entropy

Special Issue Reprint

Recent Trends and Developments in Econophysics

Edited by
Panos Argyrakis

mdpi.com/journal/entropy

MDPI

Recent Trends and Developments in Econophysics

Recent Trends and Developments in Econophysics

Editor

Panos Argyrakis

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Editor
Panos Argyrakis
Aristotle University
of Thessaloniki
Thessaloniki
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 *Entropy* (ISSN 1099-4300) (available at: https://www.mdpi.com/journal/entropy/special_issues/entropy_econophys).

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

Lastname, A.A.; Lastname, B.B. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-7258-0415-3 (Hbk)
ISBN 978-3-7258-0416-0 (PDF)
doi.org/10.3390/books978-3-7258-0416-0

Contents

Antonio Briola and Tomaso Aste
Dependency Structures in Cryptocurrency Market from High to Low Frequency
Reprinted from: *Entropy* **2022**, *24*, 1548, doi:10.3390/e24111548 . . . . . **1**

Charalampos M. Liapis, Aikaterini Karanikola and Sotiris Kotsiantis
Investigating Deep Stock Market Forecasting with Sentiment Analysis
Reprinted from: *Entropy* **2023**, *25*, 219, doi:10.3390/e25020219 . . . . . **19**

Takuya Wada, Hideki Takayasu and Misako Takayasu
Extraction of Important Factors in a High-Dimensional Data Space: An Application for High-Growth Firms
Reprinted from: *Entropy* **2023**, *25*, 488, doi:10.3390/e25030488 . . . . . **49**

Leticia Pérez-Sienes, Mar Grande, Juan Carlos Losada and Javier Borondo
The Hurst Exponent as an Indicator to Anticipate Agricultural Commodity Prices
Reprinted from: *Entropy* **2023**, *25*, 579, doi:10.3390/e25040579 . . . . . **72**

Hongli Niu, Kunliang Xu and Mengyuan Xiong
The Risk Contagion between Chinese and Mature Stock Markets: Evidence from a Markov-Switching Mixed-Clayton Copula Model
Reprinted from: *Entropy* **2023**, *25*, 619, doi:10.3390/e25040619 . . . . . **83**

Joe Scattergood and Steven Bishop
A Network Model Approach to International Aid
Reprinted from: *Entropy* **2023**, *25*, 641, doi:10.3390/e25040641 . . . . . **103**

Tamara Kyrylych and Yuriy Povstenko
Multi-Criteria Analysis of Startup Investment Alternatives Using the Hierarchy Method
Reprinted from: *Entropy* **2023**, *25*, 723, doi:10.3390/e25050723 . . . . . **121**

Frank Brennan Webb, Daniel Stimpson, Miesha Purcell and Eduardo López
Organizational Labor Flow Networks and Career Forecasting
Reprinted from: *Entropy* **2023**, *25*, 784, doi:10.3390/e25050784 . . . . . **131**

Ewa A. Drzazga-Szczęśniak, Piotr Szczepanik, Adam Zenon Kaczmarek and Dominik Szczęśniak
Entropy of Financial Time Series Due to the Shock of War
Reprinted from: *Entropy* **2023**, *25*, 823, doi:10.3390/e25050823 . . . . . **150**

Nedim Bayrakdar, Valerio Gemmetto and Diego Garlaschelli
Local Phase Transitions in a Model of Multiplex Networks with Heterogeneous Degrees and Inter-Layer Coupling
Reprinted from: *Entropy* **2023**, *25*, 828, doi:10.3390/e25050828 . . . . . **162**

Peter Tsung-Wen Yen, Kelin Xia and Siew Ann Cheong
Laplacian Spectra of Persistent Structures in Taiwan, Singapore, and US Stock Markets
Reprinted from: *Entropy* **2023**, *25*, 846, doi:10.3390/e25060846 . . . . . **197**

Francisco Yáñez Rodríguez and Alberto P. Muñuzuri
A Goodwin Model Modification and Its Interactions in Complex Networks
Reprinted from: *Entropy* **2023**, *25*, 894, doi:10.3390/e25060894 . . . . . **228**

Hua Zhong, Xiaohao Liang and Yougui Wang
Transaction Entropy: An Alternative Metric of Market Performance
Reprinted from: *Entropy* **2023**, *25*, 1140, doi:10.3390/e25081140 . . . . . **245**

entropy

MDPI

Article

Dependency Structures in Cryptocurrency Market from High to Low Frequency

Antonio Briola [1,2] and Tomaso Aste [1,2,3,*]

[1] Department of Computer Science, University College London, London WC1E 6BT, UK
[2] Center for Blockchain Technologies, University College London, London WC1E 6BT, UK
[3] Systemic Risk Center, London School of Economics, London WC2A 2AE, UK
* Correspondence: t.aste@ucl.ac.uk

Abstract: We investigate logarithmic price returns cross-correlations at different time horizons for a set of 25 liquid cryptocurrencies traded on the FTX digital currency exchange. We study how the structure of the Minimum Spanning Tree (MST) and the Triangulated Maximally Filtered Graph (TMFG) evolve from high (15 s) to low (1 day) frequency time resolutions. For each horizon, we test the stability, statistical significance and economic meaningfulness of the networks. Results give a deep insight into the evolutionary process of the time dependent hierarchical organization of the system under analysis. A decrease in correlation between pairs of cryptocurrencies is observed for finer time sampling resolutions. A growing structure emerges for coarser ones, highlighting multiple changes in the hierarchical reference role played by mainstream cryptocurrencies. This effect is studied both in its pairwise realizations and intra-sector ones.

Keywords: complex systems; network science; econophysics; economics; financial markets; cryptocurrencies

Citation: Briola, A.; Aste, T. Dependency Structures in Cryptocurrency Market from High to Low Frequency. *Entropy* **2022**, *24*, 1548. https://doi.org/10.3390/e24111548

Academic Editor: Panos Argyrakis

Received: 21 September 2022
Accepted: 26 October 2022
Published: 28 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Financial markets are complex systems [1]. The main source of complexity comes from the intricate interaction of heterogeneous actors following various strategies designed to impact at different time scales. They are highly stochastic environments with a low signal to noise ratio, dominated by strong non-stationary dynamics and characterized by feedback loops and non-linear effects [2–4]. Despite their complexity, financial systems are governed by a rather stable and partially identified framework of rules [5]. This last characteristic, jointly with the possibility to continuously monitor them across time, makes financial systems well suited for statistical characterization [6] and a good playground for the study of complex systems in general. In this paper, we analyse the behaviour of cryptocurrency market. A cryptocurrency is defined as a digital instrument for value transfer that exploits cryptography and distributed ledgers for security and decentralization [7]. As currencies, they have properties similar to fiat currencies [8]. The main differences being the exclusion of financial institutions as intermediaries [9] and not being controlled and regulated by any central authority [10]. Thanks to the above mentioned characteristics, cryptocurrency market is available 24 h a day, 7 days a week, and transactions take place between individuals with different physical locations across the globe [8]. Standard features of financial systems joined with peculiarities listed above, make cryptocurrencies highly volatile instruments. Finding assets with similar behaviours responding to endogenous or exogenous events is, hence, a challenging, but extremely valuable, exercise both from theoretical and applicative perspectives (e.g., risk management and investment). The ready access availability of large volumes of market data ease research on these instruments with respect to classical financial ones. Indeed, one of the main limits faced by research in the field of financial applications is the lack of easy access and share of high-quality data. In most cases they are sensible data, owned and managed by private financial institutions.

Cryptocurrencies are traded on digital currency exchanges (DCEs) which, differently from traditional exchanges, allow to easily access both online and historical data. Exploiting this and using instruments provided by network science, one can successfully build models able to capture and describe individual and collective behaviours in cryptocurrency market.

A network (or graph) represents components of a system as nodes (or vertices) and interactions among them as links (or edges). The number of nodes defines the size of the network. The number of links defines the sparsity (or, conversely, density) of the network. Reversible interactions between components are represented through undirected links, while non-reversible interactions are represented as directed links. Networks have been successfully used in many application domains. Some examples are social networks [11,12], security [13], epidemiology [14,15], neuroscience [16], drug design [17], management [18], and economic forecasting and modelling [6,19–25]. Many of the above cited works share the peculiarity to study networks with a size varying as a function of time. In the current research work, on the contrary, we focus on networks with a fixed size. Starting from a set of 25 liquid cryptocurrencies, we exploit the power of state-of-the-art network-based information filtering approaches (i.e., MST [26] and TMFG [27]) to build robust models capturing strong interactions among assets and pruning, at the same time, weakest ones. We hence investigate dependency structures of the networks at 6 different time horizons spanning from 15 s to 1 day. For each time horizon, we test the stability, statistical significance, and economic meaningfulness of the graphs. Such a research effort is led by two main motivations. The first one is the will to describe core dependency structures of the cryptocurrency market in a systematic way, providing a detailed characterization of the reference role played both by mainstream cryptocurrencies and by peripheral ones. The second is related to the possibility to do this at a wide range of time scales including intra-minute resolutions. Such a characterization is relevant for many reasons. Cryptocurrency market is affected by daily changes related to the introduction of new coins, collapse of existing ones, updates on existing protocols, etc. Having a stable framework able to robustly handle this intrinsic mutability, highly eases investment, and risk management decisions and provide a ductile instrument for research purposes. Such a framework should be also able to handle dynamics of cryptocurrencies showing similar characteristics and behaviours (i.e., belonging to the same sector). Dependency structures are, hence, investigated, both at an intra-sector and pairwise level. Unfortunately, there is no consensus on a unique mapping between cryptocurrencies and sectors. We adopt the taxonomy proposed by Kraken [28] digital currency exchange. Results give a deep insight into the evolutionary process of the time dependent hierarchical organization of the chosen system of cryptocurrencies. As a further step toward robustness, we compare our results with the ones achieved in the past 20 years of similar research in the field of stock market, uncovering comparable behaviours between the two systems. From an economic and financial perspective, the study of dependency structures among cryptocurrencies at different time-scales is relevant both from a theoretical and an applicative point of view. In the first case, comparing properties of time dependent hierarchical organization of the cryptocurrency market (a relatively young market) with the ones of the equity market (a consolidated market), (i) allows to measure its degree of maturity (ii) keeping track, at the same time, of the main evolutionary phases. In the second case, such an analysis is useful as a support instrument toward the achievement of different goals spanning from portfolio construction tasks (e.g., diversification purposes) to development of multi-assets trading strategies acting at different time-scales. Our contribution to the existing literature is threefold: (i) we are the first to use TMFG as an information filtering approach to model dependency structures among crypto-assets, (ii) we propose a rigorous network-based study of cryptocurrency market allowing to compare emerging dynamics to the ones observed on traditional financial markets (e.g., the "Epps effect"), and (iii) we are the first to describe the evolution of dependency structures among cryptocurrencies at time scales spanning from intra-minute to daily resolution.

The rest of the paper is organised as follows. In Section 2, we review the previous research on applications of network science to financial systems modelling. In Section 3.1,

we discuss the data acquisition and transformation pipeline. In Sections 3.3–3.5, we characterise cross-correlation between cryptocurrencies as a measure of similarity and dependency. We show how to obtain a dissimilarity measure based on cross-correlation and we review the building process and properties of MSTs, PMFGs and TMFGs. In Section 4, we present results obtained applying methods reported in Section 3. In Section 5, we conclude by discussing the economic and financial interpretation of our findings.

2. Related Work

Networks have been extensively used in order to model economic and financial systems. The work by [19] can be identified as a foundational one. It demonstrates the possibility to find a hierarchical arrangement of stocks traded in a financial market by investigating the daily time series of logarithmic price returns. A graph is obtained, exploiting information contained in the correlation matrix computed between all pairs of stocks of the portfolio by considering the synchronous time evolution of the logarithmic returns. Building on the work by [19], the paper by [29] shows that sets of stock index time series can be used to extract meaningful information about the links between different economies across the world. This goal is successfully achieved provided that the effects of the non-synchronous nature of the time series and of the different currencies used to compute the indices are properly taken into account. The work by [22] further extends the research by [19], studying modifications of the hierarchical organization of a set of stocks switching from high- to low-frequency time scales. As a first step, authors report a decrease in correlation between pairs of assets switching from coarser to finer time sampling resolutions. Such a phenomenon is known as "Epps effect" [30]. This analysis is extended, investigating both pairwise and intra-sector dynamics. They show the emergence of a more complex network structure at coarser time sampling resolutions, highlighting multiple changes in the hierarchical reference role played by sectors' representative assets. The work by [20] tests the robustness of the findings of the previously cited research works for longer periods of investigation and demonstrates that networks describing the financial domain cannot be reproduced by a random market model [31,32] and by the one-factor model [33]. Such results are also investigated in [21] which specifically shows how the topology of the networks in financial systems can be used to validate or falsify simple, although widespread, market models. This work also extends the previously cited ones introducing an analysis of the networks built on the volatility of financial time series. More recently, the work by [6] shows vulnerabilities of MST [26] in representing complex systems and proposes the usage of a planar graph, the PMFG [34], as an alternative. This research work also presents a set of methods to validate the statistical significance and robustness of achieved empirical results. The centrality role of specific financial sectors is finally investigated and the evolution of the Financial sector as a reference one is analysed over a period of 10 years. Recently, some of the network-based information filtering approaches have been sparsely applied to the cryptocurrency market. Results consistent with the ones described in our paper have been recently described by [35], adopting a different methodology. In this research, exploiting the index cohesive force [36], the author describes the changes in the hierarchical order of the most influential cryptocurrencies over a period of five years. He shows how Ethereum gradually becomes the most influential cryptocurrency at the detriment of Bitcoin. It is also useful to mention the work by [37], where, for the first time, the authors suggest a network-based approach to study the interdependencies between log-returns of cryptocurrencies, with a special focus on Bitcoin. They use the MST method in order to group assets into hierarchical clusters and they highlight the potential existence of topological properties of the cryptocurrency market. This work is extended by [38], where, the authors adopt the MST and the PMFG to study the change in cryptocurrency market's network structure before and after the COVID-19 outbreak. The last work to be mentioned is the one by [39], where the author points out how most of the studies on cryptocurrency market are focused only on daily data without considering other options. Using a range of frequencies spanning from one minute to weekly data, he shows how it is

possible to detect different profitable frequencies and underlines the relevance of analysing frequencies different from daily ones.

3. Methods

3.1. Data

The vast majority of digital currency exchanges provide a free Rest API (or a Websocket) allowing users to access both historical OHLCV (open, high, low, close, volume) data and online Limit Order Book- and trades-related data. In addition to this, there is a growing number of services providing out of the box, unified APIs which support many exchanges and merchant APIs. The work by [8] reports a comprehensive and detailed overview of the services currently available for data retrieving. In the current work, we use data from the FTX [40] digital currency exchange. They are entirely accessed through the CCXT [41] Python package. We use OHLCV data for 25 cryptocurrencies (see Table 1) sampled at time horizons $\Delta t \in [15\text{ s}, 1\text{ min}, 15\text{ min}, 1\text{ h}, 4\text{ h}, 1\text{ day}]$. For each time horizon, a sample can be defined as a "time bin". Opening and closing prices are, respectively, the first and the last price of the time bin, high and low price are, respectively, the highest and the lowest price of the time bin and can technically happen in any order, and the volume is is defined as the sum of the volumes traded in the time bin. In the rest of the paper, we will use a second-based definition of time horizons. This means that we will refer them as $\Delta t \in [15, 60, 900, 3600, 14{,}400, 86{,}400]$. Qualitatively, we will often speak about finer and coarser time horizons. In the first case, we want to indicate elements nearer to the lower bound of the set of time sampling resolutions, while, in second case, we want to indicate elements nearer to the upper bound of the set of time sampling resolutions.

Table 1. List of the 25 cryptocurrencies analysed in this paper. For each asset, the name, the symbol, the market capitalization at 29 March 2022 and the corresponding sector according to the taxonomy proposed by [42] is reported. There is no consensus on a unique mapping between cryptocurrencies and sectors. The chosen taxonomy is the one adopted by one of the main DCEs: Kraken [28]. Looking at the market capitalization column, it is worth noting that the least capitalized asset is Cream ($31.68M), while the most capitalized one is Bitcoin ($903B). Sectors' grouping is balanced. Cryptocurrencies being the only representative of a specific sector are grouped together in analyses reported in Appendices A and B.

Cryptocurrency	Symbol	Capitalization	Sector
Aave	AAVE	$2.47B	Lending
Bitcoin Cash	BCH	$7.13B	Currencies
Binance Coin	BNB	$72.17B	Centralized Exchanges
Bitcoin	BTC	$903B	Currencies
Cream	CREAM	$31.68M	Lending
Ethereum	ETH	$412B	Smart Contract Platforms
FTX Token	FTT	$7.11B	Centralized Exchanges
Helium	HNT	$2.78B	IoT
Huobi Token	HT	$1.46B	Centralized Exchanges
Hxro	HXRO	$129M	Centralized Exchanges
Litecoin	LTC	$9.11B	Currencies
Polygon	MATIC	$13.21B	Scaling
Maker	MKR	$2.10B	Lending
OMG Network	OMG	$818M	Scaling
PAX Gold	PAXG	$609M	Stablecoins
THORChain	RUNE	$3.96B	Decentralized Exchanges
Solana	SOL	$36.09B	Smart Contract Platforms
Serum	SRM	$458M	Decentralized Exchanges
SushiSwap	SUSHI	$521M	Decentralized Exchanges
Swipe	SXP	$800M	Payment Platforms
TRON	TRX	$7.24B	Smart Contract Platforms

Table 1. *Cont.*

Cryptocurrency	Symbol	Capitalization	Sector
Tether	USDT	\$81.37B	Currencies
Waves	WAVES	\$5.77B	Smart Contract Platforms
XRP	XRP	\$42.05B	Currencies
yearn.finance	YFI	\$836M	Asset Management

All the considered cryptocurrencies are liquid with a medium-to-high market capitalization. An exception is Cream, which has a low capitalisation. The only constraint in the selection process of cryptocurrencies is their historical availability on the FTX digital currency exchange. Indeed, it is worth noting that each digital currency exchange allows to access historical data only starting from the date a specific asset has been quoted on the exchange itself. The period under analysis spans between 1 January 2021 to 28 February 2022. Despite the high-quality of data, rare missing values are detected at the finest time sampling resolution (i.e., $\Delta t = 15$). In this case, they are filled using the nearest valid observation. Logarithmic returns (named in the rest of the paper as log-returns) x of closing prices p at time t for a given cryptocurrency c, are computed as follows:

$$x_c(t) = \log(p_c(t)) - \log(p_c(t - \Delta t)). \tag{1}$$

The assumption of returns' stationarity is validated for each $x_c(t)$ through the Augmented Dickey Fuller (ADF) [43] test.

3.2. Correlation-Based Filtering

Understanding how variables evolve, influencing the collective behaviour, and how the resulting system influences single variables is one of the most challenging problems in complex systems. In order to extract such an information from the set of synchronous time series discussed in Section 3.1, we proceed by determining their Pearson's correlation coefficient at each time horizon Δt. The Pearson's estimator of the correlation coefficient, for non-overlapping increments, between two synchronous data series with length $T\Delta t$ is:

$$\rho_{i,j}(\Delta t) = \frac{\frac{1}{T}\sum_{u=1}^{T}(x_i(u\Delta t) - \mu_i)(x_j(u\Delta t) - \mu_j)}{\sigma_i \sigma_j} \tag{2}$$

where $\mu_{i(j)}$ and $\sigma_{i(j)}$ are, respectively, the sample mean and the sample standard deviation of the data series $x_{i(j)}(t)$. The Pearson's correlation coefficient is a widespread measure efficient at catching similarities between the evolution process of financial assets' prices [6]. By definition, $\rho_{i,j}(\Delta t)$ has values between -1 (meaning that the two synchronous time series are completely, linearly anti-correlated) and $+1$ (meaning that the two synchronous time series are completely, linearly correlated). When $\rho_{i,j}(\Delta t) = 0$, the two synchronous time series are linearly uncorrelated. The correlation matrix $\mathbf{C}$ is $n \times n$ (where n is the number of variables) symmetric, with elements on the diagonal equal to one (i.e., $\rho_{i,i}(\Delta t) = 1$). For each time horizon Δt, $n(n-1)/2$ correlation coefficients completely characterize the correlation matrix. From a network science perspective, the correlation matrix can be considered as a fully connected graph where each asset is represented by a node and each pair of assets is joined by an undirected edge representing their correlation.

3.3. Minimum Spanning Tree (MST)

Based on the correlation matrix, we want to build an undirected graph whose topology captures dependency structures among cryptocurrencies' log-returns time series and that is greatly reduced in the number of edges with respect to a complete graph. In such a network, all the relevant relations must be represented. At the same time, the network should be kept as simple as possible. The simplest connected graph is a spanning tree. Minimum spanning trees (MSTs) [26] are largely used in multivariate analysis; they represent a class

of networks that connect all the vertices without forming cycles (i.e., closed paths of at least three nodes). MSTs are often computed with respect to a distance metric, so that minimizing the metric corresponds to linking assets that are close to each other. As a product of their building process, MSTs retain the maximum possible number of distances [19] minimizing, at the same time, the total edge distance. In [19], MSTs are computed using the Euclidean distance [44]:

$$d_{i,j} = \sqrt{2(1 - \rho_{i,j})}. \tag{3}$$

This definition is however too restrictive disfavouring negatively correlated variables that are equally important as the positive ones for the representation of the dependency structure [45]. In order to mitigate this limitation, we use the power dissimilarity measure:

$$d_{i,j} = 1 - \rho_{i,j}^2 \tag{4}$$

The work [46] provides a complete pedagogical exposition of the determination of the MST in the context of synchronous financial time series. A general approach to the construction of the MST is to connect the less dissimilar vertices while constraining the graph to be a tree as follows:

1. Make an ordered list of edges i, j, ranking them by increasing dissimilarity (first the edge expressing the highest similarity and last the edge expressing the highest dissimilarity).
2. Pop the first element of the ordered list and add it to the spanning tree.
3. If the added edge creates a cycle then remove the edge, otherwise skip to step 4.
4. Iterate the process from step 2 until all pairs have been exhausted.

Such an algorithm for the construction of the MST is known as the Prim's algorithm [47]. The resulting network has $n - 1$ edges. Considering that the system of cryptocurrencies analysed in the current paper is made of $n = 25$ assets (i.e., nodes), the resulting MST contains 24 edges (the code used to compute MSTs can be retrieved at https://github.com/shazzzm/topcorr; last access on 27 October 2022).

3.4. Planar Maximally Filtered Graph (PMFG)

The MST is a powerful method to capture meaningful relationships in a network structure describing a complex system. However, this method presents some aspects that can be unsatisfactory. The main constraint is that it has to be a tree (i.e., it cannot contain cycles). This characteristic makes impossible to represent relationships among more than two variables showing strongly correlated behaviours in their dynamics. In order to maintain the same powerful filtering properties of the MST and adding, at the same time, extra links, cycles, and cliques (i.e., complete subgraphs) in a controlled manner, it has been proposed to use the Planar Maximally Filtered Graph (PMFG) [48–51]. PMFG can be viewed as the first incremental step towards complexity after the MST. Indeed, instead of being a tree, the algorithm impose planarity. A graph is said to be planar if it can be embedded in a sphere without edges crossing. The foundational work by [6] provides a comprehensive pedagogical exposition of the determination of the PMFG. A general approach to the construction of the PMFG can be resumed as follows:

1. Make an ordered list of edges i, j, ranking them by increasing dissimilarity (first the edge expressing the highest similarity and last the edge expressing the highest dissimilarity).
2. Pop the first element of the ordered list and add it to the graph.
3. If the resulting graph is not planar, then remove the edge, otherwise skip to step 4.
4. Iterate the process from step 2 until all pairs have been exhausted.

It has been proved that the MST is always a sub-graph of the PMFG [48]. PMFG has $3 \times (n - 2)$ edges and a number of 3-cliques larger or equal to $2n - 4$. We remark that also 4-cliques can be present in this kind of graph.

3.5. Triangulated Maximally Filtered Graph (TMFG)

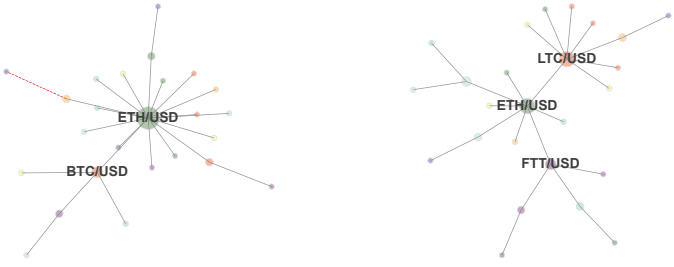
The PMFG presents two main limits: it is computational costly and it is a non-chordal graph. A graph is said to be chordal if all cycles made of four or more vertices have a chord which reduces the cycle to a set of triangles. A chord is defined as an edge that is not part of the cycle but connects two vertices of the cycle itself. In order to bypass these two constraints, the Triangulated Maximally Filtered Graph (TMFG) [27] has been proposed. A general approach to the construction of the TMFG can be resumed as follows:

1. Make an ordered list of edges i, j, ranking them by increasing dissimilarity (first the edge expressing the highest similarity and last the edge expressing the highest dissimilarity).
2. Find the 4 nodes with the lowest sum of edge weights with all other nodes in the graph and connect them forming a tetrahedron with 4 triangular faces.
3. Identify and add the node that minimize the sum of its connections to a triangle face already included in the graph, forming three new triangular faces.
4. If the graph reaches a number of edges equal to $3n - 6$, then stop, otherwise go to step 3.

Such an algorithm extracts a planar subgraph which optimises an objective function quantifying the gain of adding a new vertex to the existing tetrahedron. Compared to the PMFG, the TMFG is more efficient to be computed and is a chordal graph. The chordal structural form allows to use the filtered graph for probabilistic modeling [52,53]. A TMFG has $3 \times (n - 2)$ edges (with n representing the number of nodes) and contains both 3-cliques and 4-cliques. Considering that the system of cryptocurrencies analysed in the current paper is made of $n = 25$ assets (i.e., nodes), the resulting TMFG contains 69 edges, 88 3-cliques, and 22 4-cliques (The code used to compute TMFGs can be retrieved at https://github.com/shazzzm/topcorr; last access on 27 October 2022.).

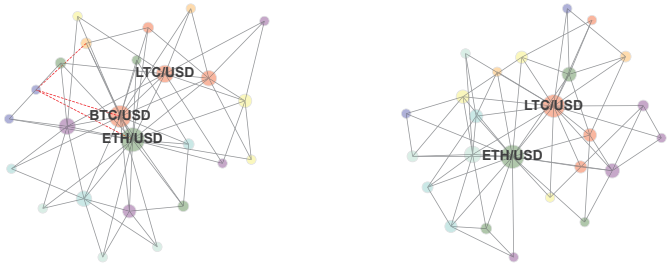
4. Results

Figures 1a and 2a report the MST and the TMFG computed at horizon $\Delta t = 15$. Figures 1b and 2b report the MST and the TMFG computed at horizon $\Delta t = 86{,}400$. Full set of MSTs computed following the procedure described in Section 3.3 is reported in Appendix A. Full set of TMFGs computed following the procedure described in Section 3.5 is reported in Appendix B.



(**a**) (**b**)

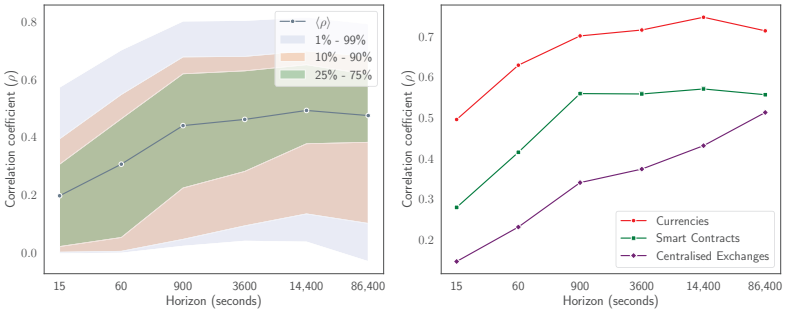
Figure 1. Minimum Spanning Tree representing log-returns time series' dependency structure computed at (**a**) 15 s and (**b**) 1 day. Only hub nodes are labelled. The adopted colour mapping scheme follows the sectors' taxonomy by [42] (see Appendix A).



(**a**) (**b**)

Figure 2. Triangulated Maximally Filtered graphs representing log-returns time series' dependency structure computed at (**a**) 15 s and (**b**) 1 day. Only hub nodes are labelled. The adopted colour mapping scheme follows the sectors' taxonomy by [42] (see Appendix B).

As a preliminary step into the study of the information level carried by the two network-based information filtering approaches, Figure 3 shows how pairwise (see Figure 3a) and the intra-sector (see Figure 3b) average Pearson's correlation coefficient $\langle \rho \rangle$ evolves as a function of time horizon Δt. Figure 3a reports the mean Pearson's correlation coefficient computed averaging over the $n(n-1)/2 = 300$ off-diagonal elements of the whole correlation matrix $\mathbf{C}$ at different time horizons. In order to give a more comprehensive view of the evolutionary dynamics of the mean pairwise correlation coefficient, we also report three meaningful percentile intervals. We observe that the average correlation coefficient $\langle \rho \rangle$ increases with time horizon Δt from a value equals to 0.19 at $\Delta t = 15$ to a value equals to 0.47 at $\Delta t = 86{,}400$. The value at $\Delta t = 15$ corresponds to the minimum average correlation coefficient across time horizons. On the other hand, the maximum average correlation coefficient does not coincide with the one computed at the maximum time horizon. It is instead detected at horizon $\Delta t = 14{,}400$, which corresponds to an intra-day resolution (i.e., 4 h). On average, the most prominent pairwise correlation weakenings are observed for most correlated pair of assets (i.e., those pairs of cryptocurrencies having a correlation coefficient included into highest percentiles).



(**a**) (**b**)

Figure 3. Evolutionary dynamics of the average correlation coefficient as a function of the time horizon Δt. (**a**) reports the horizon-related mean Pearson's correlation coefficient and three meaningful percentiles computed averaging over the $n(n-1)/2 = 300$ off-diagonal elements of the whole correlation matrix $\mathbf{C}$. (**b**) reports the horizon related mean Pearson's correlation coefficients computed averaging over the $n_s(n_s-1)/2$ correlation coefficients of the n_s assets belonging to one of three of the most relevant sectors defined by [42]: Currencies, Smart Contracts, Centralised Exchange sectors.

Figure 3b reports mean Pearson's correlation coefficient computed averaging over the $n_s(n_s-1)/2$ correlation coefficients of the n_s assets belonging to one specific sector [42] at different time horizons. Specifically, we report dynamics for Currencies, Smart Contracts, and Centralized Exchanges sectors. This choice is completed considering the relevance of the three sectors. The relevance of sectors is defined in relation to results discussed later in this section. An intra-sector scenario shows trends comparable to the ones observed in pairwise context. All the previously discussed dynamics are here more pronounced. In both cases, we observe the "Epps effect", i.e., a decrease in pair correlations at finer time sampling resolutions. This effect has been extensively studied in equity markets by [22,30]. Results reported in Figure 3 show how, also in the cryptocurrency market, the intra-sector correlation increases faster than pairwise one. The "Epps effect" is, hence, more pronounced within each sector than outside it. Going deeper, in Appendix C, we compare the probability distribution of correlation coefficients in the empirical correlation matrix **C** with the probability distribution of correlation coefficients filtered, respectively, by the MST and by the TMFG at different time horizons. We also report the probability distribution of correlation coefficients for surrogate multivariate time series obtained by randomly shuffling log-returns time series of the 25 cryptocurrencies listed in Table 1. This step is performed in order to evaluate the null hypothesis of uncorrelated returns for the considered portfolio of cryptocurrencies. Results give us the possibility to asses the statistical significance of average correlation coefficients chosen both by MST and by TMFG networks. These findings are reported in a synthetic way in Table 2. The extended count and the corresponding statistical meaning of links having a value higher than the minimum and lower than the maximum correlation coefficient detected by shuffling log-returns time series at different time horizons for the three scenarios are reported in Appendix D.

Table 2. Average absolute correlation coefficient $\langle|\rho|\rangle$ and quantiles (25–75%) computed on the empirical correlation matrix **C**, on the links filtered by MST and on the ones filtered by TMFG at different time horizons. Statistical significance of the average correlation coefficient is represented though asterisks. p-values > 0.05 are not marked. p-values ≤ 0.05 are marked as *. p-values ≤ 0.01 are marked as **. p-values ≤ 0.001 are marked as ***. The filtering power of the MST and TMFG is evident considering that the related mean correlation coefficients are always greater than the ones computed on the whole correlation coefficient matrix **C**. Results for both MST and TMFG are always robust across time horizons.

Δt	**C**			**MST**			**TMFG**		
	$\langle\|\rho\|\rangle$	25%	75%	$\langle\|\rho\|\rangle$	25%	75%	$\langle\|\rho\|\rangle$	25%	75%
15	0.20	0.02	0.31	0.35 ***	0.26	0.49	0.31 **	0.22	0.42
60	0.31	0.05	0.46	0.47 ***	0.42	0.63	0.44 ***	0.38	0.57
900	0.44 **	0.22	0.62	0.60 ***	0.57	0.76	0.57 ***	0.55	0.69
3600	0.46 **	0.28	0.63	0.62 ***	0.59	0.76	0.59 ***	0.56	0.70
14,400	0.49 *	0.38	0.65	0.65 ***	0.64	0.77	0.62 ***	0.58	0.72
86,400	0.48	0.38	0.62	0.66 **	0.64	0.77	0.61 **	0.57	0.72

Average correlation coefficients for MSTs and TMFGs are always greater than the ones computed on the empirical correlation matrix **C**. The difference between cross-horizons mean of average correlation coefficients filtered by MSTs and cross-horizons mean of average correlation coefficients in **C**, is equal to 0.16. The difference between cross-horizons mean of average correlation coefficients filtered by TMFGs and cross-horizons mean of average correlation coefficients in **C**, is equal to 0.12. Correlation coefficients filtered by TMFGs are always lower than the ones filtered by MSTs. This depends on the fact that, as reported in Section 3.5, the TMFG contains, by construction, more information than the MST. The mean difference between average correlation coefficients filtered by MSTs and the ones filtered by TMFGs, is equal to 0.03. Results reported in Table 2 confirm that the two filtering approaches prune weakest correlations among considered cryptocurrencies keeping only the strongest ones. Differently from what happens for the empirical correlation matrix

C, results for both MST and TMFG are always statistical significant across time horizons. These results enforce the evidence that both MST and TMFG carry information about strongest interactions observed in the system, disregarding most of the links consistent with the null hypothesis of uncorrelated data. It is worth noting that such an analysis does not tell much about the statistical robustness of links selected by the two network-based information filtering approaches. In order to perform such an investigation, we adopt the technique proposed by [54]. For each time horizon Δt, we sample 1000 bootstrap replicas $r = 1, \ldots, 1000$ of the empirical log-returns time series data. The length of empirical data and the one of each replica is kept equal. We compute the MST*(r) and the TMFG*(r) for each replica r. For each time sampling resolution, we map each link of the original MST and TMFG to an integer number and we count the number of links present both in the MST and TMFG and in each of the MST*(r) and TMFG*(r). Table 3 reports, for each time sampling interval Δt, the number of links of the empirical MST and TMFG with a bootstrap value larger than 95%.

Table 3. Percentage of links contained in empirical MST and TMFG at time horizon Δt with a bootstrap value larger than 95%. In the case of the MST, it is possible to notice how the robustness of the network structure decreases for coarser time sampling intervals. In the case of TMFG, on the contrary, the robustness is maintained across time horizons with low oscillations.

Δt	MST	TMFG
15	62.5%	28.9%
60	58.3%	37.7%
900	54.2%	36.2%
3600	58.3%	36.2%
14,400	41.6%	40.6%
86,400	25.0%	27.5%

Results in Table 3 show how, in the case of the MST, the robustness of the underlying network structure decreases for coarser time sampling resolutions. A consistent result has been observed by [50] in equity markets. This finding can be explained in two different ways. The first and most straightforward explanation is the statistical one and can be resumed as follows: the higher number of samples at finer time sampling resolutions implies higher statistical significance, while the lower number of samples at coarser time sampling resolutions imply lower statistical significance. A second explanation can be given looking at the structure of the networks reported in Appendix A. At finer time sampling resolutions, we observe less structured networks where numerous small-degree nodes (spokes) coexist with few anchor ones (hubs) characterised by an exceptionally high number of links. At coarser time sampling resolutions we observe more structured networks with a less imbalanced degree distribution. Such a topological change directly implies a loss in the links' statistical robustness. The case of TMFG is different. Statistical robustness of the network is maintained across horizons without significant draw-downs. Indeed, during the optimization phase of the objective function, TMFG tends to be marginally exposed to local minima, being robust to dramatic topological changes.

These last findings can be formally characterised studying the evolution of the average shortest path in MST and in TMFG as a function of time sampling resolution. Figure 4 reports the significant different behaviour in compactness' evolutionary dynamics of the two network-based information filtering approaches. In the case of MST, the minimum length of the average shortest path is equal to 2.46 and is detected at $\Delta t = 15$, while the maximum length is equal to 3.05 and is detected at $\Delta t = 86{,}400$. In the case of TMFG, we observe a strong compactness across time horizons. The minimum length of the average shortest path is equal to 1.83 and is detected at $\Delta t = 3600$, while the maximum length is equal to 1.9 at $\Delta t = 60$. In the case of MST, at the finest time sampling resolution (i.e., $\Delta t = 15$), we observe a structurally simple network with two cryptocurrencies (i.e., Ethereum and Bitcoin) acting as a hierarchical reference for the majority of other assets.

This topological structure persists switching to time horizon $\Delta t = 60$. Several changes in nodes' reference roles can be observed for networks sampled at time horizons $\Delta t = 900$ and $\Delta t = 3600$.

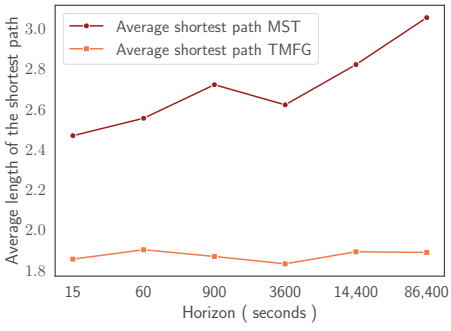


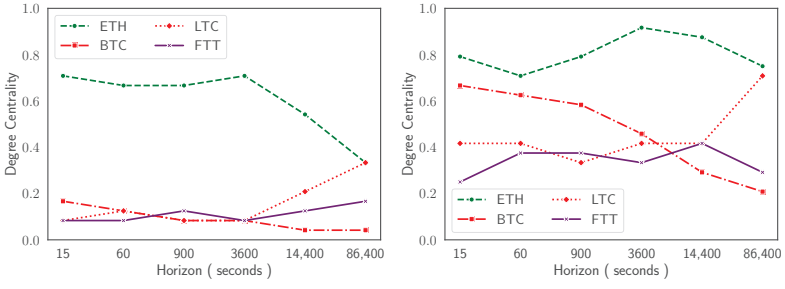
Figure 4. Average length of the shortest path in MST and TMFG as function of the time horizon at which log-returns are computed. We observe a decreasing compactness of MST networks at coarser time sampling resolutions. Instead, the compactness of the TMFG turns out to be stable across time horizons with low oscillations.

In both cases Ethereum maintains its reference role even reducing its centrality. Bitcoin, on the contrary, is gradually replaced in its role by Litecoin and FTX token (both part of the Bitcoin's cluster at time horizon $\Delta t = 60$). This structural transition is evident at $\Delta t = 14{,}400$ and fully realised at $\Delta t = 86{,}400$.

In the case of TMFG representations, it is harder to graphically detect similar dynamics. Figure 5 offers a comparative perspective between behaviours of the two network-based information filtering approaches. It shows horizon dependent evolutionary dynamics of degree centrality (i.e., measurement of the number of connections owned by a node) [55–57] for Ethereum, Bitcoin, Litecoin, and FTX Token both for MST and for TMFG. Cross-assets similarities can be detected between the two types of graphs. In the case of MSTs, degree centrality is less sensitive to minor changes in reference roles played by mainstream cryptocurrencies across time horizons, amplifying only 'extreme' ones. In the case of TMFGs, on the contrary, the same centrality measure is able to capture even small variations in network structure. This observation can be easily explained considering the amount of information the two representations are able to express. This study can be further extended looking at sectors of cryptocurrencies instead of at singular assets. Figure 6 reports the evolution of degree centrality for the three sectors Ethereum, Bitcoin, Litecoin, and FTX Token belong to: the Currencies sector, the Smart Contracts sector, and the Centralized Exchanges sector. We remark that there is no consensus on a unique mapping between cryptocurrencies and sectors. The taxonomy adopted in the current paper is described in [42] and corresponds to the one used by Kraken [28].

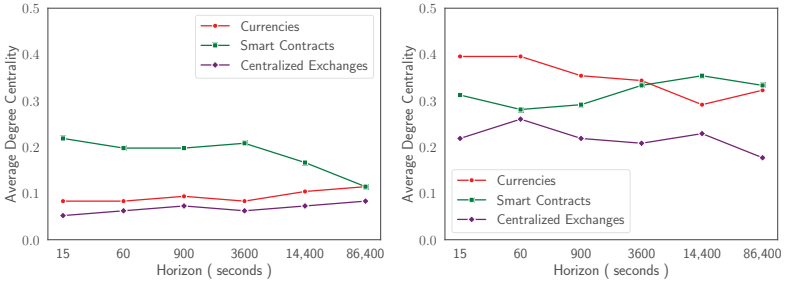
Figure 6a shows how, in the case of MST, the average degree centrality for the Smart Contracts sector strongly decreases starting from time horizon $\Delta t = 3600$, following the trend of its leading representative: Ethereum cryptocurrency. The Currencies sector, on the other hand, does not experience a decreasing trend and tends to remain stable across time horizons with low level of oscillations. In this case the loss of centrality of Bitcoin after time horizon $\Delta t = 900$, is immediately compensated by Litecoin, which reaches a hierarchical reference role at coarser time sampling resolutions. The case of Centralized Exchanges sector is different. It is stable across time horizons, without experiencing any change in intra-sector reference role dynamics and always following the behaviour of its main representative, FTX token (see Figure 5a). This last finding can be explained considering the source of the data used in the current research work. As explained in Section 3.1, we fetch data from the FTX digital currency exchange. This can cause, on the one hand an over-estimation of the role played by the exchange specific token, FTX Token, in the whole

ecology of the system under investigation, and, on the other hand, can give a potentially biased stability to the sector the asset belongs to.



