Grosso et al. [18] identified eight ways in which online sales configurators can connect with social-network software. Nowadays, with the rapid development of information technologies (e.g., web 2.0, cloud computing, and virtual reality) and manufacturing technologies (e.g., additive manufacturing), users become more actively involved in product development processes to create personalized products with higher efficiency. This emerging manufacturing paradigm is known as mass personalization, of which user experiences (e.g., emotional factors and product utility), co-creation (e.g., user participation), and product change (e.g., modular design) are regarded as three key characteristics [19].

On the other hand, research on personalized business modes has focused on enterprise applications. With the development of electronic commerce, enterprises tried to develop enterprise customization system to gain or maintain competitive advantage. The product category involves clothing, electronic products and so on. On the premise of functional analysis, the enterprise divided and designed a series of functional modules, which could form different products through the selection or combination of modules, which satisfied the different demand of market [20]. The enterprise system of personalized customization has emerged, and has achieved remarkable economic and social benefits. In western countries, enterprise Internet system construction has basically completed the transformation from the first stage to the second stage, and began to gradually transition to the third stage. In some countries outside China, to maintain the existing market position or for a new position, leaders of many high technology industries were actively developing customized businesses. This trend is especially common in the manufacturing industries in the U.S., Germany, and Japan. The customized web system established by DELL provided personalized services to users all over the world based on understanding the real needs of users. In the customized business carried out by the Internet companies in the United States, consumers designed products and reduced costs. Panasonic has carried out modular production. Users can customize standardized components and accessories to get personalized products. Buick's North American website set up customized service customization, users could not only choose the engine type, tire style, body color and other parameters could also indicate the name of the owner in writing the quasi car back, the system would eventually be calculated according to the parameters chosen by the customer, after a month, consumers could buy the car of "the one and only". In China, the construction of enterprise web system was still in the primary stage of development. The main function of enterprise web system was product introduction and enterprise propaganda. Haier has opened a personalized custom system to provide customized business to consumers, consumers could choose their own capacity, style and other customized products according to their individual needs, and customized the products they need. Personalized customization has gradually developed into a more popular business mode, and domestic personality customization has been stimulated by the existing foreign successful cases, and it also shows a trend of vigorous development. In early 2005, some enterprises have begun to explore personalized custom T-shirts, quilts and other gifts in China, such as www.tshe.com and www.productdiy.cn. Lv et al. [21] investigated a two-dimensional model involving both vertically differentiated product preferences and horizontally differentiated personalization services.

The research of personalized business modes has not formed an authoritative theoretical system, but it is still in the initial stage of active exploration. The researches on personalized business modes are mainly focused on customization technologies and web-based customization systems. On the other hand, the study of personalized business modes focus on enterprise application level for customization, and personalized business modes have been gradually become a popular business mode. However, the intention of users to participate in personalized businesses has not yet been studied explicitly.

2.2. Technology Adoption Models

After the industrial revolution in the 1900s, due to the development of modern economics and the intensification of competition, many researchers began to study consumer behavior and proposed several technology adoption models [22,23]. Mehrabian and Russell put forward the SOR model (Stimulus–Organism–Response) based on the theory of cognitive psychology. They claimed that stimulus (S) with the organism of brain (O) can lead to people's reactions (R) [24]. The theory of reasoned action (TRA) model was first proposed for social psychology [25]. According to the TRA theory, a person can make a rational and comprehensive consideration of his own factors as well as the significance and consequences of his actions based on the value judgments made by individuals in social life. Further, the TRA model was used in business areas, e.g., for predicting user acceptance of e-shopping on the web [26]. In 1989, Davis et al. proposed the TAM model (Technology Acceptance Model) [4], which was first used to explain and predict user acceptance of computer technologies. The TAM model considers whether people accept an information system or not. Basically, if an information system can help people do their jobs better, it is perceived as useful. On the other hand, if an information system is easy to use, it is perceived to be easy to use. Most of the behavioral factors in the TAM model focus on describing user behavior to accept or reject the use of new technologies. Accordingly, the TAM model defines two variables, namely perceived usefulness and perceived ease of use, to quantify user attitude to information technology, which in turn can be used to measure user acceptance of information technologies.

Although the TAM model is initially designed to explain and predict behavior of individuals on the use of information systems, it has been used in many studies, especially in E-commerce related studies. That is mainly because E-commerce is a technology-driven area that is based on new information technologies such as web, mobile computing, and recommendation. Ha et al. [27] as well as Lu and Su [28] used the TAM model to predict the acceptance and use of consumer online shopping. Gefen et al. [29] studied the MBA and senior students' willingness of buying books from Amazon. Vijayasarathy et al. also used this model to study the acceptance of online shopping by investigating 281 young people in the U.S. [30]. There are other works concentrated on the online shopping behavior of network consumers. For example, Ko et al. used the TAM model to study consumer adoption of mobile shopping for fashion products in Korea [31]. Xie and Lee [32] pointed out that in the TAM model, perceived usefulness and perceived ease of use affect users' attitude to and willingness to accept information technologies, and technologies that are easy to understand and use will be more attractive to people. Many scholars have used the TAM model to predict the acceptance and use of consumer online shopping [33–35]. Some researchers pointed out that trust has an impact on consumer acceptance intention [36,37].

Previous studies have demonstrated the applicability of the TAM model to electronic businesses, especially to online shopping. In this paper, we focus on the personalized business in the network environment. The users in this study are mainly supposed to be users in electronic business. Thus, it is a reasonable choice to consider the TAM model to model and quantify user acceptance of personal business modes. On the other hand, differing from the traditional TAM model, we integrate trust with TAM and propose to use attitude as the mediate factor, forming an augmented TAM model that is more suitable for the research issue of this paper.

3. Research Methodologies and Hypotheses

3.1. Research Model

The Technology Acceptance Model (TAM) has been widely used for predicting and explaining user behavior and IT usage [4]. According to the TAM model, individual behavior of an information system is determined using the system and the use of behavior intentions. Behavior intentions are affected by perceived usefulness and attitude. Here, attitude refers to the individual evaluation on the information system, i.e., positive or negative reactions. The use of an information system is mainly impacted by two factors, namely perceived usefulness and perceived ease of use. The perceived usefulness refers to the effect of the perceived use of the information system on improving the performance of the system. The perceived ease of use refers to the individual perceived ease of the use of the information system. When the individual perception information system is easier to use, the perceived ease of use will have a positive impact. The perceived usefulness and perceived ease of use are both affected by external factors. In addition, the perceived ease of use also positively affects the perceived usefulness.

In this paper, we first select the main factors in the TAM model as basic independent variables, i.e., perceived usefulness (*Perceived Usefulness*, PU) and perceived ease of use (*Perceived Ease of Use*, PEOU). Further, we introduce *trust* as a new independent variable. Trust has been recognized as a critical factor in online environment [29,36]; thus, it is reasonable to introduce it into the research model. The behavior intention (*behavior Intention*, BI) is designed to be the dependent variable. In the study of behavior intention, attitude has been commonly recognized as a variable that affects the willingness of behavior. Thus, we choose *attitude* as the mediating variable. On this basis, this paper puts forward the research model of consumers' willingness to participate in enterprise customization. Figure 1 summarizes the research model of this paper.

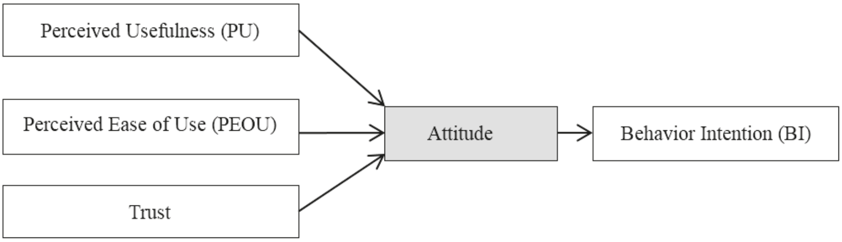


Figure 1. The research model.

3.2. Research Hypothesis

Users have a big impact on the development of personalized businesses. That means personalized businesses have to be accepted by users. Thus, enterprises need to take account of some the factors influencing users' behavior in the development of customized businesses. It has been studied before that participating in customization activities can bring some benefits for participants [38,39]. At the same time, the development of customized businesses needs to consider the difficulty for users to participate in personalized businesses. Some companies may build network-based information systems to assist users for using personalized businesses. Such systems can make users feel that customized businesses are easy to use.

The TAM model uses perceived usefulness and perceived ease of use to reflect behavior intention. Perceived ease of use refers to the individual's perceived ease of using an information system, while perceived usefulness refers to the degree to which individuals perceive the use of information systems to improve their work performance [4]. Many previous studies have shown that perceived ease of use not only affects the willingness of participating but also affects the perceived usefulness [33].

In the environment of personalized businesses, it is necessary for products or services to be easy to understand and use, otherwise consumers will not accept new products or services.

According to the research model presented in Section 3.1, there are three independent variables, namely perceived usefulness, perceived ease of use, and trust. In addition, there is one mediating variable called attitude. These variables are supposed to impact the dependent variable named behavior intention. Thus, to reveal the relationships among these factors, we first raise the following research questions, as listed in Table 1. The right column in Table 1 shows the corresponding hypotheses that are proposed to answer the research question. Q1 is to find out the influential relationship among the independent variables of the research model. Q2 is to find out the influential relationship between independent variables and the mediating variable. Q3 is to find out the influential relationship between the mediating variable and the dependent variable. Finally, Q4 aims to find out the intermediate relationship among all the factors in the research model.

Table 1. Research questions and corresponding hypotheses.

Question Number	Research Question	Corresponding Hypothesis
Q1	What relationship exists among the independent variables of the research model?	H1
Q2	What influences exist between independent variables and the mediating variable (attitude)?	H2, H3, H9
Q3	How does the mediating variable (attitude) impact the dependent variable?	H4
Q4	How does a factor in the research model play intermediate role between other factors?	H5, H6, H7, H8, H10

The details about each hypothesis are presented as follows. The objective of each hypothesis can be found in Table 1.

Perceived ease of use refers to the individual's perceived ease of using an information system. If an information system is perceived to be much accessible, it is much likely that users have a positive attitude on using the information system. In personalized businesses, perceived ease of use means that consumers can easily participate in personalized businesses, and the interaction process with the enterprise is simple and easy to use. On the other hand, the complexity of business systems and processes will hinder users to participate in personalized businesses. Therefore, we make the following hypotheses (H1 and H2).

Hypothesis 1 (H1). *Users' perceived ease of use of personalized business modes positively affects their perceived usefulness of personalized business modes.*

Hypothesis 2 (H2). *Users' perceived ease of use of personalized business modes affects their attitude to personalized business modes.*

Perceived usefulness refers to the degree to which individuals perceive the use of information systems to improve their work performance. If users think that information systems are useful, they are likely to have a positive attitude on using them. Attracting and motivating the public to participate in the process of personalized business is important to personalized businesses, because participating in personalized tailor-made activities can bring certain benefits to participants and satisfy their actual needs [38]. Some researchers summarized these factors as internal motivation and external motivation, while others divided them into personal motivation and social motivation [39]. As a result, personalized business needs to satisfy user needs, and users' perceived usefulness of personalized business modes will impact their attitude on using personalized businesses. Therefore, we make hypothesis H3, which is specified as follows.

Hypothesis 3 (H3). *Users' perceived usefulness of personalized business modes affects their attitude to personalized business modes.*

Based on the traditional TAM model, many previous works have empirically evaluated the positive influence of attitude on the willingness to participate. Users' willingness to participate in the customization of the enterprise is influenced by their subjective attitude. More active attitude may lead to high possibility of participating in personalized businesses. Based on such assumption, we make hypothesis H4.

Hypothesis 4 (H4). *Users' attitude to personalized business modes affects their behavior intention on personalized business modes.*

Venkatesh et al. [40] proposed that there must be a corresponding attitude towards real behavior before an action occurs. In other words, attitude affects user acceptance behavior [40]. User acceptance attitude refers to the overall tendency of users to participate in personalized businesses, which is impacted by cognitive tendency, emotional expression, and behavioral tendency of using personalized products or services. Users have to go through the behavioral attitude phase before generating actual participation intention. This means that perceived ease of use and perceived usefulness will first inspire users' attitude towards the customization of the enterprise. Thus, we make the following hypotheses (H5 and H6) to reflect the above observations.

Hypothesis 5 (H5). *Users' attitude to personalized business modes plays an intermediary role in perceived ease of use and behavior intention.*

Hypothesis 6 (H6). *Users' attitude to personalized business modes plays an intermediary role in perceived usefulness and behavior intention.*

According to the TAM model, perceived usefulness plays an intermediary role among perceived ease of use, attitude, and willingness to participate. If the personalized business of an enterprise is relatively simple to use, we can imagine that users' attitude on using personalized business will be stimulated, resulting in a positive attitude on participating in personalized business. Therefore, we make the following hypotheses (H7 and H8).

Hypothesis 7 (H7). *Users' perceived usefulness of personalized business modes plays an intermediary role in perceived ease of use and attitude.*

Hypothesis 8 (H8). *Users' perceived usefulness of personalized business modes plays an intermediary role in perceived ease of use and behavior intention.*

In recent years, trust has been widely studied in electronic commerce and other areas. Koo [41] and Chang et al. [42] showed that trust was strongly associated with attitude to products and services as well as attitude to purchasing behaviors. Gefen et al. pointed out that trust was very important for network merchants, and trust was not only a key factor influencing users' acceptance of information technologies, but also a key factor attracting consumers [36]. Heijden et al. studied the trust variables of consumers in Holland. They found that trust had an impact on the use intention of electronic websites [37]. Chen and Yang [43] showed that trust not only had a positive impact on users' willingness to use microblogging services. They concluded that trust can be used as an important indicator of consumer behavior in the Internet environment. In the scope of personalized businesses, we note that there is a value exchange between users and enterprises. In the process of such a value exchange, the trust between the two sides is a critical factor. In personalized businesses, the credit of the enterprise is very important. Therefore, we attempt to conduct an empirical research on whether trust has a significant impact on attitude and behavior intention. Thus, we make the following hypothesis (H9 and H10).

Hypothesis 9 (H9). *Users' trust on personalized business modes affects their attitude to personalized business modes.*

Hypothesis 10 (H10). *Users' attitude to personalized business modes plays an intermediary role in trust and behavior intention.*

4. Questionnaire and Data Collection

4.1. Questionnaire Design

Table 1 shows the design of the questionnaire indicators of our study. Most questions can be found in previous studies to ensure the validity and reliability of the questionnaire. Each item of the questionnaire is assessed using a five-point Likert scale from the bottom value "*strongly disagree*" to the top value "*strongly agree*" (see Table 2). The observation index is set up according to the environment of personalized businesses.

Table 2. Indicators of the questionnaire.