(**a**) (**b**)

Figure 5. Degree centrality computed on MST (**a**) and on TMFG (**b**) as a function of time sampling resolution. Results on the TMFG highlight the switch in the reference roles of mainstream cryptocurrencies.



(**a**) (**b**)

Figure 6. Group degree centrality computed on MST (**a**) and on TMFG (**b**) for Currencies sector, Smart Contracts sector, and Centralized Exchanges sector. Group degree centrality of a set of nodes is defined as the fraction of non-group members connected to group members. Sectors are defined following the taxonomy by [42].

5. Conclusions

We investigate how cryptocurrency market's dependency structures evolve passing from high to low frequency time sampling resolutions. Starting from the log-returns of 25 liquid cryptocurrencies traded on the FTX digital currency exchange at 6 different time horizons spanning from 15 s to 1 day, we investigate pairwise correlations demonstrating that cryptocurrency market has an "Epps effect" which is comparable to the one widely studied in the equity market. Indeed, we show that the average correlation among assets increases moving from high to low frequency time horizons and we demonstrate how this dynamic is even more evident grouping cryptocurrencies into sectors. Using the concept of power dissimilarity measure, we review the building process of two network-based information filtering approaches: MST and TMFG. If, on the one hand, MST has been historically used in the description of dependency structures of different financial markets, on the other hand, this is the very first time TMFG is used to study interactions between digital assets at different time scales. Studying topologies of MSTs at finer time sampling resolutions, we observe structurally simpler networks characterised by an hub-and-spoke configuration with statistically robust links. We observe an increase in the complexity of the networks' shape for coarser time sampling resolutions with a decrease in links' statistical robustness. Such an horizon-dependent structural change is reflected by the average path length of the networks, characterised by an increasing trend moving from

high to low frequencies. TMFG offers a different perspective for the same problem. In this case, we do not observe dramatic changes in networks' topologies across time horizons. Graphs are more compact and statistical robustness of links is maintained across time with negligible oscillations. As a consequence of this, the average path length is lower and almost constant across time horizons. Studying the relative position of assets in both MSTs and TMFGs through the usage of degree centrality measure, we outline the presence of multiple changes in the hierarchical reference role among the considered set of cryptocurrencies. These changes strongly characterise singular cryptocurrencies. We find that Ethereum acts as a hierarchical reference node for the majority of other assets and maintains this role across time, gradually losing its centrality at coarser time horizons. There is not a clear economic explanation for this result. We know that lots of other cryptocurrencies are based on the Ethereum's blockchain technology but we do not think this represents a sufficient explanation to our finding. Other cryptocurrencies play a similar role with respect to smaller clusters of assets at specific time horizons. We refer specifically to Bitcoin, Litecoin, and FTX Token. Differently from Ethereum, their role does not emerge at finer time sampling resolutions and should be considered as the result of a structured evolutionary process across time horizons. We conclude stating that sectors' dynamics captured by the chosen network-based information filtering approaches are poorly affected by the ones of their main representatives, efficiently absorbing horizon-dependent changes in cryptocurrencies dynamics. This is true especially for TMFG. Indeed, looking at the evolution of the degree centrality of the Smart Contracts and Currencies sectors, one can observe that dynamics captured by MST are strongly influenced by the ones of Ethereum and Bitcoin. This does not happen in the case of TMFG where sectors' dynamics are typically detached from the ones of specific cryptocurrencies.

Author Contributions: Conceptualization, A.B. and T.A.; methodology, A.B. and T.A.; software, A.B.; validation, A.B. and T.A.; formal analysis, A.B. and T.A.; investigation, A.B. and T.A.; writing—original draft preparation, A.B.; writing—review and editing, T.A.; visualization, A.B.; supervision, T.A.; project administration, T.A.; funding acquisition, T.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by ESRC (ES/K002309/1), EPSRC (EP/P031730/1) and EC (H2020-ICT-2018-2 825215).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are accessible for free using the CCXT [41] Python Package.

Acknowledgments: The authors acknowledge many members of the Financial Computing and Analytics Group at University College London. A special thank to Silvia Bartolucci, David Vidal-Tomás, and Yuanrong Wang. Additionally, thanks to Agne Kazakeviciute for fruitful discussions on foundational topics related to this work.

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

Abbreviations

The following abbreviations are used in this manuscript:

DCE	Digital Currency Exchange
MST	Minimum Spanning Tree
PMFG	Planar Maximally Filtered Graph
TMFG	Triangulated Maximally Filtered Graph

Appendix A

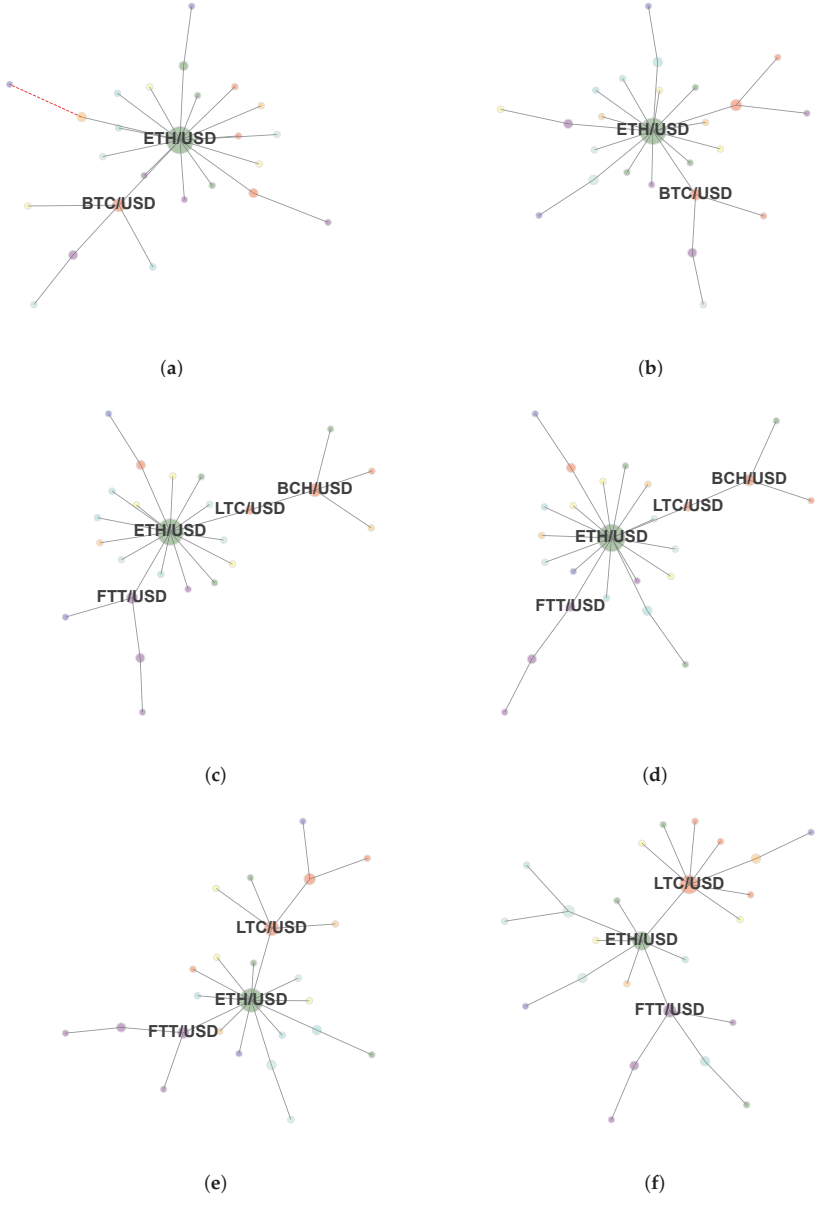


Figure A1. Minimum Spanning Tree representing log-returns time series' dependency structure computed at (**a**) 15 s, (**b**) 1 min, (**c**) 15 min, (**d**) 1 h, (**e**) 4 h, and (**f**) 1 day. The adopted colour mapping scheme follows the sectors' taxonomy by [42]: **red** → currencies, **green** → smart contract platforms, **blue** → stablecoins, **pink** → centralized exchanges, **orange** → scaling, **turquoise** → decentralized exchanges, **fuchsia** → lending, and **yellow** → all the other sectors. Dashed, red edges represent negatives linear correlations among pairs of cryptocurrencies. Only hub nodes are labelled.

Appendix B

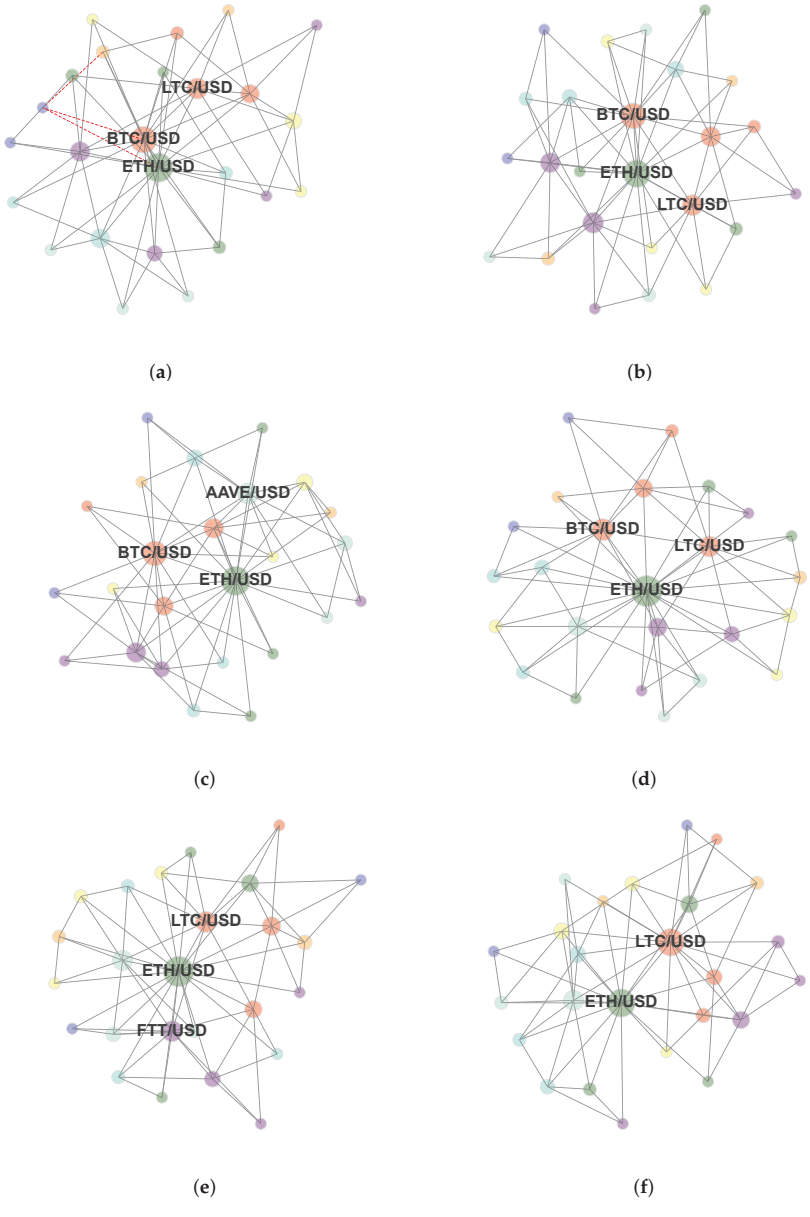


Figure A2. Triangulated Maximally Filtered Graphs representing log-returns time series' dependency structure computed at (**a**) 15 s, (**b**) 1 min, (**c**) 15 min, (**d**) 1 h, (**e**) 4 h, and (**f**) 1 day. The adopted colour mapping scheme follows the sectors' taxonomy by [42]: **red** → currencies, **green** → smart contract platforms, **blue** → stablecoins, **pink** → centralized exchanges, **orange** → scaling, **turquoise** → decentralized exchanges, **fuchsia** → lending, and **yellow** → all the other sectors. Dashed, red edges represent negatives linear correlations among pairs of cryptocurrencies. Only hub nodes are labelled.

Appendix C

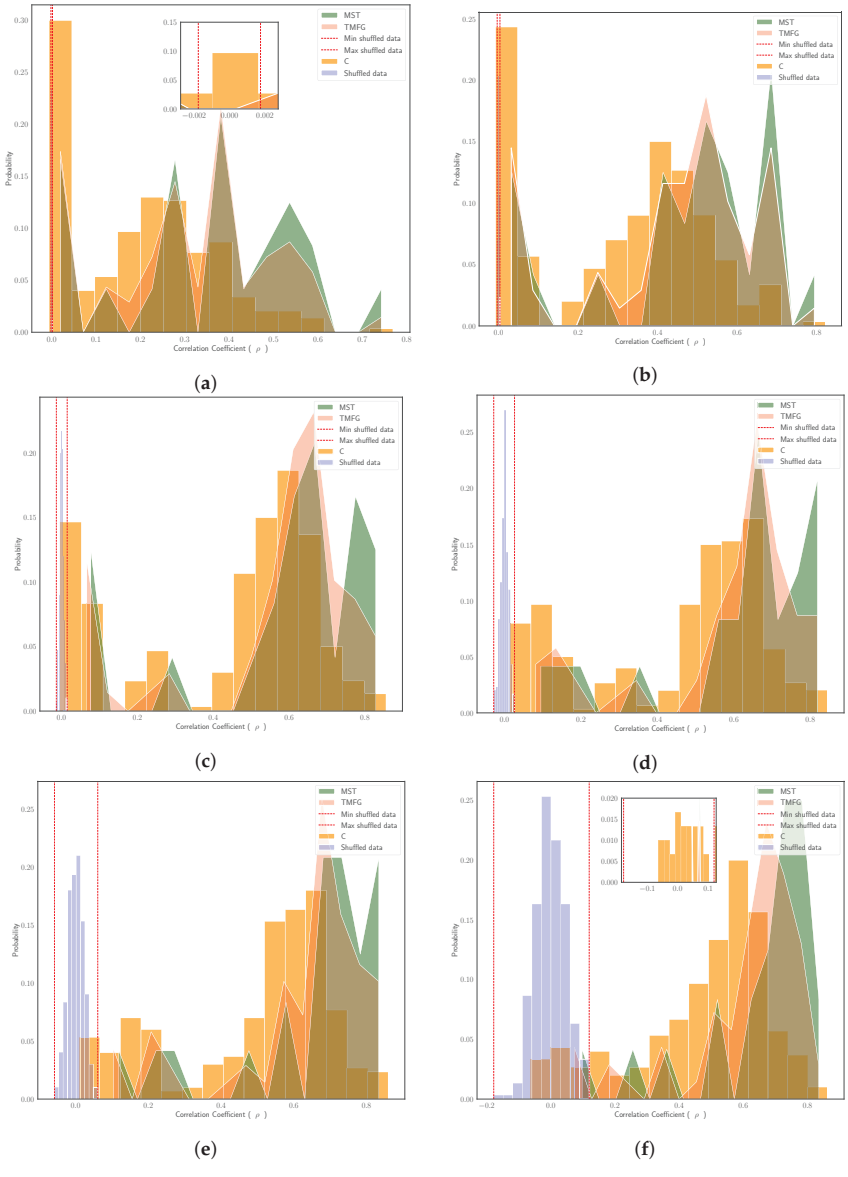


Figure A3. Probability distribution of correlation coefficients for the empirical correlation matrix **C**, MST, TMFG, and correlation matrix of shuffled log-returns time series computed at (**a**) 15 s, (**b**) 1 min, (**c**) 15 min, (**d**) 1 h, (**e**) 4 h, and (**f**) 1 day.

Appendix D

Table A1. Number of links of the empirical correlation matrix **C**, of the MST and of the TMFG having a value higher than the minimum and lower than the maximum correlation coefficient detected by shuffling log-returns time series at different time horizons. Shuffling operation is repeated 100 times. Results can be interpreted as *p*-values of the average correlation coefficient computed for **C**, for the MST and for the TMFG.

Δt	C	MST	TMFG
15	36	0	1
60	28	0	0
900	2	0	0
3600	2	0	0
14,400	16	0	0
86,400	34	1	3

References

1. Anderson, P.W. *The Economy as an Evolving Complex System*; CRC Press: Boca Raton, FL, USA, 2018.
2. Comerton-Forde, C.; Putniņš, T.J. Dark trading and price discovery. *J. Financ. Econ.* **2015**, *118*, 70–92. [CrossRef]
3. Briola, A.; Turiel, J.; Marcaccioli, R.; Aste, T. Deep reinforcement learning for active high frequency trading. *arXiv* **2021**, arXiv:2101.07107.
4. Briola, A.; Turiel, J.; Aste, T. Deep learning modeling of limit order book: A comparative perspective. *arXiv* **2020**, arXiv:2007.07319.
5. Lux, T.; Marchesi, M. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* **1999**, *397*, 498–500. [CrossRef]
6. Aste, T.; Shaw, W.; Di Matteo, T. Correlation structure and dynamics in volatile markets. *New J. Phys.* **2010**, *12*, 085009. [CrossRef]
7. Goodell, G. Tokens and Distributed Ledgers in Digital Payment Systems. *arXiv* **2022**, arXiv:2207.07530.
8. Fang, F.; Ventre, C.; Basios, M.; Kanthan, L.; Martinez-Rego, D.; Wu, F.; Li, L. Cryptocurrency trading: A comprehensive survey. *Financ. Innov.* **2022**, *8*, 1–59. [CrossRef]
9. Harwick, C. Cryptocurrency and the problem of intermediation. *Indep. Rev.* **2016**, *20*, 569–588.
10. Rose, C. The evolution of digital currencies: Bitcoin, a cryptocurrency causing a monetary revolution. *Int. Bus. Econ. Res. J.* **2015**, *14*, 617–622. [CrossRef]
11. Wasserman, S.; Faust, K. *Social Network Analysis: Methods and Applications*; Cambridge University Press: Cambridge, UK, 1994.
12. Newman, M.E.; Watts, D.J.; Strogatz, S.H. Random graph models of social networks. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 2566–2572. [CrossRef]
13. Ronfeldt, D.F.; Arquilla, J. *Networks and Netwars*; Rand: Santa Monica, CA, USA, 2001.
14. Balcan, D.; Hu, H.; Goncalves, B.; Bajardi, P.; Poletto, C.; Ramasco, J.J.; Paolotti, D.; Perra, N.; Tizzoni, M.; Van den Broeck, W.; et al. Seasonal transmission potential and activity peaks of the new influenza A (H1N1): A Monte Carlo likelihood analysis based on human mobility. *BMC Med.* **2009**, *7*, 45. [CrossRef] [PubMed]
15. Hufnagel, L.; Brockmann, D.; Geisel, T. Forecast and control of epidemics in a globalized world. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 15124–15129. [CrossRef] [PubMed]
16. Sporns, O.; Tononi, G.; Kötter, R. The human connectome: A structural description of the human brain. *PLoS Comput. Biol.* **2005**, *1*, e42. [CrossRef] [PubMed]
17. Hopkins, A.L. Network pharmacology. *Nat. Biotechnol.* **2007**, *25*, 1110–1111. [CrossRef] [PubMed]
18. Wu, L.; Waber, B.N.; Aral, S.; Brynjolfsson, E.; Pentland, A. *Mining Face-to-Face Interaction Networks Using Sociometric Badges: Predicting Productivity in an It Configuration Task*. 2008. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1130251 (accessed on 22 March 2022).
19. Mantegna, R.N. Hierarchical structure in financial markets. *Eur. Phys. J. B Condens. Matter Complex Syst.* **1999**, *11*, 193–197. [CrossRef]
20. Bonanno, G.; Caldarelli, G.; Lillo, F.; Mantegna, R.N. Topology of correlation-based minimal spanning trees in real and model markets. *Phys. Rev. E* **2003**, *68*, 046130. [CrossRef]
21. Bonanno, G.; Caldarelli, G.; Lillo, F.; Micciche, S.; Vandewalle, N.; Mantegna, R.N. Networks of equities in financial markets. *Eur. Phys. J. B* **2004**, *38*, 363–371. [CrossRef]
22. Bonanno, G.; Lillo, F.; Mantegna, R.N. High-frequency cross-correlation in a set of stocks. *Quant. Financ.* **2001**, *1*, 96–104. [CrossRef]
23. Wang, Y.; Aste, T. Dynamic Portfolio Optimization with Inverse Covariance Clustering. *arXiv* **2021**, arXiv:2112.15499.
24. Procacci, P.F.; Aste, T. Portfolio Optimization with Sparse Multivariate Modelling. *arXiv* **2021**, arXiv:2103.15232.
25. Wang, Y.; Aste, T. Sparsification and Filtering for Spatial-temporal GNN in Multivariate Time-series. *arXiv* **2022**, arXiv:2203.03991.
26. West, D.B. *Introduction to Graph Theory*; Prentice Hall: Upper Saddle River, NJ, USA, 2001; Volume 2.

27. Massara, G.P.; Di Matteo, T.; Aste, T. Network filtering for big data: Triangulated maximally filtered graph. *J. Complex Netw.* **2017**, *5*, 161–178. [CrossRef]
28. Kraken: Bitcoin & Cryptocurrency Exchange. Available online: https://www.kraken.com (accessed on 22 March 2022).
29. Bonanno, G.; Vandewalle, N.; Mantegna, R.N. Taxonomy of stock market indices. *Phys. Rev. E* **2000**, *62*, R7615. [CrossRef] [PubMed]
30. Epps, T.W. Comovements in stock prices in the very short run. *J. Am. Stat. Assoc.* **1979**, *74*, 291–298.
31. Laloux, L.; Cizeau, P.; Bouchaud, J.P.; Potters, M. Noise dressing of financial correlation matrices. *Phys. Rev. Lett.* **1999**, *83*, 1467. [CrossRef]
32. Plerou, V.; Gopikrishnan, P.; Rosenow, B.; Amaral, L.A.N.; Guhr, T.; Stanley, H.E. Random matrix approach to cross correlations in financial data. *Phys. Rev. E* **2002**, *65*, 066126. [CrossRef]
33. Campbell, J.Y.; Lo, A.; MacKinlay, C. *The Econometrics of Financial Markets*; Princeton University Press: Princeton, NJ, USA, 1997.
34. Aste, T.; Di Matteo, T.; Hyde, S. Complex networks on hyperbolic surfaces. *Phys. A Stat. Mech. Appl.* **2005**, *346*, 20–26. [CrossRef]
35. Vidal-Tomás, D. The entry and exit dynamics of the cryptocurrency market. *Res. Int. Bus. Financ.* **2021**, *58*, 101504. [CrossRef]
36. Kenett, D.Y.; Shapira, Y.; Madi, A.; Bransburg-Zabary, S.; Gur-Gershgoren, G.; Ben-Jacob, E. Index cohesive force analysis reveals that the US market became prone to systemic collapses since 2002. *PLoS ONE* **2011**, *6*, e19378. [CrossRef]
37. Zikeba, D.; Kokoszczynski, R.; Sledziewska, K. Shock transmission in the cryptocurrency market. Is Bitcoin the most influential? *Int. Rev. Financ. Anal.* **2019**, *64*, 102–125.
38. Katsiampa, P.; Yarovaya, L.; Zikeba, D. *High-Frequency Connectedness between Bitcoin and Other Top-Traded Crypto Assets during the COVID-19 Crisis*; Elsevier: Amsterdam, The Netherlands, 2021.
39. Vidal-Tomás, D. All the frequencies matter in the Bitcoin market: An efficiency analysis. *Appl. Econ. Lett.* **2022**, *29*, 212–218. [CrossRef]
40. FTX: FTX Cryptocurrency Derivatives Exchange. Available online: https://ftx.com (accessed on 22 March 2022).
41. Ccxt: CCXT—CryptoCurrency eXchange Trading Library. Available online: https://github.com/ccxt/ccxt (accessed on 22 March 2022).
42. Messari: Messari Crypto Research, Data, and Tools. Available online: https://messari.io (accessed on 22 March 2022).
43. Dickey, D.A.; Fuller, W.A. Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.* **1979**, *74*, 427–431.
44. Gower, J.C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **1966**, *53*, 325–338. [CrossRef]
45. Soramaki, K.; Cook, S.; Laubsch, A. A network-based method for visual identification of systemic risks. *J. Netw. Theory Financ.* **2016**, *2*, 67–101.
46. Mantegna, R.N.; Stanley, H.E. *Introduction to Econophysics: Correlations and Complexity in Finance*; Cambridge University Press: Cambridge, UK, 1999.
47. Prim, R.C. Shortest connection networks and some generalizations. *Bell Syst. Tech. J.* **1957**, *36*, 1389–1401. [CrossRef]
48. Tumminello, M.; Aste, T.; Di Matteo, T.; Mantegna, R.N. A tool for filtering information in complex systems. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 10421–10426. [CrossRef]
49. Aste, T.; Di Matteo, T. Dynamical networks from correlations. *Phys. A Stat. Mech. Appl.* **2006**, *370*, 156–161. [CrossRef]
50. Tumminello, M.; Di Matteo, T.; Aste, T.; Mantegna, R.N. Correlation based networks of equity returns sampled at different time horizons. *Eur. Phys. J. B* **2007**, *55*, 209–217. [CrossRef]
51. Di Matteo, T.; Pozzi, F.; Aste, T. The use of dynamical networks to detect the hierarchical organization of financial market sectors. *Eur. Phys. J. B* **2010**, *73*, 3–11. [CrossRef]
52. Turiel, J.D.; Barucca, P.; Aste, T. Simplicial persistence of financial markets: Filtering, generative processes and portfolio risk. *arXiv* **2020**, arXiv:2009.08794.
53. Barfuss, W.; Massara, G.P.; Di Matteo, T.; Aste, T. Parsimonious modeling with information filtering networks. *Phys. Rev. E* **2016**, *94*, 062306. [CrossRef] [PubMed]
54. Tumminello, M.; Coronnello, C.; Lillo, F.; Micciche, S.; Mantegna, R.N. Spanning trees and bootstrap reliability estimation in correlation-based networks. *Int. J. Bifurc. Chaos* **2007**, *17*, 2319–2329. [CrossRef]
55. Carrington, P.J.; Scott, J.; Wasserman, S. *Models and Methods in Social Network Analysis*; Cambridge University Press: Cambridge, UK, 2005; Volume 28.
56. Kleinberg, J.M.; Kumar, R.; Raghavan, P.; Rajagopalan, S.; Tomkins, A.S. The web as a graph: Measurements, models, and methods. In *International Computing and Combinatorics Conference*; Springer: Berlin/Heidelberg, Germany, 1999; pp. 1–17.
57. Maharani, W.; Gozali, A.A. Degree centrality and eigenvector centrality in twitter. In Proceedings of the 2014 8th International Conference on Telecommunication Systems Services and Applications (TSSA), Kuta, Indonesia, 23–24 October 2014; pp. 1–5.

entropy

MDPI

Article

Investigating Deep Stock Market Forecasting with Sentiment Analysis

Charalampos M. Liapis *, Aikaterini Karanikola * and Sotiris Kotsiantis

Department of Mathematics, University of Patras, 26504 Patras, Greece
* Correspondence: c.liapis@upnet.gr (C.M.L.); karanikola@upatras.gr (A.K.)

Abstract: When forecasting financial time series, incorporating relevant sentiment analysis data into the feature space is a common assumption to increase the capacities of the model. In addition, deep learning architectures and state-of-the-art schemes are increasingly used due to their efficiency. This work compares state-of-the-art methods in financial time series forecasting incorporating sentiment analysis. Through an extensive experimental process, 67 different feature setups consisting of stock closing prices and sentiment scores were tested on a variety of different datasets and metrics. In total, 30 state-of-the-art algorithmic schemes were used over two case studies: one comparing methods and one comparing input feature setups. The aggregated results indicate, on the one hand, the prevalence of a proposed method and, on the other, a conditional improvement in model efficiency after the incorporation of sentiment setups in certain forecast time frames.

Keywords: time series forecasting; deep learning; financial time series; sentiment analysis; financial BERT; multivariate; multi-step; regression; Twitter

Citation: Liapis, C.M.; Karanikola, S.; Kotsiantis, S. Investigating Deep Stock Market Forecasting with Sentiment Analysis. *Entropy* **2023**, *25*, 219. https://doi.org/10.3390/e25020219

Academic Editor: Panos Argyrakis

Received: 13 December 2022
Revised: 14 January 2023
Accepted: 20 January 2023
Published: 23 January 2023

1. Introduction

Somewhere in the course of history, the human species' need for knowledge of possible future outcomes of various events emerged. Associative norms were thus constructed between decision-making and observed data that were influenced by theoretical biases that had been inductively established on the basis of such observations. Protoscience was formed. Or not?

Even if this hypothetical description of human initiation into scientific capacities is naive or even unfounded, the bottom line is that the human species partly operates on the basis of predictions. Observing time-evolving phenomena and questioning their structure in the direction of an understanding that will derive predictions about their projected future behavior constitutes an inherent part of post-primitive human history. In response to this self-referential demand and assuming that the authors are post-primitive individuals, the core of the present work is about predicting sequential and time-dependent phenomena. This domain is called time series forecasting. Time series forecasting is, in broad terms, the process of using a model to predict future values of variables that characterize a phenomenon based on historical data. A time series is a set of time-dependent observations sampled at specific points in time. The sampling rate depends on the nature of the problem. Moreover, depending on the number of variables describing the sequentially recorded observations, a distinction is made between univariate and multivariate time series. Since there is a wide range of time-evolving problems, the field is quite relevant in modern times, with an increasing demand for model accuracy and robustness.

In addition, there are phenomena, the mathematical formalism of which is represented by time series with values which are also sub-determined by the given composition of a society of individuals. This means that the attitudes of such individuals, as they nonetheless form within the whole, are somewhat informative about aspects of the phenomenon in question. It is natural, given human nature and the consequent conceptual treatment of

the world as part of it, that these attitudes are articulated somewhere linguistically. Therefore, a hypothesis on which mathematical quantifications of the attitudes of which such linguistic representations that are signs are possible could, if valid, describe a framework for improving the modeling of the phenomena in question. For example, specific economic figures can be points in a context, the elements of which are partially shaped by what is said about them. Accordingly, it can be argued that a line of research that would investigate whether stock closing prices can be modeled in terms of their future fluctuations using relevant linguistic data collected from social networks is valid.

Thus, in this work, the incorporation of sentiment analysis in stock market forecasting is investigated. In particular, a large number of state-of-the-art methods are put under an experimental framework that includes multiple configurations of input features that incorporate quantified values of sentiment attitudes in the form of time series. These time series consist of sentiment scores extracted from Twitter using three different sentiment analysis methods. Regarding prediction methods, there are schemes that come from both the field of statistics and machine learning. Within the machine learning domain, deep learning and other state-of-the-art methods are currently in use, dominating research. Here, a large number of such widely used state-of-the-art models were benchmarked in terms of performance. Moreover, various sentiment setups of input features were tested. Two distinct case studies were investigated. In the first case study, the evaluations were organized according to methods. The subsequent comparisons followed the grouping. In the second case study, the comparisons concerned the feature setups used as inputs. Sentiment scores were tested in the context of improving the predictive capacities of the various models used. All comparisons yielded results from an extended experimental procedure that incorporated various steps. The whole setting involved a wide range of multivariate setups, which included various sentiment time series. Multiple evaluation metrics and three different time frames were used to derive multiple-view results. Below, first, a brief presentation of related literature is given. Then, the experimental procedure is thoroughly presented, which is followed by the results. Finally, Section 5 lists the extracted conclusions.

2. Related Work

The continuous and ever-increasing demand for accurate forecasts across a wide range of human activity has been a key causal factor contributing to the unabated research activity occurring within the field of time series forecasting. Thus, the prediction of time series constitutes a strong pole of interest for the scientific community. Consequently, in recent decades, this interest has been reflected in a wealth of published work and important results. In this section, a brief presentation of relevant literature is given. Due to space constraints, this presentation is more indicative than exhaustive, and its purpose is just to provide a starting point for a more thorough and in-depth review.

A trivial way to distinguish the problems associated with time series forecasting would be to divide the task into two categories with respect to the type of final output. The first category includes problems where the goal is to predict whether a future value is expected to increase or decrease over a given time horizon. This task can essentially be treated as a binary classification problem. The second category includes tasks where the goal is to accurately predict the price of a time series in a specific time frame. Here, the output can take any value within a continuous interval, and hence, the prediction process can be treated as a regression problem. One can easily imagine that the difficulty of the problems belonging to the second category is greater than that of the first and that their treatment requires more complex and precise refinements. Apparently, interesting works can be found in both categories, but the context of this paper dictates a focus on the latter.

A subclass of problems regarding focus on the direction in which a time series will move features those involving the increase or decrease of closing price values of various stocks. In particular, in [1], an ensemble technique based on tree classifiers—specifically on *random forests* and *gradient boosted decision trees*—which predicts movement in various

time frames is proposed. For the same purpose in [2], *support vector machines* (SVMs) are used in combination with sentiment analysis performed on data drawn from two forums considered to be the largest and most active mainstream communities in China. This paper is an attempt to predict stock price direction using SVMs and taking into account the so-called day-of-week effect. Adding sentiment variables results in up to 18% better predictions. Similar results, which indicate the superiority of SVMs compared to other classification algorithms, are also presented in [3], where well-known methods such as *linear discriminant analysis*, *quadratic discriminant analysis*, and *Elman backpropagation neural networks* are used for comparison. Encouraging results regarding the prediction of time series movement direction have also been achieved using hybrid methods, where modern schemes combining *deep neural network* architectures are applied to big data [4]—again—for the daily-based prediction of stock market prices. Regarding the second category, where the goal is to predict the specific future values of a time series and not merely its direction, the literature appears richer. This seems as if it is a fact rather expected if one takes into account the increased difficulty of the task and the high interest of the research community in pursuing the production of improved results. In the past decades, traditional statistical methods seemed to dominate the field of time series forecasting [5,6]. However, as expected, according to their general effectiveness, machine learning methods began to gain ground and dominate the field [7,8]. Traditional machine learning methods are incorporated in various time series forecasting tasks, such as using SVMs for economic data predictions [9] and short-term electric load forecasting [10], while architectures based on neural networks are also particularly popular. Regarding the latter—as this is probably the largest part of the literature regarding the use of machine learning in prediction problems—the use of such methods has covered a wide range of applications. Some indicative examples are the prediction of oil production [11] and traffic congestion [12] using deep *LSTM recurrent networks*, while an aggregated version of LSTMs has additionally been used for the short-term prediction of air pollution. Forecasting river water temperature using a hybrid model based on *wavelet-neural network* architecture was presented in [13], while *recurrent neural networks* (RNNs) have been deployed to forecast agricultural commodity prices in China [14]. Since the list of examples where neural network-based techniques show promise is long, the reader is urged to pursue additional personal research.

Furthermore, it is possibly worth mentioning the fact that in addition to increasingly sophisticated methods, techniques based on the theory of *ensembles* are also gaining ground. Roughly speaking, these are techniques in which the final result is derived through a process of using different models, with the prediction being formed from the combination of the individual ones. As an example, one can mention the ensemble scheme proposed in [15] for the prediction of energy consumption: it combines *support vector regression* (SVR), *backpropagation neural network* (BPNN), and *linear regression* (LR) learners. A similar endeavor is presented in [16], where an ensemble consisting of four learners, that is, *long short-term memory* (LSTM), *gate recurrent unit* (GRU), *autoencoder LSTM* (Auto-LSTM), and *auto-GRU*, is used for the prediction of solar energy production. A comparison involving over 300 individual and ensemble predictive layouts over Greek energy load data is presented in [17]. There, in addition to the large number of ensembles tested, the comparison also concerns both a number of forecast time frames as well as different modifications of the input data in various multivariate arrangements. In [18], an ensemble scheme based on *linear regression* (LR), *support vector regression* (SVR), and the *M5P regression tree* (M5PRT) is proposed to predict cases and deaths attributed to the COVID-19 pandemic regarding southern and central European countries.

With regard now to the context of this work, and given that its purpose—which is an extension of the work in [19]—is twofold, aiming, on the one hand, to compare a large number of methods and, on the other hand, to investigate the contribution of incorporating sentiment analysis into the forecasting process, it follows that a simple presentation of similarly targeted tasks seems quite essential. As for the first objective—that of comparing methods—there are several interesting works that have been carried out in recent years.

In [20], the comparison between the traditional *ARIMA* method and *LSTMs* using economic data is investigated. A similar comparison between the two methods is implemented in [21], now aiming to predict bitcoin values, while in [22], the *gated recurrent unit* (GRU) scheme is also included in the comparison. Comparative works of the *ARIMA* method with various schemes have also been carried out, such as with *neural network auto-regressive* (NNAR) techniques [23], with the *prophet* method [24], with *LSTMs* and the *XGBOOST* method [25], as well as with *wavelet neural network* (WNN) and *support vector machines* (SVM) [26]. Although, in general, modern schemes tend to perform better than ARIMA, any absolute statement would not be representative of reality. Indeed, research focused on comprehensively reviewing the use of modern methods can provide a detailed overview of the relevant work to date. Indicatively, in [27], an extensive review of the use of artificial neural networks in time series forecasting is presented, covering studies published from 2006 onwards, over a decade. A similar survey covering the period from 2005 to 2019 and focusing on deep learning techniques with applications to financial data can be found in [28]. Furthermore, regarding the experimental evaluation of modern machine learning architectures, in [29], a thorough experimental comparison is presented, concerning seven different deep learning architectures applied to 12 different forecasting problems, using more than 50,000 time series. According to the implementation of more than 38000 models, it is argued that the architectures of *LSTMs* and *CNNs* outperform all others. In [30], the comparison of a number of methods—such as *ARIMA*, *neural basis expansion analysis* (NBEATS), and probabilistic methods based on deep learning models—applied to time series of financial data is presented. Additionally, in [31], a comparison between *CNNs*, *LSTMs*, and a hybrid model of them is given, which was deployed on data concerning the forecasting of the energy load coming from photovoltaics. There, the generated results, on the one hand, indicate the dominance of the hybrid model—emphasizing the necessity to create efficient combinatorial schemes—and, on the other, show that the models' predictions improve by using a larger amount of data in the training set.

In relation to the second objective—which concerns the investigation of whether the use of information based on sentiment analysis regarding public opinion extracted from social networks favors the predictions—the available literature seems comparatively poorer but presents equally interesting results. The relationship between tweet board literature and financial market instruments is examined in [32], with results revealing a high correlation between stock prices and Twitter sentiments. In [33], using targeted topics to extract sentiment from social media, a model to predict stock price movement is presented. Moreover, the effectiveness of incorporating sentiment analysis into stock forecasting is demonstrated. In addition, ref. [34] is an attempt to capture the various relationships between news articles and stock trends using well-known machine learning techniques such as *random forest* and *support vector machines*. In [35], after assembling a financial-based sentiment analysis dictionary, a model incorporating the dictionary was developed and tested on data from the pharmaceutical market, exhibiting encouraging results. In [36], sentiment polarity is extracted by observing the logarithmic return of the ratio between the average stock price one minute before and one minute after the relevant stock's news is published. Then, using *RNNs* and *LSTMs*, the direction of the stock is successfully predicted. The exploitation of sentiment analysis techniques has also been used to predict the stock market during health crises [37] such as H1N1 and, more recently, COVID-19. Possible links between social media posts and closing stock prices at specific time horizons were found. More specifically, for COVID-19, the polarity of the posts seemed to affect the stock prices after a period of about six days.

Regarding the prediction of various stock market closing prices—which is also the thematic center of this paper—in [38], data collected from Twitter are initially analyzed in terms of their sentiment scores and are then used to predict the movement of stock prices, using *naive Bayes* and *multiclass SVM* classifiers. A similar procedure was followed in [39], where *least squares support vector regression* (LSSVR) and *backpropagation neural networks* were deployed to predict the total monthly sales of vehicles in the USA, using

additional sentiment information combined with historical sales data. Data collected from the online editions of international newspapers were used in [40] to predict the closing stock price values, incorporating both traditional methods, such as *ARIMA*, and newer ones, such as the Facebook *prophet* algorithm and *RNN* architectures that use as input both numerical values of the time series to be predicted as well as combinations of the polarity of extracted sentiments.

In [41], both traditional and modern machine learning methods such as *support vector machines*, *linear regression*, *naive Bayes*, and *long short-term memory* are used in combination with the incorporation of opinion data, current news, and past stock prices. In [42], sentiment analysis and *empirical model decomposition* are used so that complex time series can be broken down into simpler and easier to manage parts, together with an *attention* mechanism that attributes weight to the information considered most useful for the task being performed each time. A method based on the architecture of *LSTMs* that uses information derived from sentiment analysis together with multiple data sources is presented in [43]. Initially, textual data related to the stock in question are collected, and using methods based on *convolutional neural network* architectures, the polarity of investors' sentiment is extracted. This information is then combined with that of the stock's past closing prices and other technical indicators to produce the final forecast. In [44], a hybrid model that leverages deep learning architectures, such as *convolutional neural networks*, to extract and categorize investor sentiment as detected in financial forums is described. The extracted sentiments are then combined with information derived from technical financial indicators to predict future stock prices in real-world problems using *LSTM* architectures. *SVM* architectures are used on Twitter data to extract polarity in [45]. The extracted polarities are used in an incremental active learning scheme, where the continuous stream of content-changing tweets is used to predict the closing stock price of the stock market.

Sentiment analysis has also been used to predict the price of bitcoin in real time, using—and at the same time comparing—*LSTM* techniques and the classical *ARIMA* method [46], where the exploitation of the information derived from sentiment analysis has been beneficial. Similar research focused on predicting the price direction of the cryptocurrencies Bitcoin and Ethereum using sentiment analysis from data drawn from Twitter and Google Trends and given as input to a linear predictive model is presented in [47]. Interestingly, the volume of tweets affects the prediction to a greater extent than the polarity of the sentiment extracted from the tweets. Forecasting the price direction of four popular cryptocurrencies—Bitcoin, Ethereum, Ripple, and Litecoin—using machine learning techniques and data drawn from social networks is presented in [48]. Classical methods such as *neural networks* (NN), *support vector machines* (SVM), and *random forests* (RF) are compared. An interesting fact is that Twitter, roughly speaking, seems to favor the prediction of specific cryptocurrencies rather than all of them. Using sentiment analysis has also been beneficial in the field of cybersecurity. In [49], a methodology that exploits the knowledge of hacker behavior for predicting malicious events in cyberspace by performing sentiment analysis with different techniques (*VADER*, *LIWC15*, and *SentiStrength*) on data collected from hacking forums, both on the dark web and on the surface web, is presented.

The—rather diverse—list of applications in which the use of sentiment analysis techniques can improve the generated forecasts is proportional to the fields in which time series forecasting is applied since, in general, the utilization of public opinion knowledge appears to have a positive effect on the forecasting process. Some of them that have been implemented in the last five years have already been mentioned in passing, and many others can be added. Such would include predicting the course of epidemics, such as that of the Zika virus in the USA in 2016 [50] or the COVID-19 pandemic, the outcome of electoral contests [51], the prediction of the price of e-commerce products [52], and the list goes on. Given human nature and the consequent conceptual coping of the world by human subjects, sentiment analysis seems justifiably relevant in a multitude of applications. The reader is therefore encouraged to conduct additional bibliographic research.

3. Experimental Procedure

Information regarding the stages of the experimental procedure will now be presented. This presentation will be as detailed as possible given the necessary space constraints and content commitments in order not to disrupt the depictive nature of the paper.

It has already been mentioned that to some extent, the "core" of the present work consists of an experimental procedure that aims, in its most abstract scope, to check the efficiency, on the one hand, of a number of state-of-the-art algorithms and, on the other, of incorporating sentiment analysis into predictive schemas. Thus, a total of 16 *datasets* × 67 *combinations* × 30 *algorithms* × 3 *time-shifts* = *96,480 experiments* were conducted. The dataset consisted of time series containing the daily closing values of various stocks along with a multitude of 67 different sentiment score setups. Specifically, 16 datasets of stocks containing such closing price values were used over a three-year period, beginning on 2 January 2018 and ending on 24 December 2020. Generated sentiment scores from relevant textual data extracted from the Twitter microblogging platform were used. Three different sentiment analysis methods were deployed. The sentiment score time series and the closing values were subjected to a 7-day and a 14-day rolling mean strategy, yielding a total of 12 distinct features. Various combinations of the created features resulted in a total of 67 distinct input setups per algorithm. The calculated sentiment scores along with the closing values were then tested under both univariate and multivariate forecasting schemes. Lastly, 30 state-of-the-art methods were investigated. Below, a more thorough presentation of the aforementioned experimental setting follows.

3.1. Datasets

Starting with data, the process of collecting and creating the sets used will now be addressed.

3.1.1. Overview

To begin with, Table 1 contains the names of the aforementioned datasets along with their corresponding abbreviations. These initial data included time series containing closing values for 16 well-known listed companies. All sets comprise three-year period data for dates ranging from 2 January 2018 to 24 December 2020.

Table 1. Stock datasets.

No	Dataset	Stocks
1	AAL	American Airlines Group
2	AMD	Advanced Micro Devices
3	AUY	Yamana Gold Inc.
4	BABA	Alibaba Group
5	BAC	Bank of America Corporation
6	ET	Energy Transfer L.P.
7	FCEL	FuelCell Energy Inc.
8	GE	General Electric
9	GM	General Motors
10	INTC	Intel Corporation
11	MRO	Marathon Oil Corporation
12	MSFT	Microsoft Corporation
13	OXY	Occidental Petroleum Corporation
14	RYCEY	Rolls-Royce Holdings
15	SQ	Square
16	VZ	Verizon Communications

Essentially, the initial features were four: that is, the closing prices of each stock and three additional time series containing relative sentiment scores for the given period. Subsequently, and after applying 7- and 14-day rolling averages, a total of 14 features were extracted. Thus, for each share, the final input settings were composed by introducing

altered features derived from stock values and a sentiment analysis process applied to an extended corpus of tweets. Figure 1 depicts a—rather abstractive—snapshot of the whole process from data collection to the creation of the final input setups.

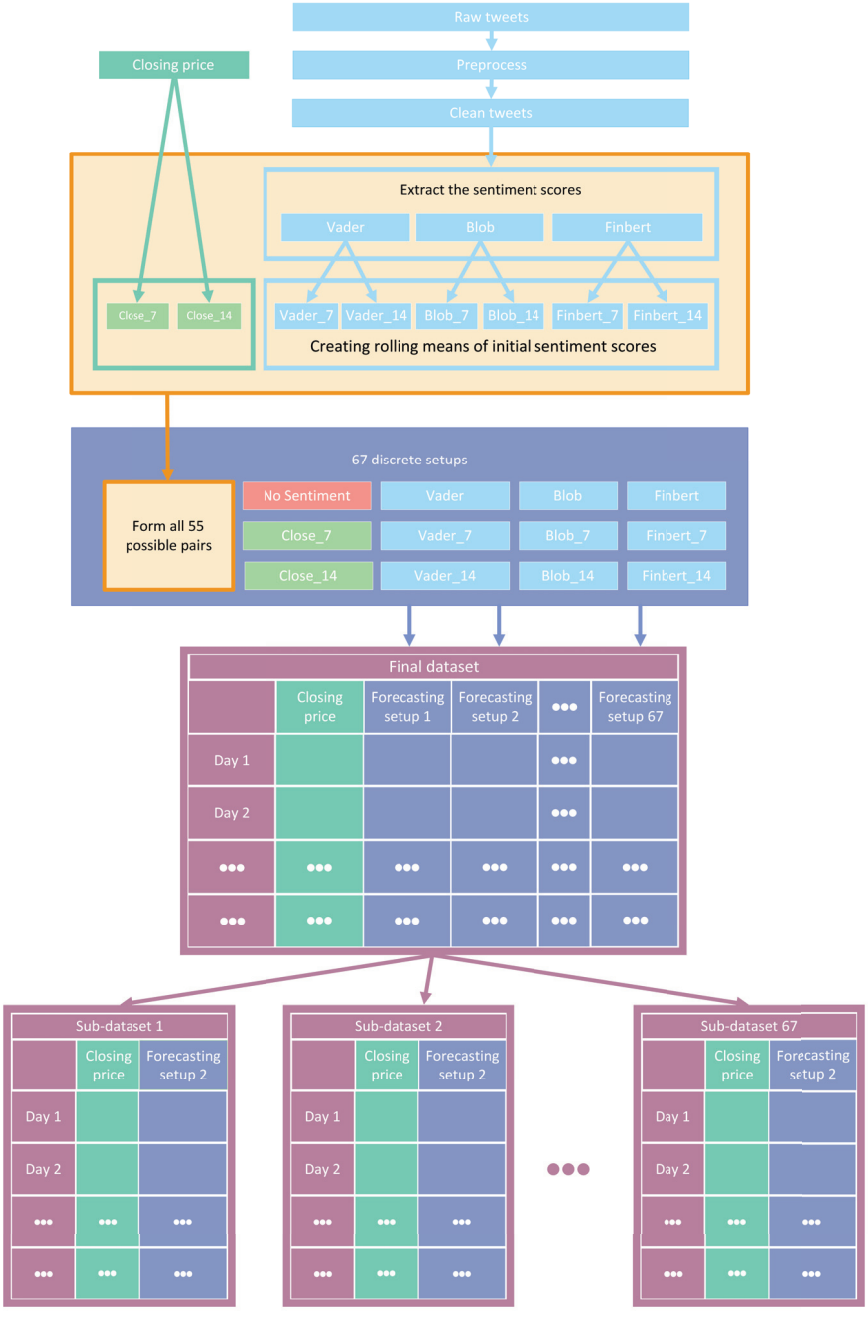


Figure 1. Feature setups: creation pipeline.

3.1.2. Tweets and Preprocessing

A large part of the process involved deriving sentiment scores related to stocks. Using the *Twitter Intelligence Tool* (TWINT) [53], a large number of stock-related posts written in English were downloaded from Twitter and grouped by day. TWINT is an easy-to-use yet sophisticated Python-based Twitter scraping tool. After a comprehensive search for stock-related remarks that were either directly or indirectly linked to shares under consideration, a sizable amount of text data containing daily attitudes toward stocks were created. Then, the collected textual sets underwent the various preprocessing procedures necessary in order to be passed on to the classification modules for extracting their respective sentiment scores.

Regarding preprocessing tweets, initially, irrelevant hyperlinks and URLs were removed using the *Re* Python library [54]. Each tweet was then converted to lowercase and split into words. Then, unwanted phrases from a manually produced list and various numerical strings were also dismissed. After performing the necessary joins to restore each text to its original structure, each tweet was tokenized in terms of its sentences using the *NLTK* [55,56] library. Lastly, using the *String* [57] module, punctuation removal was applied. The whole text-preprocessing step is schematically presented in Figure 2.

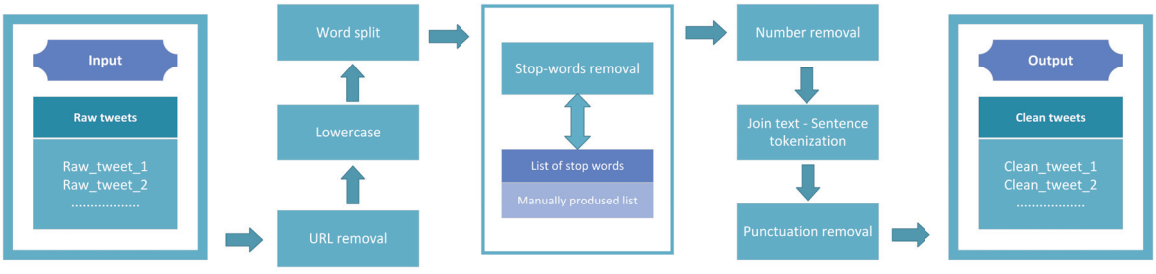


Figure 2. Preprocessing.

3.1.3. Sentiment Analysis

The subsequent process involved extracting sentiment scores from the gathered yet cleaned tweets. To perform the sentiment quantification step, three different sentiment analysis methods were utilized.

Specifically, the procedure included extracting sentiment scores from *TextBlob* [58], using the *Vader* sentiment analysis tool [59], and incorporating *FinBERT* [60]. FinBERT is a financial-based fine-tuning of the *BERT* [61] language representation model. Using each of the above methods, daily sentiment scores were extracted for each stock. The daily mean was then extracted, forming the final collection, which constituted the sentiment-valued time series of every corresponding method. Then, 7- and 14-day moving averages were applied to the previously extracted sentiment score time series. This resulted in the extraction of nine sentiment time series, which, together with the application of the aforementioned procedure to the closing price time series, led to the final number of 12 generated time series used as features. Various combinations of the above features, along with the univariate case scenario, resulted in 67 different study cases. These data constituted the distinct experimental procedures that run for every algorithm. The use of three different methods of sentiment analysis has already been mentioned. Below, a rough description of these methods is given. For further information, the reader is advised to refer to the respective papers.

- TextBlob: The *TextBlob* module is a Python-based library for performing a wide range of manipulations over text data. The specific TextBlob method used in this work is a *rule-based* sentiment-analysis scheme. That is, it works by simply applying manually created rules. This is how the value attributed to the corresponding sentiment score is calculated. An exemplified snapshot of the process would be counting the number of times a term of interest appears within a given section. This would modify the

projected sentiment score values in line with the way the phrase is assessed. Here, within this experimental setup and by exploiting TextBlob's *sentiment* property, a real number within the $[-1, 1]$ interval representing the sentiment polarity score was generated for each tweet. The algorithm's numerical output was then averaged using the individual scores of each tweet to obtain a single sentiment value representing the users' daily attitudes;

- Vader: *Vader* is also a straightforward *rule-based* approach for realizing general sentiment analysis. In the context of this work, the Vader sentiment analysis tool was used in order to extract a compound score produced by a normalization of sentiment values that the algorithm calculates. Specifically, given a string, the procedure outputs four values: negative, neutral, and positive sentiment values, as well as the aforementioned composite score used. A normalized average of all compound scores for each day was generated the usual way. The resulting time series contained daily sentiment scores that ranged within the $[-1, 1]$ interval;
- FinBERT: Regarding *FinBERT*, in this work, the implementation contained in [62] was utilized. Specifically, the model that was trained on *PhraseBank* presented in [63] was used. Again, first, the daily scores regarding sentiment attitudes were extracted to eventually form a daily average time series. Generally, the method is a pre-trained *natural-language-processing* (NLP) model for sentiment analysis. It is produced by simply fine-tuning the pre-trained *BERT* model over financial textual data. BERT, meaning *bidirectional encoder representations from transformers*, is an implementation of the *transformers* architecture used for natural language processing problems. The technique is basically a pre-trained representational model based on transfer learning principles. Given textual data, multi-layer deep representations are trained with a bidirectional attention strategy so that the various different contexts of each linguistic token constitute the content of the token's embedding. Regardless of data references—here financial—the model can be fine-tuned in any domain by only using a single additional layer that addresses the specific tasks.

3.2. Algorithms

In this section, the methods, algorithmic schemes, and architectures employed in the experiments are listed. Additional details are given on the implementation framework and the tools used.

Regarding the algorithms used, a total of 30 different state-of-the-art methods and method variations were compared. The number of 30 methods used results from the supplementation of the set of well-known core methods with their variations. Further details can be found in the cited *tsAI* library [64], using which the implementation was carried out. However, it is this multitude of methods that apparently makes a detailed presentation practically impossible. Nevertheless, the reader is urged to track the cited papers. Table 2 contains the main algorithms utilized during the experimental procedure along with a corresponding citation. There, among others, one can notice that in addition to a multitude of state-of-the-art methods, implementations involving combinations of the individual architectures were also used. Note that in addition to the corresponding papers, information regarding the variations of the basic algorithms employed can be searched, inter alia, in notebook files taken from the library implementations.

In order to carry out the experiments, the Python library *tsAI* [64] was used. The *tsAI* module is "an open-source deep learning package built on top of Pytorch and Fastai focused on state-of-the-art techniques for time series tasks like classification, regression, forecasting" [64], and others. Here, the forecasting procedure was essentially treated as a predictive regression problem. In the experiments, the initial parameters of the respective methods from the library were preserved with the implementation environment being kept fixed for all algorithmic schemes. Thus, all algorithms compared were utilized in the most basic configuration. That way, one can gain additional insight regarding implementing high-level yet low-code programming and data analysis in real-world tasks. Of the data,

20% were used as the test set. Regarding prediction time horizons, three forecast scenarios were implemented: one single-step and two multi-step. In particular, with regard to multi-step forecasts, and leaving aside the single-step predictions, estimates were provided for a seven-day window on the one hand and a fourteen-day window on the other. The results were evaluated according to the metrics presented in the following paragraph.

Table 2. Algorithms.

No.	Abbreviation	Algorithm [1]
1	FCN	Fully Convolutional Network [65]
2	FCNPlus	Fully Convolutional Network Plus [66]
3	IT	Inception Time [67]
4	ITPlus	Inception Time Plus [68]
5	MLP	Multilayer Perceptron[65]
6	RNN	Recurrent Neural Network [69]
7	LSTM	Long Short-Term Memory [70]
8	GRU	Gated Recurrent Unit [71]
9	RNNPlus	Recurrent Neural Network Plus [69]
10	LSTMPus	Long Short-Term Memory Plus [69]
11	GRUPlus	Gated Recurrent Unit Plus [69]
12	RNN_FCN	Recurrent Neural—Fully Convolutional Network [72]
13	LSTM_FCN	Long Short-Term Memory—Fully Convolutional Network [73]
14	GRU_FCN	Gated Recurrent Unit—Fully Convolutional Network [74]
15	RNN_FCNPlus	Recurrent Neural—Fully Convolutional Network Plus [75]
16	LSTM_FCNPlus	Long Short-Term Memory—Fully Convolutional Network Plus [75]
17	GRU_FCNPlus	Gated Recurrent Unit—Fully Convolutional Network Plus [75]
18	ResCNN	Residual—Convolutional Neural Network [76]
19	ResNet	Residual Network [65]
20	RestNetPlus	Residual Network Plus [77]
21	TCN	Temporal Convolutional Network [78]
22	TST	Time Series Transformer [79]
23	TSTPlus	Time Series Transformer Plus [80]
24	TSiTPlus	Time Series Vision Transformer Plus [81]
25	Transformer	Transformer Model [82]
26	XCM	Explainable Convolutional Neural Network [83]
27	XCMPlus	Explainable Convolutional Neural Network Plus [84]
28	XceptionTime	Xception Time Model [85]
29	XceptionTimePlus	Xception Time Plus [86]
30	OmniScaleCNN	Omni-Scale 1D-Convolutional Neural Network [87]

[1] Methods and method variations used.

3.3. Metrics

Regarding performance evaluation, six metrics were used. The use of the different metrics serves the necessity of having not only a presentation of the conclusions of a large comparison of methods and feature and sentiment setups but also a number of diverse extractions in terms of evaluation aspects that can be used in future research. This is exactly because each of the metrics exposes the results in different aspects, and therefore, an investigation would be incomplete if it focused on just one of them. Thus, regarding evaluating results, each one of the six performance indicators utilized has advantages and disadvantages. The metrics used are:

- the *Mean Absolute Error* (MAE);
- the *Mean Absolute Percentage Error* (MAPE);
- the *Mean Squared Error* (MSE);
- the *Root Mean Squared Error* (RMSE);
- the *Root Mean Squared Logarithmic Error* (RMSLE);

- the *Coefficient of Determination* R^2.

In what follows, a rather detailed description of aspects of the aforementioned well-known evaluation metrics is given. The presentation aspires to provide details and some insight regarding the interpretation of the metrics. Below, the actual values are denoted by y_{a_i} and the forecasts are denoted by y_{p_i}.

3.3.1. MAE

First is MAE:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_{p_i} - y_{a_i}| \tag{1}$$

MAE stands for the arithmetic *mean of the absolute errors*, and it is a very straightforward metric and easy to calculate. By default, in terms of the difference between the prediction and the observation, the values share the same weights. The absence of exponents in the analytic form ensures good behavior, which is displayed even when outliers are present. The target variable's unit of measurement is the one expressing the results. MAE is a scale-dependent error metric; that is, the scale of the observation is crucial. This means that it can only be used to compare methods in scenarios where every scheme incorporates the same specific target variable rather than different ones.

3.3.2. MAPE

Next is MAPE:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_{p_i} - y_{a_i}|}{|y_{a_i}|} \tag{2}$$

MAPE is the *mean absolute percentage error*. It is a relative and not an absolute error measure. MAPE is common when evaluating the accuracy of forecasts. It is the average of the absolute differences between the prediction and the observations divided by the absolute value of the observation. A multiplication by 100 can afterwards convert this output to a percentage. This error cannot be calculated when the actual value is zero. Instead of being a percentage, in practice, it can take values in $[0, \infty)$. Specifically, when the predictions contain values much larger than the observations, then the MAPE output can exceed 100%. Conversely, in cases where both the prediction and the observation contain low values, the output of the metric may deviate greatly from 100%. This, in turn, can lead to a misjudgment of the model's predictive capabilities, believing them to be limited when, in fact, the errors may be low. MAPE attributes more weight to cases where the predicted value is higher than the actual one. These cases produce larger errors. Hence, using this metric is best suitable for methods with low prediction values. Lastly, MAPE, being not scale-dependent, can be used to evaluate comparisons of a variety of different time series and variables.

3.3.3. MSE

The next metric is MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_{p_i} - y_{a_i})^2 \tag{3}$$

MSE stands for *mean squared error*. It constitutes a common forecast evaluation metric. The mean squared error is the average of the squares of the differences between the actual and predicted values. Its unit of measurement is the square of the unit of the variable of interest. Looking at the analytical form, first, the square of the differences ensures the non-negativity of the error. At the same time, it makes information about minor errors usable. It is obvious, at the same time, that larger deviations entail larger penalties, i.e., a higher MSE. Thus, outliers have a big influence on the output of the error; that is, the existence of such extreme values has a significant impact on the measurements and, consequently, the evaluation. Furthermore, and in a sense the other way around, when differences are less than 1, there is a risk of overestimating the predictive capabilities of the model. Given the error's differentiability, as one can observe, it can easily be optimized.

3.3.4. RMSE

Moving on to RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{p_i} - y_{a_i}\right)^2} \tag{4}$$

RMSE stands for *root mean squared error*. It is a common metric for evaluating differences between estimated values and observations. To compute it, apparently, one just calculates the root of the mean squared error. From the numerical formulation, one can think of the metric as an abstraction that captures the representation of something of an average distance between the actual values and the predictions. That is, if one ignores the denominator, then one can observe the formula as being the Euclidean distance. The subsequent interpretation of the metric as a kind of normalized distance comes out of the act of division by the number of observations. Here also, the existence of outliers has a significant impact on the output. In terms of interpreting error values, the RMSE is expressed in the same units as the target variable and not in its square, as in the MSE, making its use straightforward. Finally, the metric is scale-dependent; hence, one can only use it to evaluate various models or model variations given a particular fixed variable.

3.3.5. RMSLE

The next metric is also an error. The formula for RMSLE is as follows:

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log(y_{p_i}+1) - \log(y_{a_i}+1)\right)^2} \tag{5}$$

RMSLE stands for Root Mean Squared Logarithmic Error. The RMSLE metric seems as if it is a modified version of the MSE. Using this modification is preferred when predictions display significant deviations. RMSLE uses logarithms of both the observations and predicted values while ensuring non-zero values in the logarithms through the appropriate simple unit additions appearing in the formula. This modified version is resistant to the existence of outliers and noise, and it smooths the penalty that the MSE imposes in cases in which predictions deviate significantly from observations. The metric cannot be used when there are negative values. RMSLE can be interpreted as a relative error between observations and forecasts. This can be made evident by simply applying the following property to the radicand term of the square root:

$$\log(y_{p_i}+1) - \log(y_{a_i}+1) = log\left(\frac{y_{p_i}+1}{y_{a_i}+1}\right) \tag{6}$$

Since RMSLE gives more weight to cases where the predicted value is lower than the actual value, it is quite a useful metric for types of predictions where similar conditions require special care for the reliability of the application in real-world conditions, where lower forecasts may lead to specific problems.

3.3.6. R^2

The last metric is the coefficient of determination R^2:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_{i=1}^{n}\left(y_{p_i} - y_{a_i}\right)^2}{\sum_{i=1}^{n}\left(y_{p_i} - \bar{y}\right)^2} \tag{7}$$

The coefficient of determination R^2 is not an error evaluation metric. It is the ratio depicted in the above equation. This metric is essentially not a measure of model reliability. R^2 is a measure of how good a fit is: a quantification of how well a model *fits* the data. Its values typically range from 0 to 1. A rather simple interpretation would be this: the closer to 1 the value of the metric is, the better the model fits the observations, i.e., the predictions are closer, in terms of their values, to the observations. Thus, the value 0 corresponds to cases where the explanatory variables do not explain the variance of the dependent variable

at all. Conversely, the value 1 corresponds to cases where the explanatory variables fully explain the dependent variable. However, this interval does not strictly constitute the set of values of the metric. There are conditions in which R^2 could take negative values. Observing the formula, one can identify the above as permissible. In such cases, the model performs worse in fitting the data than a simple horizontal line, essentially being unable to follow the trend. Lastly, values outside the above range indicate either an inadequate model or other flaws in its implementation.

4. Results

Returning to the dual objective of this work, the two case studies whose results will be presented in this chapter were:

- On the one hand, the comparison of a large number of time series forecasting contemporary algorithms;
- On the other hand, the investigation of whether knowledge of public opinion, as reflected in social networks and quantified using three different sentiment analysis methods, can improve the derived predictions.

Accordingly, the presentation of the results of the experimental process is split into two distinct parts. In what follows, both various statistical analysis and visualization methods are incorporated. However, it should be noted that the number of comparisons performed yielded a quite large volume of results. Specifically, as already pointed out, in each case, the performance of the 30 predictive schemes and the 67 different feature setups was investigated over three different time frames (1, 7, and 14 day shifts). Note that these three time-shifting options have no—or at least no intended—financial consequences. Here, the primary goal in designing the framework was to forecast the stock market over short time frames, such as a few days. Then, an expansion was made to investigate the performance of both methods and feature setups over longer periods of time. Each of these schemas was evaluated with six different metrics, while the process was repeated for each of the datasets. Consequently, it becomes clear that the complete tables with the numerical results cannot contribute satisfactorily to the understanding of the conclusions drawn. Below, following a necessary brief reminder of the process, results are presented.

As has already been mentioned, during the procedure, for each of the stocks, the following strategy was followed: each of the thirty algorithms to be compared was "ran" 67 times, each time accepting as input one of the different feature setups. This was repeated three times, once for each of the three forecast time frames. In each of the above runs, the six metrics used in the evaluation of the results were calculated. The comparison of the algorithms was performed by using Friedman's statistical tests in terms of feature setups for each of the time shifts. Thus, given setups and stocks, the ranking of the methods per evaluation metric was extracted according to the use of the Friedman test [88]. Therefore, regarding this case study, a total of $67 \times 6 \times 3 = 1206$ *statistical tests* were executed. In a similar way, the Friedman rankings of input feature setups were estimated in terms of metrics and time shifts, given the various algorithms and stocks. Here, a total of $30 \times 6 \times 3 = 540$ *statistical tests* were performed. An additional abstraction of the results was derived as follows: For each of the 30 methods, the average rank achieved by each method in terms of feature setups and shares was calculated. So, for each metric and each of the three time frames, a more comprehensive display of the information was obtained based on the average value of the different setups. In an identical way, in the case of checking the effectiveness of features, the average value of the 30 algorithms for each of the 67 different input setups was calculated in each case. In both cases, the ranking was calculated based on the positions produced by the Friedman test, while at the same time, with the Nemenyi post hoc test [89] that followed, every schema was checked pair-wise for significant differences. The results of the Nemenyi post hoc tests are shown in the corresponding Critical Difference diagrams (CD-diagrams), in which methods that are not significantly different are joined by black horizontal lines. Two methods are considered not significantly different when the difference between their mean ranks is less than the CD value.

Next, organized in both cases based on time frames, the results concerning the comparison of the forecast algorithms are presented, which are followed by those regarding the feature setups.

4.1. Method Comparison

The presentation begins with results concerning the investigation of methods. The results are presented per forecast time shift. In each case, the Friedman Ranking results for all six metrics are listed. To save space, only methods that occupy the top ten positions of the ranking are listed. Full tables are available at: shorturl.at/FTU06 (accessed on 15 January 2023). The CD diagrams follow. There, we can visually observe which of the methods exhibit similar behavior and which differ significantly. Finally, box plots of results per metric are presented, again for the best 10 methods. The box plots present in a graphical and concise manner information concerning the distribution of the aforementioned data, that is, in our case, the average values of the sentiment setups per algorithm for all stocks. In particular, one can derive information about the maximum and minimum value of the data, the median, as well as the 1st and 3rd quartile values isolated by 25% and 75% of the observations, respectively.

4.1.1. Time Shift 1

With respect to the one-day forecasts, Table A1 lists the Friedman Ranking results for the top 10 scoring methods per metric. Although there is no single method that dominates all metrics and significant reorderings are also observed in the table positions, the TCN method achieves the best ranking in three out of six metrics (MAPE, R2, and RMSLE) and is always in the top four. Furthermore, from the box plots, it is evident that TCN has by far the smallest range of values.

Apart from this, in all metrics, GRU_FCN is always in the top five. It is also observed that LSTM_FCN and LSTMPlus behave equally well. The latter shows a drop in the MAPE metric, but in all other cases, it is in the top three, while in two metricsm it ranks first. It should also be noted that the LSTMPlus method ranks first in two metrics, namely MAE and RMSE. In terms of R^2 and RMSLE, it occupies the second position of the ranking, while regarding MSE, LSTMPlus ranks third. However, at the same time, according to MAPE, the method is not even in the top ten. Thus, as will be seen in the following, TCN is the consistent choice.

The results produced by Friedman's statistical test, in terms of the six metrics, are presented in Table A1, while the corresponding CD diagrams and box plots are depicted in Figures 3 and 4.

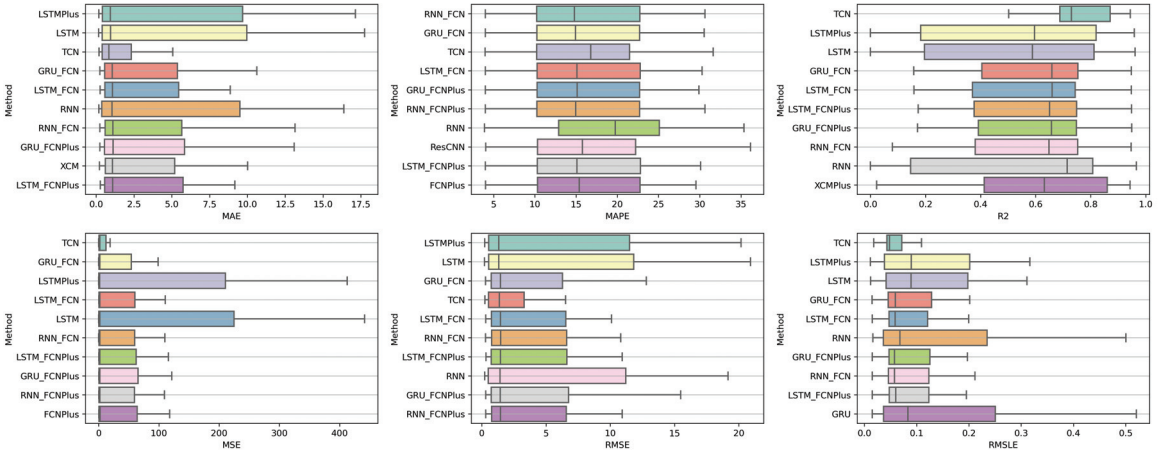


Figure 3. Box Plots: Methods—Shift 1.

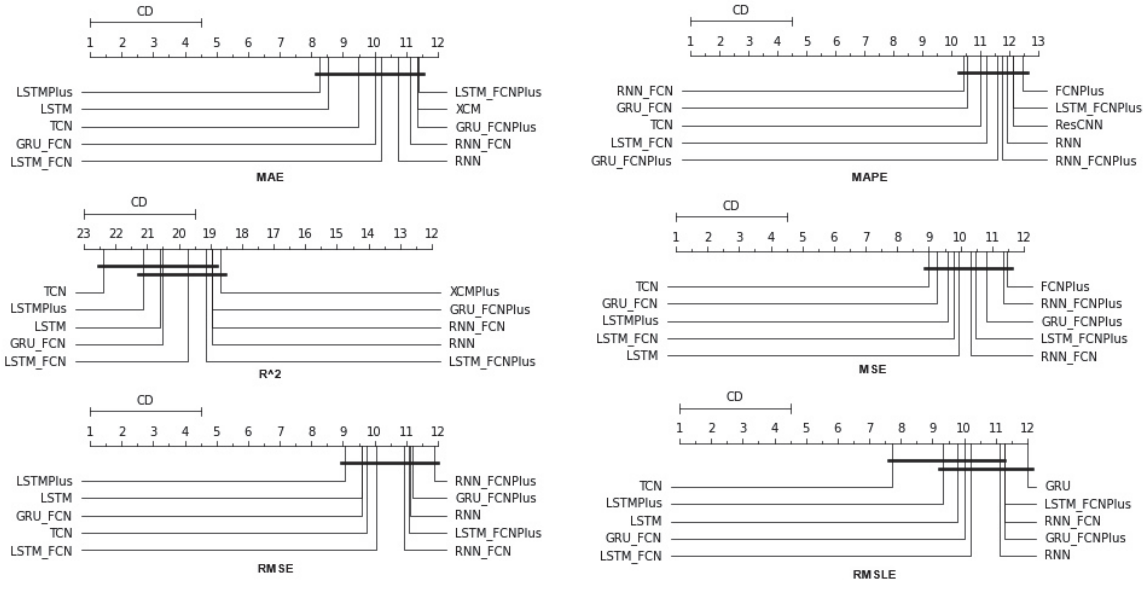


Figure 4. CD Diagrams: Methods—Shift 1.

4.1.2. Time Shift 7

At the one-week forecast time frame, the algorithms that occupy the top positions in the ranking produced by the statistical control appear to have stabilized. The corresponding ranking produced by the Friedman statistical test regarding the ten best methods with respect to the six metrics is presented in Table A2. In all metrics, the TCN method ranks first. From the CD diagrams, it can be seen that in all metrics—except for R2—this superiority is also validated by the fact that this method differs significantly from the others. Box plots show the method also having the smallest range around the median. Figures 5 and 6 contain the relevant results in the form of box plots and CD-diagrams.

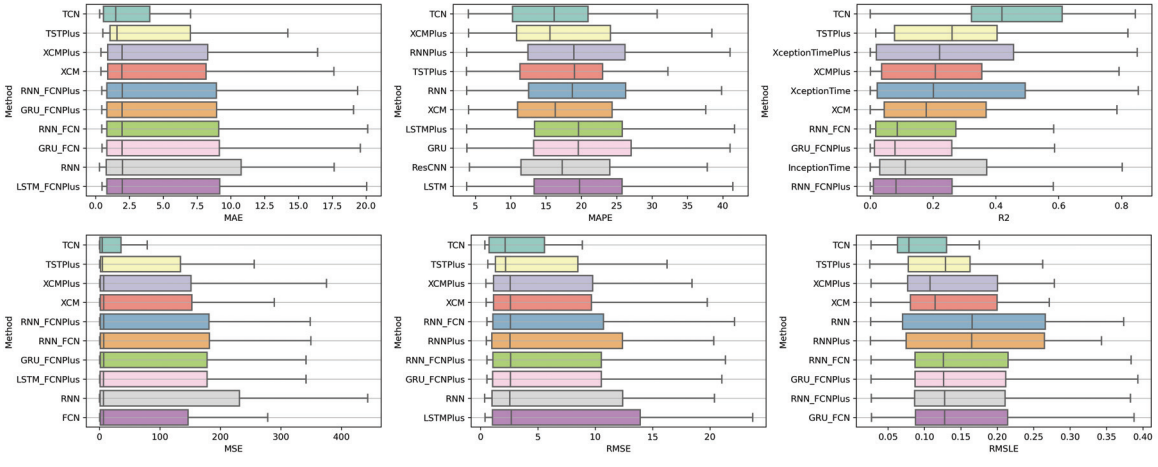


Figure 5. Box Plots: Methods—Shift 7.

Other methods that clearly show some dominance over the rest in terms of given performance ratings are, on the one hand, TSTPlus, which ranks second in all metrics

except MAPE, and, on the other hand, XCMPlus and XCM, which are mostly found in the top five. In general, the same methods can be found in similar positions in all metrics, with minor rank variations. In addition, the statistical correlations between the methods are shown in the CD diagram plots.

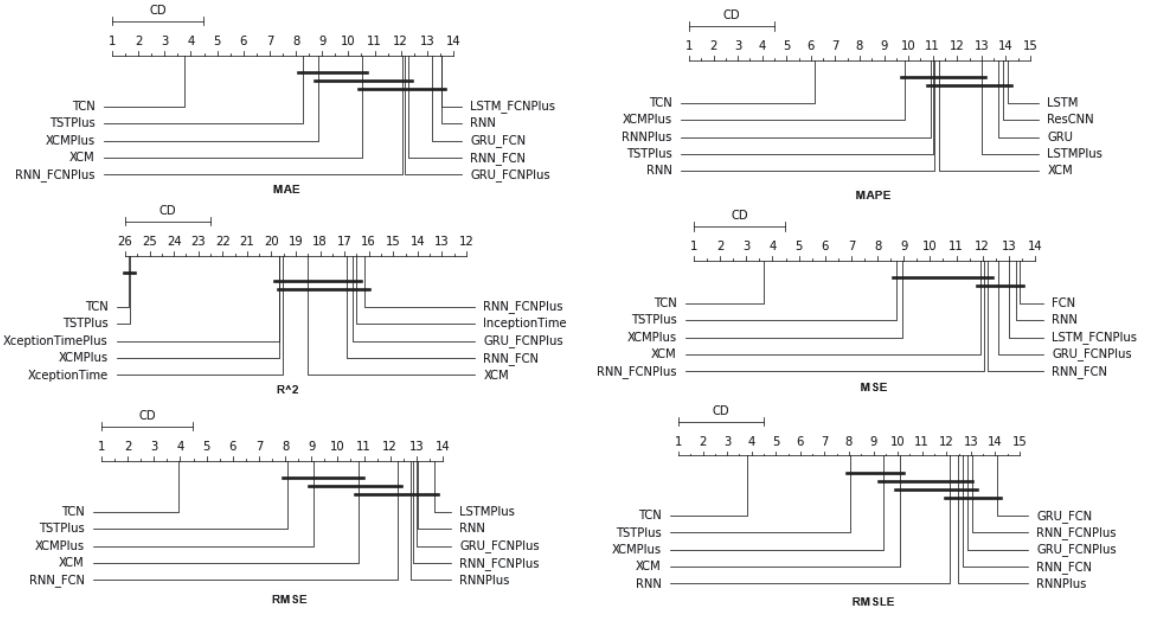


Figure 6. CD Diagrams: Methods—Shift 7.

4.1.3. Time Shift 14

In the forecast results with a two-week shift, a relative agreement can be seen in the top-ranking algorithms with those of the one-week frames. The ranking produced by the Friedman statistical test for the ten best methods with respect to the six metrics is presented in Table A3.

Once more, TCN ranks first in all metrics. TSTPlus again ranks second in all metrics except for R2, where it ranks third. In almost all cases, XCMPlus and RNNPlus appear in the top five. Likewise, as in the previous time shift, there is a relative agreement in the methods appearing in the corresponding positions regarding all metrics. Moreover, according to the above, an argument regarding the general superiority of the TCN method in this particular scenario is easily obtained. An obvious predominance of the TCN method is established. The corresponding CD diagrams and box plots for the 10 best performing algorithms are seen in Figures 7 and 8.

4.2. Feature Setup Comparison

Now, we are moving on to the findings of the second case study, which concern, on the one hand, the investigation of whether the use of sentiment analysis contributes to the improvement of the extracted predictions and, on the other hand, the identification of specific feature setups whose use improves the model's predictive ability.

Again, the results of the experimental procedure will be presented separately for the three forecast time frames. Likewise, due to the volume of results, only the 10 most promising feature setups will be listed. These were again derived based on the Friedman classification of the averages calculated for each of them, taking into account the predictions in the use of the 30 forecast methods used. The full rankings of all 67 setups can be found

at shorturl.at/alqwx (accessed on 13 December 2022). For the presentation below, again, the corresponding CD diagrams and box plots were used.

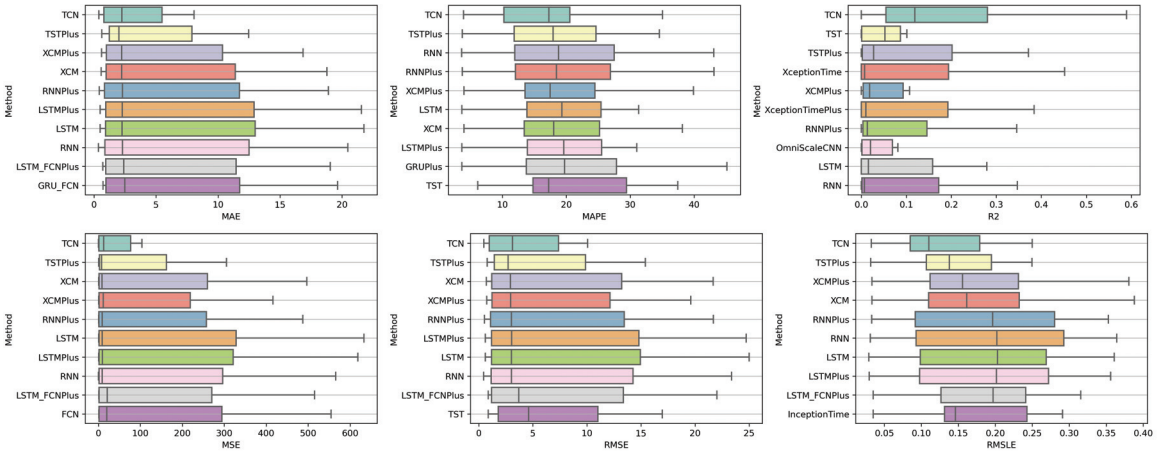


Figure 7. Box Plots: Methods—Shift 14.

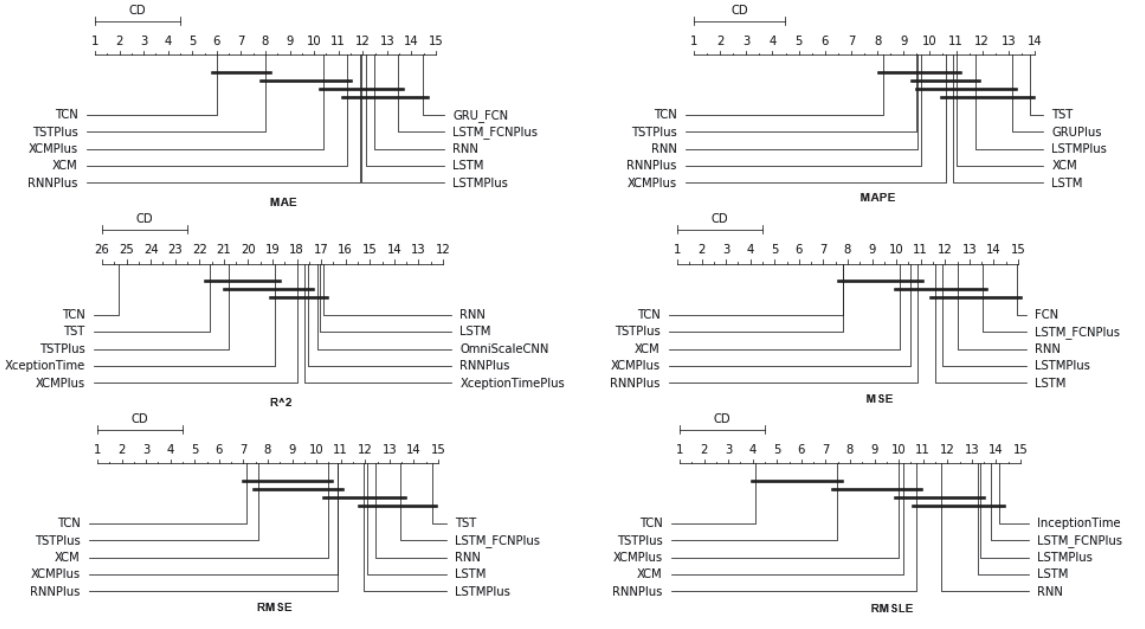


Figure 8. CD Diagrams: Methods—Shift 14.

4.2.1. Time Shift 1

Starting with the results concerning one-day depth forecasting, one notices that the univariate version, in which the forecasts are based only on the stock price of the previous days, ranks first only in the case of the R^2 metric. In fact, in three metrics, the univariate version is not even in the top twenty of the ranking (See Figures 9 and 10).