Construct	Question Code	Measurement Problem
PU	PU1	If I were to adopt personalized business modes, it would enable me to purchase favorite products
	PU2	If I were to adopt personalized business modes, it would enable me to broaden my understanding of the product or service
	PU3	If I were to adopt personalized business modes, the effect of purchasing goods or services would improve
	PU4	If I were to adopt personalized business modes, it would satisfy my personal consumption needs
PEOU	PEOU1	Learning how to participate in personalized business would be easy for me
	PEOU2	Participation in enterprise personalized business takes less time and effort
	PEOU3	In the process of enterprise personalized business, enterprises provide clear and easy communication
Trust	TR1	I'll think products or services purchased by personalized business modes will be trustworthy
	TR2	I believe that companies will take into account customer needs in the process of customization
	TR3	I believe the expected effect of customized products or services is predictable
	TR4	I believe that the after-sale service of personalized business modes is guaranteed
	TR5	I believe the organizers will provide meticulous service to help me solve all kinds of problems in the process of customization
Attitude	AT1	I think it's a good idea to take part in customization
	AT2	I like to experience personalized business modes
	AT3	Participating in personalized business modes will bring me a pleasant experience
	AT4	I'll be positive about personalized business modes
BI	BI1	I'll intend to experience personalized business as soon as possible.
	BI2	I intend to continue to participate in the personalized business of the enterprise
	BI3	I will recommend personalized business modes to my friends
	BI4	I have an urge to participate in personalized business

The questionnaire consists of two parts: (1) demographic questionnaire; and (2) questionnaire on independent variable, mediating variable, and dependent variable. The explanations of these parts are as follows:

1. Demographic questionnaire: This part mainly surveys the gender, age, educational background, income, etc.
2. Questionnaire on independent variables, mediating variables and dependent variables: This part of questionnaire is the core of the whole questionnaire. This study contains five latent variables: perceived usefulness, perceived ease of use, trust, attitude, and behavior intention. After a preliminary investigation, we make the formal index of investigation, as shown in Table 2.

4.2. Data Collection

We conduct an online survey to verify our research model. In the survey, total 210 questionnaires were issued, and 208 questionnaires were received within one week. Seventeen questionnaires that contained inconsistent answers or incomplete information filling were removed from the dataset. Consequently, 191 valid questionnaires were collected.

In the collected dataset, the proportions of males and females are 40.3% and 59.7%, respectively. The age of surveyed users ranges from 18 to 40 years old, and most of the users, i.e., 80.6%, are between 18 and 25 years old. Ninety percent of users have undergraduate or graduate degrees.

4.3. Reliability and Validity

We first use the SPSS tool to validate the reliability and validity of the collected data. The internal-consistency reliability reflects the stability of individual measurement items across replications from the same information source. This kind of reliability is assessed by computing Cronbach's α, whose coefficients for the eight constructs are over 0.6, indicating a reasonable level of internal consistency among the items [42]. The validity test is to examine the authenticity of the subjects. The analysis results of the reliability and validity of the collected data are shown in Table 3. It shows that the Cronbach's α of all variables are over 0.8, and the overall Cronbach's α is 0.935, indicating that the scale system is highly reliable. All the values of KMO of the variables are greater than 0.7, except the value of KMO of PU. The significant level of Bartlett's test is less than 0.05, which shows that the factor analysis method is applicable to the questionnaire.

Table 3. Reliability and validity analysis on variables.

Variable	Cronbach's α	KMO Value	Bartlett's Test		
			Approximate Chi-Square	Freedom	Significance
PU	0.841	0.800	313.315	6	<0.001
PEOU	0.833	0.687	235.547	3	<0.001
Trust	0.875	0.848	462.343	10	<0.001
Attitude	0.885	0.825	414.888	6	<0.001
BI	0.859	0.810	340.999	6	<0.001

Cronbach's α = 0.935.

5. Data Analysis

In this section, we perform data analysis on the collected data to evaluate the hypotheses. In Section 5.1, we use the principal component analysis and maximum likelihood method to carry out factor analysis to ensure the design rationality of model variables. In Section 5.2, we perform correlation analysis to find out the correlations among the independent variables, the mediator variable, and the dependent variable. Correlation analysis refers to the analysis on two or more correlated variable elements to measure the closeness of the variable factors. In Section 5.3, we conduct regression analysis to find out the causal relationship between factors of the research model, including PU, PEOU, trust, attitude, and BI. In Section 5.4, we measure the mediating effects of attitude and PU in the research model. In Section 5.5, we test the interferential effects of the control variables based on different types of sample data.

5.1. Factor Analysis

Factor analysis aims to find out the number of factors that affect the observed variables, as well as the correlations between each factor and each of the observed variables in an attempt to reveal the inherent structure of a relatively large set of variables. In this subsection, we will use the SPSS software to perform factor analysis on the collected dataset. Particularly, we use the maximum likelihood method and the principal component analysis method to analyze the exploratory factors.

5.1.1. Independent Variables

Exploratory factor analysis on independent variables is shown in Table 4. The cumulative contribution rate of the three factors is 70%, indicating that the content of the questionnaire can be well explained by the three factors. The orthogonal rotation is performed by using the maximum variance rotation method. The results are shown in Table 5.

Table 4. Variance of independent variables.

Component	Initial Eigenvalue			Sum of Squares of Extracted Load		
	Total	Percentage Variance	Cumulative Percentage	Total	Percentage Variance	Cumulative Percentage
1	5.518	45.984	45.984	5.518	45.984	45.984
2	1.524	12.700	58.684	1.524	12.700	58.684
3	1.357	11.311	69.995	1.357	11.311	69.995
4	0.615	5.128	75.123			
5	0.580	4.833	79.955			
6	0.491	4.090	84.046			
7	0.453	3.776	87.822			
8	0.358	2.987	90.809			
9	0.347	2.895	93.704			
10	0.295	2.460	96.163			
11	0.267	2.229	98.392			
12	0.193	1.608	100.000			

Table 5. Factor loading matrix.

Questions	Component		
	PU	PEOU	Trust
Q101: If I were to adopt personalized business modes, it would enable me to purchase favorite products	0.816		
Q102: If I were to adopt personalized business modes, it would enable me to broaden my understanding of the product or service	0.827		
Q103: If I were to adopt personalized business modes, the effect of purchasing goods or services would improve	0.691		
Q104: If I were to adopt personalized business modes, it would satisfy my personal consumption needs	0.791		
Q201: Learning how to participate in personalized business would be easy for me.		0.854	
Q202: Participation in enterprise personalized business takes less time and effort		0.904	

Table 5. *Cont.*

Questions	Component		
	PU	PEOU	Trust
Q203: In the process of enterprise personalized business, enterprises provide clear and easy communication		0.697	
Q301: I'll think products or services purchased by personalized business modes will be trustworthy			0.773
Q302: I believe that companies will take into account customer needs in the process of customization			0.750
Q303: I believe the expected effect of customized products or services is predictable			0.774
Q304: I believe that the after-sale service of personalized business modes is guaranteed			0.810
Q305: I believe the organizers will provide meticulous service to help me solve all kinds of problems in the process of customization			0.730

5.1.2. Mediating Variable

The mediating variable is explored by factorial analysis to explain the total variance of the variable. The results are shown in Table 6. The cumulative contribution rate of one factor is 74.397%, which can well explain the contents of the original questionnaire.

Table 6. Variance of the mediating variable.

Component	Initial Eigenvalue			Sum of Squares of Extracted Load		
	Total	Percentage Variance	Cumulative Percentage	Total	Percentage Variance	Cumulative Percentage
1	2.976	74.397	74.397	2.976	74.397	74.397
2	0.431	10.780	85.177			
3	0.344	8.610	93.787			
4	0.249	6.213	100.000			

From the load matrix of the mediating variable, as shown in Table 7, we can see that the factor scores are greater than 0.8. This means that these factors can be classified into one class, which is exactly the attitude.

Table 7. Load matrix of the attitude factor.

Questions	Component
	Attitude
Q401: I think it's a good idea to take part in customization	0.832
Q402: I like to experience personalized business modes	0.864
Q403: Participating in personalized business modes will bring me a pleasant experience	0.893
Q404: I'll be positive about personalized business modes	0.860

5.1.3. Dependent Variable

The main variance of the dependent variables is explained by the factor analysis on the dependent variables. As shown in Table 8, the cumulative contribution rate of one factor is 70.475%, which can well explain the contents of the original questionnaire.

Table 8. Variance of the behavior intention variable.

Component	Initial Eigenvalue			Sum of Squares of Extracted Load		
	Total	Percentage Variance	Cumulative Percentage	Total	Percentage Variance	Cumulative Percentage
1	2.819	70.475	70.475	2.819	70.475	70.475
2	0.503	12.570	83.045			
3	0.351	8.782	91.827			
4	0.327	8.173	100.000			

The results of the variable factor load are shown in Table 9, which shows that the factor scores are greater than 0.8. Thus, they can be classified into one class, and this class is named as behavior intention according to the characteristics of the question.

Table 9. Load matrix of the behavior intention factor.

Questions	Component
	Behavior Intention
Q501: I'll intend to experience personalized business as soon as possible	0.857
Q502: I intend to continue to participate in the personalized business of the enterprise	0.832
Q503: I will recommend personalized business modes to my friends	0.850
Q504: I have an urge to participate in personalized business	0.819

5.2. Correlation Analysis

Correlation analysis is a statistical method that aims to reveal whether there is a relationship between variables. If there is a relationship, we need to quantify the strength of this relationship. As correlation analysis generally employs the Pearson coefficient to express the linear relationship among variables, in this paper we also use the Pearson correlation analysis method to analyze the correlation between PU, PEOU, trust, attitude, and BI. In particular, we use r to represent the Pearson correlation coefficient. A positive value of r indicates that the two variables tested are positive correlated, while a negative value of r means the two variables have a negative correlation. When $-1 < r < 1$, the greater the absolute value of r, the greater the correlation between the two variables. Table 10 shows the mean value of each variable according to the Pearson analysis.

Table 10. Overall results of the correlation analysis.

	PU	PEOU	Trust	Attitude	BI
Overall Mean	4.045	3.506	3.699	3.952	3.598
Standard Deviation	0.684	0.795	0.705	0.673	0.815

As shown in Table 10, the mean value of behavior intention is 3.598, indicating high tendency of users' willingness to participate in personalized business. The mean value of perceived usefulness is 4, indicating that perceived usefulness is more sensitive to the perception of the user. The mean value of perceived ease of use is 3.506, which is consistent with the relatively low sensitivity of users to technology in the age of information technology.

5.2.1. Independent Variables and Attitude

Table 11 shows the correlation analysis results between the three independent variables and the mediator variable (Attitude) is analyzed.

Table 11. Correlation analysis (two-tailed) between independent variables and attitude.

Independent Variable	Index	Mediation Variable (Attitude)
PU	Pearson correlation Significant (two-tailed)	0.604 <0.001
PEOU	Pearson correlation Significant (two-tailed)	0.407 <0.001
Trust	Pearson correlation Significant (two-tailed)	0.682 <0.001

In Table 11, a value less than 0.001 means that the actual value is too small to be correctly shown in the SPSS software. For such values, SPSS always output 0.000 by default.

Table 11 shows that the correlation coefficients of perceived usefulness, perceived ease of use, and attitude towards trust are 0.604, 0.407 and 0.682, respectively. The Pearson correlation coefficients of the three independent variables and the mediator variables are all positive. The significant is at the 0.001 level, meaning that the correlation between attitude and PU, PEOU, and trust is positive. The correlation between perceived ease of use and attitude is not tighter than that between the other two factors, indicating that in the digital age most users are not sensitive to the ease-of-use of personalized business. On the other hand, both perceived usefulness and trust have a high impact on attitude. Thus, Hypothesis 2, 3 and 9 are validated.

5.2.2. Attitude and Behavior Intension

The correlation analysis between attitude and behavior intention is shown in Table 12. The correlation coefficient of attitude towards behavior intention is 0.681. The significance probability of attitude and behavior intention is less than 0.001, which is significant at the 0.001 level. We can see from the table that attitude and behavior intention is positively correlated. The correlation coefficient of attitude towards behavior intention is higher than that between perceived usefulness and perceived ease of use. This shows that attitude has a high impact on behavior intention. As attitude and behavior intention are both subjective, we can see that users' attitude towards personalized business directly affects behavior intention. Thus, Hypothesis 4 is validated.

Table 12. Correlation analysis between attitude and behavior intention.

Independent Variable	Index	Dependent Variable
Attitude	Pearson correlation Significant (two-tailed)	0.681 <0.001

5.2.3. Perceived Ease of Use and Perceived Usefulness

The correlation analysis between perceived ease of use and perceived usefulness and is shown in Table 13. The correlation coefficient of perceived ease of use for perceived usefulness is 0.401. The significance probability is less than 0.001, which is significant at the 0.001 level. The correlation coefficient of perceived ease of use for perceived usefulness shows that users' usefulness assessment for personalized business is partly derived from the irritation of usability. The less the difficulty of personalized business is, the less time and effort users need to participate in personalized business. Consequently, we can see from the table that perceived ease of use and perceived usefulness are positively correlated. This indicates that Hypothesis 1 is validated.

Table 13. Correlation analysis between perceived ease of use and perceived usefulness.

Independent Variable	Index	Dependent Variable (PU)
PEOU	Pearson correlation Significant (two-tailed)	0.401 <0.001

5.3. Regression Analysis

In this section, we use SPSS software to explain whether there is a causal relationship between the variables. We use the regression analysis as the basic tool for revealing causal relationships. Regression analysis can explore the magnitude of the influence between variables. It can also explore the direction of the impact between variables. If two or more variables have a causal relationship, the regression equation is expected to be established between them. In our study, we use the hierarchical regression method to detect the mediating effect between the argument and the dependent variable.

5.3.1. Perceived Usefulness and Attitude

We first perform regression analysis on perceived usefulness and attitudes. The results are shown in Table 14, which shows that the standardized regression coefficient of perceived usefulness is 0.604. The *F* value is 108.459, showing that the *F* test is passed. Table 4 shows that the regression effect is significant and the regression coefficient is positive. It indicates that users' perceived usefulness for personalized business has a significant positive impact on attitude when other factors remain unchanged, meaning that the original Hypothesis 3 is validated. The value of the Adjusted R^2 is 0.361, which shows that the explanatory power of perceived usefulness to attitude is 36.1%. This means that perceived usefulness has a significant positive effect on attitude of users to participate in personalized business. When perceived usefulness has a high value, users are much likely to have a positive attitude to participate in personalized businesses.

Table 14. Regression analysis on perceived usefulness and attitude.

Variable	Standardized Coefficient	
	β	Significance
PU	0.604	<0.001
F	108.459	
R^2	0.365	
Adjusted R^2	0.361	
N	191	

5.3.2. Perceived Ease of Use and Attitude

In this section, we report the results of regression analysis on perceived ease of use and attitude. The results are shown in Table 15. We can conclude from the results that perceived ease of use has a positive effect on attitude. The standardized regression coefficient of perceived ease of use is 0.407. The *F* value is 37.546, showing that the *F* test is passed. Table 15 shows that the regression effect is significant and the regression coefficient is positive. It shows that users' perceived ease of use for personalized business has a significant positive impact on attitude. The value of the Adjusted R^2 is 0.161, which shows that perception is easy to use and its explanatory power to attitude is 16.1%. In summary, Hypothesis 2 is validated.

Perceived ease of use positively affects the attitude of users participating in personalized businesses, because the less difficulty of personalized business will make users easier to use and participate in personalized businesses, yielding a positive attitude.

Table 15. Regression analysis on perceived ease of use and attitude.

Variable	Standardized Coefficient	
	β	Significance
PEOU	0.407	<0.001
F	37.546	
R^2	0.166	
Adjusted R^2	0.161	
N	191	

5.3.3. Trust and Attitude

Next, we perform regression analysis on trust and attitude. The results are shown in Table 16. In Table 16, we can conclude that trust has a positive influence on attitude. The standardized regression coefficient of trust is 0.4682. The *F* value is 164.128, showing that the *F* test is passed. Table 16 shows that the regression effect is significant and the regression coefficient is positive. It shows that users' trust in personalized business has a significant positive impact on attitude in the absence of other factors. The value of the Adjusted R^2 is 0.462, showing that the explanatory power of trust to attitude is 46.2%. From this, we can see that Hypothesis 9 is validated.

Table 16. Regression analysis on Trust and attitude.