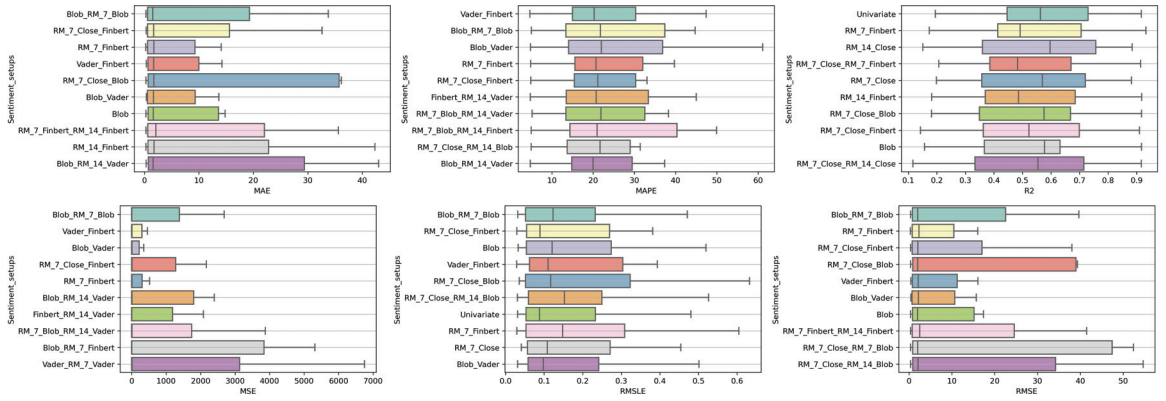


Figure 9. Box Plots: Features—Shift 1.

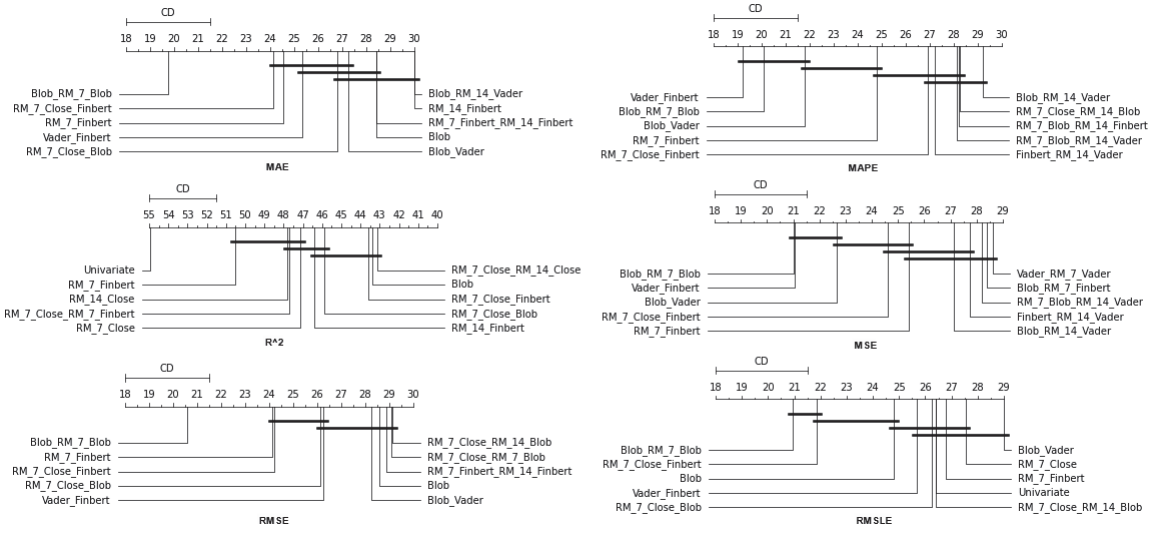


Figure 10. CD Diagrams: Features—Shift 1.

Another interesting observation would be that even though there are rerankings of the sentiment setups in terms of their performance on the six metrics, the Blob_RM_7_Blob setup—that is, the setup incorporating Blob and Rolling Mean 7 Blob along with the closing values time series—although it does not score well in the ranking regarding R^2, it is, on the one hand, at the top ranking in four metrics, that is, MAE, MSE, RMSE, RMSLE, and, on the other hand, second in MAPE. Moreover, from the results, it becomes evident that an argument in favor of using sentiment analysis in multivariate time series layouts, even in the case where the forecasts concern one-day depth, is, at least, relevant. At the same time, using smoothed versions of both the sentiment time series and those containing the closing stock price values appears to be beneficial in general.

4.2.2. Time Shift 7

Regarding the time frame of one week, one can notice that the use of the univariate version is marginally ranked first in three metrics, namely, the R^2, RMSE and RMSLE, while in two metrics, the Vader sentiment setup appears to be superior, actually being, at the

same time, in second place regarding the MAPE and RMSE metrics and fifth regarding the RMSLE (Figures 11 and 12).

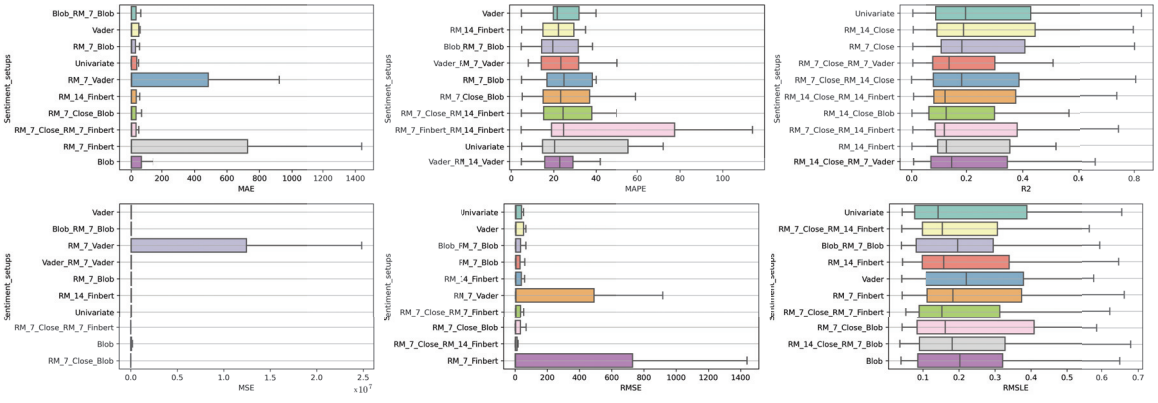


Figure 11. Box Plots: Features—Shift 7.

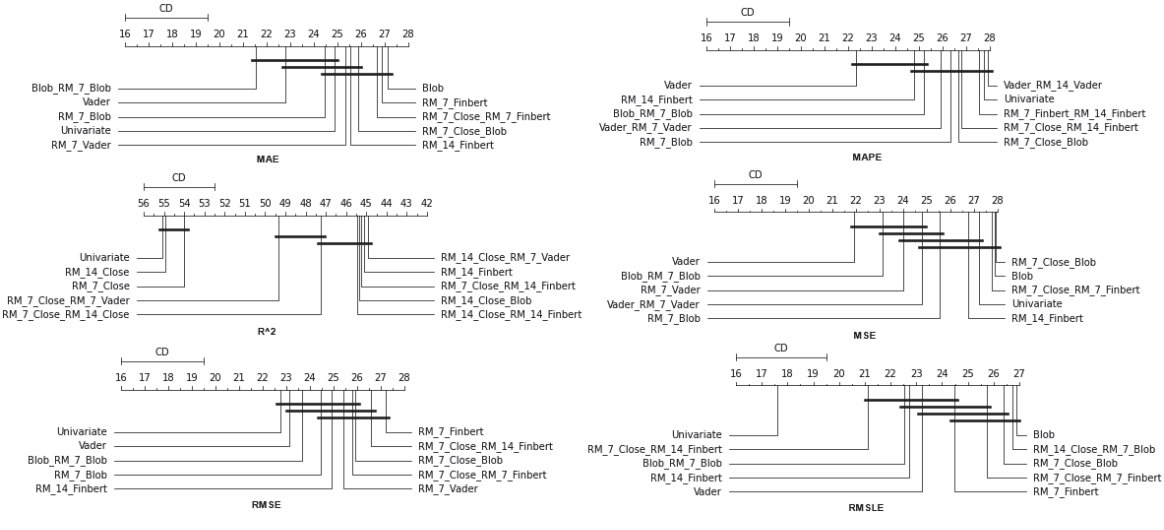


Figure 12. CD Diagrams: Features—Shift 7.

It is also notable that Blob_RM_7_Blob, which appeared to perform particularly well during the one-day shift, remains in the top three rankings in five of the six metrics. More generally, once again, one notices that there are rearrangements, especially in the central positions of the table. However, given the small differences in performance between the different setups, this should not be considered unreasonable. Overall, the picture still points in favor of using multivariate inputs containing sentiment data.

4.2.3. Time Shift 14

Finally, regarding the two-week time frame, a first observation is that in relation to the R^2, a feature setup that does not contain sentiment data dominates. This pattern is also present in the previous time shifts (See Figures 13 and 14).

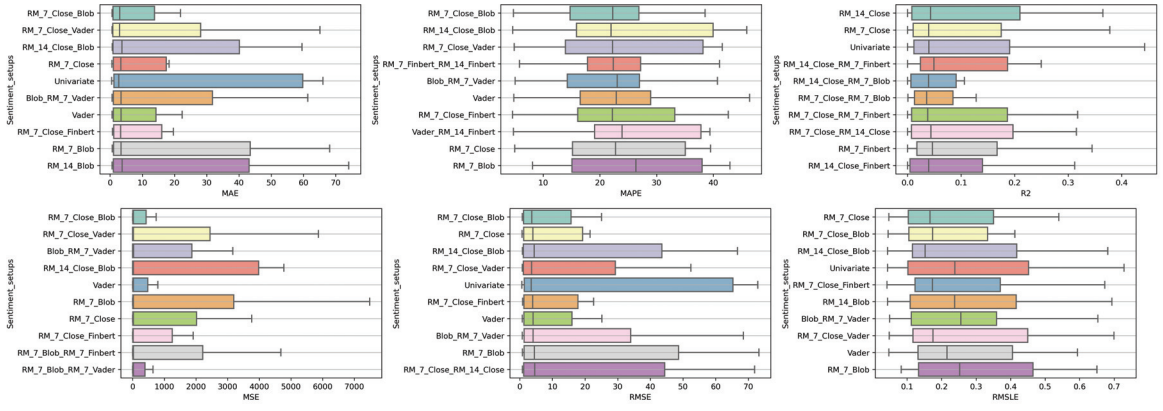


Figure 13. Box Plots: Features—Shift 14.

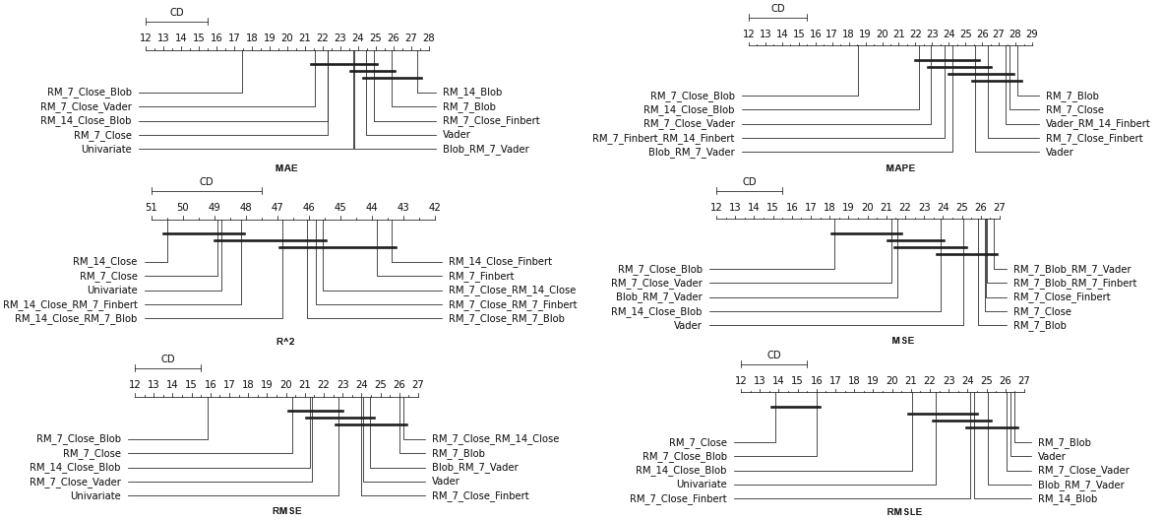


Figure 14. CD-Diagrams: Features—Shift 14.

In addition, although there are metrics in which the univariate version is in the top ten, in these cases, the difference in its performance with those in the first positions is quite significant. This is easily seen from the CD diagrams: there are no connections with setups that appear in the top positions. At the seven-day time lag, it was observed that the univariate version prevailed in three cases. However, as one examines the 14-day time shift, one notices that the superiority of methods that use sentiment data is reinforced.

At the same time, combinations containing the closing price appear in the first positions of the table more often than in the previous two setups. Furthermore, it is observed that the setup that dominates four of the six metrics is RM_7_Close_Blob. These metrics are MAE, MAPE, MSE, and RMSE. The RM_7_Close_Blob feature setup is the one that incorporates both a smoothed version of the closing values as well as sentiment scores. Thus, the use of weighted averages in the original time series along with the incorporation of sentiment scores is mostly shown to be optimal regardless of the individual choice of a specific layout. Methodologically, the utilization of both has an improving effect.

5. Conclusions

Some general conclusions drawn from the whole experimental procedure will now be addressed. The discussion will follow the binary separation of the preceding case studies.

5.1. Methods

The first case study of the paper consisted of a comparison of 30 methods for time series forecasting. Within the above-discussed experimental context, the extracted results are such as to safely allow a conclusion regarding the superiority of the TCN method over the rest. This is the case because, in the vast majority of comparisons, it excels, being, for the most part, at the top of the Friedman ranking. In particular, the only cases where it does not outperform all the rest are found in the single-day time frame predictions. In fact, from the CD diagrams, one can extract the additional fact that in many cases, the superiority of the aforementioned method is marked by a significant difference. Furthermore, in addition to the TCN method, other methods whose predictive capacities can be considered significant were identified. TSTPlus is one of them, as it produces significant results, particularly over longer time horizons. XCMPlus is another.

In Figure 15, one can see the relative rankings of these three methods per time shift. The values in Figure 15 correspond to the values of Tables A1–A3. Regarding the one-day forecast window, LSTMPlus is an additional option, as is the combination of GRU and FCN. However, an additional point to note here is that the individual method differences are less clear in their significance. On the contrary, there can also be conclusions regarding methods whose behavior was not evaluated, on average, as satisfactory. In particular, specific methods that are always ranked last in all scenarios were identified. Specifically, TSiTPlus ranks last in all three scenarios across all metrics. In addition to this, there are methods, such as Transformer Model, XceptionTime, and XceptionTimePlus, which are always at the bottom of the table in the vast majority of cases. In conclusion, given the limitations and further prerequisites developed throughout this paper, TCN can be easily recommended.

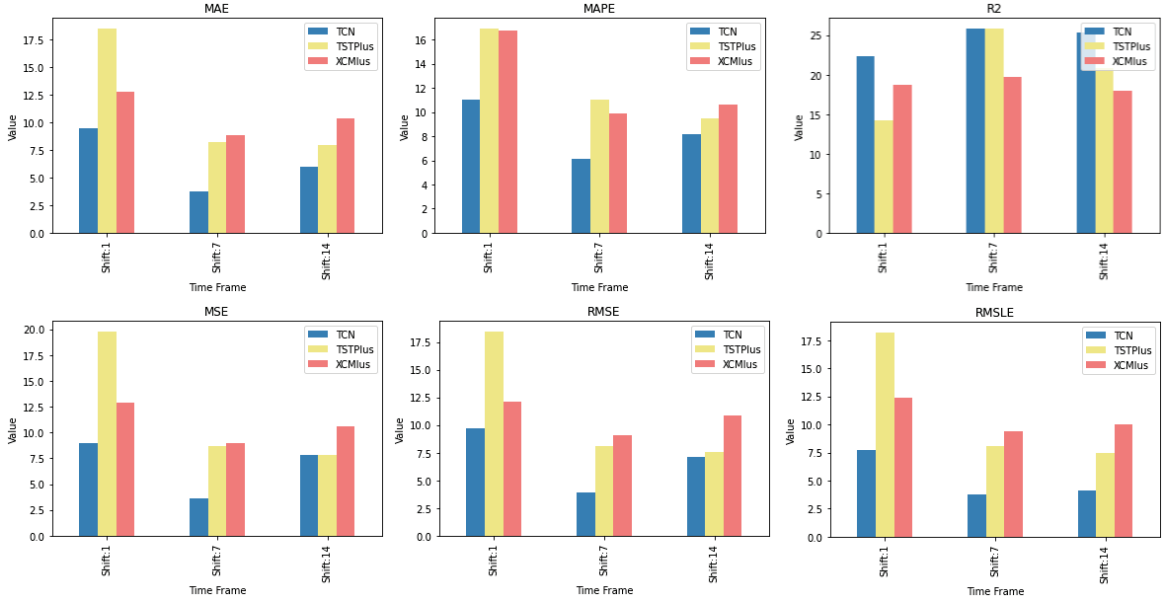


Figure 15. TCN, TSTPlus and XCMPlus relative rankings.

5.2. Feature and Sentiment Setups

In relation to the second case study, the consideration of the results also points in some important directions. Of these, the main conclusion drawn seems to be that the use of information derived from both smoothed versions of the initial time series and sentiment analysis shows, in most cases, to have a beneficial effect on the derived forecasts. Not using sentiments in the feature setup of the inputs dominates the rest only in a small number of cases, and, as confirmed by the CD diagrams, only in two of them is this difference significant.

Moreover, the answer to whether the use of sentiment setups specifically leads to the extraction of more accurate forecasts, as evidenced by the individual layouts of the weighted results, seems to be that, in general, sentiment analysis improves forecasts. Of course, it is also reasonable to investigate whether there is a specific sentiment setup that outperforms the rest. This would also lead to an assessment of the performance of the three sentiment analysis methods used. However, the answer to this question needs further investigation. However, even with the possibility of further inquiries within the framework of the experimental setup presented here, it is still not certain that firm conclusions will be drawn. Here, while such setups can be found for each time horizon, there is not one that dominates all three.

In order, however, to illustrate a relative ranking of the three sentiment analysis methodologies used, regardless of the particular variation involved, an additional table was created. All variations of each method were placed under a corresponding class. The Friedman-aligned ranks [90] were then calculated. Hence, in order to draw a clearer picture of the way the three employed approaches to sentiment analysis performed, three sentiment classes were formed, one matching each of the previously described sentiment analysis methods. The arithmetic mean of all the sentiment setups that solely contain different variations of a particular sentiment analysis algorithm, that is, only one of the three incorporated, is used to represent the corresponding class concerning each metric. In other words, each class represents a sentiment analysis method, and each class corresponds to six sentiment setups that contain variations exclusively of the technique in question. Specifically, a representative value of a class, as it pertains to a particular *method*, is formed by the following setups: *method*, RM7*method*, RM14*method*, *method* + RM7*method*, *method* + RM14*method*, and RM7*method* + RM14*method*. The sum is then divided by six, which is apparently the number of setups, and this result is the output value to be depicted. This way, setups produced either by combining the various sentiment analysis methods or by using the target variable in variants containing rolling means are excluded in order to compare only the relative performances of the three individual techniques and their variations.

Figure 16 illustrates these relative rankings of the three sentiment analysis methods per time shift. One can observe the relative performances in terms of individual wins with respect to each metric and time shift: the *Blob* and *Vader* classes top the ranking seven times each, while the *Finbert* class only has four wins. Again, a conclusion in terms of an obvious generality regarding a specific algorithm does not appear. Nevertheless, the identification of groups of such setups, even at the level of a specific time frame, can be particularly useful, with the methodology for the selection of individual setups needing more investigation.

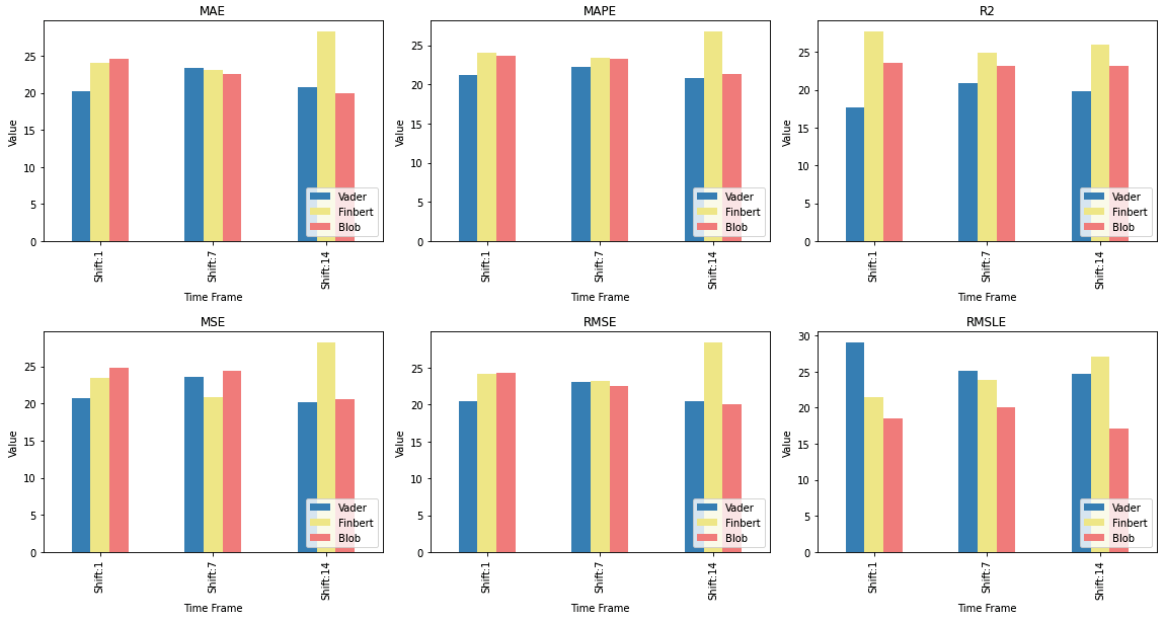


Figure 16. Sentiment rankings.

Author Contributions: Conceptualization, C.M.L. and S.K.; methodology, C.M.L.; software, C.M.L.; validation, C.M.L., A.K. and S.K.; formal analysis, C.M.L. and A.K.; investigation, C.M.L. and A.K.; resources, S.K.; data curation, A.K.; writing—original draft preparation, C.M.L. and A.K.; writing—review and editing, C.M.L.; visualization, A.K.; supervision, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: URLs of the full Friedman Ranking results. (i) Methods rankings: shorturl.at/FTU06 (accessed on 15 January 2023). (ii) Feature setup rankings: shorturl.at/alqwx (accessed on 15 January 2023).

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

Appendix A

Table A1. Friedman results: Algorithms—Shift 1.

	MAE		MAPE		R^2	
	Method	**Friedman Score**	**Method**	**Friedman Score**	**Method**	**Friedman Score**
1st	LSTMPlus	8.266667	RNN_FCN	10.4	TCN	22.4
2nd	LSTM	8.533333	GRU_FCN	10.53333	LSTMPlus	21.13333
3rd	TCN	9.466667	TCN	11	LSTM	20.6
4th	GRU_FCN	10	LSTM_FCN	11.2	GRU_FCN	20.53333
5th	LSTM_FCN	10.2	GRU_FCNPlus	11.6	LSTM_FCN	19.73333
6th	RNN	10.73333	RNN_FCNPlus	11.73333	LSTM_FCNPlus	19.13333
7th	RNN_FCN	11.13333	RNN	11.93333	GRU_FCNPlus	18.93333
8th	GRU_FCNPlus	11.33333	ResCNN	12.13333	RNN_FCN	18.93333
9th	XCM	11.33333	LSTM_FCNPlus	12.13333	RNN	18.93333
10th	LSTM_FCNPlus	11.4	FCNPlus	12.46667	XCMPlus	18.66667

Table A1. *Cont.*

	MSE		RMSE		RMSLE	
	Method	**Friedman Score**	**Method**	**Friedman Score**	**Method**	**Friedman Score**
1st	TCN	9	LSTMPlus	9.066667	TCN	7.733333
2nd	GRU_FCN	9.266667	LSTM	9.6	LSTMPlus	9.333333
3rd	LSTMPlus	9.6	GRU_FCN	9.6	LSTM	9.8
4th	LSTM_FCN	9.8	TCN	9.733333	GRU_FCN	10
5th	LSTM	9.933333	LSTM_FCN	10.06667	LSTM_FCN	10.2
6th	RNN_FCN	10.33333	RNN_FCN	10.93333	RNN	11.13333
7th	LSTM_FCNPlus	10.46667	LSTM_FCNPlus	11.06667	GRU_FCNPlus	11.26667
8th	GRU_FCNPlus	10.8	RNN	11.13333	RNN_FCN	11.26667
9th	RNN_FCNPlus	11.33333	GRU_FCNPlus	11.2	LSTM_FCNPlus	11.26667
10th	FCNPlus	11.46667	RNN_FCNPlus	11.86667	GRU	12

Table A2. Friedman results: Algorithms—Shift 7.

	MAE		MAPE		R^2	
	Method	**Friedman Score**	**Method**	**Friedman Score**	**Method**	**Friedman Score**
1st	TCN	3.733333	TCN	6.133333	TCN	25.86667
2nd	TSTPlus	8.266667	XCMPlus	9.866667	TSTPlus	25.8
3rd	XCMPlus	8.866667	RNNPlus	10.93333	XceptionTimePlus	19.7
4th	XCM	10.53333	TSTPlus	11	XCMPlus	19.66667
5th	RNN_FCNPlus	12.06667	RNN	11.06667	XceptionTime	19.53333
6th	GRU_FCNPlus	12.13333	XCM	11.26667	XCM	18.53333
7th	RNN_FCN	12.26667	LSTMPlus	13	RNN_FCN	16.9
8th	GRU_FCN	13.2	GRU	13.66667	GRU_FCNPlus	16.66667
9th	RNN	13.53333	ResCNN	13.86667	InceptionTime	16.5
10th	LSTM_FCNPlus	13.53333	LSTM	14.06667	RNN_FCNPlus	16.16667
	MSE		**RMSE**		**RMSLE**	
	Method	**Friedman Score**	**Method**	**Friedman Score**	**Method**	**Friedman Score**
1st	TCN	3.666666667	TCN	3.933333333	TCN	3.8
2nd	TSTPlus	8.733333333	TSTPlus	8.066666667	TSTPlus	8.066666667
3rd	XCMPlus	8.933333333	XCMPlus	9.066666667	XCMPlus	9.4
4th	XCM	11.93333333	XCM	10.8	XCM	10.06666667
5th	RNN_FCNPlus	12.06666667	RNN_FCN	12.26666667	RNN	12.13333333
6th	RNN_FCN	12.2	RNNPlus	12.8	RNNPlus	12.46666667
7th	GRU_FCNPlus	12.6	RNN_FCNPlus	12.86666667	RNN_FCN	12.66666667
8th	LSTM_FCNPlus	13	GRU_FCNPlus	13	GRU_FCNPlus	12.86666667
9th	RNN	13.26666667	RNN	13.06666667	RNN_FCNPlus	13.06666667
10th	FCN	13.4	LSTMPlus	13.66666667	GRU_FCN	14.06666667

Table A3. TFriedman results: Algorithms—Shift 14.

	MAE		MAPE		R^2	
	Method	**Friedman Score**	**Method**	**Friedman Score**	**Method**	**Friedman Score**
1st	TCN	6	TCN	8.2	TCN	25.33333333
2nd	TSTPlus	8	TSTPlus	9.466666667	TST	21.6
3rd	XCMPlus	10.4	RNN	9.533333333	TSTPlus	20.8
4th	XCM	11.33333333	RNNPlus	9.666666667	XceptionTime	18.9
5th	RNNPlus	11.86666667	XCMPlus	10.6	XCMPlus	17.96666667
6th	LSTMPlus	11.93333333	LSTM	10.86666667	XceptionTimePlus	17.7
7th	LSTM	12.13333333	XCM	11	RNNPlus	17.53333333
8th	RNN	12.46666667	LSTMPlus	11.73333333	OmniScaleCNN	17.13333333
9th	LSTM_FCNPlus	13.46666667	GRUPlus	13.13333333	LSTM	17.03333333
10th	GRU_FCN	14.46666667	TST	13.8	RNN	16.93333333
	MSE		**RMSE**		**RMSLE**	
	Method	**Friedman Score**	**Method**	**Friedman Score**	**Method**	**Friedman Score**
1st	TCN	7.8	TCN	7.133333333	TCN	4.133333333
2nd	TSTPlus	7.8	TSTPlus	7.6	TSTPlus	7.466666667
3rd	XCM	10.13333333	XCM	10.46666667	XCMPlus	10
4th	XCMPlus	10.6	XCMPlus	10.86666667	XCM	10.2
5th	RNNPlus	10.86666667	RNNPlus	10.86666667	RNNPlus	10.73333333
6th	LSTM	11.6	LSTMPlus	11.93333333	RNN	11.73333333
7th	LSTMPlus	11.86666667	LSTM	12.06666667	LSTM	13.26666667
8th	RNN	12.53333333	RNN	12.4	LSTMPlus	13.33333333
9th	LSTM_FCNPlus	13.53333333	LSTM_FCNPlus	13.46666667	LSTM_FCNPlus	13.8
10th	FCN	14.93333333	TST	14.73333333	InceptionTime	14.13333333

Appendix B

Appendix B.1

Please use the abbreviation table below to read the corresponding results of the Friedman Ranks.

Table A4. Feature Setups and Abbreviations.

No.	Abbreviation	Feature Setup
1	U	Univariate
2	B	Blob
3	V	Vader
4	F	Finbert
5	RM7C	Rolling Mean 7 Closing Value
6	RM14C	Rolling Mean 14 Closing Value
7	RM7B	Rolling Mean 7 Blob
8	RM14B	Rolling Mean 14 Blob
9	RM7V	Rolling Mean 7 Vader
10	RM14V	Rolling Mean 14 Vader
11	RM7F	Rolling Mean 7 Finbert
12	RM14F	Rolling Mean 14 Finbert

Appendix B.2

Table A5. Friedman results: feature setups—Shift 1.

	MAE		MAPE		R^2	
	Feature Setup	**Friedman Score**	**Feature Setup**	**Friedman Score**	**Feature Setup**	**Friedman Score**
1st	B_RM7B	19.73333	V_F	19.2	U	54.93333
2nd	RM7C_F	24.13333	B_RM7B	20.06667	RM7F	50.53333
3rd	RM7F	24.53333	B_V	21.8	RM14C	47.8
4th	V_F	25.33333	RM7F	24.8	RM7C_RM7F	47.73333
5th	RM7C_B	26.8	RM7C_F	26.93333	RM7C	47.13333
6th	B_V	27.26667	F_RM14V	27.2	RM14F	46.4
7th	B	28.4	RM7B_RM14V	28.13333	RM7C_B	45.86667
8th	RM7F_RM14F	28.4	RM7B_RM14F	28.2	RM7C_F	43.6
9th	RM14F	30	RM7C_RM14B	28.26667	B	43.4
10th	B_RM14V	30	B_RM14V	29.2	RM7C_RM14C	43.13333
	MSE		**RMSE**		**RMSLE**	
	Feature Setup	**Friedman Score**	**Feature Setup**	**Friedman Score**	**Feature Setup**	**Friedman Score**
1st	B_RM7B	21	B_RM7B	20.6	B_RM7B	20.93333
2nd	V_F	21.06667	RM7F	24.13333	RM7C_F	21.86667
3rd	B_V	22.66667	RM7C_F	24.2	B	24.8
4th	RM7C_F	24.6	RM7C_B	26.13333	V_F	25.66667
5th	RM7F	25.4	V_F	26.26667	RM7C_B	26.26667
6th	B_RM14V	27.13333	B_V	28.26667	RM7C_RM14B	26.4
7th	F_RM14V	27.73333	B	28.6	U	26.4
8th	RM7B_RM14V	28.2	RM7F_RM14F	28.86667	RM7F	26.8
9th	B_RM7F	28.4	RM7C_RM7B	29.06667	RM7C	27.53333
10th	V_RM7V	28.6	RM7C_RM14B	29.13333	B_V	29

Table A6. Friedman results: feature setups—Shift 7.

	MAE		MAPE		R^2	
	Feature Setup	**Friedman Score**	**Feature Setup**	**Friedman Score**	**Feature Setup**	**Friedman Score**
1st	B_RM7B	21.53333	V	22.33333	U	55.06667
2nd	V	22.8	RM14F	24.8	RM14C	54.93333
3rd	RM7B	24.46667	B_RM7B	25.2	RM7C	54
4th	U	24.86667	V_RM7V	25.93333	RM7C_RM7V	49.33333
5th	RM7V	25.33333	RM7B	26.33333	RM7C_RM14C	47.23333
6th	RM14F	25.53333	RM7C_B	26.66667	RM14C_RM14F	45.46667
7th	RM7C_B	25.86667	RM7C_RM14F	26.8	RM14C_B	45.33333
8th	RM7C_RM7F	26.66667	RM7F_RM14F	27.53333	RM7C_RM14F	45.26667
9th	RM7F	26.86667	U	27.73333	RM14F	45.13333
10th	B	27.13333	V_RM14V	27.93333	RM14C_RM7V	44.93333
	MSE		**RMSE**		**RMSLE**	
	Feature Setup	**Friedman Score**	**Feature Setup**	**Friedman Score**	**Feature Setup**	**Friedman Score**
1st	V	21.93333	U	22.73333	U	17.6
2nd	B_RM7B	23.13333	V	23.13333	RM7C_RM14F	21.13333
3rd	RM7V	24	B_RM7B	23.66667	B_RM7B	22.53333
4th	V_RM7V	24.8	RM7B	24.46667	RM14F	22.73333
5th	RM7B	25.53333	RM14F	24.93333	V	23.2
6th	RM14F	26.73333	RM7V	25.4	RM7F	24.46667
7th	U	27.2	RM7C_RM7F	25.8	RM7C_RM7F	25.73333
8th	RM7C_RM7F	27.73333	RM7C_B	25.93333	RM7C_B	26.4
9th	B	27.86667	RM7C_RM14F	26.6	RM14C_RM7B	26.73333
10th	RM7C_B	27.93333	RM7F	27.2	B	26.86667

Table A7. Friedman results: feature setups—Shift 14.

	MAE		MAPE		R^2	
	Feature Setup	**Friedman Score**	**Feature Setup**	**Friedman Score**	**Feature Setup**	**Friedman Score**
1st	RM7C_B	17.46667	RM7C_B	18.53333	RM14C	50.5
2nd	RM7C_V	21.53333	RM14C_B	22.2	RM7C	48.9
3rd	RM14C_B	22.26667	RM7C_V	22.93333	U	48.76667
4th	RM7C	22.26667	RM7F_RM14F	23.73333	RM14C_RM7F	48.16667
5th	U	23.73333	B_RM7V	24.2	RM14C_RM7B	46.83333
6th	B_RM7V	23.8	V	25.6	RM7C_RM7B	46.06667
7th	V	24.46667	RM7C_F	26.33333	RM7C_RM7F	45.76667
8th	RM7C_F	24.86667	V_RM14F	27.4	RM7C_RM14C	45.56667
9th	RM7B	25.86667	RM7C	27.66667	RM7F	43.83333
10th	RM14B	27.33333	RM7B	28.13333	RM14C_F	43.36667
	MSE		**RMSE**		**RMSLE**	
	Feature Setup	**Friedman Score**	**Feature Setup**	**Friedman Score**	**Feature Setup**	**Friedman Score**
1st	RM7C_B	18.26667	RM7C_B	15.86667	RM7C	13.8
2nd	RM7C_V	21.26667	RM7C	20.33333	RM7C_B	16
3rd	B_RM7V	21.6	RM14C_B	21.26667	RM14C_B	21.06667
4th	RM14C_B	23.86667	RM7C_V	21.4	U	22.33333
5th	V	25.06667	U	22.8	RM7C_F	24.13333
6th	RM7B	25.86667	RM7C_F	24	RM14B	24.33333
7th	RM7C	26.2	V	24.06667	B_RM7V	25.06667
8th	RM7C_F	26.26667	B_RM7V	24.46667	RM7C_V	26.06667
9th	RM7B_RM7F	26.33333	RM7B	26	V	26.26667
10th	RM7B_RM7V	26.66667	RM7C_RM14C	26.2	RM7B	26.46667

References

1. Basak, S.; Kar, S.; Saha, S.; Khaidem, L.; Dey, S.R. Predicting the direction of stock market prices using tree-based classifiers. *N. Am. J. Econ. Financ.* **2019**, *47*, 552–567. [CrossRef]
2. Ren, R.; Wu, D.D.; Liu, T. Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine. *IEEE Syst. J.* **2019**, *13*, 760–770. [CrossRef]
3. Huang, W.; Nakamori, Y.; Wang, S.Y. Forecasting stock market movement direction with support vector machine. *Comput. Oper. Res.* **2005**, *32*, 2513–2522. [CrossRef]
4. Zhong, X.; Enke, D. Predicting the daily return direction of the stock market using hybrid machine learning algorithms. *Financ. Innov.* **2019**, *5*, 24. [CrossRef]
5. Abraham, B.; Ledolter, J. (Eds.) *Statistical Methods for Forecasting*; Wiley Series in Probability and Statistics, John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1983. [CrossRef]
6. Armstrong, J.S.; Collopy, F.L. Integration of Statistical Methods and Judgment for Time Series Forecasting: Principles from Empirical Research. *Forecast. Model. eJournal* **1998**, 269–293.
7. Bontempi, G.; Ben Taieb, S.; Le Borgne, Y.A. *Machine Learning Strategies for Time Series Forecasting*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 138. [CrossRef]
8. Masini, R.P.; Medeiros, M.C.; Mendes, E.F. Machine learning advances for time series forecasting. *J. Econ. Surv.* **2021**, *37*, 76–111. [CrossRef]
9. Cao, L.; Tay, F. Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Trans. Neural Netw.* **2003**, *14*, 1506–1518. [CrossRef] [PubMed]
10. Yang, A.; Li, W.; Yang, X. Short-term electricity load forecasting based on feature selection and Least Squares Support Vector Machines. *Knowl.-Based Syst.* **2019**, *163*, 159–173. [CrossRef]
11. Sagheer, A.; Kotb, M. Time series forecasting of petroleum production using deep LSTM recurrent networks. *Neurocomputing* **2019**, *323*, 203–213. [CrossRef]
12. Zhao, Z.; Chen, W.; Wu, X.; Chen, P.C.Y.; Liu, J. LSTM network: A deep learning approach for short-term traffic forecast. *IET Intell. Transp. Syst.* **2017**, *11*, 68–75.
13. Graf, R.; Zhu, S.; Sivakumar, B. Forecasting river water temperature time series using a wavelet–neural network hybrid modelling approach. *J. Hydrol.* **2019**, *578*, 124115. [CrossRef]
14. Kurumatani, K. Time series forecasting of agricultural product prices based on recurrent neural networks and its evaluation method. *SN Appl. Sci.* **2020**, *2*, 1434.