Variable	Standardized Coefficient	
	β	Significance
Trust	0.682	<0.001
F	164.128	
R^2	0.465	
Adjusted R^2	0.462	
N	191	

Trust is users' psychological dependence on enterprises, and the attitude of participation is also users' psychological tendency. Thus, it is understandable that trust has a positive impact on attitude. In other words, when users trust a company a lot, they are much likely to have a positive attitude to accept personalized businesses offered by the company.

5.4. Mediating Effects Analysis

In this section, we use the hierarchical regression method to verify the mediating effect of attitude between different factors. We first present the mediating effect analysis of perceived usefulness and behavior intention in Section 5.4.1. Then, we discuss the mediating effect analysis of perceived ease of use and behavior intention in Section 5.4.2. Section 5.4.3 presents the mediating effect analysis of trust and behavior intention. Section 5.4.4 discusses the mediating effect of perceived usefulness on perceived ease of use and attitude. Finally, in Section 5.4.5, we present the mediating effect of perceived usefulness on perceived ease of use and behavior intention.

In the following subsections, we first perform the regression analysis of the direct variables on the mediating variable, and then perform the regression analysis of the direct variables on the target variable. Finally, we select the direct variables and the mediating variable as the independent variables to conduct the regression analysis on the target variable. If the target variable is weakened by the direct variables, we can conclude that the mediating effect exists.

5.4.1. Attitude on Perceived Usefulness and Behavior Intention

In this study, we take perceived usefulness as the independent variable and behavior intention as the dependent variable. Then, we conduct regression analysis on perceived usefulness and behavior intention. The results are shown in Table 17. As shown in the table, the Adjusted R^2 is 0.276, meaning that the explanatory power of perceived usefulness to behavior intention is 27.6%. This is also validated by the *F* test and the *T* test.

Table 17. Regression analysis on perceived usefulness and behavior intention.

Independent Variable	R^2	Adjusted R^2	*F*	Sig.	Standardized Coefficient β	*T*	Sig	VIF
PU	0.280	0.276	73.387	<0.001	0.529	8.567	<0.001	1

Next, we take perceived usefulness and attitude as the independent variables and behavior intention as the dependent variable. The results of the regression analysis on perceived usefulness, attitude, and behavior intention are shown in Table 18. Here, the Adjusted R^2 is 0.481, which is validated by the *F* test and the *T* test. Based on Tables 17 and 18, we can see that the Adjusted R^2 in Table 18 increases by 20.5% compared with that in Table 17. In addition, the standardized regression coefficient β of perceived usefulness is changed from 0.529 in Table 16 to 0.185 in Table 18. Although both values of β in the two tables are statistically significant at the 0.001 level, the fact of $0.185 < 0.529$ indicates that the effect of perceived usefulness on behavior intention decreases. The value of VIF for perceived usefulness and attitude is less than 10, meaning that there is no serious collinearity. As a result, attitude plays a partial mediating role between perceived usefulness and behavior intention; thus Hypothesis 6 is established.

Table 18. Regression analysis on perceived usefulness, attitude, and behavior intention.

Independent Variable	R^2	Adjusted R^2	*F*	Sig.	Standardized Coefficient β	*T*	Sig	VIF
PU Attitude	0.486	0.481	88.913	<0.001	0.185 0.570	2.816 8.690	<0.001 0.005	1.574 1.574

According to the results, we can see that users first have a stage of attitude before they participate in personalized businesses. That is, perceived usefulness first stimulates users' attitude towards personalized businesses. A positive attitude will lead to high willingness to participate in personalized businesses. The results of analysis also show that perceived usefulness can directly affect users' willingness to participate in personalized businesses.

5.4.2. Attitude on Perceived Ease of Use and Behavior Intention

The regression analysis on perceived ease of use and behavior intention is carried out by setting perceived ease of use as the independent variable and behavior intention as the dependent variable. The results are shown in Table 19. In this regression analysis, the Adjusted R^2 is 0.168, indicating that the explanatory power of perceived ease of use to behavior intention is 16.8%, which is validated by the *F* test and the *T* test.

Table 19. Regression analysis on perceived ease of use and behavior intention.

Independent Variable	R^2	Adjusted R^2	*F*	Sig.	Standardized Coefficient β	*T*	Sig	VIF
PEOU	0.172	0.168	39.325	<0.001	0.415	6.271	<0.001	1

Next, we use perceived ease of use and attitude as independent variables. Behavior intention is used as the dependent variable. The results of the regression analysis of perceived ease of use, attitude and behavior intention are shown in Table 20. The Adjusted R^2 is 0.482, which is validated by the

F test and the *T* test. In Tables 18 and 20, we can see that the Adjusted R^2 in Table 20 increases by 31.4% compared with that in Table 19. In addition, the standardized regression coefficient β of perceived ease of use is changed from 0.415 in Table 19 to 0.165 in Table 20. As 0.165 < 0.415, we can see that the effect of perceived ease of use on behavior intention is weakened. Further, the value of VIF for perceived ease of use and attitude is less than 10, showing that there is no serious collinearity. As a result, attitude plays a partial mediating role between perceived ease of use and behavior intention, which shows that Hypothesis 5 is established.

Table 20. Regression analysis on attitude, perceived ease of use, and behavior intention.

Independent Variable	R^2	Adjusted R^2	*F*	Sig.	Standardized Coefficient β	*T*	Sig	VIF
PEOU Attitude	0.487	0.482	88.274	<0.001	0.165 0.614	2.884 10.743	0.004 <0.001	1.199 1.199

The above results show that users need to take a stage of behavior attitude before they take the actual behavior. Perceived ease of use first motivates users' attitude towards personalized businesses, and a positive acceptance attitude can make users increase their behavior intention. Attitude plays a partly intermediary role, indicating that perceived ease of use can directly affect users' behavior intention. When personalized businesses are easy to understand and to use, it will lower the cost of users to participate in personalized businesses and does not require users to evaluate the quality of customized products or services. To this end, the mediating effect of users' participation attitude is weakened.

5.4.3. Attitude on Trust and Behavior Intention

The regression analysis on the mediating effects of attitude on trust and behavior information is carried out by setting behavior information as the dependent variable and trust as the independent variable. The results are shown in Table 21. In this regression analysis, the Adjusted R^2 is 0.397. This means that trust's interpretation power on BI is 39.7%, which is validated by the *F* test and the *T* test.

Table 21. Regression analysis on behavior information and trust.

Independent Variable	R^2	Adjusted R^2	*F*	Sig.	Standardized Coefficient β	*T*	Sig	VIF
trust	0.397	0.394	124.293	<0.001	0.630	11.149	<0.001	1

Next, we use trust and attitude as the independent variables and conduct regression analysis on trust, attitude, and behavior information. The results are shown in Table 22. The Adjusted R^2 is 0.515, which is validated by the *F* test and the *T* test. From Tables 21 and 22, we can see that the Adjusted R^2 in Table 22 increases by 11.6% compared with that in Table 21. Meanwhile, the standardized regression coefficient β of trust is changed from 0.630 in Table 21 to 0.309 in Table 22. Both values of β are statistically significant at the 0.001 level, but, as 0.309 < 0.630, we can infer that the effect of trust on behavior intention decreases. The value of VIF for trust and attitude is also less than 10, showing that there is no serious collinearity. As a result, attitude plays a partial mediating role between trust and behavior intention, indicating that Hypothesis 10 is established.

Table 22. Regression analysis on behavior information, trust, and attitude.

Independent Variable	R^2	Adjusted R^2	*F*	Sig.	Standardized Coefficient β	*T*	Sig	VIF
trust attitude	0.515	0.510	99.994	<0.001	0.309 0.471	4.450 6.787	<0.001 <0.001	1.868 1.868

Due to psychological similarity, users' trust in personalized businesses is closely connected with users' subjective attitude. Therefore, trust is very likely to affect users' behavior intention through user attitude. The analyzing results show that attitude plays a partial mediating role. A possible reason is that firms or brands trusted by users can reduce user perceived risks for personalized businesses, which can also urge users participate in personalized businesses.