15. Khairalla, M.A.E.; Ning, X.; Al-Jallad, N.T.; El-Faroug, M.O. Short-Term Forecasting for Energy Consumption through Stacking Heterogeneous Ensemble Learning Model. *Energies* **2018**, *11*, 1605. [CrossRef]
16. Alkandari, M.; Ahmad, I. Solar power generation forecasting using ensemble approach based on deep learning and statistical methods. *Appl. Comput. Inform.* **2020**. [CrossRef]
17. Liapis, C.M.; Karanikola, A.; Kotsiantis, S.B. Energy Load Forecasting: Investigating Mid-Term Predictions with Ensemble Learners. In Proceedings of the AIAI, Crete, Greece, 17–20 June 2022.
18. Liapis, C.M.; Karanikola, A.C.; Kotsiantis, S.B. An ensemble forecasting method using univariate time series COVID-19 data. In Proceedings of the 24th Pan-Hellenic Conference on Informatics, Athens, Greece, 20–22 November 2020.
19. Liapis, C.M.; Karanikola, A.; Kotsiantis, S.B. A Multi-Method Survey on the Use of Sentiment Analysis in Multivariate Financial Time Series Forecasting. *Entropy* **2021**, *23*, 1603. [CrossRef] [PubMed]
20. Siami-Namini, S.; Tavakoli, N.; Namin, A.S. A Comparison of ARIMA and LSTM in Forecasting Time Series. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; pp. 1394–1401.
21. Çıbıkdiken, A.; Karakoyun, E. Comparison of ARIMA Time Series Model and LSTM Deep Learning Algorithm for Bitcoin Price Forecasting. In Proceedings of the 13th multidisciplinary academic conference in Prague, Hamburg, Germany, 27–30 August 2018.
22. Yamak, P.T.; Yujian, L.; Gadosey, P.K. A Comparison between ARIMA, LSTM, and GRU for Time Series Forecasting. In Proceedings of the ACAI, Sanya, China, 20–22 December 2019.
23. Maleki, A.; Nasseri, S.; Aminabad, M.S.; Hadi, M. Comparison of ARIMA and NNAR Models for Forecasting Water Treatment Plant's Influent Characteristics. *KSCE J. Civ. Eng.* **2018**, *22*, 3233–3245.
24. Satrio, C.B.A.; Darmawan, W.; Nadia, B.U.; Hanafiah, N. Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET. *Procedia Comput. Sci.* **2021**, *179*, 524–532.
25. Paliari, I.; Karanikola, A.; Kotsiantis, S.B. A comparison of the optimized LSTM, XGBOOST and ARIMA in Time Series forecasting. In Proceedings of the 2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA), Chania Crete, Greece, 12–14 July 2021; pp. 1–7.
26. Zhang, Y.; Yang, H.L.; Cui, H.; Chen, Q. Comparison of the Ability of ARIMA, WNN and SVM Models for Drought Forecasting in the Sanjiang Plain, China. *Nat. Resour. Res.* **2019**, *29*, 1447–1464.
27. Tealab, A. Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Comput. Inform. J.* **2018**, *3*, 334–340. [CrossRef]
28. Sezer, O.B.; Gudelek, M.U.; Ozbayoglu, A.M. Financial Time Series Forecasting with Deep Learning: A Systematic Literature Review: 2005–2019. *arXiv* **2020**, arXiv:abs/1911.13288.
29. Lara-Benítez, P.; Carranza-García, M.; Santos, J.C.R. An Experimental Review on Deep Learning Architectures for Time Series Forecasting. *Int. J. Neural Syst.* **2021**, *31*, 2130001. [CrossRef] [PubMed]
30. Karanikola, A.; Liapis, C.M.; Kotsiantis, S. A Comparison of Contemporary Methods on Univariate Time Series Forecasting. In *Advances in Machine Learning/Deep Learning-Based Technologies: Selected Papers in Honour of Professor Nikolaos G. Bourbakis—Volume 2*; Tsihrintzis, G.A., Virvou, M., Jain, L.C., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 143–168. [CrossRef]
31. Wang, K.; Qi, X.; Liu, H. A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network. *Appl. Energy* **2019**, *251*, 113315.
32. Rao, T.; Srivastava, S. Analyzing Stock Market Movements Using Twitter Sentiment Analysis. In Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, Sydney, Australia, 31 July–3 August 2012.
33. Nguyen, T.H.; Shirai, K.; Velcin, J. Sentiment analysis on social media for stock movement prediction. *Expert Syst. Appl.* **2015**, *42*, 9603–9611.
34. Kalyani, J.; Bharathi, H.N.; Jyothi, R. Stock trend prediction using news sentiment analysis. *arXiv* **2016**, arXiv:abs/1607.01958.
35. Shah, D.; Isah, H.; Zulkernine, F.H. Predicting the Effects of News Sentiments on the Stock Market. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 4705–4708.
36. Souma, W.; Vodenska, I.; Aoyama, H. Enhanced news sentiment analysis using deep learning methods. *J. Comput. Soc. Sci.* **2019**, *2*, 33–46.
37. Valle-Cruz, D.; Fernandez-Cortez, V.; Chau, A.L.; Sandoval-Almazán, R. Does Twitter Affect Stock Market Decisions? Financial Sentiment Analysis During Pandemics: A Comparative Study of the H1N1 and the COVID-19 Periods. *Cogn. Comput.* **2021**, *14*, 372–387.
38. Sharma, V.; Khemnar, R.K.; Kumari, R.A.; Mohan, B.R. Time Series with Sentiment Analysis for Stock Price Prediction. In Proceedings of the 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur, India, 28–29 September 2019; pp. 178–181.
39. Pai, P.F.; Liu, C. Predicting Vehicle Sales by Sentiment Analysis of Twitter Data and Stock Market Values. *IEEE Access* **2018**, *6*, 57655–57662.
40. Mohan, S.; Mullapudi, S.; Sammeta, S.; Vijayvergia, P.; Anastasiu, D. Stock Price Prediction Using News Sentiment Analysis. In Proceedings of the 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), Newark, CA, USA, 4–9 April 2019; pp. 205–208.
41. Mehta, P.; Pandya, S.; Kotecha, K. Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. *PeerJ Comput. Sci.* **2021**, *7*, e476. [CrossRef]

42. Jin, Z.; Yang, Y.; Liu, Y. Stock closing price prediction based on sentiment analysis and LSTM. *Neural Comput. Appl.* **2019**, *32*, 9713–9729. [CrossRef]
43. Wu, S.H.; Liu, Y.; Zou, Z.; Weng, T.H. S_I_LSTM: Stock price prediction based on multiple data sources and sentiment analysis. *Connect. Sci.* **2021**, *34*, 44–62.
44. Jing, N.; Wu, Z.; Wang, H. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Syst. Appl.* **2021**, *178*, 115019.
45. Smailovic, J.; Grcar, M.; Lavra, N.; Znidarsic, M. Stream-based active learning for sentiment analysis in the financial domain. *Inf. Sci.* **2014**, *285*, 181–203.
46. Raju, S.M.; Tarif, A.M. Real-Time Prediction of BITCOIN Price using Machine Learning Techniques and Public Sentiment Analysis. *arXiv* **2020**, arXiv:abs/2006.14473.
47. Abraham, J.; Higdon, D.W.; Nelson, J.; Ibarra, J. Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis. *SMU Data Sci. Rev.* **2018**, *1*, 1.
48. Valencia, F.; Gómez-Espinosa, A.; Valdés-Aguirre, B. Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning. *Entropy* **2019**, *21*, 589. [PubMed]
49. Deb, A.; Lerman, K.; Ferrara, E. Predicting Cyber Events by Leveraging Hacker Sentiment. *Information* **2018**, *9*, 280. [CrossRef]
50. Masri, S.; Jia, J.; Li, C.; Zhou, G.; Lee, M.C.; Yan, G.; Wu, J. Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic. *BMC Public Health* **2019**, *19*, 761.
51. Chauhan, P.; Sharma, N.; Sikka, G. The emergence of social media data and sentiment analysis in election prediction. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *12*, 2601–2627. [CrossRef]
52. Tseng, K.K.; Lin, R.F.Y.; Zhou, H.; Kurniajaya, K.J.; Li, Q. Price prediction of e-commerce products through Internet sentiment analysis. *Electron. Commer. Res.* **2018**, *18*, 65–88. [CrossRef]
53. Twintproject. Twintproject/Twint: An Advanced Twitter Scraping & OSINT Tool. Available online: https://github.com/twintproject/twint (accessed on 7 October 2021).
54. Van Rossum, G. *The Python Library Reference, Release 3.8.2*; Python Software Foundation: Wolfeboro Falls, NH, USA, 2020.
55. Bird, S. NLTK: The Natural Language Toolkit. *arXiv* **2004**, arXiv:cs.CL/0205028.
56. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python*; Packt Publishing Ltd.: Birmingham, UK, 2009.
57. String—Common String Operations. Available online: https://docs.python.org/3/library/string.html (accessed on 7 October 2021).
58. Simplified Text Processing. Available online: https://textblob.readthedocs.io/en/dev/ (accessed on 7 October 2021).
59. Hutto, C.J.; Gilbert, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1–4 June 2014.
60. Araci, D. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv* **2019**, arXiv:abs/1908.10063.
61. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:abs/1810.04805.
62. ProsusAI. ProsusAI/finBERT: Financial Sentiment Analysis with Bert. Available online: https://github.com/ProsusAI/finBERT (accessed on 7 October 2021).
63. Malo, P.; Sinha, A.; Korhonen, P.J.; Wallenius, J.; Takala, P. Good debt or bad debt: Detecting semantic orientations in economic texts. *J. Assoc. Inf. Sci. Technol.* **2014**, *65*, 782–796. [CrossRef]
64. timeseriesAI. Timeseriesai/Tsai: Time Series Timeseries Deep Learning Machine Learning Pytorch FASTAI: State-of-the-Art Deep Learning Library for Time Series and Sequences in Pytorch/Fastai. Available online: https://github.com/timeseriesAI/tsai (accessed on 7 October 2021).
65. Wang, Z.; Yan, W.; Oates, T. Time series classification from scratch with deep neural networks: A strong baseline. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 1578–1585.
66. Oguiza, I. tsAI Models: FCNPlus. Available online: https://timeseriesai.github.io/tsai/models.fcnplus.html (accessed on 7 October 2021).
67. Fawaz, H.I.; Lucas, B.; Forestier, G.; Pelletier, C.; Schmidt, D.F.; Weber, J.; Webb, G.I.; Idoumghar, L.; Muller, P.A.; Petitjean, F. InceptionTime: Finding AlexNet for Time Series Classification. *arXiv* **2020**, arXiv:abs/1909.04939.
68. Oguiza, I. tsAI Models: InceptionTimePlus. Available online: https://timeseriesai.github.io/tsai/models.inceptiontimeplus.html (accessed on 7 October 2021).
69. Oguiza, I. tsAI Models: RNNS. Available online: https://timeseriesai.github.io/tsai/models.rnn.html (accessed on 7 November 2022).
70. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
71. Chung, J.; Gülçehre, C.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:abs/1412.3555.
72. Oguiza, I. tsAI Models: RNN_FCN. Available online: https://timeseriesai.github.io/tsai/models.rnn_fcn.html (accessed on 7 November 2022).
73. Karim, F.; Majumdar, S.; Darabi, H.; Chen, S. LSTM Fully Convolutional Networks for Time Series Classification. *IEEE Access* **2018**, *6*, 1662–1669. [CrossRef]
74. Elsayed, N.; Maida, A.; Bayoumi, M.A. Deep Gated Recurrent and Convolutional Network Hybrid Model for Univariate Time Series Classification. *arXiv* **2019**, arXiv:abs/1812.07683.

75. Oguiza, I. tsAI Models: RNN_FCNPlus. Available online: https://timeseriesai.github.io/tsai/models.rnn_fcnplus.html (accessed on 7 November 2022).
76. Zou, X.; Wang, Z.; Li, Q.; Sheng, W. Integration of residual network and convolutional neural network along with various activation functions and global pooling for time series classification. *Neurocomputing* **2019**, *367*, 39–45. [CrossRef]
77. Oguiza, I. tsAI Models: ResNetPlus. Available online: https://timeseriesai.github.io/tsai/models.resnetplus.html (accessed on 7 November 2022).
78. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:abs/1803.01271.
79. Zerveas, G.; Jayaraman, S.; Patel, D.; Bhamidipaty, A.; Eickhoff, C. A Transformer-based Framework for Multivariate Time Series Representation Learning. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Singapore, 14–18 August 2021.
80. Oguiza, I. tsAI Models: TSTPlus. Available online: https://timeseriesai.github.io/tsai/models.tstplus.html (accessed on 7 November 2022).
81. Oguiza, I. tsAI Models: TSIT. Available online: https://timeseriesai.github.io/tsai/models.tsitplus.html (accessed on 7 November 2022).
82. Oguiza, I. tsAI Models: Transformermodel. Available online: https://timeseriesai.github.io/tsai/models.transformermodel.html (accessed on 7 November 2022).
83. Fauvel, K.; Lin, T.; Masson, V.; Fromont, E.; Termier, A. XCM: An Explainable Convolutional Neural Network for Multivariate Time Series Classification. *arXiv* **2021**, arXiv:abs/2009.04796.
84. Oguiza, I. tsAI Models: XCMPlus. Available online: https://timeseriesai.github.io/tsai/models.xcmplus.html (accessed on 7 November 2022).
85. Rahimian, E.; Zabihi, S.; Atashzar, S.F.; Asif, A.; Mohammadi, A. XceptionTime: A Novel Deep Architecture based on Depthwise Separable Convolutions for Hand Gesture Classification. *arXiv* **2019**, arXiv:abs/1911.03803.
86. Oguiza, I. tsAI Models: XceptionTimePlus. Available online: https://timeseriesai.github.io/tsai/models.xceptiontimeplus.html (accessed on 7 November 2022).
87. Tang, W.; Long, G.; Liu, L.; Zhou, T.; Blumenstein, M.; Jiang, J. Omni-Scale CNNs: A simple and effective kernel size configuration for time series classification. *arXiv* **2022**, arXiv:2002.10061.
88. Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701. [CrossRef]
89. Dunn, O.J. Multiple Comparisons among Means. *J. Am. Stat. Assoc.* **1961**, *56*, 52–64.
90. Hodges, J.L.; Lehmann, E.L. Rank Methods for Combination of Independent Experiments in Analysis of Variance. *Ann. Math. Stat.* **1962**, *33*, 403–418. [CrossRef]

entropy

MDPI

Article

Extraction of Important Factors in a High-Dimensional Data Space: An Application for High-Growth Firms

Takuya Wada [1], Hideki Takayasu [2,3] and Misako Takayasu [1,2,*]

[1] Department of Mathematical and Computing Science, School of Computing, Tokyo Institute of Technology, Yokohama 226-8502, Japan
[2] Institute of Innovative Research, Tokyo Institute of Technology, Yokohama 226-8502, Japan; takayasu@csl.sony.co.jp
[3] Sony Computer Science Laboratories, Tokyo 141-0022, Japan
* Correspondence: takayasu.m.aa@m.titech.ac.jp

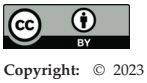
Abstract: We introduce a new non-black-box method of extracting multiple areas in a high-dimensional big data space where data points that satisfy specific conditions are highly concentrated. First, we extract one-dimensional areas where the data that satisfy specific conditions are mostly gathered by using the Bayesian method. Second, we construct higher-dimensional areas where the densities of focused data points are higher than the simple combination of the results for one dimension, and then we verify the results through data validation. Third, we apply this method to estimate the set of significant factors shared in successful firms with growth rates in sales at the top 1% level using 156-dimensional data of corporate financial reports for 12 years containing about 320,000 firms. We also categorize high-growth firms into 15 groups of different sets of factors.

Keywords: variable selection; feature selection; high-growth firms; Bayesian method; big data

Citation: Wada, T.; Takayasu, H.; Takayasu, M. Extraction of Important Factors in a High-Dimensional Data Space: An Application for High-Growth Firms. *Entropy* **2023**, *25*, 488. https://doi.org/10.3390/e25030488

Academic Editor: Panos Argyrakis

Received: 6 February 2023
Revised: 2 March 2023
Accepted: 8 March 2023
Published: 10 March 2023



1. Introduction

We consider the general problem of extracting areas in a high-dimensional data space where points that satisfy specific conditions are concentrated. Generally, as factors associated with a specific condition are often unknown, we use the most available factors and examine their relevance to a particular condition [1]. However, the majority of the factors used are irrelevant or redundant, resulting in problems such as reduced accuracy of the analysis and increased analysis time [1,2]. Therefore, we are reducing the number of variables, a process called variable selection. Variable selection has various advantages, such as accuracy increase, analysis time reduction, and overfitting avoidance [2–4]. Many models have been proposed for this variable selection and used in various fields [4–6]. In recent years, machine learning models have been used to improve the accuracy of variable selection. For example, Genuer used random forests [7] to select significant variables in high-dimensional classification problems [8]. Grandvalet proposed a model that automatically performs relevance judgments and feature selection on support vector machines [9] and showed its effectiveness in facial expression recognition tasks [10]. However, machine learning models also have disadvantages; for example, generally their results are difficult to understand logically due to the complexity of these models and their black-box structure [11,12]. In addition, to the best of our knowledge, a general method for exhaustively extracting areas where the data that satisfy specific conditions are highly concentrated has not been established in the study of big data.

In this paper, we propose a new method based on a non-black-box model to solve this general problem. We use indicators calculated using the Bayesian method and Szymkiewicz-Simpson coefficient as evaluation measures for variable selection and extraction of pairs of variables, respectively. The Bayesian method is a data analysis method that uses existing information [13,14]. This point differs from the likelihood method and gives the advantage

of more flexible model assumptions and facilitating statistical inference even for complex problems [15,16]. We use the Bayesian method, which is used in various fields, including ecology and seismology [14,15,17–19], to construct the posterior distribution of a specific indicator. Then we use the lower limit of the confidence interval as a new indicator for the evaluation measure. As a basic tool of data analysis, we introduce the Szymkiewicz–Simpson coefficient, which quantitatively evaluates the degree of overlap between two sets [20].

In this study, we analyze the factors that contribute to a firm's high growth as an example of the application of this model. Firm growth is significant and attracts the attention of investors and banks [21,22]. Demirgüç-Kunt clarified that a firm's growth is related to the financial and legal system [23]. Baum extracted venture growth factors with structural equation modeling and data on 17 predictor variables [24]. We analyze the factors of a firm's growth using machine learning models in recent years. Van Witteloostuijn and Kolkman analyzed the factors that contribute to a firm's growth using random forests [25]. Among them, the phenomenon of high growth is heterogeneous, and Delmar showed that it can be classified into seven groups via cluster analysis [26]. Coad forecasted high-growth firms with Lasso [27], a machine learning model [28]. We identify high-growth patterns using our model and verify them with Delmar's and Coad's results.

The remainder of this paper is organized as follows. Section 2 explains the dataset and defines each firm's growth rate in sales and high-growth firms. Section 3 describes the mathematical basis used in this methodology and methods. We first determine the posterior distribution of the probability that a firm will grow high within a particular area using the Bayesian method and then define the existence probability of high-growth firms. We also provide proof of the formulas used in Section 4 and the subsequent sections. Section 4 describes step-by-step the results of the method and classifies the high-growth firms into 15 groups based on different factors. In Section 5, we discuss the advantages, considerations, concerns, comparison with previous studies, and indicators of analysis. Finally, Section 6 describes our results and the potential applications of our method.

2. Data

In this study, we use the corporate financial dataset provided by TEIKOKU DATABANK, Ltd. (TDB). In Japan, companies often ask a third-party corporate credit research organization to obtain information about a firm when they are looking for new business partners to expand sales or to check the business condition of existing business partners. TDB is one of the largest corporate credit research providers in Japan and has been providing corporate credit research for more than a century [29]. In this study, we use 12 years of data from 2005 to 2016 with sales data existing for the next three years contained in this corporate financial dataset. The data include about 320,000 firms with 1.7 million data points. The first 10 years of the 12 years of data are used for the analysis, and the remaining 2 years are used for validation. Note that the dataset is not complete, and some financial items are missing in some firms. In such cases, we simply neglect missing items in our analysis. As a result, the number of firms in each financial item becomes equal to the total number of firms minus the number of missing data for the item.

We focus on the rate of increase in sales for each firm, which is defined by the following equation:

$$\text{Growth rate in sales} = \frac{\text{Current sales after 3 years}}{\text{Current sales}} \tag{1}$$

In this paper, we define high-growth firms as ones whose growth rate is in the top 1% of all firms in each analysis or verification data. Specifically, a high-growth firm has a growth rate of 4.913 times or higher for the analysis data and 4.428 times or higher for the validation data. We use our method to extract the conditions commonly satisfied by these high-growth firms in financial items. We exclude financial items that have a very strong correlation (correlation coefficient of higher than 0.95) with the current sales used in

the definition of growth rate in sales to avoid false correlations. We consider 156 financial items, such as the capital and current ratio in general.

To verify whether high-growth firms are dense not by coincidence, we randomly shuffle the 10 years of data from 2005 to 2014 for comparison. Namely, we create five sets of randomly shuffled data by using the command "shuffle" in Python for each of the 156 financial items with pseudorandom numbers generated by PCG64 [30].

We apply our method explained in the following Section 3 to the 10 years of real data and the five sets of randomly shuffled data.

3. Method

In this section, we explain the definition of the existence probability of high-growth firms used in the analysis and show how to calculate the existence probability of high-growth firms when the conditions are independent (in Section 3.1). We describe the analytical procedure of our method (in Section 3.2).

3.1. Mathematical Basis

Let q be the existence probability of high-growth firms in a specific area J, a be the number of high-growth firms, and b be the number of non-high-growth firms. The probability of occurrence conditioned by q, $f(a, b|q)$, fulfills the following equation:

$$f(a, b|q) = \binom{a+b}{a} q^a (1-q)^b \tag{2}$$

Then, using Bayesian analysis with the prior distribution $\pi(q)$, the posterior distribution $\pi(q|a, b)$ of q conditioned by a and b is given as follows:

$$\pi(q|a, b) = \frac{\pi(q) f(a, b|q)}{\int_0^1 \pi(q) f(a, b|q) dq} \tag{3}$$

Here, we use the conjugate prior $\pi(q) \propto q^\alpha (1-q)^\beta$, which is a beta distribution with parameters $\alpha + 1$ and $\beta + 1$, for the prior distribution of binomial distribution to reduce computational effort during the analysis. In addition, we condition that $E[q|a = 0, b = 0] = r$ and $\alpha + \beta = 0$; that is the expectation of probability q in the case of no sample data is equal to $r = 0.01$. Then, $\alpha = -\beta = 2r - 1$, and $\pi(q|a, b)$ is obtained as follows:

$$\pi(q|a, b) = \frac{\Gamma(a + b + 2)}{\Gamma(a + 2r)\Gamma(b - 2r + 2)} q^{a+2r-1} (1-q)^{b-2r+1} \tag{4}$$

From this posterior distribution, we estimate the lower bound of the probability of the existence of high-growth firms with a 99% confidence interval. That is, we regard the existence probability in the area J with a and b by the value of y, which is determined by solving the following equation, the inverse of the regularized incomplete beta function.

$$r = \frac{\Gamma(a + b + 2)}{\Gamma(a + 2r)\Gamma(b - 2r + 2)} \int_0^y q^{a+2r-1} (1-q)^{b-2r+1} dq \tag{5}$$

We apply this 99% confidence value for the extraction of one to multi-dimensional areas.

Here, we prove the basic equation, which is used in the following sections for the extraction of two- or higher-dimensional areas. We consider particular conditions 1 to n and let A_1 to A_n be flag variables that specify these conditions. For example, $A_n = 1$ implies that condition n is fulfilled. In addition, let X be a flag variable that indicates whether high growth has occurred. We assume that A_1 to A_n are independent of each other and also independent under the condition of $X = 0$, namely, for non-high growth cases. The

probabilities of satisfying the conditions from 1 to n, $P(A_1 = 1, A_2 = 1, \ldots, A_n = 1)$, and from 1 to n under $X = 0$, $P(A_1 = 1, A_2 = 1, \ldots, A_n = 1|X = 0)$, are given as follows:

$$P(A_1 = 1, A_2 = 1, \ldots, A_n = 1) = \prod_{i=1}^{n} P(A_i = 1) \tag{6}$$

$$P(A_1 = 1, A_2 = 1, \ldots, A_n = 1|X = 0) = \prod_{i=1}^{n} P(A_i = 1|X = 0) \tag{7}$$

Then, under the conditions from 1 to n fulfilled, the existence probability of high-growth firms $P(X = 1|A_1 = 1, A_2 = 1, \ldots, A_n = 1)$ can be calculated using these equations with Bayes' formula as follows:

$$P(X = 1|A_1 = 1, A_2 = 1, \ldots, A_n = 1) = 1 - \frac{\prod_{i=1}^{n} 1 - P(X = 1|A_i = 1)}{(1 - P(X = 1))^{n-1}} \tag{8}$$

3.2. Method

We consider the financial data as a distribution of points in a 156-dimensional space with 156 financial items as variables and then search for areas with high concentrations of points of high-growth firms. Figure 1 shows an image of this model if it were two-dimensional.

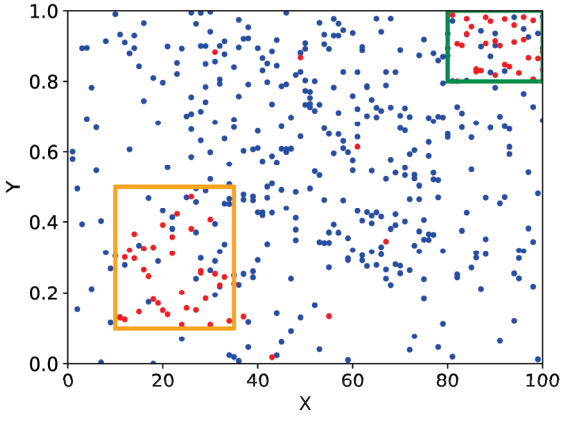


Figure 1. Schematic of our method if it were two-dimensional. The red dots represent high-growth firms, the blue dots represent non-high-growth firms, and the orange and green boxes are the areas to be extracted as high density areas.

Our analysis involves five steps:

Step1 Extraction of one-dimensional areas for each financial item;
Step2 Reduction of areas containing similar data points;
Step3 Extraction of two-dimensional areas;
Step4 Extraction of higher-dimensional areas;
Step5 Grouping.

3.2.1. Step1. Extraction of One-Dimensional Areas for Each Financial Item

In Step1, we extract high-concentration areas of high-growth firms in one dimension. A schematic of this step is presented in Figure 2. First, we project the points in a 156-dimensional space onto a single coordinate axis. Second, we segment the data into non-overlapping intervals, including at least 5% of the data. Third, in each separated area, we calculate the existence probability of high-growth firms using Equation (5) with the numbers of high-growing and non-high-growth firms. Subsequently, we extract areas

where the existence probability is higher than 0.01 with the 99% confidence. It should be noted that the proportion of high-growth firms in each financial item depends on the number of missing data, and there are items whose whole proportions of high-growth firms exceed 0.01. For such financial items, we set the threshold values of extraction by the value of the whole proportion for each item instead of 0.01. The schematic of the procedure up to this point is presented in Figure 2a. In this case, four areas are extracted: $[a, b)$, $[b, c)$, $[f, g)$, and $[h, i)$.

When multiple areas are extracted in one dimension, we check the possibility of combining the areas. The schematic of this procedure is presented in Figure 2b. For the extracted areas that are adjacent to each other, they are combined, as schematically shown by the interval $[a, c)$ in Figure 2b. If there is an unselected area in between, as shown by the interval $[g, h)$ in Figure 2a, the existence probability of high-growth firms in the connected area is calculated by using Equation (5). If it exceeds 0.01, these areas are merged, as shown by the interval $[f, i)$ in Figure 2b. This process is continued until no more areas can be combined.

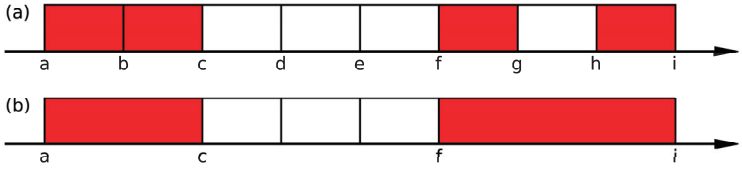


Figure 2. Schematic of Step1. (**a**) We divide the axis into non-overlapping segmented areas where at least 5% of the data points are included. For each area, we calculate the existence probability using Equation (5), and if it is higher than 0.01, the area is colored in red. In this case, four areas are extracted: $[a, b)$, $[b, c)$, $[f, g)$, and $[h, i)$. (**b**) We merge neighboring areas into one area as shown for $[a, c)$, and $[f, i)$ if the merged area's existence probability is higher than 0.01.

3.2.2. Step2. Reduction of Areas Containing Similar Data Points

In Step2, we reduce overlapping areas, which are extracted in Step1 based on the similarities defined below. Let us denote the set of firms in area A extracted from financial item $\tilde{a}$ as $\tilde{A}$. Note that the whole space is 156-dimensional, and this area is defined by the restricted range only for item $\tilde{a}$; thus, all other items can take any value in this set. If another item $\tilde{b}$ is similar to item $\tilde{a}$, then firms $\tilde{B}$ in the extracted area B may overlap with $\tilde{A}$. For a quantitative evaluation of such overlap, we introduce the Szymkiewicz–Simpson coefficient defined as follows:

$$Szymkiewicz - Simpson\ coefficient = \frac{|\tilde{A} \cap \tilde{B}|}{min(|\tilde{A}|, |\tilde{B}|)} \tag{9}$$

We calculate this indicator for all combinations of two areas chosen from the areas extracted in Step1 and observe the cumulative distribution function of this indicator. From the shape of the distribution, we introduce a threshold value of this indicator and delete these areas with higher indicators than the threshold. Detailed processes are discussed in Section 4.2.

3.2.3. Step3. Extraction of Two-Dimensional Areas

In Step3, we extract the two-dimensional areas where the existence probability of high-growth firms is higher. Subsequently, we calculate the existence probability of high-growth firms for all two-dimensional areas characterized by the direct product of the two conditions chosen from the areas after Step2. When the probability of a two-dimensional area estimated by using Equation (5) is less than that calculated using Equation (8), which assumes the independence of two financial items, then the two-dimensional area is aborted.

3.2.4. Step4. Extraction of Higher-Dimensional Areas

In Step4, we extract high-dimensional areas where the existence probability of high-growth firms is higher than the value of independent direct products estimated using Equation (8). For the two-dimensional area chosen in Step3 with the highest existence probability of high-growth firms, we add another one-dimensional condition that is chosen from Step2 and not already used in two-dimensional conditions. For all conditions in Step2, we calculate the existence probabilities of the combined three-dimensional areas using Equation (5) and choose the case that provides the highest existence probability of high-growth firms. If this probability is higher than the value estimated using Equation (8) and the existence probability of high-growth firms before adding the condition, then we assume that the new three-dimensional area's density of high-growth firms is significantly higher than the case of independent direct products. Thus, we adopt this three-dimensional area. If this condition is not fulfilled, then the two-dimensional area condition is kept two-dimensional. We proceed to the process for the 2nd candidate of the two-dimensional area chosen in Step3 and repeat the same procedure, followed by the 3rd and 4th, etc., to all two-dimensional candidates. For the newly adopted three-dimensional area, we add another one-dimensional condition chosen from Step2 as before and construct four-dimensional areas. We find the case that provides the highest existence probability of high-growth firms. Similarly, if the probability estimated using Equation (5) is higher than the value of Equation (8), we adopt the four-dimensional area. These processes of finding higher-dimensional areas are completed if there remains no combination of a higher-dimensional area that satisfies a certain condition; that is the probability of high-growth firms estimated using Equation (5) exceeds the value of Equation (8), and the existence probability of high-growth firms is higher than before the condition is added.

For the areas obtained in these processes, we verify whether the existence probability of high-growth firms is also increased in the data for validation. The procedure is used to add conditions in the same order as the conditions for the areas obtained in these processes until the existence probability of high-growth firms stops increasing. Using this procedure, we examine the validity of the obtained higher-dimensional areas and select high-dimensional areas that are non-local and have a high existence probability of high-growth firms. For the selected areas, the following process is followed to determine the areas of focus:

1. Remove high-dimensional areas that have the same set of conditions.
2. Remove similar high-dimensional areas where all firms in the area match, despite not being under the same conditions.
3. If the inclusion relationship is established, remove the area with the smallest number of firms.

3.2.5. Step5. Grouping

In Step5, we define groups of the high-dimensional areas selected in Step4 using hierarchical clustering using the Ward method [31] with the measure of the dissimilarity between areas given as follows:

$$dissimilarity = 1 - \frac{|\grave{A} \cap \grave{B}|}{min(|\grave{A}|, |\grave{B}|)} \tag{10}$$

where $\grave{A}$ and $\grave{B}$ are groups of high-growth firms belonging to areas A and B, respectively. We set the dissimilarity threshold to a value where most high-dimensional areas in the same group contain the same condition. The detailed process is discussed in Section 4.5.

4. Results

We define the abbreviated names for commonly used financial items, conditions, and units in Table 1.

Table 1. Abbreviated names of items, units, and indicators.

Abbreviated Name	Item Name
OIR	Ordinary income to revenue ratio
CLR	Current liabilities to revenue ratio
OITC	Ordinary income to total capital ratio
LR	Liabilities to revenue ratio
NIR	Net income to revenue ratio
CACL	Current assets to current liabilities ratio
LACL	Liquid assets to current liabilities ratio
CGSR	Cost of goods sold to revenue ratio
GPE	Gross profit per employee
TCR	Total capital to revenue ratio
FAR	Fixed assets to revenue ratio
FAFL	Fixed assets to fixed liability ratio
NOLR	Non-operating loss to revenue ratio
IR	Inventories to revenue ratio
CAR	Current assets to revenue ratio
APR	Accounts payable to revenue ratio
ARR	Accounts receivable to revenue ratio
PPER	Property, plant and equipment to revenue ratio
NCLR	Not current liabilities to revenue ratio
DR	Depreciation to revenue ratio
CFS	Compared to all firms in the same industry
IC	Industry comparison
DT	After discounting and transferring
DA	In data for analysis
DV	In data for verification
NAE-nD	Number of areas extracted in n-D
NDEHA	Number of dimensions of each high-dimensional areas
NC	Number of conditions
NF	Number of firms
NHF	Number of high-growth firms
EPHF	Existence probability of high-growth firms
M	Months
T	Thousands of yen

4.1. Extraction of One-Dimensional Areas for Each Financial Item

Step1 extracted 197 areas of 143 financial items. The top five areas with the highest existence probability of high-growth firms are presented in Table 2.

Table 2. Top five areas in the 197 areas of 143 financial items extracted in Step1. The extracted areas are from lower to upper limits. The lower and upper limits are denoted by percentage points within the financial item. The existence probability of high-growth firms (EPHF) in an area is calculated using the number of high-growth firms in the area, the number of all firms in the area, and Equation (5). The abbreviated names used in this table are defined in Table 1.

Item Name	Lower Limit	Upper Limit	NHF	NF	EPHF
OIR (CFS)	0.0%	6.7%	4219	96,108	0.042
OIR (IC)	0.0%	6.7%	4206	96,248	0.042
CLR (M)	91.9%	100.0%	4309	117,124	0.036
OITC (IC)	0.0%	6.6%	3509	95,920	0.035
LR (M)	90.8%	100.0%	4650	132,567	0.035

The areas with the first and second highest existence probability of high-growth firms have a value of about 0.042. This implies that they are more than four times more densely populated with high-growth firms than normal ones. Two areas were extracted for each of

the 54 financial items. The distribution of the existence probability of high-growth firms and details of the areas extracted for one example of those financial items are presented in Figure 3 and Table 3.

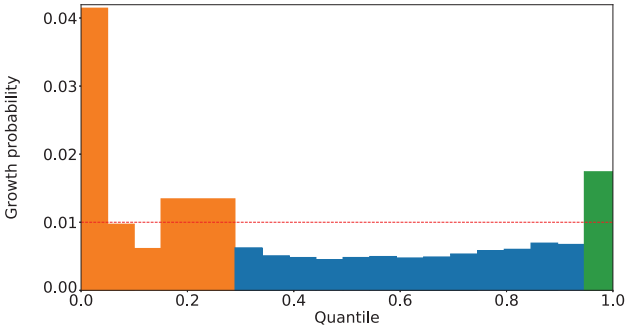


Figure 3. Existence probability of high-growth firms in each of the segmented areas, projected on the axis of the ratio of net income to sales (before amortization and after tax, %). The horizontal axis is the quantile from the beginning to the end of the segmented area, and the vertical axis is the existence probability of high-growth firms within the segmented area. The red dashed line represents 0.01, the percentage of high-growth firms in the overall area. For this financial item, the orange and green areas were extracted as the areas with densely populated high-growth firms, and the blue area was not extracted because it was not densely populated with high-growth firms. For the orange area, two areas were initially extracted: the 0–5.0% and 15.0–28.9% areas. These two areas and the areas in between where the existence probability of high-growth firms is low were merged into one area, as shown in Figure 2b.

Table 3. Two areas extracted in the ratio of net income to sales (before amortization and after tax, %), orange and green, respectively, in Figure 3. The extracted areas are from lower to upper limits, which are denoted by percentage points within the financial item. The existence probability of high-growth firms of an area is calculated using the number of high-growth firms in the area, the number of all firms in the area, and Equation (5). The abbreviated names used in this table are defined in Table 1.

Area	Lower Limit	Upper Limit	NHF	NF	EPHF
orange	0.0%	28.9%	7203	417805	0.017
green	94.6%	100.0%	1439	77645	0.017

These orange and green areas are where high-growth firms are about 1.7 times more dense than normal ones. These areas are the two edges of the financial items, and it is thought that firms grow high due to different factors.

For validation, we performed the same one-dimensional extraction on five random data. We extracted 11, 11, 12, 13, and 13 areas, respectively. No multiple areas were extracted within a single financial item. The area with the highest existence of high-growth firms in these areas was about 1.08 times more dense than normal ones. These areas are used in Step2.

4.2. Reduction of Areas Containing Similar Data Points

Similar areas were deleted in Step2 for the 197 areas of 143 financial items extracted in Step1. The result of calculating Equation (9) for all combinations of the 197 areas is presented in Figure 4.

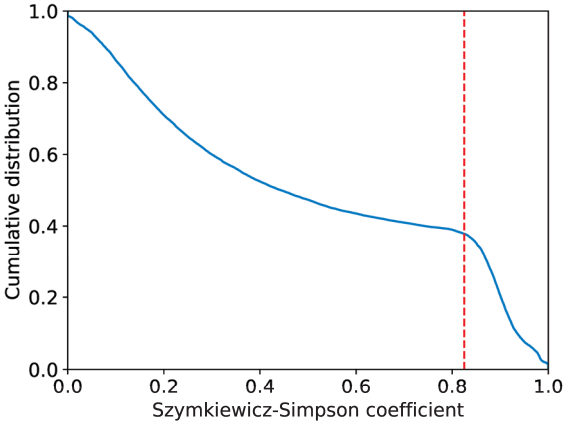


Figure 4. Cumulative distribution function of the values calculated for all combinations by using Equation (9). The horizontal axis is the value of the Szymkiewicz–Simpson coefficient, and the vertical axis is the cumulative distribution. The red dashed line represents 0.825, where the shape of the cumulative distribution function changes. This value was used as the threshold value.

Figure 4 shows that the cumulative distribution function changes its slope around when the value of the Szymkiewicz–Simpson coefficient is 0.825. This value was used as the threshold value. In the combination of areas where the value of the Szymkiewicz–Simpson coefficient is greater than this value, the area with the smallest existence probability of high-growth firms was deleted. For example, the combination of an area with a turnover of current debt (months) of 7.44 or higher and an area with an increase/decrease in an investment of less than 0 (thousands of yen) resulted in a Szymkiewicz–Simpson coefficient value of 0.916. Therefore, we compared the existence probability of high-growth firms and removed the area with an investment volume of less than 0 (thousands of yen), which was a lower area. We finally extracted 67 areas of 51 financial items.

For the five random data, the highest Szymkiewicz–Simpson coefficient was about 0.24 in the combination obtained from the areas of financial items obtained in each. Considering that this is smaller than the threshold value of 0.825 in the data for analysis and that no similarity exists among the financial items and among the areas as the data were randomly shuffled, none of the areas were removed. The 11, 11, 12, 13, and 13 areas obtained in Step1 were used in Step3–Step5.

4.3. Extraction of Two-Dimensional Areas

The 67 areas of 51 financial items extracted in Step2 were used to extract the two-dimensional areas. We checked all possible combinations, and the top five two-dimensional areas with the highest existence probability of high-growth firms are presented in Table 4.

In the two-dimensional area where the existence of high-growth firms is in the first and second places, high-growth firms are about 20 times more dense than normal ones. Table 4 displays how many times the existence of high-growth firms is compared to when the two conditions are independent (Column Ratio), and these five areas are about five times as high. Therefore, some synergy must exist in the combination of these conditions. Figure 5 presents the extracted two-dimensional area of the first rank.

Table 4. Top five two-dimensional areas in Step3. The existence probability of high-growth firms of an area is calculated using the number of high-growth firms in the area, the number of all firms in the area, and Equation (5). The ratio in this table is the existence probability of high-growth firms in two dimensions divided by the existence probability of high-growth firms calculated given that the two conditions are independent using Equation (8). The abbreviated names used in this table are defined in Table 1.

Item Name	EPHF(1D)	Item Name	EPHF(1D)	EPHF(2D)	Ratio
CLR (M)	0.036	CACL (%)	0.013	0.199	5.142
CLR (M)	0.036	LACL (%)	0.012	0.196	5.232
CSGR (%)	0.025	GPE (T)	0.020	0.177	5.147
TCR (M)	0.022	FAR (M)	0.019	0.174	5.647
FAFL (%)	0.024	FAR (M)	0.023	0.173	4.702

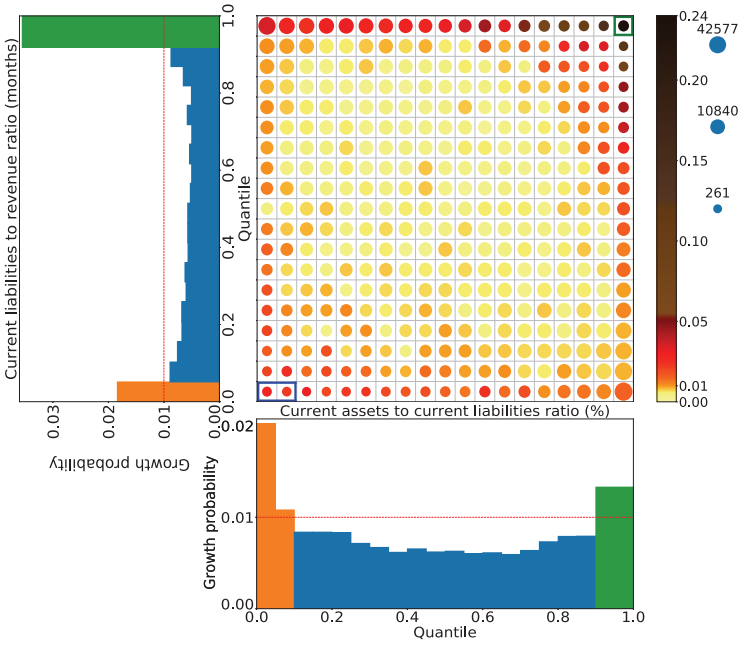


Figure 5. Extracted two-dimensional area of the first rank. The vertical and horizontal axes are divided by the current liabilities to revenue ratio (months) and the current assets to current liabilities ratio (%), respectively. The size of the circle represents the number of firms in the area, and the radius is scaled in a logarithmic scale. The colors of the circles represent the proportion of high-growth firms in the area. It is drawn in the order of yellow, orange, red, brown, and black, starting from the lowest to the highest. The green box at the right top is the area extracted as the two-dimensional area with the highest concentration of high-growth firms. The blue box at the left bottom is the area that was not extracted because the existence probability of high-growth firms in this area is lower than that of high-growth firms using Equation (8) if the two conditions are independent.

The green box area at the right top in Figure 5 is the area that satisfies the green areas in the turnover of current debt and the current ratio in the one-dimensional axes. It is 20 times more densely populated with high-growth firms than normal ones. It was also extracted as a two-dimensional area with the highest existence probability of high-growth firms. Meanwhile, the blue box area in Figure 5 is the area that satisfies the orange areas

in the turnover of the current debt and the current ratio in the one-dimensional axes. The existence probability of high-growth firms in this area is 0.014. This value is lower than that of high-growth firms when the two conditions are independent, as calculated using Equation (8). Therefore, this area was not extracted as a two-dimensional area.

We obtain 2211 two-dimensional areas using the 67 conditions used for the 67 areas extracted in Step2. Among them, we extracted 1036 areas that are more densely concentrated with high-growth firms than that when the conditions were independent.

For the five random data, we check whether high-growth firms are densely populated in the two-dimensional areas using the conditions used for the areas extracted in Step2. The number of areas extracted as areas where the existence probability of high-growth firms is higher than that of high-growth firms calculated using Equation (8), under the condition that the two conditions are independent were 3, 4, 6, 7, and 9. Even in the area with the highest concentration of high-growing firms in any of the random data, the concentration of highest-growing firms is about 1.7 times the normal concentration. It was also about 1.5 times higher than when all conditions were independent, indicating no strong synergistic effect. These two-dimensional areas extracted as densely populated with high-growth firms in the random data are used in the analysis in step 4.

4.4. Extraction of Higher-Dimensional Areas

For the 1036 two-dimensional areas extracted in Step3, we extract 1036 high-dimensional areas by repeatedly adding the 67 conditions used in the 67 areas extracted in Step2. The top two high-dimensional areas that are extracted are presented in Tables 5 and 6.

Table 5. Eight-dimensional area with the first highest existence probability of high-growth firms among the extracted high-dimensional areas. The ratio in this table is the existence probability of high-growth firms in the n-dimensional area divided by that of high-growth firms calculated under conditions where the n-conditions are independent using Equation (8); n is the number of conditions in the row (Column NC). The abbreviated names used in this table are defined in Table 1.

NC	Item Name (Threshold)	NHF	EPHF	Ratio
1	NOLR (%) (≤0)	3715	0.031	1.000
2	IR (M) (≤0)	2447	0.066	1.276
3	CAR (M) (≥10.2)	664	0.161	2.182
4	OITC (IC) (≤2)	217	0.408	4.194
5	GPE (T) (≤2727)	112	0.530	4.984
6	FAR (M) (≥10.19)	40	0.673	5.699
7	APR (M) (≤0)	33	0.748	5.382
8	CLR (M) (≥7.44)	26	0.779	4.837
9	OIR (CFS) (≤2)	26	0.779	4.133

Table 6. Seven-dimensional area with the second highest existence probability of high-growth firms among the extracted high-dimensional areas. The ratio in this table is the existence probability of high-growth firms in the n-dimensional area divided by that of high-growth firms calculated under conditions where the n-conditions are independent using Equation (8); n is the number of conditions in the row (Column NC). The abbreviated names used in this table are defined in Table 1.

NC	Item Name (Threshold)	NHF	EPHF	Ratio
1	ARR (DT) (M) (≤0.25)	4336	0.028	1.000
2	Revenue to total capital ratio (IC) (≤3)	1790	0.058	1.604
3	OITC (IC) (≤2)	550	0.215	3.548
4	PPER (M) (≤0.16)	136	0.475	6.658
5	NCLR (M) (≤0)	92	0.606	7.117
6	Investment and financing returns (%) (≤0.02)	60	0.683	7.272
7	DR (%) (≤0)	37	0.771	7.322
8	OIR (CFS) (≤2)	36	0.765	5.690

The existence probability of high-growth firms decreased when the 8th and 9th conditions were added to the areas in Tables 5 and 6. Therefore, the areas with the 7th and 8th dimensions in Tables 5 and 6 were extracted as areas with a high concentration of high-growth firms. The existence probability of high-growth firms in these high-dimensional areas is about 0.77. This implies that high-growth firms in these areas are about 77 times more dense than normal ones. They are also about 5–7 times higher than that when all conditions were independent. Therefore, we can assume that some synergistic effects occur in the combinations of these conditions. As shown in Tables 5 and 6, we extract the high-dimensional areas from the 1036 two-dimensional areas obtained in Step3. The distribution of the existence probability of high-growth firms in the high-dimensional areas finally obtained is presented in Figure 6.

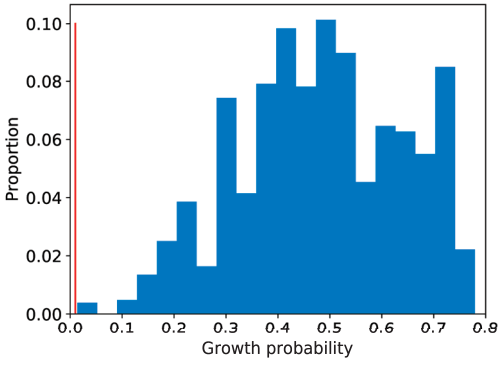


Figure 6. Distribution of the existence probability of high-growth firms in the high-dimensional areas. The vertical axis and horizontal axes are the proportion of 1036 areas and the existence probability of high-growth firms, respectively. The red line represents 0.01, the percentage of high-growth firms in the overall area.

As shown in Figure 6, 90% of the 1036 high-dimensional areas were able to extract areas where the high-growth firms are dense at 30 times or higher than the normal density. We have also extracted four areas where the high-growth firms are dense at less than three times the normal density, and all of these areas were two-dimensional ones. Subsequently, areas with a small number of data are called local ones. These areas became localized at the two-dimensional level, and no further high-dimensional areas could be extracted. Our method searched the entire area exhaustively, and the extracted areas include the local ones.

For the 1036 high-dimensional areas obtained in these processes, we verified whether the existence probability of high-growth firms is also increased in the data for validation. The verification procedure is to add conditions in the same order as the conditions for the areas obtained in these processes until the existence probability of high-growth firms stops to increase. As specific examples, the results of the verification in the areas of Tables 5 and 6 are presented in the Tables 7 and 8, respectively.

In the validation for both areas, the existence probability of high-growth firms decreased when the 5th condition was added. Thus, we confirmed the robustness of the results up to the four-dimensional area in these areas. In this validation, the existence probability of high-growth firms in the one-dimensional area in both validation results was almost the same as that when the data for analysis were used. The existence probability of high-growth firms in the four-dimensional area when the data for verification were used was about 0.33 and 0.21 for Tables 7 and 8, respectively. Although these values are lower than when using the data for analysis, we can assume that high-growth firms are concentrated at a high density, which cannot be considered coincidental. The reason for the lower existence probability of high-growth firms in the four-dimensional area, com-

pared to that for analysis, and the failure of these areas to maintain robustness in the five-dimensional area can be attributed to the fact that the data for verification are one-fifth the number of data for analysis. That is the number of high-growth firms in the area at the four-dimensional area is about 15.7% and 14.0% in Tables 7 and 8 for validation compared to that for analysis. Thus, the number of high-growth firms in the area is reduced, and the results are no longer stable and robust in high dimensions. The same verification was conducted for the remaining 1034 high-dimensional areas. The distribution of the number of dimensions for which the existence probability of high-growth firms was maximized in the data for analysis and verification was checked (Figure 7).

Table 7. Validation result for the high-dimensional area of Table 5 with the highest existence probability of high-growth firms. We add conditions in the same order as in Table 5 until the existence probability of high-growth firms stops to increase. The abbreviated names used in this table are defined in Table 1.

NC	Item Name	Threshold	NHF	EPHF
1	NOLR (%)	$\leq$0	551	0.031
2	IR (M)	$\leq$0	379	0.064
3	CAR (M)	$\geq$0.122	106	0.074
4	OITC (IC)	$\leq$2	34	0.334
5	GPE (T)	$\leq$2727	10	0.204

Table 8. Validation result for the high-dimensional area of Table 6 with the second-highest existence probability of high-growth firms. We add conditions in the same order as in Table 6 until the existence probability of high-growth firms stops to increase. The abbreviated names used in this table are defined in Table 1.

NC	Item Name	Threshold	NHF	EPHF
1	ARR (DT) (M)	$\leq$0.25	680	0.029
2	Revenue to total capital ratio (IC)	$\leq$3	233	0.046
3	OITC (IC)	$\leq$2	79	0.183
4	PPER (M)	$\leq$0.16	19	0.208
5	NCLR (M)	$\leq$0	11	0.207

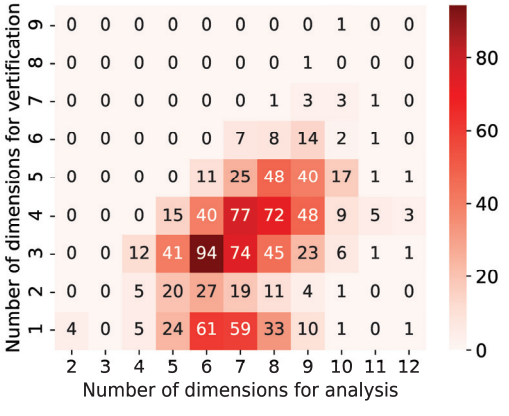


Figure 7. Distribution of the number of dimensions for which the existence probability of high-growth firms was maximized in the data for analysis and verification. The vertical and horizontal axes are the number of dimensions in verification data and analysis data, respectively. The numbers represent the number of areas with each dimension in the analysis and validation data. The colors indicate that the darker the red color, the higher the value, i.e., the greater the number of areas.

The numbers in Figure 7 represent the number of areas with each dimension in the analysis and validation data. For example, 77 with a vertical axis of 4 and a horizontal axis of 7 indicates that 77 areas have been extracted in seven-dimensional areas for analysis and verified to four-dimensional areas. Specifically, the area in Table 5 is contained in 72 with a vertical axis of 4 and a horizontal axis of 8, and that in Table 6 is contained in 77 with a vertical axis of 4 and a horizontal axis of 7 in Figure 7. Figure 7 presents that many high-dimensional areas of more than three dimensions are robust for verification. In addition, we can observe a relationship whereby the areas with higher dimensionality for analysis also maintain a higher dimensionality for validation. There was also a 10-dimensional area for which robustness was confirmed up to nine dimensions for verification. The details of this area are provided in Table 9.

Table 9. Ten-dimensional area for which robustness was confirmed in up to nine dimensions for verification. The abbreviated names used in this table are defined in Table 1.

NC	Item Name (Threshold)	NHF (DA)	EPHF (DA)	NHF (DV)	EPHF (DV)
1	Revenue (T) (≤108,917)	9577	0.032	1253	0.041
2	NOLR (%) (≤0)	3156	0.056	450	0.065
3	CAR (M) (≥10.2)	1046	0.134	143	0.123
4	OITC (IC) (≤2)	338	0.278	51	0.234
5	PPER (M) (≤0.16)	137	0.471	25	0.317
6	LR (M) (≥14.13)	78	0.562	17	0.358
7	NCLR (M) (≤0)	63	0.672	11	0.371
8	Revenue to total capital ratio (IC) (≤ 3)	59	0.691	11	0.404
9	IR (M) (≤0)	37	0.694	10	0.464
10	LACL (%)(≤41.45)	18	0.700	3	0.144
11	OIR (CFS) (≤2)	18	0.700		

The area in Table 9 is the area where the high-growth firms are about 70 times more densely populated than usual for the analysis. This area maintains robustness up to nine dimensions. In the data for verification, the high-growth firms are about 46 times denser than usual in this nine-dimensional area. We also extracted high-dimensional areas that can retain such robustness.

There are 165 areas where the increase in the existence probability of high-growth firms stops at one-dimensional areas for validation, despite that for analysis they are high-dimensional areas with six or more dimensions. In addition, in about half of the 1036 high-dimensional areas, an increase in the existence probability of high-growth firms stopped at three dimensions or less in the data for verification. Therefore, our method exhaustively searches the entire range and extracts local areas.

In the following, we focus on somewhat larger areas wherein the number of high-growth firms includes more than 1% (145 firms) of the total number of high-growth firms in the four-dimensional area in the data for analysis. There were 160 such high-dimensional areas. The areas in Tables 5 and 9 are included in these 160 areas, but the area in Table 6 is not. The distributions of the number of dimensions with the maximum existence probability of high-growth firms in the 1036 high-dimensional areas and the 160 non-local high-dimensional areas for verification are presented in Figure 8a,b.

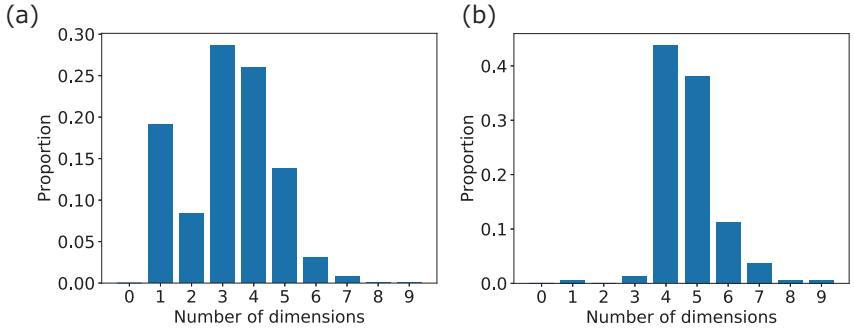


Figure 8. Distribution of the number of dimensions with the maximum existence probability of high-growth firms for verification. The vertical axis and horizontal axes are the proportion of 1036 areas in (**a**) and 160 areas in (**b**) and the number of dimensions, respectively. (**a**) In the 1036 high-dimensional areas. (**b**) In the 160 high-dimensional areas, which include more than 145 high-growth firms.

As shown in Figure 8, the distribution of the number of dimensions that maximizes the existence probability of high-growth firms in the data for validation has changed significantly by narrowing down from 1036 high-dimensional areas to 160 high-dimensional areas, which include more than 145 high-growth firms. In most of the 160 areas, the number of dimensions in which the existence probability of high-growth firms is maximized in data for verification is four-dimensional or higher. Therefore, in these 160 areas, the robustness can be assumed to be up to four-dimensional. Focusing on these 160 areas, 1–3 in Section 3.2.4 of the method are performed on these areas. The first corresponds to 40 areas, the second to zero areas, and the third to two areas. We finally focused on the 118 four-dimensional areas.

We extracted high-dimensional areas from each of the 29 two-dimensional areas extracted by the five random data. Consequently, we extracted seven three-dimensional areas and 22 two-dimensional areas. The results using the random data are presented in Table 10.

Table 10. Results using random data. EPHF represents the value of the existence probability of high-growth firms in the area where the existence probability of high-growth firms is the highest among the extracted high-dimensional areas.

Data	NAE-1D	NAE-2D	NDEHA	EPHF
1	11	3	2, 2, 3	0.0140
2	11	7	2, 2, 2, 2, 2, 2, 3	0.0226
3	12	9	2, 2, 2, 2, 2, 2, 2, 3, 3	0.0161
4	13	4	2, 2, 2, 2	0.0167
5	13	6	2, 2, 2, 3, 3, 3	0.0152

Table 10 shows that we did not extract any high-dimensional areas in any random data. The area with the highest existence probability of high-growth firms among all the random data was the area where high-growth firms were 2.3 times more densely populated than usual. A comparison of the results with the data for analysis indicates that the high-growth firms are much more densely populated than in the random data. Considering that the random data extracted a maximum of only nine areas, the data for analysis, which extracted 1036 high-dimensional areas, showed that the high-growth firms were densely concentrated in many areas. Therefore, we can assume that strong relations exist between high-growing factors of firms and financial items.

4.5. Grouping

We define groups of the 118 four-dimensional areas selected in Step4 via hierarchical clustering with the ward method, Step5. The result is presented in Figure 9.

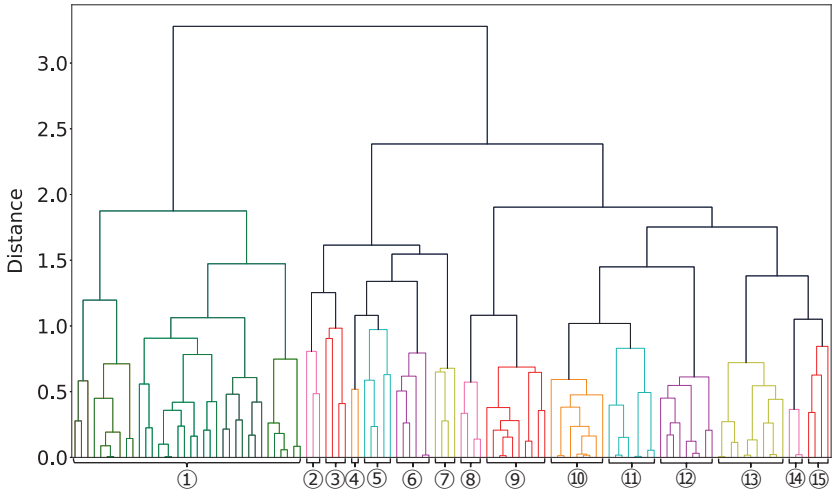


Figure 9. Dendrogram of the result of hierarchical clustering for the 118 four-dimensional areas. The vertical and horizontal axes are the dissimilarity defined using Equation (10) and the result of grouping the 118 four-dimensional areas, respectively. We divided the 118 four-dimensional areas into 15 groups (Groups ① to ⑮). Four-dimensional areas belonging to the same group have a common color. For example, Group ① has green. Groups ② to ⑮ are cyclically painted in six colors.

We set the dissimilarity threshold used for grouping in Figure 9 to a value that has a condition common to most of the grouped four-dimensional areas. Thus, the threshold was set to 1, except for the one group on the left, which is grouped because 34 of the 36 four-dimensional areas have the same condition. Finally, we divided the 118 four-dimensional areas into 15 groups. The conditions common to each of the 15 groups are presented in Table 11. We focused on Groups ①, ②, ⑫, and ⑭, which are characteristic among the 15 groups.

Here, 34 of the 36 four-dimensional areas in Group ① have the common condition of small gross profit per capita (less than 2727). The small value indicates that the firms in these 34 areas have small sales and poor operating efficiency. The remaining two four-dimensional areas have the condition that the total capital (compared to all firms in the same industry) is small (smaller than three) and the turnover of total capital (month) is large (larger than 17.79). The total capital (compared to all firms in the same industry) is the value evaluated by TDB and takes the value 0–10. The small value indicates that the total capital is very small compared to other firms in the same industry. The turnover of total capital (month) is the value of total capital divided by sales. Specifically, a large value of that indicates that sales are smaller than the total capital, given that the total capital is very small. Therefore, these two areas extract firms with very small sales and low efficiency. Therefore, the 36 four-dimensional areas in Group ① extract firms with small sales and low operating efficiency. These firms are considered to have improved their operations and increased their sales significantly after three years.

Table 11. Conditions common to each of the 15 groups. The abbreviated names used in this table are defined in Table 1. If the variables common to a group include those with an alphabet in front of the variable name, all four-dimensional areas in the group have in common that one or more of them are satisfied. For example, all four-dimensional areas in Group ⑤ contain the condition of the turnover of current assets and one or more of either (a) or (b). If the variables common to a group include variables with an alphabet with tilde in front of the variable name, all four-dimensional areas in the group have in common that two or more conditions are satisfied in them. For example, all four-dimensional areas in Group ⑮ contain two or more of the conditions ã, b̃, or c̃.

Group	Item name	Threshold
①	GPE (T)	≤2727
②	ARR (DT) (M)	≥3.86
③	PPER (M) (a) Financial account to revenue ratio (%) (b) DR (%)	≤0.00 ≤0.00 ≤0.00
④	Cash and deposits to revenue ratio (days) OITC (IC)	≥130.33 ≤2.00
⑤	CAR (M) (a) Interest coverage ratio (times) (b) Capital to revenue ratio (M)	≥10.20 ≤−8.49 ≤−0.81
⑥	OITC (IC) (a) CACL (%) (b) NCLR (M) (c) Capital to revenue ratio (M)	≤2.00 ≤78.68 ≤0.00 ≤−0.81
⑦	CAR (M) Non-operating income to revenue ratio (%)	≥10.20 ≥4.62
⑧	CAR (M) DR (%)	≥10.20 ≤0.00
⑨	PPER (M) (a) TCR (M) (b) Revenue to total capital ratio (IC) (c) OIR (CFS)	≤0.16 ≥17.79 ≤3.00 ≤2.00
⑩	IR (M)	≤0.00
⑪	CAR (M) ($\tilde{a}$) Non-operating income to revenue ratio (%) ($\tilde{b}$) OITC (IC) ($\tilde{c}$) APR (M)	≥10.2 ≤0.00 ≤2.00 ≤0.00
⑫	ARR (DT) (M)	≤0.25
⑬	CAR (M) (a) Financial account to revenue ratio (%) (b) Investment and financing returns (%)	≥10.20 ≤0.03 ≤0.02
⑭	CAR (M) IR (M) Non-operating income to revenue ratio (%)	≥10.20 ≤0.00 ≤0.05
⑮	($\tilde{a}$) Total capital (CFS) ($\tilde{b}$) Investment and financing returns (%) ($\tilde{c}$) Capital to revenue ratio (M)	≤3 ≤0.02 ≥8.53

Next, we focus on Group ② and Group ⑫. These two groups are characterized by different areas of the single variable of the trade receivables (discounted and transferred) turnover periods (months) as shown in Figure 10. Therefore, there is no firm that belongs to both Group ② and Group ⑫.

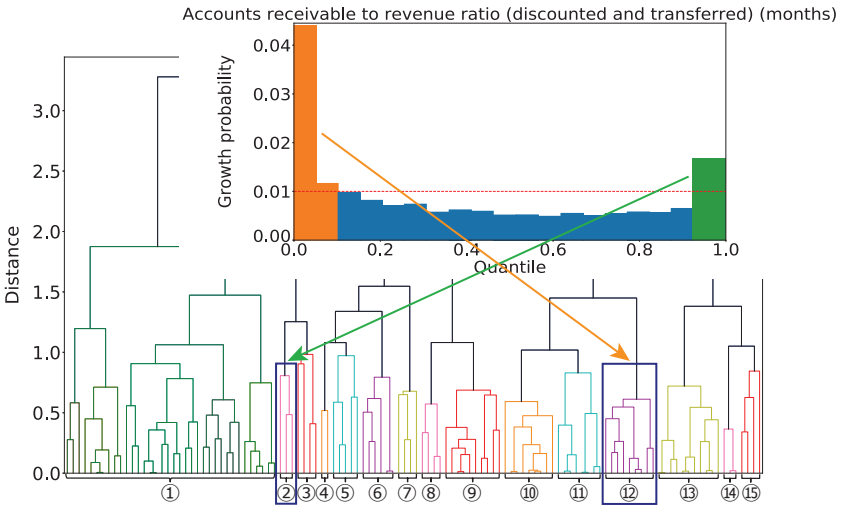


Figure 10. An example of the relation between the groups and a financial item. Group ② is characterized by the green area of the item, the accounts receivable to revenue ratio (discounted and transferred) (months), on the other hand, Group ⑫ is characterized by the orange area.

We consider what type of firms each group is extracting. Group ② has in common the condition that the value of the trade receivables (discounted and transferred) turnover periods (months) is large. This large value implies that the ratio of trade receivables to sales is significant. That is, a firm takes a long time to convert its receivables into cash; thus, firms with insufficient working capital are extracted. In addition, the conditions that the ratio of ordinary income to total assets (industry comparison), turnover of total capital (industry comparison), and ratio of ordinary income to net sales (compared to all firms in the same industry) are bad are extracted together. Thus, we have extracted firms in Group ② that do not have enough working capital and whose profitability is worse. These firms could have improved their operations to afford working capital, which would have led to higher sales. Group ⑫ has in common the condition that the value of the trade receivables (discounted and transferred) turnover periods (months) is small. This small value indicates that, in contrast to Group ②, firms in Group ⑫ can afford working capital. In addition, the conditions that the ratio of ordinary income to total assets (industry comparison) and the ratio of ordinary income to net sales (compared to all firms in the same industry) are bad are extracted together. Therefore, firms in Group ⑫ with low profitability were able to use their surplus working capital to increase sales after three years.

Finally, we focused on Group ⑭. The shared conditions are presented in Table 11. That is, these conditions include the absence of inventories, almost no non-operating income, and very large current assets. In Japan, current assets generally comprise of the following three elements [32]:

- Liquid assets: Short-term fixed deposits, securities, trade notes receivable, trade accounts receivable;
- Inventories: Assets expected to sell on to earn revenue from sales of goods, products, etc.;
- Others: Short-term loans receivable.

Short-term fixed deposits are those with a maturity of one year or less from the closing date. Securities are those with a maturity of one year or less or those held for the short term for trading purposes. Trade notes receivable are promissory notes received as payment for transactions with customers. Trade accounts receivable are accounts receivable from customers for business transactions. Liquid assets are the collective category of these four

assets. Inventories are assets that decrease in quantity in the short term that are sold to earn revenue. Short-term loans receivable are loans with a maturity of one year or less from the closing date. Current assets are collectively liquid assets, inventories, and short-term loans receivable. Shared conditions indicate that Group ⑭ firms have large short-term fixed deposits, trade notes receivable, trade accounts receivable, and short-term loans receivable. Therefore, these firms have more assets that can be cashed in within a year. In addition, the conditions of small revenues, small gross profit per employee, and small ordinary income to revenue ratio are extracted together. Hence, we can assume that the firms in Group ⑭ are financially robust and have increased their operating efficiency by making capital investments, developing human resources, and increasing employment, resulting in a significant increase in sales after three years.

5. Discussion

We discussed the advantages of using our method. In this study, we first extracted one-dimensional areas, then deleted similar ones, and finally combined the conditions characterizing those areas to extract higher-dimensional crowded data satisfying specific rules. Our method has two advantages. The first is the possibility of extracting combinations of synergistic conditions. In the one-dimensional area extracted in this study, high-growth firms in the most densely populated area were about four times more densely populated than usual, and the average was about 1.7 times more densely populated than usual. However, by combining the conditions, our method can extract areas where the density of high-growth firms is much higher than when the conditions were independent. For example, the two-dimensional area with the highest existence probability of high-growth firms in Table 4 is five times more densely populated with high-growth firms than that when the conditions are independent. Further, the high-dimensional area with the second highest existence probability of high-growth firms in Table 6 is seven times more densely populated with high-growth firms than that when all conditions are independent. Thus, our method can exhaustively extract combinations that seem to have synergistic effects.

Second, our method can also extract local areas and robust high-dimensional ones. In this study, we focused on somewhat larger areas to analyze universal factors, but we also extracted local areas. For example, we extracted the areas on the left side in Figure 6 where the existence probability of high-growth firms is lower than other extracted high-dimensional areas. We also extracted the high-dimensional areas at the bottom in Figure 7 that can only validate up to low dimensions due to insufficient data for verification. Contrary to this study, we can use our method if we want to focus on local and specific cases, rather than universal ones. In addition, we can extract localized areas and areas with robustness. For example, we extracted the high-dimensional area with strong robustness (Table 9). We can use our method when we want to focus on something universal, as in this study.

We discussed some of the considerations for this study. After the extraction of high-dimensional areas, we selected four dimensions as the number of dimensions that could withstand verification. First, we discussed regarding the extraction of high-dimension areas. Meanwhile, we extracted the areas of seven or more dimensions, in the data for verification, more than half of all extracted areas where the increase in the existence probability of high-growth firms stopped at three dimensions (see Figure 8a). There are two reasons for this. The first one is that there were cases where the number of firms was small in the initially extracted areas because our method performed an exhaustive search that includes local areas. The second one is that the increase in the existence probability of high-growth firms tends to stop since the data for verification is one-fifth of the data for analysis in terms of the number of data. Therefore, if it is not a local area, we can increase the number of dimensions that allow verification by increasing the data for verification to about the same number as that for analysis. Second, we discussed the number of dimensions that we used. While increasing the number of dimensions that allow verification by increasing the data for verification, considering that the area tends to be localized is necessary. In

this study, to focus on areas where firms universally tend to high growth, we focused on 160 four-dimensional ones where more than 1% of the total number of high-growth firms existed. Considering that we initially extracted 1036 high-dimensional areas, clearly that our method can easily extract localized areas. Therefore, determining to what dimensionality the results should be validated and used as universal results is necessary.

We also discussed some concerns when using our method. In this study, we first extracted one-dimensional areas, then deleted similar ones, and finally combined the conditions characterizing those areas to extract higher-dimensional crowded with data satisfying specific rules. However, if the densification occurs in the way shown in the following Figure 11a,b, we miss dense areas.

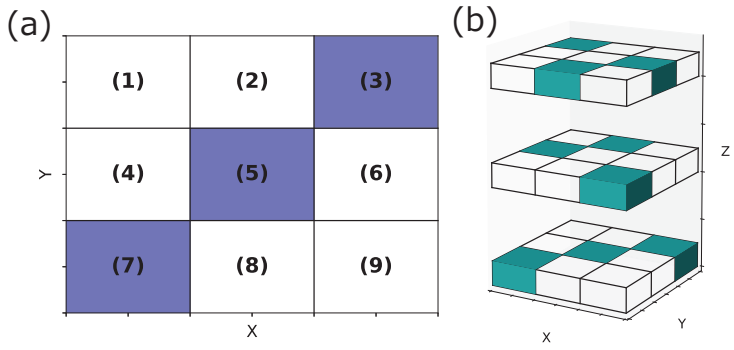


Figure 11. Examples of missing dense areas with this method. The colored areas are where data that satisfy specific conditions are densely distributed. (**a**) Example of missing in two dimensions. (**b**) Example of missing in three dimensions.

In Figure 11a,b, we divided each axis into three parts. Data satisfying specific conditions were densely populated in the colored areas in these figures. In Figure 11a, the case of the missing dense area is when the existence probability of data satisfying specific conditions in areas (1)∼(3), (4)∼(6), and (7)∼(9) is equal. In this case, when projected onto the Y-axis, we cannot extract the area on the Y-axis. Thus, we cannot extract the two-dimensional areas (3), (5), and (7). We also consider the case where (1)∼(3) > (4)∼(6) > (7)∼(9) in terms of the density of data satisfying specific conditions between areas (1)∼(3), (4)∼(6) and (7) ∼(9). We consider areas (7)∼(9) as the areas where data satisfying the specified conditions are not dense on the Y-axis, and we cannot extract area (7). The possibility exists that a similar phenomenon may occur in the third dimension and beyond. In the case of Figure 11b, as in the previous case, if the existence probability of data satisfying specific conditions is equal in the three divisions in any of the X-, Y-, and Z-axis directions, we cannot extract the colored areas in Figure 11b.

We can consider a possible method to address this concern to start focusing on two or higher dimensions, rather than focusing on one dimension. In a pair that selects two from all variables, we can address this by dividing the area, calculating the existence probability of data satisfying specific conditions in each area, and extracting the areas with a higher density of data satisfying certain conditions than normal ones. In Figure 11a, we can extract areas (3), (5), and (7) by calculating the existence probability of data satisfying specific conditions in each of areas (1)∼(9). In Figure 11b, we can extract the colored areas by calculating the existence probability of data satisfying specific conditions in each of the 27 areas. Meanwhile, since this method requires considering all variable partitions and calculating the probability in each of them, we predicted a significant increase in computational cost. Specifically, we considered the case where we divide each financial item by 5% as in this study and searched in two dimensions, as shown in Figure 11a, to avoid missing anything in dense areas. In this case, we divided each financial item

by a maximum of 20 and considered the 12,090 combinations of selecting two from all 156 financial items. Therefore, it is necessary to calculate the existence probability of data satisfying specific conditions in a maximum of 400 areas in each combination, totaling a maximum of about 4.8 million areas. We also considered the case of focusing on three dimensions, as shown in Figure 11b. We considered the 620,620 combinations of selecting three from all 156 financial items. Therefore, it is necessary to calculate the existence probability of data satisfying specific conditions in a maximum of 8000 areas in each combination, totaling a maximum of about 3 billion areas. Thus, the computational cost increases exponentially as we increase the number of dimensions that we begin to focus on. Therefore, we consider this method of addressing this problem when only a few variables exist. However, even if we searched exhaustively for a specific dimension, the same problem can occur above that dimension and beyond. Specifically, Figure 11b shows an example where a miss occurs in some three-dimensional areas, regardless of whether one starts looking at a one-dimensional or two-dimensional area. Therefore, we must discuss which dimension to examine exhaustively and which dimension and beyond to ignore invisible relationships.

We compared some popular existing methods with our method for comparison. In high-dimensional areas, when data satisfying specific conditions are concentrated in multiple areas, we call the problem of extracting all areas the multimodality problem. In the special case that there is only one highly concentrated area in the whole space, we call it a unimodality problem. For unimodality problems, we can extract the dense area by using popular methods such as multiple regression analysis or support vector machines. However, these methods are not suitable for the analysis of high-growth firms in this study, as we showed in Section 4, there are at least 15 dense areas in the 156-dimensional space. In addition, other popular methods, neural networks [33], are black-box methods, making it impossible to interpret the results in terms of important financial items. Random forests are also popular in big data analysis; however, they are unsuitable for the present problem of extracting important factors in the form of sets of variables. Our method can extract the sets of important factors for multimodality problems and is suitable for the analysis of high-growth firms.

We also compared the factors extracted in this study to Coad's previous study [28]. In that study, they used cluster analysis, which is strong for multimodality problems, to analyze the important factors of high-growth firms. Although the high-growth firms in the previous study are about 2% of the total data, we note that the definitions of high-growth firms and the variables used are very different. The previous study found that firms with low inventories, higher previous employment growth, and higher short-term liabilities are more likely considered high-growth firms. As previous employment growth is excluded from the financial item of this study, we analyzed other results. We identified the factor of low inventory as a universal factor in Group ⑩ and Group ⑭ of this study (see Figure 9 and Table 11). We extracted the factor of higher short-term liabilities in the high-dimensional area of Table 9. Therefore, we can assume that we have extracted the same results as in the previous studies.

We also compared the factors extracted in this study to that of Delemar's previous study [26]. In that study, they used Lasso, which is strong for unimodality problems, to analyze the important factors involved in forecasting high-growth firms. We note that the definition of high-growth firms differs from the previous study and the variables used are also very different. After comparing the results with this assumption, we extracted similar results to the previous study for increasing employment. In the previous study, increasing employment was part of the factors for the seven clusters of high-growth firms. The firms in Group ⑭ in this study are financially robust and have increased their operating efficiency by making capital investments, developing human resources, and increasing employment. Therefore, we believe that the result extracted in this study is similar to the previous one. The previous study focused on revenue growth. However, in this study, we extracted the areas that focused on this as localized areas, with the number of high-growth firms being

less than 100 in any two-dimensional ones. The study was different from previous studies that extracted revenue growth as universal.

Finally, we analyzed the indicators used in our method. For the 15 groups extracted using our method, we found the poor operating efficiency for most groups. The possible reason is that we used the top 1% of all firms in sales growth rate as the definition of high-growth firms. Firms with approximately four times or higher sales after three years often have either a pattern; that is, firms with poor operating efficiency have succeeded in improving their sales or sales are small from the start. Thus, we may need to change the definition of high-growth firms. In addition, we measured firm growth in this study using the absolute one in sales over three years. As sales are not a perfect indicator [26], some studies used the number of employees [21,34] and both the number of employees and sales [35]. Therefore, discussing which items we should use as a measure of growth and what should be the definition of a high-growing firm is necessary.

6. Conclusions

We introduced a new non-black-box method of extracting multiple areas in a high-dimensional big data space where data points that satisfy specific conditions are highly concentrated. We analyzed high-growth firms in all industries as an example of the applications in this study. We categorized the high-growth firms into 15 groups of different sets of factors. Conducting factor analysis of high-growth firms in specific industries or firms that have gone bankrupt by using this method is feasible. In addition, this method is not limited to corporate data and can be applied to various fields of analysis, including the use of medical data for predicting diseases based on genetic changes.

Author Contributions: Conceptualization, H.T. and M.T.; methodology, H.T.; software, T.W.; validation, H.T., M.T. and T.W.; formal analysis, T.W.; investigation, T.W.; resources, M.T.; data curation, T.W.; writing—original draft preparation, T.W.; writing—review and editing, H.T. and M.T.; supervision, M.T.; project administration, M.T.; funding acquisition, M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Center for TDB Advanced Data Analysis and Modeling, Tokyo Institute of Technology for academic research purposes. TEIKOKU DATABANK, Ltd. supported our research by providing the data related to Japanese business firms.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. The data were obtained from TEIKOKU DATABANK, Ltd. (chuo-ku, Tokyo 104-8685) and are available from the authors with the permission of TEIKOKU DATABANK, Ltd.

Acknowledgments: We thank Takaya Ohsato (TEIKOKU DATABANK, Ltd.) for the discussions and TEIKOKU DATABANK, Ltd., Center for TDB Advanced Data Analysis and Modeling at Tokyo Institute of Technology, for providing data and financial support.

Conflicts of Interest: TEIKOKU DATABANK, Ltd. did not participate in the research or the preparation of the manuscript, except for the data collection.

References

1. Dash, M.; Liu, H. Feature selection for classification. *Intell. Data Anal.* **1997**, *1*, 131–156. [CrossRef]
2. Kira, K.; Rendell, L.A. A practical approach to feature selection. In *Machine Learning Proceedings 1992*; Elsevier: Amsterdam, The Netherlands, 1992; pp. 249–256.
3. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
4. Saeys, Y.; Inza, I.; Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef]
5. Jain, A.; Zongker, D. Feature selection: Evaluation, application, and small sample performance. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 153–158. [CrossRef]
6. Liu, H.; Li, J.; Wong, L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Inform.* **2002**, *13*, 51–60.
7. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

8. Genuer, R.; Poggi, J.M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236. [CrossRef]
9. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1999.
10. Grandvalet, Y.; Canu, S. Adaptive scaling for feature selection in SVMs. *Adv. Neural Inf. Process. Syst.* **2002**, *15*.
11. Shah, A.D.; Bartlett, J.W.; Carpenter, J.; Nicholas, O.; Hemingway, H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *Am. J. Epidemiol.* **2014**, *179*, 764–774. [CrossRef]
12. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
13. Antoniak, C.E. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Stat.* **1974**, *2*, 1152–1174. [CrossRef]
14. Beaumont, M.A.; Rannala, B. The Bayesian revolution in genetics. *Nat. Rev. Genet.* **2004**, *5*, 251–261. [CrossRef]
15. Pella, J.; Masuda, M. Bayesian methods for analysis of stock mixtures from genetic characters. *Fish. Bull.* **2001**, *99*, 151.
16. Martinez, E.Z.; Achcar, J.A. Trends in epidemiology in the 21st century: Time to adopt Bayesian methods. *Cad. Saúde Pública* **2014**, *30*, 703–714. [CrossRef]
17. Ellison, A.M. Bayesian inference in ecology. *Ecol. Lett.* **2004**, *7*, 509–520. [CrossRef]
18. Yazdani, A.; Kowsari, M. Bayesian estimation of seismic hazards in Iran. *Sci. Iran.* **2013**, *20*, 422–430.
19. Yamada, K.; Takayasu, H.; Takayasu, M. Estimation of economic indicator announced by government from social big data. *Entropy* **2018**, *20*, 852. [CrossRef]
20. Vijaymeena, M.; Kavitha, K. A survey on similarity measures in text mining. *Mach. Learn. Appl. Int. J.* **2016**, *3*, 19–28.
21. Evans, D.S. The relationship between firm growth, size, and age: Estimates for 100 manufacturing industries. *J. Ind. Econ.* **1987**, *35*, 567–581. [CrossRef]
22. Lang, L.; Ofek, E.; Stulz, R. Leverage, investment, and firm growth. *J. Financ. Econ.* **1996**, *40*, 3–29. [CrossRef]
23. Demirgüç-Kunt, A.; Maksimovic, V. Law, finance, and firm growth. *J. Financ.* **1998**, *53*, 2107–2137. [CrossRef]
24. Baum, J.R.; Locke, E.A.; Smith, K.G. A multidimensional model of venture growth. *Acad. Manag. J.* **2001**, *44*, 292–303. [CrossRef]
25. Van Witteloostuijn, A.; Kolkman, D. Is firm growth random? A machine learning perspective. *J. Bus. Ventur. Insights* **2019**, *11*, e00107. [CrossRef]
26. Delmar, F.; Davidsson, P.; Gartner, W.B. Arriving at the high-growth firm. *J. Bus. Ventur.* **2003**, *18*, 189–216. [CrossRef]
27. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
28. Coad, A.; Srhoj, S. Catching Gazelles with a Lasso: Big data techniques for the prediction of high-growth firms. *Small Bus. Econ.* **2020**, *55*, 541–565. [CrossRef]
29. Teikoku Databank Ltd. Our Profile and History. 2022. Available online: https://www.tdb-en.jp/company/profile.html (accessed on 31 January 2023).
30. O'Neill, M.E. PCG: A Family of Simple Fast Space-Efficient Statistically Good Algorithms for Random Number Generation. ACM Transactions on Mathematical Software. 2014. Available online: https://www.pcg-random.org/pdf/toms-oneill-pcg-family-v1.02.pdf (accessed on 30 January 2023).
31. Ward, J.H., Jr. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [CrossRef]
32. Sakurai, H. *Financial Accounting Lecture*, 22nd ed.; Chuokeizai-Sha Holdings, Inc.: Chiyoda-ku, Tokyo, 2021; pp. 91, 92, 139, 140. (In Japanese)
33. Haykin, S. *Neural Networks: A Comprehensive Foundation*; Prentice Hall PTR: Upper Saddle River, NJ, USA, 1998.
34. Evans, D.S. Tests of alternative theories of firm growth. *J. Political Econ.* **1987**, *95*, 657–674. [CrossRef]
35. Davidsson, P. Continued entrepreneurship: Ability, need, and opportunity as determinants of small firm growth. *J. Bus. Ventur.* **1991**, *6*, 405–429. [CrossRef]

entropy

MDPI

Article

The Hurst Exponent as an Indicator to Anticipate Agricultural Commodity Prices

Leticia Pérez-Sienes [1], Mar Grande [1,2], Juan Carlos Losada [1] and Javier Borondo [1,2,3,*]

[1] Grupo de Sistemas Complejos, ETS Ingeniería Agronómica, Alimentaria y de Biosistemas, 28040 Madrid, Spain
[2] AgrowingData, Navarro Rodrigo 2 AT, 04001 Almería, Spain
[3] Departamento de Gestión Empresarial, Universidad Pontificia de Comillas ICADE, Alberto Aguilera 23, 28015 Madrid, Spain
* Correspondence: jborondo@comillas.edu

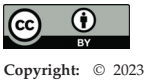
Abstract: Anticipating and understanding fluctuations in the agri-food market is very important in order to implement policies that can assure fair prices and food availability. In this paper, we contribute to the understanding of this market by exploring its efficiency and whether the local Hurst exponent can help to anticipate its trend or not. We have analyzed the time series of the price for different agri-commodities and classified each day into persistent, anti-persistent, or white-noise. Next, we have studied the probability and speed to mean reversion for several rolling windows. We found that in general mean reversion is more probable and occurs faster during anti-persistent periods. In contrast, for most of the rolling windows we could not find a significant effect of persistence in mean reversion. Hence, we conclude that the Hurst exponent can help to anticipate the future trend and range of the expected prices in this market.

Keywords: efficient market; time series; agri-food; Hurst; market; prices; agriculture

Citation: Pérez-Sienes, L.; Grande, M.; Losada, J.C.; Borondo, J. The Hurst Exponent as an Indicator to Anticipate Agricultural Commodity Prices. *Entropy* **2023**, *25*, 579. https://doi.org/10.3390/e25040579

Academic Editor: Panos Argyrakis

Received: 1 March 2023
Revised: 16 March 2023
Accepted: 23 March 2023
Published: 28 March 2023



1. Introduction

Financial markets are extremely complex systems with a large number of interacting units, and anticipating their evolution is far from straightforward. Thus, their study has attracted the attention of researchers over the past decades.

An active and relevant topic of discussion among researchers is whether or not the financial market prices display long memory properties. The importance of this question lies in its consequences for market theories and its predictability. The fact that a market presents a long time memory implies that prices do not follow a random walk, as there is autocorrelation, and they are therefore predictable. On the other hand, if there is no memory, the Efficient Market Hypothesis (EMH) [1,2] cannot be rejected. The EMH was introduced by Fama in 1970 and states that new information is immediately reflected in the asset prices and therefore show martingale behavior. According to this theory, price changes are not related to the historical behavior of price volatility, but represent a response to new information, and since this arrives randomly, the evolution of prices is unpredictable. In the current literature there are papers supporting both hypotheses. Several authors have shown evidence of markets that present long time memory [3–9] whereas other authors have found evidence supporting the EMH [10–12].