5.4.4. Perceived Usefulness on Perceived Ease of Use and Attitude

First, we perform the regression analysis on perceived ease of use and attitude by setting attitude as the dependent variable and perceived ease of use as the independent variable. The analysis results are shown in Table 23. The Adjusted R^2 is 0.166, that is, the perceived ease of use interpretation of attitude is 11.6%, which is validated by the *F* test and the *T* test.

Table 23. Regression analysis on perceived ease of use and attitude.

Independent Variable	R^2	Adjusted R^2	*F*	Sig.	Standardized Coefficient β	*T*	Sig	VIF
PEOU	0.166	0.161	37.546	<0.001	0.407	6.128	<0.001	1

Next, we set perceived usefulness and perceived ease of use as independent variables to measure their mediating effects on attitude. The results of regression analysis are shown in Table 24. The Adjusted R^2 is 0.397, which is validated by the *F* test and the *T* test. In Tables 23 and 24, we can see that the Adjusted R^2 in Table 24 increased by 23.1% compared with the Adjusted R^2 in Table 23. Meanwhile, the standardized regression coefficient β of perceived ease of use is changed from 0.407 in Table 23 to 0.197 in Table 24. Both values of the standardized regression coefficient β are statistically significant at the 0.001 level. However, the fact $0.197 < 0.407$ indicates that the effect of perceived ease of use on attitude has been weakened. The VIF value for perceived ease of use and Perceived usefulness is less than 10, meaning that there is no serious collinearity. As a result, perceived usefulness plays a partial mediating role between perceived ease of use and attitude, assuming that Hypothesis 7 is established.

Table 24. Regression analysis on perceived usefulness, perceived ease of use, and attitude.

Independent Variable	R^2	Adjusted R^2	*F*	Sig.	Standardized Coefficient β	*T*	Sig	VIF
PEOU PU	0.397	0.391	61.920	<0.001	0.197 0.525	3.184 8.495	0.002 <0.001	1.191 1.191

The results show that perceived usefulness is partly mediated by perceived ease of use and attitude. When personalized businesses are relatively easy for users to use, users can be stimulated to participate in personalized businesses, leading to a positive attitude towards personalized businesses. However, if personalized businesses of an enterprise do not bring expected benefits to users, users may not have high behavior intention even if personalized businesses are easy to use.

5.4.5. Perceived Usefulness on Perceived Ease of Use and Behavior Intention

First, we perform the regression analysis on perceived ease of use and behavior information by setting behavior information as the dependent variable and perceived ease of use as the independent variable. The analysis results are shown in Table 25. The Adjusted R^2 is 0.168, showing that the perceived ease of use interpretation of behavior information is 16.8%, which is validated by the *F* test and the *T* test.

Table 25. Regression analysis on perceived ease of use and attitude.

Independent Variable	R^2	Adjusted R^2	*F*	Sig.	Standardized Coefficient β	*T*	Sig	VIF
PEOU	0.172	0.168	39.325	<0.001	0.415	6.271	<0.001	1

Next, we set perceived usefulness and perceived ease of use as the independent variables to measure their mediating effects on behavior intention. The results of regression analysis are shown in Table 26. The Adjusted R^2 is 0.322, which is validated by the *F* test and *T* test. From Tables 25 and 26 we can see that the Adjusted R^2 in Table 26 increases by 15.4% compared with that in Table 25. Meanwhile, the standardized regression coefficient β of perceived ease of use is changed from 0.415 in Table 25 to 0.242 in Table 26. Both values of the standardized regression coefficient β are statistically significant at the 0.001 level. However, the fact of $0.242 < 0.415$ indicates that the effect of perceived ease of use on behavior intention has been weakened. The VIF value for perceived usefulness and perceived ease of use is less than 10, meaning that there is no serious collinearity. As a result, perceived usefulness plays a partial mediating role between perceived ease of use and attitude, assuming that Hypothesis 7 is established.

Table 26. Regression analysis on perceived usefulness, perceived ease of use, and attitude.

Independent Variable	R^2	Adjusted R^2	F	Sig.	Standardized Coefficient β	*T*	Sig	VIF
PEOU	0.329	0.322	46.057	<0.001	0.242	3.711	<0.001	1.191
PU					0.432	6.623	<0.001	1.191

Perceived usefulness plays a partial mediating role between perceived ease of use and behavior intention. This result has two implications. First, perceived ease of use can reduce the cost of users to participate in personalized businesses, which can lead to high behavior intention. Second, even if personalized businesses are easy to use, it will not always satisfy user needs unless personalized businesses are demonstrated to be useful for users.

5.5. Control-Variable Analysis

User participation in personalized business may be related to some user attributes such as gender and educational background. Thus, in this section we conduct a comparative study to measure the influence of user attributes on behavior intention.

The control variables in this study include gender and educational background. We use the SPSS software (SPSS Inc., Chicago, IL, USA) to test the interferential effects of these control variables in different types of sample data. The results of non-standardized coefficient are shown in Table 27. We can see that only gender has a significant moderating effect on perceived usefulness and behavior intention, while others are not influential. Compared with the female group, the perceived usefulness of the male group has a higher impact on behavior intention. The fact $0.672 > 0.398$ shows that male users pay more attention to the usefulness of personalized businesses, while females are greatly influenced by psychological trust and perceived ease of use, which will reduce the importance of perceived usefulness.

Table 27. Analysis results of control variables including gender and education background.

Control Variable	Relationship	Group	Standardized Coefficient		Z
			β	Number of People	
Gender	PU→BI	male female	0.672 0.398	77 114	2.619
	Attitude→BI	male female	0.701 0.663	77 114	1.142
	PEOU→BI	male female	0.309 0.495	77 114	−0.085
	Trust→BI	male female	0.675 0.589	77 114	0.958
Education Background	PU→BI	bachelor degree and below master and above	0.576 0.377	116 75	1.72
	Attitude→BI	bachelor degree and below master and above	0.641 0.609	116 75	0.348
	PEOU→BI	bachelor degree and below master and above	0.446 0.364	116 75	0.651
	Trust→BI	bachelor degree and below master and above	0.700 0.626	116 75	0.879

5.6. Summary of Hypothesis Validation

By combing the results described in Sections 5.1–5.4, we present the summary of the hypothesis validation in Table 28.

Table 28. Summary of hypothesis validation.

Number	Hypothesis	Validation Result
H1	Perceived ease of use positively affects perceived usefulness	established
H2	Perceived ease of use positively influences attitudes	established
H3	Perceived usefulness positively influences attitude	established
H4	Attitudes positively influence behavior intention	established
H5	Attitude plays an intermediary role between perceived ease of use and behavior intention	partial intermediary
H6	Attitude plays an intermediary role between perceived usefulness and behavior intention	partial intermediary
H7	Perceived usefulness plays a mediating role between perceived ease of use and attitude	partial intermediary
H8	Perceived usefulness plays an intermediary role between perceived ease of use and behavior intention	partial intermediary
H9	Trust positively influences attitude	established
H10	Attitude plays an intermediary role between trust and behavior intention	partial intermediary

The validation result column in Table 28 shows the final validation results of each hypothesis, from which we can see that H1–H4 and H9 are well established. These hypotheses correspond to research questions Q1–Q3, as presented in Table 1. This implies that the independent variables of the research model we propose in Section 3.1 are influential to user acceptance of personalized business modes. Among all the factors, perceived usefulness and perceived ease of use are two ones reflecting the advantages of personalized businesses. It is understandable that a useful and easy-to-use personalized business mode can attract users' interests and finally lead to users' participation behavior in personalized businesses. The factor of trust depends on both the properties of personalized businesses and specific user attributes, but we can see that trust has a positive impact on behavior intention. In other words, if users trust a specific company, they are very likely to participate in the personalized business modes provided by the company.