In nature we find several examples of physical systems that do present long time memory properties, such as radiation or rainfall. Thus, researchers from econophysics, inspired by the idea that the financial system may share the same properties, have also attempted to detect trends and patterns in financial time series that can help to anticipate its trend. This discipline is known as Technical Analysis [13,14].

Due to the high importance of long time memory for the predictability of time series, there is a clear need for a method that identifies the existence or not of this memory.

Currently, the Hurst exponent (H) is the most widely used and accepted test to measure long-term memory properties [10,15]. H is a measure for long term memory and fractality of a time series that quantifies the degree of persistence of similar change patterns. This analysis was originally introduced by Hurst in 1951 to study the storage capacity of reservoirs in the Nile River, taking into account the cyclical trends of the flow, drought periods, and floods. This work was popularized and extended to other disciplines in the 1960s by Benoit Mandelbrot [16–18], who claimed that this methodology was superior to the autocorrelation, the variance analysis, and to the spectral analysis. Since then, several other methods of Hurst calculation have been developed. The best known ones are Rescaled Range [15], Detrended Fluctuation Analysis [19,20], wavelet transforms [21], and Generalized Hurst Exponent [22].

H ranges between 0 and 1, and provides information on whether the series presents long-term or not. If $H = 0.5$, then each step is independent of the past values of the series. Thus, there is no memory and the series is equivalent to white noise. Under this setting, the time series is unpredictable and the EMH is fulfilled. When $H \leq 0.5$, the series is anti-persistent. In this scenario, the series is expected to display 'mean-reversion'. This fact implies that increments are generally followed by a decrease, while drops are followed by an increment. Finally, when $H \geq 0.5$, the series is persistent. In a persistent regimen, the series is more likely to maintain the trend in a broader range than what is expected by pure random walk. Thus, a rise in the previous step will most likely be followed by another rise, while a fall will be followed by another fall.

Long-term memory is an important feature of market dynamics with implications for its predictability. As a consequence, the Hurst exponent has been widely applied to study the stock, currencies markets, and, more recently, to cryptocurrencies [23,24]. For instance, Di Matteo et al. [12] show how H serves as an index to classify mature and emergent markets. In the same line of research, Bianchi et al. [25], used the Hurst–Hölder exponents to detect periods of efficiency and inefficiency in stock markets [26]. Other researchers have analyzed how to use H to find the most profitable trading pairs [27], concluding that H performs better when compared with the classical methods. Despite the wide use of H in the stock market, there is still limited research on its applicability to the agri-commodities market [28–30]. In this paper we will study the agri-commodities market, and more particularly the evolution of prices for four horticultural products. Understanding this market is becoming increasingly important to make the agri-food industry sustainable [31,32], as price crises result in a waste of food.

The stock market and the agri-commodities market have some similarities, but at the same time also have some important differences. On the one hand, both of them represent a market that is driven by demand. Matia et al. [33,34] showed that the two markets share several properties, although they also found some differences. The cumulative distribution of returns can be adjusted to a power law for both markets. In addition, the returns for the stocks and commodities market exhibit a multifractal behavior. On the other hand, there are differences in the nature of these two markets. In contrast with the stock market, in agri-commodities markets, commodities represent a physical product that has to be stored and transported, and in some cases it is even a fresh and perishable product. Moreover, for agri-commodities we can expect slower changes and response to the demand, since the market is very conditioned by the supply of each product.

In this paper, we will explore the applicability of H to the agri-food market in order to anticipate the trend and range of the future price. In particular, we focus on fresh vegetables because price crises have a big impact on them, since they are perishable products that can not be stored. Thus, when co-ops fail to anticipate the price and can not market their production, it results in tons of wasted food. The effect of the long memory properties on the agri-commodities markets has still attracted little attention from researchers. Thus, there is a gap in the current scientific literature, which misses to fully understand the behavior of such markets. In Ref. [35] the authors analyze the auto-correlations and cross-correlations of the volatility time series for the Brazilian stock and commodity markets.

They found auto-correlations in the commodity market, which in fact are stronger than that observed for the stock market. In another study—see Ref. [30]—the authors computed the Hurst exponent for several commodity price series, and found that most commodity prices are consistent with the underlying assumption of a geometric Brownian motion. We will contribute to understanding the dynamics and properties of the market by analyzing the evolution of H over time, and evaluating whether the value of H can provide useful information to anticipate the future trend of the price. In addition, we will compare the results obtained when computing H for different time windows.

The present paper is organized as follows. In Section 2, we will explain the methodology followed to compute the local Hurst exponent of the series and the mean reversion. Next, in Section 2.3, we describe our data. In Section 3, we expose our results. Finally, in Section 4, we present our conclusions and discuss the importance of our results.

2. Materials and Methods

2.1. Hurst Exponent

The Hurst exponent (H) is used in time series analysis and fractal analysis as a measure of the long-term memory of a time series. In other words, H measures how chaotic or unpredictable a time series is. In the literature, we can find several methods to calculate H, such as re-scaled Range (RS) [15], Detrended Fluctuation Analysis (DFA) [19,20], wavelet transforms [21], and Generalized Hurst Exponent (GHE) [22].

In this work, we use the GHE algorithm in order to measure the long-term memory of the price time series of different agri-commodities. This method is based on the scaling behavior of the statistic:

$$K_q(\tau) = \frac{\langle |X(t+\tau) - X(t)|^q \rangle}{\langle |(X(t))^q| \rangle}, \tag{1}$$

which is given by

$$K_q(\tau) \propto \tau^{qH}, \tag{2}$$

where τ is the time scale and can vary between 1 and τ_{max}, H is the Hurst exponent, $< \cdot >$ denotes the sample average on time t, and q represents the order of the moment considered.

H is then calculated by taking logarithms in relation (2) for different values of τ. In this paper, we work with $\tau = 2^n$ $(n = 0, 1, \ldots, log_2(N) - 2)$, and $q = 1$, as H_1 is the closest estimation to the classical Hurst exponent [12].

H ranges between 0 and 1, where $H = 0.5$ means that there is no memory and the series is equivalent to white noise. When $H \leq 0.5$, the series is considered anti-persistent and is expected to display 'mean-reversion'. Finally, when $H \geq 0.5$, the series is considered persistent and is more likely to maintain the trend in a broader range than what is expected by a pure random walk.

In order to prevent H from using future values of the time series, we calculate a local Hurst exponent with reference to a rolling window of 4, 8, 16, 32, and 52 weeks that ends the day of measurement. This method ensures that we use only past data to determine H.

Note that in order to compute H, we have coded the described method using Python.

2.2. Days to Mean Reversion

Mean reversion (MR), or reversion to the mean, is a theory used in finance that suggests that a measure of interest such as the price of a commodity or asset eventually reverts to its long-term average levels. Thus, this theory assumes that a variable that deviates far from its long-term trend will return, reverting to its average value. This concept has been used to define many investment strategies that seek to purchase or sell financial products whose recent market price differs greatly from their historical average [36].

In this work we are going to test whether this reversion is more likely to occur during anti-persistent periods rather than during persistent periods in the price time series. Thus, for each day i, we compute the number of days (d) that the price (p) of an agri-commodity lasts to revert to its average value (m) as

$$d = j^* - i$$

where j^* is the minimum j that satisfies

$$\begin{cases} p_j \leq m_i + \epsilon, & \text{if } p_i \geq m_i \\ p_i \geq m_i - \epsilon, & \text{if } p_i < m_i \end{cases}$$

with $j \geq i$.

The average value (m) of each price time series have been calculated with reference to a rolling window of 4, 8, 16, 32, and 52 weeks, thus ensuring that we use only past values of the series to determine m.

2.3. Dataset

In this work, we analyze the price time series of four agri-commodities (aubergine, zucchini, green pepper, and cucumber) in the South region of Spain. We have focused on these four products because they are representative of the European vegetables market, as they represent a significant percentage of the total imports and exports.

The units of the prices are measured in EUR per kilogram. Each time series consists of daily prices collected over five years, and we must note that a typical week consists of six observations, since the market is closed on Sundays.

Figure 1 shows the evolution of the price of each commodity during the period 2015–2019. As it can be seen, all products present high volatility and despite showing a seasonal component, the noise component is still very important. In addition, we have included in the figure the moving averages for different rolling windows.

In Table 1, the Hurst exponent, average price, and standard deviation of each time series are shown. All of the products present a global Hurst below 0.45, i.e., antipersistent. Green pepper is the most antipersistent time series ($H = 0.18$), and also the commodity with the highest and least volatile price.

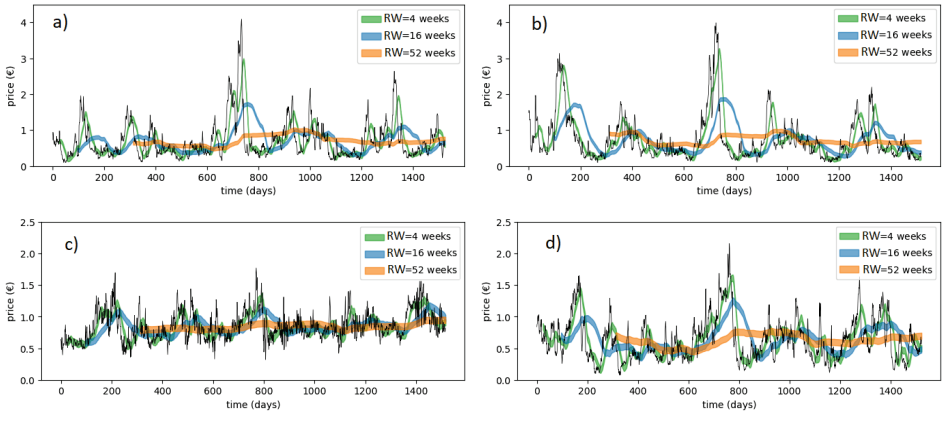


Figure 1. The figure shows the evolution of the price and its corresponding moving average for different rolling windows *RW* for (**a**) aubergine, (**b**) zucchini, (**c**) pepper, and (**d**) cucumber.

Table 1. Total Hurst, H, and mean and standard deviation of price for the four time series shown in Figure 1.

	Aubergine	**Zucchini**	**Green Pepper**	**Cucumber**
H	0.36	0.40	0.18	0.32
$\bar{p} \pm \sigma_p$	0.70 ± 0.53	0.75 ± 0.63	0.85 ± 0.22	0.64 ± 0.33
$p_{\max} - p_{\min}$	3.99	3.89	1.45	2.08

3. Results

The main goal of our paper is to analyze if H can help to anticipate the future trend of the price of four different agri-commodities, and discuss the difference in performance when considering different rolling windows (RW). To this end, we will analyze whether the probability of MR in the short term is more likely during anti-persistent days than on persistent days or not.

3.1. Evolution of the Hurst Exponent

To achieve this goal we begin by computing H as described in the methods section, over our time series for different values of RW (4, 8, 16, 32, and 52 weeks). Thus, for each day of our series we have computed the value of H for the mentioned windows. The evolution of H, for aubergine over time and its distribution, can be found in Figure 2. Panel B of the figure shows the distribution of H, which approximately follows a normal distribution. The mean value of H depends on the commodity and RW, but for most cases (except for green pepper) is close to 0.5. The exact values for each combination can be found in Tables 2 and 3.

We found that for the four products, H varies over time, alternating periods of persistence, anti-persistency, and neutral regimens. Changes over days tend to be relatively smooth, and when the series enters one of the three regimens it keeps for a while. This effect is illustrated in panel B of Figure 2, which shows the case of aubergine.

3.2. Mean Reversion

In the second step, we compute for each day and RW the days to MR—this is the number of days left before the price will return or cross the mean. The days to MR follow a heterogeneous distribution, where for all the RWs over 50% of the observations revert to the mean in less than 20 days, while a small fraction take over 100 days. This can be observed in Figure 3a), which shows the probability mass functions (PMF) of each RW for aubergine.

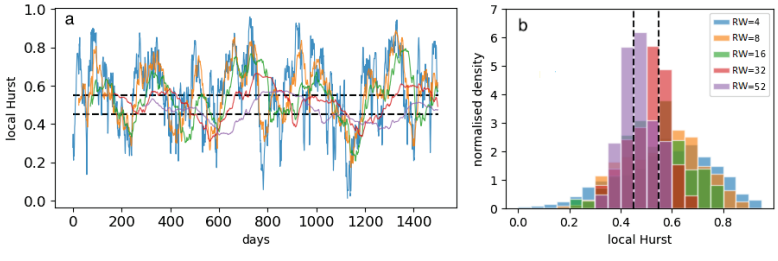


Figure 2. (**a**) The evolution of the local Hurst exponent for aubergine and different rolling windows. (**b**) Histogram of aubergine local Hurst exponent values. In both panels the dashed lines reflect the selected thresholds of 0.45 and 0.55.

Table 2. This table shows the median of days to MR and mean local Hurst ($\bar{H}$) for the studied rolling windows $RW = 4, 8, 15, 32, 52$ weeks for aubergine and zucchini. The global value of H of each product is also shown in the top row.

	Aubergine ($H = 0.36$)		**Zucchini ($H = 0.40$)**	
RW	**Days to MR**	$\bar{H} \pm \sigma_H$	**Days to MR**	$\bar{H} \pm \sigma_H$
4	10	0.57 ± 0.18	10	0.57 ± 0.20
8	11	0.56 ± 0.14	12	0.57 ± 0.15
16	15	0.54 ± 0.12	22	0.55 ± 0.12
32	20	0.58 ± 0.08	24	0.52 ± 0.10
52	14	0.46 ± 0.06	17	0.49 ± 0.07

Table 3. This table shows the median of days to MR and mean local Hurst ($\bar{H}$) for the studied rolling windows $RW = 4, 8, 16, 32, 52$ weeks for green pepper and cucumber. The global Hurst of each product is also shown in the top row.

	Green Pepper ($H = 0.18$)		**Cucumber ($H = 0.32$)**	
RW	**Days to MR**	$\bar{H} \pm \sigma_H$	**Days to MR**	$\bar{H} \pm \sigma_H$
4	3	0.32 ± 0.16	11	0.52 ± 0.19
8	4	0.28 ± 0.10	11	0.53 ± 0.12
16	5	0.26 ± 0.07	14	0.50 ± 0.07
32	5	0.25 ± 0.05	15	0.46 ± 0.05
52	4	0.22 ± 0.04	14	0.41 ± 0.06

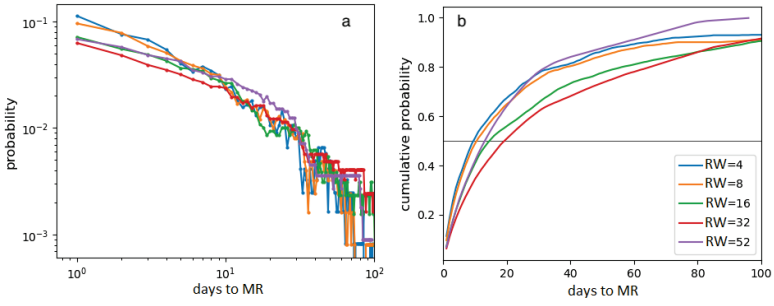


Figure 3. (**a**) Probability mass distribution of days to MR for the aubergine time series and for various RW. (**b**) The corresponding cumulative probability distributions.

When exploring the relation between H and days to MR, we find that the product (pepper) with a significantly smaller value of H, and in an anti-persistent regime (0.22–0.32), generally reverts to the mean more rapidly. For example, when conducting the analysis with $RW = 16$ weeks, the price of pepper returns to the mean in an average of 5 days, while for the other products this value ranges between 11 and 22 days.

Next, we analyze if the probabilities that the price will revert to the mean are significantly different for persistent and anti-persistent periods for each of the rolling windows. To this end, we classify each observation of the time series into persistent ($H \geq 0.55$), neutral ($0.45 < H < 0.55$), and anti-persistent ($h \leq 0.45$). Thus, we classify each day into one of the three different groups. We calculate the cumulative probability distribution of the days to MR for the persistent group, the anti-persistent group, and the total population and compare them for all the RWs. Figure 4 shows these distributions for aubergine.

We find that for all the RWs, except for 52 weeks, the cumulative distribution of days to MR significantly differs for the two regimens, the probability of MR being higher for anti-persistent days. In agreement, with this observation, the curve for the total population lies in-between both. Thus, anti-persistent days revert to the mean more quickly than persistent ones, which at the same time revert with lower probability than the global population. For the mentioned RW of 52 weeks (1 year), we observe the contrary effect, where the probability of MR in d or less days is always smaller for the anti-persistent regimen, the persistent group and the total population exhibiting a very similar behavior.

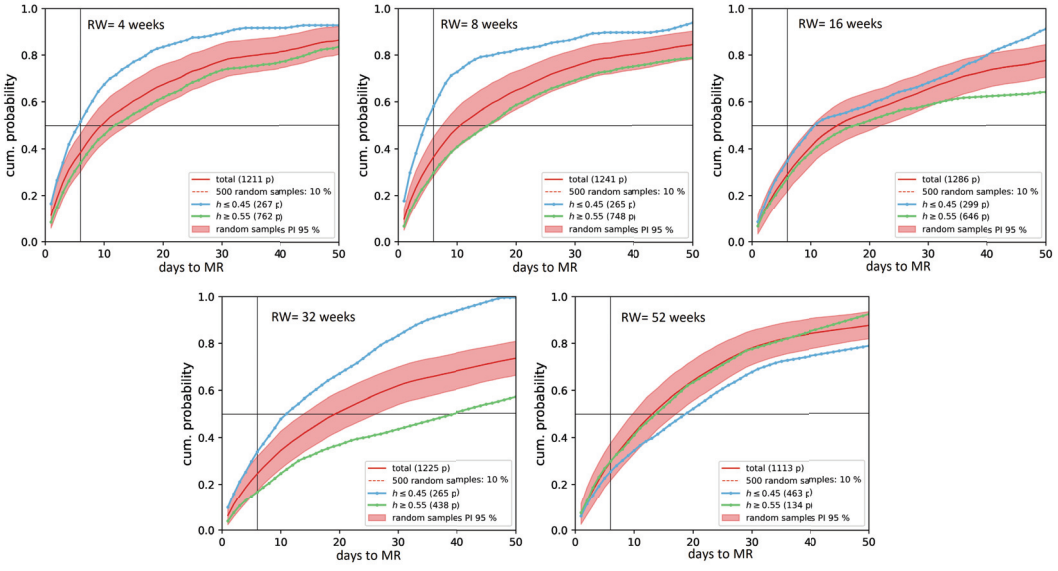


Figure 4. Cumulative probability of days to MR for the price time series of aubergine for different rolling windows RW. The blue curve represents the antipersistent regime, while the green curve represents the persistent regime. The horizontal line shows the 50% of probability, and the vertical one marks 6 days. The figure also shows the 95% level of confidence for the 500 random sub-samples.

To test whether the observed behavior for the persistent and antipersistent groups could be random, we take random samples of the total population and compare their behavior to both groups. We do so, because the persistent and antipersistent groups are subsamples of the total population. Thus, there is a chance that by choosing a random subsample of similar size we find a similar effect. In such cases the effect we have observed would not be significant. Thus, we have randomly selected 500 subsamples and computed the cumulative distribution of days to MR for each one of them. In Figure 4, we have included the 2.5% and 97.5% percentiles so that they can be compared with the curves of the two groups. We find that for all the RWs, except 32 weeks, the effect of the persistent group is not likely to be significant as the curves lie in-between the two percentiles. However, for $RW = 32$ weeks the probability of MR in 10 or less days for persistent days is below 30% and for 20 days it is below 40%. This means that we are around 70% confident that the price will stay at the current side of the border of the mean during the following 10 days, and 60% that it will also stay in the next 20 days. Hence, when computed with a RW of 32 weeks, $H \geq 0.55$ seems to be a good and informative indicator to anticipate the range in which the price will move in the medium-long term.

On the other hand, the anti-persistent group differs from the total population and the random samples for RWs of 4, 8, and 32 weeks, but not for 16 weeks. When analyzing our data, we can see that the most informative RWs are the shorter ones: 4 and 8, as the gap between the antipersistent group and the total population is the largest. This effect is

especially relevant in the case of 8 weeks, where the probability of MR in one week is 58%, while the value for the total population is under 40%. Thus, an indicator detecting points where $H \leq 0.45$ would be informative for movements in the short term, as we will have 58% of probabilities that the price will return to its average. Hence, if the actual price is below the average we can anticipate an uptrend, and if the price is below we can expect a drop in the prices for the following week.

3.3. Paired t-Test

To further test the effect of persistence and anti-persistence in the expected days to MR, we adopt a quasi-experimental design, through which we can compare persistent and antipersistent observations against a control group. To this end, we perform two dependent *t*-tests for paired samples. The first one is to measure the effect of persistence and the second one is to measure the effect of anti-persistence. We use a dependent paired t-test, because our observations are extracted from a time series of prices, and thus can not be considered independent. In particular, we control for the commodity and the week, since the time series of prices have a marked seasonal component. Thus, for the persistence experiment we match the number of days to MR of each persistent observation to the number of days to MR on a randomly selected observation of the same week and commodity in the control group (all non antiperspirant observations). The anti-persistence experiment is designed analogously.

The results of both tests for the RWs of 4, 8, 16, and 32 weeks are summarized in Figure 5. As it can be seen for all the RWs except for RW = 16, the MR occurs significantly faster for antipersistent observations. The most informative RW is the one of 8 weeks, where for antipersistent days we can expect MR to happen on average 12 days faster. In contrast, the effect of persistence is not so evident in our data, and we only find a significant effect for RW = 16, where in persistent periods MR takes on average 12 days longer. Note that we have not included the RW = 52 case, as there were not enough paired observations for the results to be reliable.

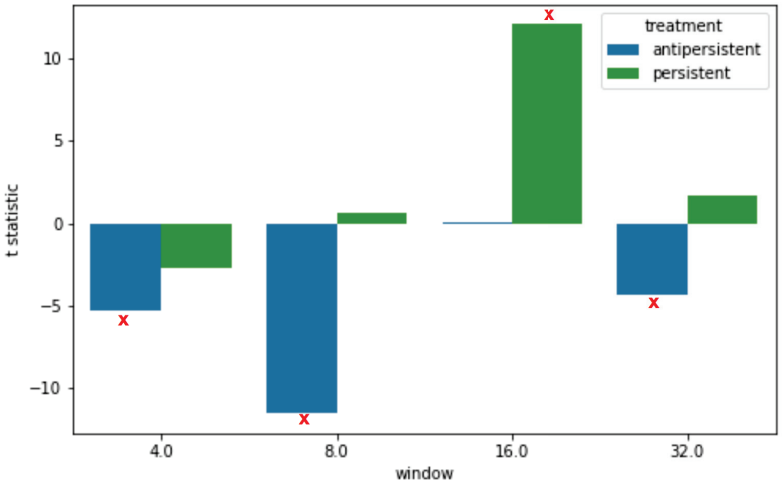


Figure 5. Results of the paired t-test for the persistence ($H \geq 0.55$) and antipersistence ($H \leq 0.45$) experiments. The figure shows the t-statistic, measured in days, for both experiments and the different RWs (RW = 4, 8, 16, 32 weeks). The significant differences have been marked with a red cross (p-value < 0.01).

4. Discussion

In this paper we contribute to understand the agri-food market by exploring whether the market presents long term memory properties or not for several agri-commodities.

To this end, we have computed the local Hurst exponent for several values of RW and measured the relation between its evolution and the probability that the price will revert to the mean. We found that in general for antipersistent days MR is more probable in the short term and occurs faster than for persistent and neutral days. The only value of RW for which we could not find a significant effect was $RW = 16$, while the most informative one was $RW = 8$.

The fact that MR is more probable and happens faster during antipersistent periods means that H can be a good indicator to anticipate the future motion and help actors operating in this market, such as co-ops or supermarkets. More particularly, it is important to discuss its implications to anticipate the price and operate in this market in the short-term. For $RW = 8$ weeks we found that for antipersistent days almost 60% of the times MR will happen in one week or less, a magnitude significantly larger than what is expected for the full population, where the chances of MR in one week are below 45%. This fact, shows that H can be helpful to operate in such a market as it provides information to anticipate the future trend of the price. If the price is below the average, the operator will know that there is a high chance that the price will go up, while when the price is above the average a downtrend is very probable. We focus on the one week resolution, because anticipating MR in the long term is not so useful in this market. For example, knowing that MR will happen during the following two months, but not knowing when, means that the operator has to trade daily tons of fresh products with a high uncertainty on when the price movement will happen.

On the other hand, market indicators related to persistence can be related to the fact that MR is not very probable or will happen slowly. Thus, this kind of indicator is more useful to operate in the market in the long term. The fact that there is a low probability of MR for the following days is not too informative, as price time series present autocorrelation and the price from one day to another usually does not present big differences. In contrast, knowing that MR is not probable in the long term will help to anticipate the range in which the price will oscillate in the following months. When the price is above the average, knowing MR is not probable, and it is useful to know the lower barrier that the price will not cross in the following weeks and months. Likewise, when the price is below the average we have an upper frontier that the price is unlikely to cross. Thus, this information helps the actors of the market to negotiate long term contracts, which is very common between supermarkets and co-ops, where the second commits to provide a minimum quantity of tons during the following months to the second one for a fixed price.

A relevant research topic that we plan to explore in future work is the development of a methodology to find the optimal rolling window to use when computing the local Hurst exponent for each price time series. Thus, we aim to analyze a wider variety of products that follow different dynamics, and find their corresponding best rolling window.

Author Contributions: Conceptualization, J.C.L. and J.B.; Methodology, J.C.L. and J.B.; Validation, M.G., J.C.L. and J.B.; Formal analysis, L.P.-S. and M.G.; Investigation, L.P.-S.; Data curation, M.G.; Writing—original draft, J.B.; Writing—review & editing, L.P.-S., M.G., J.C.L. and J.B.; Visualization, L.P.-S. and M.G.; Funding acquisition, J.C.L. and J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Spanish Ministry of Science and Innovation under Contract No. PID2021-122711NB-C21 and DIN2018-010114, and by DG of Research and Technological Innovation of the Community of Madrid (Spain) under Contract No. IND2022/TIC-23716.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used for this study are publicly available from Observatorio de Precios y Mercados, Junta Andalucia, Spain at https://www.juntadeandalucia.es/agriculturaypesca/observatorio/servlet/FrontController?action=Static&url=introduccion.jsp and agroprecios.com at https://www.agroprecios.com/es/precios-subasta/.

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

References

1. Fama, E.F. Efficient capital markets: A review of theory and empirical work. *J. Financ.* **1970**, *25*, 383–417. [CrossRef]
2. Fama, E.F. Efficient capital markets: II. *J. Financ.* **1991**, *46*, 1575–1617. [CrossRef]
3. Greene, M.T.; Fielitz, B.D. Long-term dependence in common stock returns. *J. Financ. Econ.* **1977**, *4*, 339–349. [CrossRef]
4. Hampton, J. Rescaled range analysis: Approaches for the financial practitioners, Part 3. *Neuro Vest J.* **1996**, *4*, 27–30.
5. Lillo, F.; Farmer, J.D. The Long Memory of the Efficient Market. *Stud. Nonlinear Dyn. Econom.* **2004**, *8*, 1–19. [CrossRef]
6. Barkoulas, J.T.; Baum, C.F. Long-term dependence in stock returns. *Econ. Lett.* **1996**, *53*, 253–259. [CrossRef]
7. Wright, J.H. Long memory in emerging market stock returns. *FRB Int. Financ.* **2000**, pii: Discussion Paper No. 650. Available online: https://www.federalreserve.gov/econres/ifdp/long-memory-in-emerging-market-stock-returns.htm (accessed on 27 March 2023).
8. Kasman, S.; Turgutlu, E.; Ayhan, A.D. Long memory in stock returns: Evidence from the major emerging central European stock markets. *Appl. Econ. Lett.* **2009**, *16*, 1763–1768. [CrossRef]
9. Cheong, C. Estimating the hurst parameter in financial time series via heuristic approaches. *J. Appl. Stat.* **2010**, *37*, 201–214. [CrossRef]
10. Lo, A.W. Long-Term Memory in Stock Market Prices. *Econometrica* **1991**, *59*, 1279–1313. [CrossRef]
11. Lo, A.W.; MacKinlay, A.C. *Long-Term Memory in Stock Market Prices. A Non-Random Walk Down Wall Street*, 1st ed.; Princeton University Press: Princeton, NJ, USA, 1999.
12. Di Matteo, T.; Aste, T.; Dacorogna, M.M. Long term memories of developed and emerging markets: Using the scaling analysis to characterize their stage of development. *J. Bank. Financ.* **2005**, *29*, 827–851. [CrossRef]
13. Brown, D.P.; Jennings, R.H. On technical analysis. *Rev. Financ. Stud.* **1989**, *2*, 527–551. [CrossRef]
14. Park, C.H.; Irwin, S.H. The Profitability of Technical Analysis: A Review. 2004. AgMAS Project Research Report No. 2004-04. Available online: http://dx.doi.org/10.2139/ssrn.603481 (accessed on 27 March 2023).
15. Hurst, H.E. Long Term storage capacity of reservoirs. *Trans. Am. Soc. Civ. Eng.* **1951**, *116*, 770–799. [CrossRef]
16. Mandelbrot, B.B. When can price be arbitraged efficiently? A limit to the validity of the random walk and martingale models. *Rev. Econ. Stat.* **1971**, *53*, 225–236. [CrossRef]
17. Mandelbrot, B. Statistical methodology for nonperiodic cycles from covariance to R/S analysis. *Ann. Econ. Soc. Meas.* **1972**, *1*, 259–290.
18. Mandelbrot, B.; Wallis, J.R. Robustness of the rescaled range R/S in the measurement of noncyclic long-run statistical dependence. *Water Resour.* **1969**, *5*, 967–988. [CrossRef]
19. Peng, C.K.; Buldyrev, S.V.; Havlin, S.; Simons, M.; Stanley, H.E.; Goldberger, A.L. Mosaic organization of DNA nucleotides. *Phys. Rev. E* **1994**, *49*, 1685. [CrossRef] [PubMed]
20. Hu, K.; Ivanov, P.C.; Chen, Z.; Carpena, P.; Stanley, H.E. Effect of trends on detrended fluctuation analysis. *Phys. Rev. E* **2001**, *64*, 011114. [CrossRef]
21. Simonsen, I.; Hansen, A.; Nes, O.M. Determination of the Hurst exponent by use of wavelet transforms. *Phys. Rev. E* **1998**, *58*, 2779. [CrossRef]
22. Barabasi, A.L.; Vicsekt, T. Multifractality of self affine fractals. *Phys. Rev. A* **1991**, *44*, 2730–2733. [CrossRef]
23. Caporale, G.M.; Gil-Ana, L.; Plastun, A. Persistence in the cryptocurrency market. *Res. Int. Bus. Financ.* **2018**, *46*, 141–148. [CrossRef]
24. Dimitrova, V.; Fernández-Martínez, M.; Sánchez-Granero, M.A.; Trinidad Segovia, J.E. Some comments on Bitcoin market (in)efficiency. *PLoS ONE* **2019**, *14*, e0219243. [CrossRef] [PubMed]
25. Bianchi, S.; Pianese, A. Time-varying Hurst-Hölder exponents and the dynamics of (in)efficiency in stock markets. *Chaos Solitons Fractals* **2018**, *109*, 64–75. [CrossRef]
26. Cajueiro, D.O.; Tabak, B.M. Ranking efficiency for emerging markets. *Chaos Solitons Fractals* **2004**, *22*, 349. [CrossRef]
27. Ramos-Requena, J.P.; Trinidad-Segovia, J.E.; Sánchez-Granero, M.A. Introducing Hurst exponent in pair trading. *Physica A* **2017**, *488*, 39–45. [CrossRef]
28. Corazza, M.; Malliaris, A.G.; Nardelli, C. Searching for fractal structure in agricultural future markets. *J. Future Mark.* **1997**, *17*, 433–473. [CrossRef]
29. Barkoulus, J.; Labys, W.C.; Onochie, J. Fractional dynamics in international commodity prices. *J. Future Mark.* **1997**, *17*, 161–189. [CrossRef]
30. Turvey, C.G. A note on scaled variance ratio estimation of the Hurst exponent with application to agricultural commodities prices. *Physica A A Stat. Mech. Its Appl.* **2007**, *377*, 155–165. [CrossRef]
31. Allen, P. *Together at the Table: Sustainability and Sustenance in the American Agrifood System*; Penn State Press: University Park, PA, USA, 2004.
32. Borsellino, V.; Schimmenti, E.; El Bilali, H. Agri-food markets towards sustainable patterns. *Sustainability* **2020**, *12*, 2193. [CrossRef]
33. Matia, K.; Amaral, L.A.N.; Goodwin, S.; Stanley, H.E. Different scaling behaviors of commodity spot and future prices. *Phys. Rev. E* **2002**, *66*, 045103. [CrossRef]

34. Matia, K.; Ashkenazy, Y.; Stanley, H.E. Multifractal properties of price fluctuations of stocks and commodities. *Eutophys. Lett.* **2003**, *61*, 422–428. [CrossRef]
35. Siqueira, E.L., Jr.; Stošić, T.; Bejan, L.; Stošić, B. Correlations and cross-correlations in the Brazilian agrarian commodities and stocks. *Phys. A Stat. Mech. Its Appl.* **2010**, *389*, 2739–2743. [CrossRef]
36. Poterba, J.M.; Summers, L.H. Mean reversion in stock prices: Evidence and Implications. *J. Financ. Econ.* **1998**, *22*, 27–59. [CrossRef]

entropy

MDPI

Article

The Risk Contagion between Chinese and Mature Stock Markets: Evidence from a Markov-Switching Mixed-Clayton Copula Model

Hongli Niu [1,*], Kunliang Xu [1] and Mengyuan Xiong [2]

[1] School of Economics and Management, University of Science and Technology Beijing, Beijing 100083, China
[2] School of Economics and Management, Hanjiang Normal University, Shiyan 442000, China
* Correspondence: niuhongli@ustb.edu.cn

Abstract: Exploring the risk spillover between Chinese and mature stock markets is a promising topic. In this study, we propose a Markov-switching mixed-Clayton (Ms-M-Clayton) copula model that combines a state transition mechanism with a weighted mixed-Clayton copula. It is applied to investigate the dynamic risk dependence between Chinese and mature stock markets in the Americas, Europe, and Asia–Oceania regions. Additionally, the conditional value at risk (CoVaR) is applied to analyze the risk spillovers between these markets. The empirical results demonstrate that there is mainly a time-varying but stable positive risk dependence structure between Chinese and mature stock markets, where the upside and downside risk correlations are asymmetric. Moreover, the risk contagion primarily spills over from mature stock markets to the Chinese stock market, and the downside effect is stronger. Finally, the risk contagion from Asia–Oceania to China is weaker than that from Europe and the Americas. The study provides insights into the risk association between emerging markets, represented by China, and mature stock markets in major regions. It is significant for investors and risk managers, enabling them to avoid investment risks and prevent risk contagion.

Keywords: risk contagion; Chinese stock market; mature stock markets; Markov-switching; Clayton copula

Citation: Niu, H.; Xu, K.; Xiong, M. The Risk Contagion between Chinese and Mature Stock Markets: Evidence from a Markov-Switching Mixed-Clayton Copula Model. *Entropy* **2023**, *25*, 619. https://doi.org/10.3390/e25040619

Academic Editor: Panos Argyrakis

Received: 20 February 2023
Revised: 27 March 2023
Accepted: 1 April 2023
Published: 6 April 2023

1. Introduction

As global stock markets become increasingly interconnected, the risk of contagion is becoming more prominent [1,2]. Measuring this contagion effectively is crucial for China and other emerging economies, which may be more vulnerable to international risk contagion, to improve their risk supervision [3,4]. Numerous scholars have studied the risk contagion between markets [5–7]. With the rapid development of the Chinese stock market, the largest emerging market in the world, there is growing interest in investigating the risk contagion between it and more mature markets, and several models have been developed for empirical analysis [8,9]. Traditional models have limitations in depicting the dynamic and asymmetric structures and are constrained by their ability to only show linear correlations. Consequently, scholars have turned to copula-based models to enrich research in this field, and the advantages of copula-based models over traditional models have been confirmed [10–13].

The motivation behind this work is twofold. Firstly, most existing copula-based models that evaluate risk contagion tend to focus on measuring individual tail correlations or positive dependence, which limits the analysis of the contagion mechanism from a comprehensive perspective [14,15]. While it is important to examine positive dependence, which occurs when two stock markets rise or fall in tandem, it is also crucial to consider negative dependence structures, where one market rises while the other falls, which may offer opportunities for hedging investment risks or realizing arbitrage. Overemphasizing risk contagion under one dimension may lead to a distorted perception of international

markets. Therefore, it is necessary to develop a tool that provides a more comprehensive assessment of the dependence structure between markets. Secondly, the Chinese stock market has a unique profile with its late start, rapid development, and high volatility, and most existing conclusions and guidelines drawn from mature markets may not provide reliable references for its development. As emerging countries are often in a passive position in international risk contagion, investigating the risk contagion mechanism between the Chinese market and different mature markets has important reference significance [16–18].

Thus, to comprehensively investigate the risk contagion between Chinese and mature markets in three representative regions on a global scale (Asia–Oceania, Europe, and the Americas [19]), we attempt to construct a novel Markov-switching mixed-Clayton (Ms-M-Clayton) copula model. This model considers four types of tail correlations simultaneously and calculates the conditional value-at-risk (CoVaR) in different routes [20,21].

We start with a Clayton copula model that describes the upside correlation of two random variables under a positive dependence structure. We first rotate it by 90°, 180°, and 270° and weight the four individual models as a mixed-Clayton (M-Clayton) copula. The M-Clayton model can capture both upside and downside correlations between two markets under both positive and negative dependence structures. Then, to capture correlations in a time-varying manner, we introduce a two-state switching mechanism following the Markov chain. Using the estimated results of the Ms-M-Clayton copula model, we calculate the CoVaRs under the four dimensions to measure the markets' risk spillover. The empirical results indicate that there is dynamic and generally stable positive dependence between the Chinese and mature markets, with the downside risk correlation being stronger than the upside correlation in most cases. Additionally, the risk contagion is primarily manifested in a spillover from mature markets to the Chinese market. Furthermore, the risk spillover from Asia-Oceania to China is weaker than that from Europe and the Americas, implying that Japanese and Australian markets may be potential choices for Chinese market investors to diversify investment risks. Overall, this study reveals the risk contagion effects between emerging markets, represented by China, and major mature markets. Our findings have practical and policy implications for investors and supervisors to mitigate the adverse effects of risk contagion.

This study makes several contributions to the literature on risk contagion between Chinese and mature markets. Firstly, a novel Ms-M-Clayton copula model is formulated and applied to dynamically measure the asymmetrical dependence structure between Chinese and mature markets in three global risk regions, providing a more comprehensive perspective on the risk contagion patterns between economies. Secondly, by calculating the CoVaR in four relevant scenarios based on the results of the Ms-M-Clayton copula, we quantify and compare the risk dependence and contagion between Chinese and different mature markets, respectively. Thirdly, we provide detailed explanations for the time-varying risk dependence structure and contagion. Based on the empirical results, we provide targeted insights for both emerging and mature economies on how they can defuse risk contagion and stay safe by monitoring objects with high-risk dependence.

The remainder of this paper is arranged as follows: Section 2 sorts out the existing research on the risk contagion and the related measurement methods. Section 3 introduces the construction of marginal distribution model and Ms-M-Clayton copula model. Section 4 summarizes the datasets. Section 5 reports the empirical experiments and results. Finally, this work is concluded in Section 6.

2. Literature Review

Despite numerous studies exploring risk contagion, the definition is still controversial [22,23]. It is commonly believed that the risk contagion is driven by heterogeneous factors such as the investors' behaviors and expectations [24], the information bias [25], the market supervision [26], and the completeness of financial system [27].

As one of the most representative emerging economies, the Chinese market is gradually becoming international, especially since its accession to the WTO. Thus, based on

different but not mutually independent definitions of risk contagion, plenty of scholars have discussed the risk contagion between Chinese and various markets by using the traditional econometric methods, such as the granger causality, generalized autoregressive conditional heteroskedasticity (GARCH), vector autoregressive model, etc. [28–30], but most of the methods fail to depict nonlinear dependence and capture the asymmetric relationships dynamically. Moreover, they have poor ability to measure the tail correlation reflecting the extreme risk contagion.

To overcome these shortcomings, various copula-based methods [31] are proposed to capture the dynamic and asymmetric dependencies between series. Chang [32] constructed a mixed copula of Gumbel and Clayton copula to investigate the asymmetry between the upside and downside risk correlations of crude oil spot and futures. Huang et al. [33] proposed the rotated Gumbel and Clayton copulas, which provide a flexible perspective to measure the asymmetric risk correlation. Hussain and Li [34] found that the Chinese market has stronger dependence with Asia and Europe than the US by employing stochastic copulas. Luo et al. [35] measured the multiscale financial risk association among nine stock markets by introducing empirical mode decomposition into copulas, revealing that the high-frequency fluctuation is the major contributor of contagion. Although scholars have extended copula models on the measurement of asymmetric tail correlations, most of them are still time-invariant and only suitable for depicting static relationships.

More recently, time-varying mechanisms, such as parameter autocorrelation equations and state transition probabilities, are introduced to the invariant copulas, allowing dynamic and periodic dependence analyses [32,36–38]. Huang et al. [39] verified the superiority of the time-varying parameter (TVP) copulas compared to traditional methods in constructing the minimum-risk portfolios from G7 countries' markets. Chang [32] documented the non-fixed dependence between inflation rate and REIT return by constructing a Markov-switching GRG copula, while Wang et al. [40] highlighted that the negative dependence reflects the reversal effect, which is crucial to revisit the dependence structure between markets. Thus, they constructed a dependence-switching copula based on multiple Clayton copulas to examine the risk relevance between stock and foreign exchange markets. Ji et al. [41] identified the conditional dependence between energy and agricultural commodity markets and confirmed the significance of negative dependence. However, on one hand, most of the dynamic dependence-switching copulas methods are still limited to capture the positive dependence; on the other hand, the literature utilizing TVP copulas to investigate the risk contagion between Chinese and mature markets remains to be enriched.