Regarding Hypotheses 5–8 and 10, which aim to answer research question Q4, we can see that all selected factors have partial intermediary effect on user acceptance of personalized business modes. Partial intermediary effect means that the mediating factor does have effect on the target factor, but the final behavior of user acceptance of personalized business modes are also influenced by other factors. Thus, enterprises should pay more attention to the positive effect of mediating factors and build a framework to make all factors work for the advancement of personalized businesses.

6. Discussion

6.1. Research Implications

(1) The study of this paper is based on the background of personalized business modes in the digital age, where electronic commerce and online shopping has become a part of people's daily life. Due to the variety of users, it is necessary to exploit personalized business modes. This paper aims to answer several research questions by modeling and quantifying user acceptance of personalized business modes based on questionnaire data. We propose a research model based on the integration of the TAM model, trust, and attitude. This model augments the application of the widely used TAM model and offers referential values for other related research. In addition, we present empirical results on user acceptance of participating in personalized businesses. These results can provide new research insights for advancing personalized business modes, e.g., designing operational schemes for personalized businesses.

(2) This paper studies users' behavior intention in personalized business modes, and is valuable for enterprises to realize the importance of developing personalized business modes. Enterprises need to pay more attention to personalized business modes. With the rapid development of economics and social businesses, users' business demands have changed a lot. More and more people want to personalize their business modes so that they can get better experiences in business and learning activities. User acceptance of personalized business modes reflects the broad market prospects for customization. Thus, enterprises need to keep reforming their development models to meet user needs.

(3) We introduce trust as a new factor influencing personalized businesses. Trust is an important indicator of consumer behavior in the Internet environment, and we incorporate trust into the research model to model user acceptance of personalized business modes. The empirical research in this paper shows that trust will affect users' attitude and behavior intention on personalized business modes. Therefore, a sound trust mechanism is needed for advancing personal business modes. To attract more users to participate in personalized businesses, one key issue is to eliminate users' concerns about personalized business modes. On the one hand, companies can strive to improve their brand images, thereby enhancing users' brand loyalty. On the other hand, enterprises should abide by the commitment to users, to ensure that users participate in customized businesses in a timely manner to obtain convenient services.

(4) We find that attitude plays an intermediary role among perceived usefulness, perceived ease of use, trust and behavior intention. Perceived usefulness and perceived ease of use can affect users' behavior intention by using attitude as the mediating factor. Further, they can directly affect users' behavior intention. Therefore, enterprises should pay more attention to perceived ease of use and perceived usefulness when starting personalized businesses. On the one hand, users' profit in personalized business modes should be enhanced. It is necessary in personalized businesses to let users get tangible benefits from personalized businesses. Users may expect personalized business modes to facilitate their own consumption, and to improve their quality of consumption as well. There are also some other aspects that may bring extra benefits to users exploiting personalized businesses.

However, enterprises have to carefully design the specific schemes and systems to better assist users to carry out personalized businesses. On the other hand, enterprises need to improve their services and reduce the difficulty of public participation. In the empirical study, consumers' perceived ease of use of personalized businesses has a significant impact on users' attitude and behavior intention to participate in personalized businesses. Thus, it is better for enterprises to establish a good channel for communication and feedbacks in the process of personalized businesses. Through such a channel, users can communicate with service providers and express their comments on business modes. Enterprises can also provide support for users to self-define customization details, such as user interface, modules, and processes, to offer better experiences for users when carrying out personalized businesses.

(5) The data used in this paper mainly come from a survey on young people. Young people live, learn and entertain in the digitalized environment since their childhood. Their familiarity and dependence on the Internet are significantly higher than other old people. Our study shows that the demand of young consumers for products is no longer satisfied with traditional business modes, which calls for personalized business modes. The characteristics of young consumers also make this study practicable and applicable, e.g., on electricity consumption for youth groups.

(6) The study in this paper finds that attitude has a central role on user acceptance of personalized business modes. Thus, it is very important for enterprises to consider users' attitude when advancing personalized businesses. Enterprises should not only improve their personalized services or products, but also focus on the brand image of personalized businesses in the market. Our study has shown that users' trust on brands will lead to a positive attitude towards personalized business modes, which will in turn lead to participation behavior in personalized businesses.

6.2. Suggestions

Based on the empirical analysis in this study, we further make the following suggestions for enterprises to better develop personal business modes:

(1) Enterprises need to pay more attention to personalized business modes.

With the rapid development of economics and social businesses, users' business demands have changed a lot. More and more people want to personalize their business modes so that they can get better experiences in business and learning activities. User acceptance of personalized business modes reflects the broad market prospects for customization. Thus, enterprises need to keep reforming their development models to meet user needs.

(2) A sound trust mechanism is needed for advancing personal business modes.

The empirical research in this paper shows that trust will affect users' attitude and behavior intention on personalized business modes. To attract more users to participate in personalized businesses, one key issue is to eliminate users' concerns about personalized business modes. On the one hand, companies can strive to improve their brand images, thereby enhancing users' brand loyalty. On the other hand, enterprises should abide by the commitment to users, to ensure that users participate in customized businesses in a timely manner to obtain convenient services.

(3) Users' profit in personalized business modes should be enhanced.

It is necessary in personalized businesses to let users get tangible benefits from personalized businesses. Users may expect personalized business modes to facilitate their own consumption, and to improve their quality of consumption as well. Some other aspects might bring extra benefits to users exploiting personalized businesses. However, enterprises have to carefully design the specific schemes and systems to better assist users to carry out personalized businesses.

(4) Enterprises need to improve their services and reduce the difficulty of public participation.

In the empirical study, consumers' perceived ease of use of personalized businesses has a significant impact on users' attitude and willingness to participate in personalized businesses. Thus, it is better for enterprises to establish a good channel for communication and feedbacks in the process of personalized businesses. Through such a channel, users can communicate with service providers and express their comments on business modes. Enterprises can also provide supports for users to self-define customization details, such as user interface, modules, and processes, to offer better experiences for users when carrying out personalized businesses.

7. Conclusions

In this paper, we analyze the major factors that influence user acceptance of personalized business modes. In particular, we propose a research model based on the TAM (Technology Acceptance Model) model to study user acceptance of personalized business modes via some tailor-made independent variables, including perceived usefulness, perceived ease of use, and trust. We also introduce a mediating factor, i.e., attitude, to reflect the influence of independent variables on the dependent variable named behavior intention. Based on the research model, we use the structural-equation method to conduct an empirical analysis on questionnaire data from the Internet. We present comprehensive results from various aspects including reliability and validity, factor analysis, correlation analysis, regression analysis, mediating effects analysis and control-variable analysis. The results show that perceived usefulness, trust, and perceived ease of use have significant influence on user acceptance of personalized business modes, and perceived ease of use impacts perceived usefulness. In addition, gender plays an adjusting role between perceived usefulness and perceived ease of use.

Acknowledgments: We would like to thank the editors and anonymous reviewers for their suggestions and comments to improve the quality of the paper. This paper is partially supported by the National Science Foundation of China (No. 71273010 and No. 61672479) and the Doctor Start-up Fund of Anhui University. We also thank Jianfei Wang for his advice on the revision of this paper.

Author Contributions: Jie Zhao designed this study and drafted the paper in English; Suping Fang conceived and designed the experiments; and Peiquan Jin revised the paper critically for important intellectual contents.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. China Internet Network Information Center (CNNIC). *The Thirty-Ninth China Statistical Report on the Development of the Internet*. Available online: https://www.cnnic.net.cn/hlwfzyj/hlwxzbg/201601/P020160122469130059846.pdf (accessed on 19 October 2017).
2. Choi, J.; Lee, D. The influence of purchasing context and reversibility of choice on consumer responses toward personalized products and standardized products. *Proc. Psychol. Rep.* **2016**, *118*, 510–526. [CrossRef] [PubMed]
3. Liu, S.; Liu, F. Network sales and customization system design of ceramic products. *Comput. Integr. Manuf. Syst.* **2002**, *8*, 1548–1550.
4. Davis, F. Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIS Q.* **1989**, *13*, 319–340. [CrossRef]
5. Meyer, A.; Ferdows, K. Integration of information systems in manufacturing. *Int. J. Oper. Prod. Manag.* **1985**, *5*, 5–12. [CrossRef]
6. Helander, M.; Jiao, J. Research on E-product development (ePD) for mass customization. *Technovation* **2002**, *22*, 717–724. [CrossRef]
7. Walsh, J.; Roche, D.; Foping, F. NitroScript: A PHP template engine for customizing of E-commerce applications. In Proceedings of the 2012 International Conference for Internet Technology and Secured Transactions, London, UK, 10–12 December 2012.

8. Asenov, D.; Müller, P. Customizing the visualization and interaction for embedded domain-specific languages in a structured editor. In Proceedings of the 2013 IEEE Symposium on Visual Languages and Human Centric Computing, San Jose, CA, USA, 15–19 September 2013.
9. Satyen, U.; Love, W. How to customize digital avionics but spend less. In Proceedings of the 16th DASC AIAA/IEEE Conference on Digital Avionics System, Irvine, CA, USA, 30 October 1997.
10. Fogliatto, F.; Silveira, G.; Borenstein, D. The mass customization decade: An updated review of the literature. *Int. J. Prod. Econ.* **2012**, *138*, 14–25. [CrossRef]
11. Wang, Y.; Ma, H. Industry 4.0: A way from mass customization to mass personalization production. *Adv. Manuf.* **2017**, *5*, 311–320. [CrossRef]
12. Yang, D.; Miao, R.; Wu, H.; Zhou, Y. Product configuration knowledge modeling using ontology web language. *Expert Syst. Appl.* **2009**, *36*, 4399–4411. [CrossRef]
13. Yang, D.; Dong, M. A constraint satisfaction approach to resolving product configuration conflicts. *Adv. Eng. Inf.* **2012**, *26*, 592–602. [CrossRef]
14. Wu, D.; Zhang, G.; Lu, J.A. Fuzzy preference tree-based recommender system for personalized business-to-business E-services. *IEEE Trans. Fuzzy Syst.* **2015**, *23*, 29–43. [CrossRef]
15. Kim, H.; Lee, Y.; Yim, H.; Cho, N. Design and implementation of a personalized business activity monitoring system. In Proceedings of the International Conference on Human-Computer Interaction, Beijing, China, 22–27 July 2007.
16. Jiao, J.; Helander, M. Development of an electronic configure-to-order platform for customized product development. *Comput. Ind.* **2006**, *57*, 231–244. [CrossRef]
17. Yang, Y.; Zhang, X.; Liu, F.; Xie, Q. An internet-based product customization system for CIM. *Robot. Comput. Integr. Manuf.* **2005**, *21*, 109–118. [CrossRef]
18. Grosso, C.; Forza, C.; Trentin, A. Supporting the social dimension of shopping for personalized products through online sales configurators. *J. Intell. Inf. Syst.* **2017**, *49*, 9–35. [CrossRef]
19. Zheng, P.; Yua, S. User-experience based product development for mass personalization: A case study. *Procedia CIRP* **2017**, *63*, 2–7. [CrossRef]
20. Tseng, M.; Jiao, J.; Merchant, M. Design for mass customization. *CIRP Ann. Manuf. Technol.* **1996**, *45*, 153–156. [CrossRef]
21. Lv, H.; Wan, Y.; Wu, F. Effect of online personalization services on complementary firms. *Electron. Commer. Res. Appl.* **2017**, *24*, 12–22. [CrossRef]
22. Lai, P. The literature review of technology adoption models and theories for the novelty technology. *J. Inf. Syst. Technol. Manag.* **2017**, *14*, 21–38.
23. Tarhini, A.; Arachchilage, N.; Masa'deh, R.; Abbasi, M. A critical review of theories and models of technology adoption and acceptance in information system research. *Int. J. Technol. Diffus.* **2015**, *6*, 58–77. [CrossRef]
24. Mehrabian, A.; Russell, J. The basic emotional impact of environments. *Percept. Motor. Skills* **1974**, *38*, 283–301. [CrossRef] [PubMed]
25. Fishbein, M.; Ajzen, I. *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*; Reading, M.A., Ed.; Addison-Wesley: Boston, MA, USA, 1975.
26. Shih, H. An empirical study on predicting user acceptance of e-shopping on the Web. *Inf. Manag.* **2004**, *41*, 351–368. [CrossRef]
27. Ha, S.; Stoel, L. Consumer e-shopping acceptance: Antecedents in a technology acceptance model. *J. Bus. Res.* **2009**, *62*, 565–571. [CrossRef]
28. Lu, H.; Su, P. Factors affecting purchase intention on mobile shopping web site. *Int. Res.* **2009**, *19*, 442–458. [CrossRef]
29. Gefen, D.; Karahanna, E.; Straub, D. Inexperience and experience with online stores: The importance of TAM and trust. *IEEE Trans. Eng. Manag.* **2003**, *50*, 307–321. [CrossRef]
30. Vijayasarathy, L. Predicting consumer intentions to use online shopping: The case for an augmented technology acceptance model. *Inf. Manag.* **2004**, *41*, 747–762. [CrossRef]
31. Ko, E.; Kim, E.; Lee, E. Modeling consumer adoption of mobile shopping for fashion products in Korea. *Psychol. Mark.* **2009**, *26*, 669–687. [CrossRef]
32. Xie, K.; Lee, Y. Social media and brand purchase: Quantifying the effects of exposures to earned and owned social media activities in a two-stage decision making model. *J. Manag. Inf. Syst.* **2015**, *32*, 204–238. [CrossRef]

33. Marakarkandy, B.; Yajnik, N.; Dasgupta, C. Enabling internet banking adoption: An empirical examination with an augmented technology acceptance model (TAM). *J. Enterp. Inf. Manag.* **2017**, *30*, 263–294. [CrossRef]
34. Wallace, L.; Sheetz, S. The adoption of software measures: A technology acceptance model (TAM) perspective. *Inf. Manag.* **2014**, *51*, 249–259. [CrossRef]
35. Yu, J.; Ha, I.; Choi, M.; Rho, J. Extending the TAM for a T-commerce. *Inf. Manag.* **2005**, *42*, 965–976. [CrossRef]
36. Gefen, D.; Karahanna, E.; Straub, D. Trust and TAM in online shopping: An integrated model. *MIS Q.* **2003**, *27*, 51–90. [CrossRef]
37. Heijden, H. Factors influencing the usage of websites: The case of a generic portal in The Netherlands. *Inf. Manag.* **2003**, *40*, 541–549. [CrossRef]
38. Hars, A.; Ou, S. Working for free? Motivations of participating in open source projects. *Int. J. Electron. Commer.* **2002**, *6*, 25–39.
39. Ciffolilli, A. Phantom authority, self-selective recruitment and retention of members in virtual communities: The case of Wikipedia. *First Monday* **2003**, *8*, 57–72. [CrossRef]
40. Venkatesh, V.; Thong, J.; Xu, X. Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Q.* **2012**, *36*, 157–178.
41. Koo, D. An investigation on consumer's internet shopping behavior explained by the technology acceptance model. *Asia Pac. J. Inf. Syst.* **2003**, *13*, 141–170.
42. Chang, K.; Yang, W.; Park, Y. An analysis on trust factors of B2C electronic commerce. *Inf. Policy* **2002**, *9*, 3–17. [CrossRef]
43. Che, H.; Yang, C. Examining wechat users' motivations, trust, attitudes, and positive world-of-mouth: Evidence from China. *Comput. Hum. Behav.* **2014**, *41*, 104–111.

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Sustainability Editorial Office
E-mail: sustainability@mdpi.com
www.mdpi.com/journal/sustainability

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34
Fax: +41 61 302 89 18

www.mdpi.com

ISBN 978-3-03928-009-4