Several studies further quantify the degree of directional risk contagion by calculating VaR and CoVaR based on the risk association captured by copula-based models, proving the function of copula-CoVaR paradigm in measuring risk contagion. Reboredo and Ugolini [42] used the CoVaR-copula method to investigate the systemic risk contagion level in European sovereign debt markets as well as the asymmetric downside and upside spillover between precious metals [43]. Xiao [44] developed a MSGARCH-EVT-copula model and computed the CoVaR to investigate the risk spillovers of Chinese market to major East Asian markets, reporting that the downside and upside spillovers are generally different between the turbulent and calm periods. Jiang et al. [19] constructed a vine-copula-GARCH-MIDAS model and computed the CoVaR to estimate the risk spillovers among multiple stock markets. Sun et al. [45] verified that the GARCH-Copula-CoVaR method is suitable for evaluating the risk contagion of international commodity markets. Therefore, it is essential to assess the risk spillovers in different routes, which can help to understand the risk contagion mechanism between markets.

In summary, although copula models provide a more flexible perspective for depicting the non-linear risk dependence between markets, most of them focus on single tail correlation or in the positive dependence structure. Positive and negative risk dependency structures provide novel insights into financial risk contagion [40,41]. In particular, tail correlations in negative dependency structures are helpful to identify risk-hedging opportunities, so it is essential to enrich research in this field. As one of the most representative

emerging markets that are vulnerable in global risk contagion, the risk contagion between Chinese and global mature markets is still controversial [44]. Therefore, we formulate a Ms-M-Clayton copula model and compute the CoVaR to analyze the risk spillovers, which not only enriches the application of dependence-switching copula models but also helps to revisit the risk contagion between Chinese and mature markets around the world.

3. Methodology

3.1. Marginal Distribution Modeling

Prior to the Copula modeling that is used to capture risk dependence between markets, a marginal distribution modeling is necessary to be applied to the original financial return time series, i.e., extracting the components that can be described by econometric models and treating the residuals as risks that cannot be depicted by models. Then, the residuals are used as the input of copula model to describe the risk dependence between markets. Considering the autocorrelation, volatility clustering and leptokurtosis of financial return series, the AR-GARCH is one of the most commonly used models to describe the financial time series [46]. Moreover, compared with the normal distribution, the generalized error distribution (GED) fits the financial time series better as it captures the thick-tailed properties well. Therefore, the AR(m)-GARCH(p,q) model with GED process is employed for marginal distribution modeling, which is written as:

$$\begin{cases} r_t = \phi_0 + \sum_{i=1}^{m} \phi_i r_{t-i} + \varepsilon_t \\ \varepsilon_t = \sigma_t e_t, \varepsilon_t | I_{t-1} \sim GED(0, \sigma_t^2, v) \\ \sigma_t^2 = \omega + \sum_{h=1}^{p} \alpha_h e_{t-h}^2 + \sum_{k=1}^{q} \beta_k \sigma_{t-k}^2 \end{cases} \tag{1}$$

where ϕ_0 is the unconditional mean of the return series, and ϕ_i is and autoregressive parameter, m denotes the lag order, and error item ε_t follows the GED process with freedom v and conditional variance σ_t^2. σ_t^2 is expressed by the GARCH model, in which e_{t-h}^2 denotes the ARCH component, and σ_{t-k}^2 denotes the GARCH component. The following restrictions: (1) $\omega > 0, \alpha_h \geq 0, \beta_h \geq 0$ and (2) $\sum_{h=1}^{p} \alpha_h + \sum_{k=1}^{q} \beta_k < 1$ need to be met to ensure a stationary GARCH process. Following the GED, the conditional probability density function of ε_t is given as:

$$f(x, v) = \frac{v e^{-\frac{1}{2}|\frac{x}{\lambda}|^v}}{2^{-\frac{2}{v}} \lambda \Gamma\left(\frac{1}{v}\right)} \tag{2}$$

in which λ is the tail-thickness parameter defined as:

$$\lambda = \left[2^{-\frac{2}{v}} \Gamma\left(\frac{1}{v}\right) \Gamma\left(\frac{3}{v}\right)\right]^{\frac{1}{2}} \tag{3}$$

where $\Gamma(\cdot)$ is the Gamma function. The freedom parameter $v > 2$ when GED follows a thick-tailed distribution; the $v > 2$ when GED follows a thin-tailed distribution; and $v = 2$ when GED follows a normal distribution. In general, the volatility clustering in financial returns series can be effectively described by the GARCH family models with the lag order of 1 [47].

3.2. Markov-Switching Mixed-Clayton Copula Function

The copula model is a connecting function for multivariate marginal distributions defined in $[0, 1]^n$. For example, a bivariate joint distribution function with the marginal distributions of $F_X(x)$ and $F_Y(y)$ can be defined as:

$$F_{XY}(x, y) = C(F_X(x), F_Y(y)) \tag{4}$$

If the marginal distributions $F_X(x)$ and $F_Y(y)$ are continuous and their joint distribution function is given, the corresponding copula model $C(u,v)$ with $u = F_X(x)$ and $v = F_Y(y)$ is uniquely determined as:

$$C(u,v) = H\Big(F^{-1}(u), F^{-1}(v)\Big) \tag{5}$$

Moreover, the joint density function can be obtained by

$$f_{XY}(x,y) = c(u,v) f_X(x) f_Y(y) \tag{6}$$

where $c(u,v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}$ is the copula density function, and $f_X(x)$ and $f_Y(y)$ are the marginal densities of variables x and y. Therefore, a distribution function with N variables is composed of N univariate marginal distributions and a copula function capturing the dependence structure between the distributions.

The copula theory and method provide a flexible perspective to measure the tail dependence. To further analyze the asymmetric risk correlations, we build a mixed-Clayton (M-Clayton) copula by combining four basic Clayton copulas with 0°, 90°, 180°, and 270° rotation, respectively, under non-fixed weights. Among the rotated Clayton copulas, the Clayton copula and 180° rotated Clayton copula are used to measure the positive dependence reflected by the lower–lower tail and higher–higher tail correlation, while the 90° and 270° rotated Clayton copulas are used to measure the negative dependence reflected by the lower–upper tail and upper–lower tail correlation. The two copulas are defined as:

$$C_1(u,v,\alpha_1,\alpha_3) = 0.5C_{c0}(u,v;\alpha_1) + 0.5C_{c180}(u,v;\alpha_3) \tag{7}$$

$$C_2(u,v;\alpha_2,\alpha_4) = 0.5C_{c90}(u,v;\alpha_2) + 0.5C_{c270}(u,v;\alpha_4) \tag{8}$$

where

$$\begin{cases} C_{c0}(u,v;\alpha_1) = \left(u^{-\alpha_1} + v^{-\alpha_1} - 1\right)^{-\frac{1}{\alpha_1}} \\ C_{c90}(u,v;\alpha_2) = u - \left[u^{-\alpha_2} + (1-v)^{-\alpha_2} - 1\right]^{-\frac{1}{\alpha_2}} \\ C_{c180}(u,v;\alpha_3) = u + v - 1 + \left[(1-u)^{-\alpha_3} + (1-v)^{-\alpha_3} - 1\right]^{-\frac{1}{\alpha_3}} \\ C_{c270}(u,v;\alpha_4) = v - \left[(1-u)^{-\alpha_4} + v^{-\alpha_4} - 1\right]^{-\frac{1}{\alpha_4}} \end{cases} \tag{9}$$

Thus, a M-Clayton copula can be obtained by weighting C_1 and C_2 copulas as:

$$C_M(u,v,\theta) = \omega C_1(u,v;\alpha_1,\alpha_3) + (1-\omega)C_2(u,v;\alpha_2,\alpha_4) \tag{10}$$

where the $\theta = (\alpha_1,\alpha_2,\alpha_3,\alpha_4) \in (0,+\infty)$, denoting the parameters of the four separate copulas, the greater the $\alpha_1,\alpha_2,\alpha_3,\alpha_4$, the stronger the correlation. $\omega \in [0,1]$ is used to determine the weights of the C_1 and C_2.

Affected by time-varying fundamental information, the correlation and possible structural changes between financial markets is usually not static. Therefore, a state-switching mechanism assumed to be subject to Markov chain is introduced to further capture the dynamic correlation and potential dependence structural changes. We assume that there are two dependence states between financial markets [48], and the Ms-M-Clayton copula can be expressed as:

$$C_{Ms-M}\left(u,v;\theta^{S_t}\right) = \omega^{S_t} C_1\left(u,v;\alpha_1^{S_t},\alpha_3^{S_t}\right) + \left(1-\omega^{S_t}\right) C_2\left(u,v;\alpha_2^{S_t},\alpha_4^{S_t}\right) \tag{11}$$

where S_t denotes the state variable and is assumed as the following Markov transition probability [48]:

$$\begin{cases} P_{11} = P(S_t = 1|S_{t-1} = 1) = \frac{\exp(\pi_1)}{1+\exp(\pi_1)} \\ P_{12} = P(S_t = 2|S_{t-1} = 1) = \frac{1}{1+\exp(\pi_1)} \\ P_{21} = P(S_t = 1|S_{t-1} = 2) = \frac{1}{1+\exp(\pi_2)} \\ P_{22} = P(S_t = 2|S_{t-1} = 2) = \frac{\exp(\pi_2)}{1+\exp(\pi_2)} \end{cases} \tag{12}$$

3.3. Markov-Switching Mixed-Clayton Copula Function

In the Ms-M-Clayton copula, the correlations of lower–lower tail, lower–higher tail, higher–higher tail, and higher–lower tail are provided as follows [49]:

$$\begin{cases} \lambda_{MS-M}^{LL} = \lim\limits_{\alpha \to 0} \mathrm{P}(V \le \alpha | U \le \alpha) = 0.5\omega^{s_t} 2^{-\frac{1}{\alpha_1}} \\ \lambda_{MS-M}^{LU} = \lim\limits_{\alpha \to 0} \mathrm{P}(V \ge 1-\alpha | U \le \alpha) = 0.5(1-\omega^{s_t}) 2^{-\frac{1}{\alpha_2}} \\ \lambda_{MS-M}^{UU} = \lim\limits_{\alpha \to 1} \mathrm{P}(V \ge \alpha | U \ge \alpha) = 0.5\omega^{s_t} 2^{-\frac{1}{\alpha_3}} \\ \lambda_{MS-M}^{UL} = \lim\limits_{\alpha \to 1} \mathrm{P}(V \le 1-\alpha | U \ge \alpha) = 0.5(1-\omega^{s_t}) 2^{-\frac{1}{\alpha_4}} \end{cases} \tag{13}$$

3.4. Parameter Estimation Method

We employ the maximum-likelihood (ML) function [50] as the basis for estimating parameters. Given that there are 12 parameters to be estimated, and a traditional approach, such as the interior-point method, easily falls into local optimum, we apply the genetic algorithm (GA) that performs well in global optimization of high-dimensional parameters to exact the solution of the model [51].

Referring to Equation (6), the joint probability density function of the Ms-M-Clayton copula model with variables x and y is given as:

$$f_{XY}(x,y) = \sum_{S_t=1}^{2} f_X(x) f_Y(y) c\left(u, v, \theta^{S_t}\right) P(S_t) \tag{14}$$

where $P(S_t)$ is the prediction probability of S_t at time $t-1$. $P(S_t = 1)$ and $P(S_t = 2)$ are defined as [52]:

$$P(S_t = 1) = P_{11} * \left[\frac{c_{t-1}^1 P(S_{t-1} = 1)}{c_{t-1}^1 P(S_{t-1} = 1) + c_{t-1}^2 P(S_{t-1} = 2)} \right] + P_{21} * \left[\frac{c_{t-1}^2 P(S_{t-1} = 2)}{c_{t-1}^1 P(S_{t-1} = 1) + c_{t-1}^2 P(S_{t-1} = 2)} \right] \tag{15}$$

$$P(S_t = 2) = 1 - P(S_t = 1) \tag{16}$$

where c_{t-1}^1 and c_{t-1}^2 represent the conditional probability density functions of the copula function in state 1 and state 2, respectively, at time $t-1$. Then the logarithmic likelihood function of the copula model is expressed as:

$$\ln L = \sum_{t=1}^{T} \ln c\left(u, v; \theta^{S_t}\right) P(S_t) + \sum_{t=1}^{T} \ln f_X(x) + \sum_{t=1}^{T} \ln f_Y(y) \tag{17}$$

3.5. VaR and CoVaR

This work employs the value-at-risk (VaR) to measure the downside and upside risks, which indicates the maximum loss that an investor may suffer within a certain time horizon and significant level by holding a long or a short position. For return series r_t, we calculate the VaR based on its marginal distribution. With a given tail probability α, the $VaR_D^{\alpha,t}$

and $VaR_U^{\alpha,t}$ at time t is calculated by $P\left(r_t \le VaR_D^{\alpha,t}\right) = \alpha$ and $P\left(r_t \ge VaR_U^{\alpha,t}\right) = 1-\alpha$ respectively, which is formulated as:

$$\begin{cases} VaR_D^{\alpha,t} = \mu_t + \sigma_t \cdot F_v^{-1}(\alpha) \\ VaR_U^{\alpha,t} = \mu_t + \sigma_t \cdot F_v^{-1}(1-\alpha) \end{cases} \tag{18}$$

where μ_t and σ_t represent the conditional mean and standard deviation determined by the marginal distribution model, and $F_v^{-1}(\alpha)$ is the α-quantile of GED.

The conditional VaR (CoVaR) is used to capture the risk spillover between markets [42]. The CoVaR is calculated based on the measurement of copula model, reflecting the VaR of a market conditional on the extreme volatility in another market. Let r_t^i and r_t^j denote the return series of market i and j, and the CoVaR in four different market statuses can be expressed as follows:

$$\begin{cases} P\left(r_t^i \le CoVaR_{iD|jD}^{\beta,t} \middle| r_t^j \le VaR_{jD}^{\alpha,t}\right) = \beta \\ P\left(r_t^i \ge CoVaR_{iU|jD}^{\beta,t} \middle| r_t^j \le VaR_{jD}^{\alpha,t}\right) = \beta \\ P\left(r_t^i \le CoVaR_{iD|jU}^{\beta,t} \middle| r_t^j \ge VaR_{jU}^{\alpha,t}\right) = \beta \\ P\left(r_t^i \ge CoVaR_{iU|jU}^{\beta,t} \middle| r_t^j \ge VaR_{jU}^{\alpha,t}\right) = \beta \end{cases} \tag{19}$$

where $CoVaR_{iD|jD}^{\beta,t}$ and $CoVaR_{iU|jD}^{\beta,t}$ represent the downside and upside VaRs of market i conditional on the extreme downside movement of market j given a confidence level β, while $CoVaR_{iD|jU}^{\beta,t}$ and $CoVaR_{iU|jU}^{\beta,t}$, respectively, represent downside and upside VaR of market i conditional on the extreme upside movement of market j given a confidence level β.

For example, the first row in Equation (19) can be written as:

$$\frac{F_{r_t^i r_t^j}\left(CoVaR_{iD|jD}^{\beta,t}, VaR_{jD}^{\alpha,t}\right)}{F_{r_t^j}\left(VaR_{jD}^{\alpha,t}\right)} = \beta \tag{20}$$

Therefore, the CoVaR requires the joint distribution function of r_t^i and r_t^j, and it can be represented by a copula function as Equation (4). Thus, Equation (19) can be written as:

$$\begin{cases} C\left(F_{r_t^i}\left(CoVaR_{iD|jD}^{\beta,t}\right), \alpha\right) = \alpha\beta \\ C\left(F_{r_t^i}\left(CoVaR_{iU|jD}^{\beta,t}\right), \alpha\right) = \alpha - \alpha\beta \\ F_{r_t^i}\left(CoVaR_{iD|jU}^{\beta,t}\right) - C\left(F_{r_t^i}\left(CoVaR_{iD|jU}^{\beta,t}\right), 1-\alpha\right) = \alpha\beta \\ F_{r_t^i}\left(CoVaR_{iU|jU}^{\beta,t}\right) - C\left(F_{r_t^i}\left(CoVaR_{iU|jU}^{\beta,t}\right), 1-\alpha\right) = \alpha - \alpha\beta \end{cases} \tag{21}$$

Hence, the value of $F_{r_t^i}\left(CoVaR_{iD|jD}^{\beta,t}\right)$ can be inferred by inverting the copula function for given values of α and β, which is denoted as $\hat{F}_{r_t^i}\left(CoVaR_{iD|jD}^{\beta,t}\right)$, and the value of CoVaR can be inferred by inverting the marginal distribution function of r_t^i as $CoVaR_{iD|jD}^{\beta,t} = F_{r_t^i}^{-1}\left(\hat{F}_{r_t^i}\left(CoVaR_{iD|jD}^{\beta,t}\right)\right)$. Similarly, the other three types of CoVaR can be obtained. To validate the significance of the risk contagion, the Kolmogorov–Smirnov (K-S) test [20] is employed to implement the significance test.

4. Data and Descriptive Statistics

This work adopts the China Securities Index 300 (CSI300), an important financial index jointly released by the Shanghai and Shenzhen Stock Exchanges on 8 April 2005 to represent

the Chinese stock market. It consists of 300 stocks, accounting for approximately 70% of the total market capitalization of the Shanghai and Shenzhen stock markets. Compared with other stock indexes in China, the issuers of the constituent stocks in CSI300 are mostly mature companies that have the characteristics of strong resistance to manipulation, lower volatility, and strong liquidity. Therefore, it comprehensively reflects the performance of the Chinese stock market. According to [19], three risk areas, including Asia–Oceania, Europe, and the Americas, can be identified in risk contagion. Therefore, the S&P500 and GSPTSE indexes are selected to represent the Americas market, the DAX30 and FTSE100 indexes are selected to represent the European market, and the Nikkei225 and ASX200 indexes are selected to represent the Asia–Oceania market. The monthly price time series collected from Wind database are used for empirical analyses because: (1) it covers less noises than the daily and weekly prices and is widely employed in copula modeling, and (2) it contains more trend information than the yearly prices but does not suffer from manipulation [14,20,48]. The period is from July 2005, when CSI300 is officially released, to December 2020, with 186 data points containing multiple economic cycles and economic events. The logarithmic returns series r_t reflecting the level of price changes are calculated as: $r_t = (lnP_t - lnP_{t-1}) \times 100\%$, where P_t denotes the price at the end of month t.

Figure 1 reports the prices and returns of the selected stock indexes. First, the stock market volatility in the same region is relatively similar, but those in different regions are quite different. Second, due to the global emergencies during the sample period, such as the global financial crisis, the European debt crisis, and the COVID-19 epidemic, the markets experienced several large fluctuations simultaneously, implying the potential risk contagion between Chinese and mature markets. Third, the volatility of Chinese market is significantly higher than mature markets, which may be caused by the large gap between Chinese and mature stock markets in terms of the completeness of risk supervision and the professionalism of market participants.

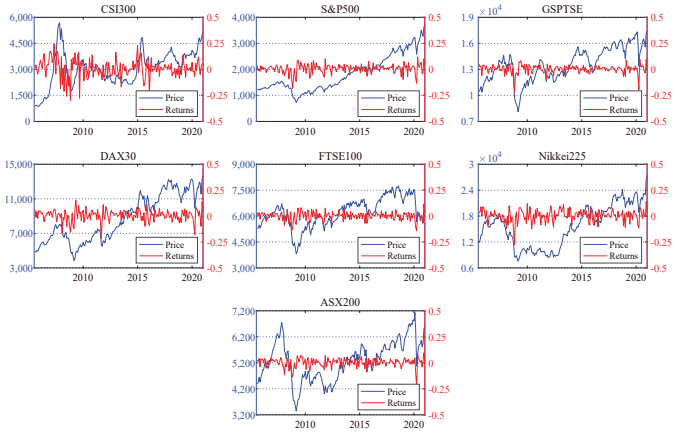


Figure 1. Monthly prices and returns of the selected indexes.

Table 1 reports the descriptive statistics of the return series, in which their average values are all positive. The CSI300 has the highest monthly average return with 0.0096, followed by the S&P500 and the DAX30, while the FTSE100 has the lowest monthly average return. The CSI300 has the highest volatility, with the standard deviation of 0.0858, followed by the Nikkei225. The lowest standard deviation 0.0402 is observed in the FTSE100. Moreover, the skewness statistics are all less than 0, suggesting that all the return series are featured as a long tail to the left, and there are more extreme negative returns. The skewness values of the Nikkei225 and the ASX200 are larger than others, and that of the CSI300 is closer to 0. Meanwhile, the Nikkei225 and the ASX200 have the highest kurtosis, implying the leptokurtosis feature in Asia–Oceania market is more prominent.

The Jarque-Bera (J-B) test confirms that all return series are not normally distributed but featured as leptokurtosis. The Pearson correlation coefficients between CSI300 and other indexes proves a weak but positive correlation between Chinese and mature markets, and the correlations between Chinese market and the Americas, Asia–Oceania, and the European markets decreases in turn.

Table 1. Descriptive statistics of monthly returns.

	CSI300	S&P500	GSPTSE	DAX30	FTSE100	Nikkei225	ASX200
Mean	0.0096	0.0062	0.0030	0.0059	0.0013	0.0046	0.0023
Max.	0.2463	0.1194	0.0997	0.1550	0.1155	0.1401	0.0949
Min.	−0.2991	−0.1856	−0.2168	−0.2131	−0.1413	−0.2722	−0.2380
Std.	0.0858	0.0436	0.0404	0.0543	0.0402	0.0570	0.0428
Skew.	−0.524	−0.888	−1.649	−0.809	−0.668	−0.896	−1.470
Kurt.	4.691	5.245	9.930	5.010	4.236	5.347	7.963
J-B.	30.685 [a]	63.540 [a]	456.478 [a]	51.620 [a]	25.654 [a]	67.595 [a]	257.900 [a]
Pearson.	1.000	0.402 [a]	0.403 [a]	0.374 [a]	0.307 [a]	0.361 [a]	0.380 [a]

Note: superscript a represent the significant levels at 1%.

5. Empirical Results

This study uses Eviews 9 to perform a marginal distribution estimation and output the residual series and MATLAB 2018 to fit copula models.

5.1. Marginal Distribution Estimation

A diagnostic test on stationarity, autocorrelation, and heteroscedasticity needs to be conducted before marginal distribution modeling. The results are reported in Table A1 (seen in Appendix A), showing that all the return series are stationary by ADF, PP, and KPSS tests. According to the Ljung–Box test, only the CSI300 have autocorrelation. The Q2(P) and ARCH(P) statistics ensure the presence of ARCH effects in all series except the DAX30. Thus, AR-GARCH is suitable to fit the marginal distribution.

Considering the significance of parameters and the results of diagnostic test, the results of marginal distribution are provided in Panel A of Table A2 (see Appendix A). Most coefficients are significant at 5% level. Panel B of Table A2 reports the diagnostic results for the residuals, in which the autocorrelation and conditional heteroscedasticity are effectively overcome. Then, the standard residues are employed to conduct the risk dependence analyses with copula models.

5.2. Dynamic and Asymmetric Dependence Measured by MS-M-Clayton Copula

The M-Clayton copula model is first employed to measure both positive and negative dependence structures (Wang et al., 2013; Ji et al., 2018), and Table 2 reports the results, in which all parameters are significant at the 1% level. It is worth noting that the weight parameter ω across different pairwise returns is various, indicating that the existence of negative dependence between Chinese and mature stock markets. Therefore, how to recognize the occurrence of different risk dependence structures and correlations has become an urgent problem to be clarified.

Table 3 further reports the estimated results of the MS-M-Clayton copula model, where the model outperforms the invariant M-Clayton copula in terms of the logarithmic likelihood values. Most of the estimated parameters are significant at the 10% level, meaning that there are not only both positive and negative dependence structures but also dependence-switching between Chinese and mature stock markets. Overall, the risk dependence structures and correlations are different in each dependence state. Taking the CSI300-S&P500 as an example, the P_{22} of 0.864 is significant and higher than P_{11}, meaning that state 2 is the dominant dependence structure. Similarly, for the CSI300-GSPTSE, CSI300-FTSE100, and CSI300-Nikkei225 pairs, state 2 plays a dominant role, while state 1 is dominant in CSI300-DAX30 and CSI300-ASX200 pairs.

Table 2. M-Clayton copula estimates of CSI300 with mature stock indexes.

	CSI300-S&P500	CSI300-GSPTSE	CSI300-DAX30	CSI300-FTSE100	CSI300-Nikkei225	CSI300-ASX200
α_1	0.682 [a]	2.194 [a]	0.889 [a]	0.892 [a]	0.488 [a]	0.969 [a]
α_2	2.68×10^{-7} [a]	8.67×10^{-8} [a]	1.69×10^{-7} [a]	3.24×10^{-8} [a]	49.415 [a]	5.23×10^{-7} [a]
α_3	0.674 [a]	0.513 [a]	2.587 [a]	7.08×10^{-9} [a]	0.506 [a]	4.705 [a]
α_4	3.0967 [a]	5.58×10^{-8} [a]	1.18×10^{-7} [a]	9.049 [a]	5.946 [a]	2.45×10^{-7} [a]
ω	0.924 [a]	0.594 [a]	0.571 [a]	0.923 [a]	0.947 [a]	0.443 [a]
Log-L	−13.473	−8.808	−10.369	−10.453	−10.075	−9.776

Note: superscript a represent the significant levels at 1%.

Table 3. MS-M-Clayton copula estimates of CSI300 with mature stock indexes.

Copula	CSI300-S&P500	CSI300-GSPTSE	CSI300-DAX30	CSI300-FTSE100	CSI300-Nikkei225	CSI300-ASX200
α_1^{S1}	3.176	0.262 [a]	0.652 [a]	1.622	0.527 [a]	0.243 [a]
α_2^{S1}	1.82×10^{-10} [a]	9.12×10^{-8} [a]	25.801 [a]	4.11×10^{-9} [a]	54.908 [a]	2.68×10^{-8} [a]
α_3^{S1}	0.601 [c]	5.594 [a]	0.998 [a]	3.278 [a]	2.66×10^{-9} [a]	0.141
α_4^{S1}	3.043	8.28×10^{-8} [a]	3.95×10^{-9} [a]	1.70×10^{-9} [a]	3.62×10^{-9} [a]	1.770
α_1^{S2}	0.584 [a]	1.637 [a]	0.275	0.653 [a]	0.461 [a]	1.627
α_2^{S2}	1.95×10^{-10} [a]	3.23×10^{-8} [a]	1.11×10^{-10} [a]	3.28×10^{-9} [a]	3.24×10^{-10} [a]	0.094
α_3^{S2}	0.616 [a]	1.39×10^{-7} [a]	20.078 [a]	1.93×10^{-10} [a]	0.776 [a]	5.363 [a]
α_4^{S2}	3.077	8.03×10^{-9} [a]	5.89×10^{-10} [a]	6.832 [a]	6.844 [a]	1.37×10^{-8} [a]
ω^{S1}	0.656 [b]	0.984 [a]	0.978 [a]	0.827 [a]	0.676 [a]	1.000 [a]
ω^{S2}	0.999 [a]	0.676 [a]	0.415 [c]	0.817 [a]	0.971 [a]	0.932 [c]
P_{11}	0.517	0.943 [a]	0.846 [a]	0.941 [a]	0.943 [a]	0.881 [a]
P_{22}	0.864 [a]	0.974 [a]	0.710 [a]	0.971 [a]	0.993 [a]	0.727 [a]
Log-L	−13.626	−11.276	−12.487	−12.369	−11.522	−11.777

Note: superscript a, b, and c represent the significant levels at 1%, 5%, and 10%, respectively.

Table 4 reports the tail correlation coefficients based on the constructed copula. Specifically, the values of λ_{UU} are larger than that of λ_{LL} between CSI300 and S&P500, DAX30, and Nikkei225, meaning that the upside risk correlation triggered by positive factors is stronger than the downside risk correlation triggered by negative factors, while the opposite relationship occurs between CSI300 and GSPTSE, FTSE100, and ASX200. Moreover, compared with the Americas and European mature markets, the downside risk correlation between Chinese and Asia–Oceania markets manifesting in synchronized decline is the lowest, which is usually paid special attention in practice. Although the negative dependence is not in dominant in the dominant state, it is still asymmetric. Specifically, the upper–lower tail correlation between CSI300 and S&P500, FTSE100, and Nikkei225 is stronger than the lower–upper tail correlation, indicating the probability of extreme rises in Chinese market when extreme declines occur in the three mature markets. The opposite situation can be found between CSI300 and DAX30. As for the main dependence state between CSI300 and GSPTSE and ASX200 returns, the negative dependence correlation is not observed. Therefore, during the period of smooth economic operation denoted by the main state, except for monitoring the positive risk spillover, Chinese investors and managers should pay close attention to investment opportunities in the declines of S&P500, FTSE100, and Nikkei225 while managing exposure carefully in the rises of DAX30.

Table 4. Tail correlation coefficients between CSI300 and mature stock indexes.

	State 1				State 2			
	λ_{LL}	λ_{LU}	λ_{UU}	λ_{UL}	λ_{LL}	λ_{LU}	λ_{UU}	λ_{UL}
CSI300-S&P500	0.264	0.000	0.103	0.137	0.152	0.000	0.162	0.001
CSI300-GSPTSE	0.035	0.000	0.435	0.000	0.221	0.000	0.000	0.000
CSI300-DAX30	0.169	0.011	0.244	0.000	0.017	0.000	0.200	0.000
CSI300-FTSE100	0.270	0.000	0.334	0.000	0.141	0.000	0.000	0.083
CSI300-Nikkei225	0.091	0.160	0.000	0.000	0.108	0.000	0.199	0.013
CSI300-ASX200	0.029	0.000	0.004	0.000	0.304	0.000	0.410	0.000

Figure 2 provides the trajectories of P^{S_1} and P^{S_2}, in which the state transitions are observed in the risk dependence between Chinese and most mature markets. For CSI300-S&P500, there is no state-switching, and state 2 is dominant during the entire sample period, implying the stable dependence and risk correlation between the two markets. For CSI300-GSPTSE, the state transitions occur concentrated in the periods from 2013 to 2015, corresponding to cyclical financial market bubbles and the post-COVID-19 [3], in which the secondary state should be paid more attention because more investment opportunities appear with a stronger upside tail correlation and a downside tail correlation close to 0. The state transitions of CSI300-DAX30 appear periodically around 2009 (may be affected by European debt crisis) with weak persistence [53]. In the secondary state, the upside tail correlation is significant, while the downside correlation decreases to near 0, increasing the investment motivation. For CSI300-FTSE100, state 2 with apparent downside risk correlation is dominant in most of the period. However, state 1 with both upside and downside risk correlations switches to be the main dependence structure temporarily around 2009 (European debt crisis) and since the COVID-19 epidemic [53]. For CSI300-Nikkei225, state 1 with reversal correlation was the main state before 2009 and in 2012, corresponding to the global financial crisis and the Asian financial turmoil led by the exchange rate system, respectively [54]. However, state 2 with positive dependence structure plays a dominant role in most of the period, especially in recent years. For CSI300-ASX200, state 1 with a relatively low tail correlation is dominant. The state-switching process occurs around 2012 and 2015 temporarily, which is accompanied by an increase in positive risk correlation caused by regional financial turmoil [54]. Moreover, in the comparison between markets in different regions, the Asia–Oceania markets have the relatively low risk association, especially the downside risk correlation that is paid much attention in practice, with the Chinese market.

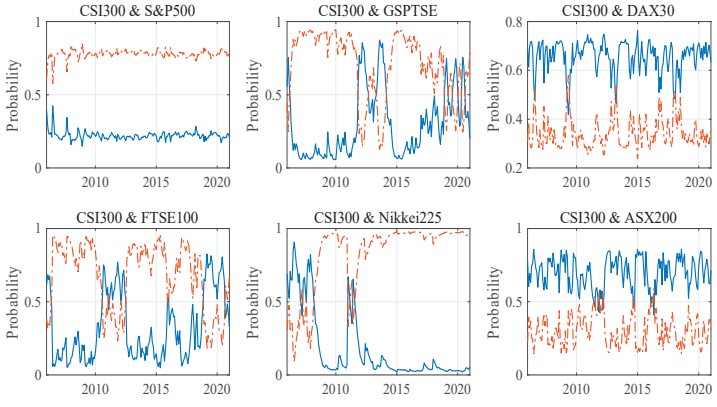


Figure 2. State transition probabilities between Chinese and mature markets (The blue line represents state 1, and the orange line represents state 2).

5.3. Comparative Analysis

5.3.1. Static Dependence Measured by Invariant Copula Models

To explain the similarity and differences between our findings and previous research, we first employ seven commonly used invariant copulas, including the Gaussian, Student's t, Gumbel, 180° rotated Gumbel, Clayton, 180° rotated Clayton, and SJC copulas [53] to measure the risk dependence between Chinese and mature markets. The estimated results of invariant copulas are reported in Table 5.

Table 5. Invariant copula estimates of the CSI300 with mature stock indexes.

Copula	CSI300-SP500	CSI300-GSPTSE	CSI300-DAX30	CSI300-FTSE100	CSI300-Nikkei225	CSI300-ASX200
Gaussian						
ρ	0.372 [a]	0.348 [a]	0.361 [a]	0.266 [a]	0.314 [a]	0.349 [a]
Log-L	−12.496	−10.781	−11.660	−6.077	−8.637	−10.853
Student's t						
ρ	0.372 [a]	0.347 [a]	0.375 [a]	0.285 [a]	0.314 [a]	0.356 [a]
v	99.899 [a]	99.983 [a]	7.966 [a]	7.527 [a]	99.320 [a]	8.910 [a]
Log-L	−12.474	−10.596	−12.438	−7.069	−8.634	−11.710
Gumbel						
δ	1.223 [a]	1.183 [a]	1.260 [a]	1.141 [a]	1.183 [a]	1.228 [a]
Log-L	−6.541	−4.251	−8.307	−2.477	−4.729	−6.603
180° rotated Gumbel						
δ	1.326 [a]	1.281 [a]	1.329 [a]	1.259 [a]	1.256 [a]	1.316 [a]
Log-L	−15.498	−11.794	−14.612	−10.533	−10.484	−14.249
Clayton						
ρ	0.630 [a]	0.543 [a]	0.594 [a]	0.523 [b]	0.505 [a]	0.593 [a]
Log-L	−16.645	−13.235	−14.419	−12.023	−12.126	−14.634
180° rotated Clayton						
ρ	0.307 [c]	0.263	0.355 [c]	0.130	0.237	0.310 [c]
Log-L	−4.233	−3.046	−5.244	−0.671	−2.478	−4.320
SJC						
λ_U	2.83×10^{-7}	4.77×10^{-7}	5.57×10^{-8}	1.85×10^{-7}	4.21×10^{-7}	4.96×10^{-7}
λ_L	0.380	0.404	0.366 [a]	0.346	0.320	0.354
Log-L	−16.508	−11.868	−14.474	−12.044	−11.464	−14.735

Note: superscript a, b, and c represent the significant levels at 1%, 5%, and 10%, respectively.

According to the logarithmic likelihood values, it is found that the Clayton copula performs the best with significant estimated parameters, followed by the 180° rotated Gumbel copula, the Student's t copula, and the Gaussian copula, successively, and the Gumbel copula and 180° rotated Clayton copula perform the worst. In the SJC copula measuring asymmetric positive dependence, the lower tail correlations are larger than the upper ones, but most parameters are not significant. The results suggest a positive but asymmetric risk dependence between Chinese and mature markets, and the downside correlation is stronger than the upside correlation. Overall, the results are generally consistent with the findings drawn from M-Clayton and MS-M-Clayton copulas but fail to capture the negative dependence structure and the upside correlations between CSI300 and S&P500, DAX30, and Nikkei225 effectively. Moreover, the static copulas are unable to capture the time-varying or dependence-switching characteristics of the correlations.

5.3.2. Dynamic Dependence Measured by Time-Varying Parameter Copula

To assess the dynamic risk dependence correlation between Chinese and mature markets, Table 6 further reports the estimated results of four TVP copulas, in which most of the estimated parameters are significant at the 10% level. It can be found that TVP copulas perform better than the corresponding invariant copulas. Specifically, the TVP-180° rotated Gumbel copula describing the lower–lower tail correlation effectively captures the risk dependence between Chinese and mature markets, and the TVP-SJC copula also proves

that the lower–lower correlation is more significant. The results confirm the positive risk dependence structure and the prominent downside risk correlation between Chinese and mature markets. The effectiveness of time-varying mechanism in depicting the dynamic risk correlation is also verified. Although TVP copulas provide an analytical view on dynamic risk correlation, a significant difference between them and the proposed MS-M-Clayton copula is that the potential negative dependence structure is not effectively depicted.

Table 6. TVP copula estimates of the CSI300 with mature stock indexes.

Copula	CSI300-S&P500	CSI300-GSPTSE	CSI300-DAX30	CSI300-FTSE100	CSI300-Nikkei225	CSI300-ASX200
TVP-Gaussian						
ψ_0	0.255 [a]	1.162 [a]	0.091 [a]	0.711 [a]	0.321 [a]	1.601 [a]
ψ_1	0.270 [a]	0.258 [a]	−0.169 [a]	0.023 [a]	−0.049 [a]	−0.739 [a]
ψ_2	1.227 [a]	−1.423 [a]	2.042 [a]	−0.635 [a]	1.117 [a]	−1.762 [a]
Log-L	−13.283	−10.836	−13.771	−6.078	−8.661	−11.657
TVP-180° Rotated Gumbel						
ω_L	2.435 [a]	1.144 [a]	2.800 [a]	1.557 [a]	0.993 [a]	−0.429 [a]
α_L	−0.815 [a]	−0.311 [a]	−0.768 [a]	−0.557 [a]	−0.324 [a]	0.683 [a]
β_L	−2.851 [a]	−0.766 [a]	−5.030 [a]	−1.291 [a]	−0.263 [a]	0.340 [a]
Log-L	−18.779	−11.955	−17.436	−10.738	−10.512	−14.862
TVP-Gumbel						
ω_U	2.338 [a]	2.903 [a]	3.366 [a]	3.106 [a]	−0.654 [a]	−0.608 [a]
α_U	−0.916 [a]	−0.744 [a]	−1.018 [a]	−1.363 [a]	0.942 [a]	0.620 [a]
β_U	−2.541 [a]	−6.164 [a]	−6.821 [a]	−4.265 [a]	−0.110 [a]	1.100 [a]
Log-L	−9.867	−9.881	−15.758	−7.732	−5.015	−8.784
TVP-SJC						
ω_U	−14.830 [a]	−14.363 [a]	−15.343 [a]	−15.242 [a]	−14.593 [a]	−14.490 [a]
α_U	−0.012 [a]	−0.002 [b]	-8.391×10^{-4} [c]	−0.002 [a]	−0.002 [a]	-5.799×10^{-4} [b]
β_U	−0.003 [a]	7.327×10^{-5}	4.025×10^{-6}	-1.465×10^{-6}	-1.469×10^{-5}	-1.642×10^{-4}
ω_L	2.792 [a]	0.459 [a]	5.150 [a]	4.447 [a]	−0.181 [a]	−2.235 [a]
α_L	−5.960 [a]	−2.988 [a]	−18.110 [a]	−15.809 [a]	−1.477 [a]	1.071 [a]
β_L	−4.505 [a]	−1.209 [a]	−4.230 [a]	−4.203 [a]	−0.858 [a]	3.771 [a]
Log-L	−19.595	−12.379	−17.450	−12.804	−11.371	−14.976

Note: superscript a, b, and c represent the significant levels at 1%, 5%, and 10%, respectively.

5.4. Asymmetric Risk Spillover Measurement by VaR, CoVaR and Nomalized CoVaR

To provide implications for risk supervision and portfolio risk management, we studied the extreme risk spillovers between Chinese and mature stock markets in different routes by VaR and CoVaR based on the information from marginal distribution and Ms-M-Clayton copula model. We set α and β equal to 0.05 for downside CoVaR and 0.95 for the upside CoVaR calculation. Table 7 reports the summary statistics of the VaR and the CoVaR, and Figure 3 shows the dynamic trajectories for intuitive observation.

For stock index pairs except CSI300-FTSE100, the absolute values of upside VaR and CoVaR are larger than those of the downside, respectively, meaning that the upside risk is larger than the downside risk in Chinese market. Moreover, the VaR and CoVaR show phased extreme fluctuations, which may be related to the macroeconomic uncertainties, such as the periods around 2008, 2013, and 2015. For the positive risk contagion (3 and 6 rows in Table 7), the absolute values of CoVaR are all greater than that of VaR when measuring either upside or downside risks, indicating the synergistic risk spillover from mature markets to the Chinese market. In the measurement of negative risk contagion (4–5 rows in Table 7), the absolute values of CoVaR are generally smaller than that of VaR, implying the weak existence of reverse risk spillovers. Overall, the positive risk contagion from mature markets to the Chinese market are more significant than the negative contagion. It is noteworthy that the downside risk contagion between Chinese and Asia–Oceania markets is relatively weak, suggesting that the Asia–Oceania market can be considered as a potential choice for investors in the Chinese market to diversify their investment portfolios.

Table 8 further reports the hypothesis testing results by K-S test, and the statistics are generally significant at 10% level, rejecting the null hypothesis that VaR is equal to CoVaR.

Table 7. Summary statistics of the VaR and the CoVaR (The $CoVaR^{\beta,t}_{CSI300(D)|Other(D)}$ and $CoVaR^{\beta,t}_{CSI300(U)|Other(D)}$ denote the downside and upside VaRs of the CSI300 conditional on the extreme declines of mature markets, respectively; the $CoVaR^{\beta,t}_{CSI300(D)|Other(U)}$ and $CoVaR^{\beta,t}_{CSI300(U)|Other(U)}$ denote the downside and upside VaRs of the CSI300 conditional on the extreme rises of mature markets, respectively).

	CSI300-S&P500	CSI300-GSPTSE	CSI300-DAX30	CSI300-FTSE100	CSI300-Nikkei225	CSI300-ASX200
$VaR^{\alpha,t}_{CSI300,D}$			−12.258 (4.483)			
$VaR^{\alpha,t}_{CSI300,U}$			14.720 (4.227)			
$CoVaR^{\beta,t}_{CSI300(D)\|Other(D)}$	−19.060 (6.337)	−18.840 (6.335)	−18.791 (6.271)	−19.021 (6.228)	−18.275 (6.143)	−18.369 (6.126)
$CoVaR^{\beta,t}_{CSI300(D)\|Other(U)}$	−7.773 (3.309)	−12.474 (5.122)	−8.736 (3.646)	−11.459 (4.664)	−10.170 (4.875)	−9.026 (3.706)
$CoVaR^{\beta,t}_{CSI300(U)\|Other(D)}$	13.260 (3.827)	14.060 (5.324)	10.639 (3.344)	16.491 (5.574)	12.942 (5.057)	11.247 (3.468)
$CoVaR^{\beta,t}_{CSI300(U)\|Other(U)}$	21.470 (6.056)	19.090 (4.497)	21.704 (6.116)	18.901 (4.802)	21.119 (5.988)	20.147 (5.650)

Note: this table reports the means and the standard errors (in parentheses) of VaR and CoVaR.

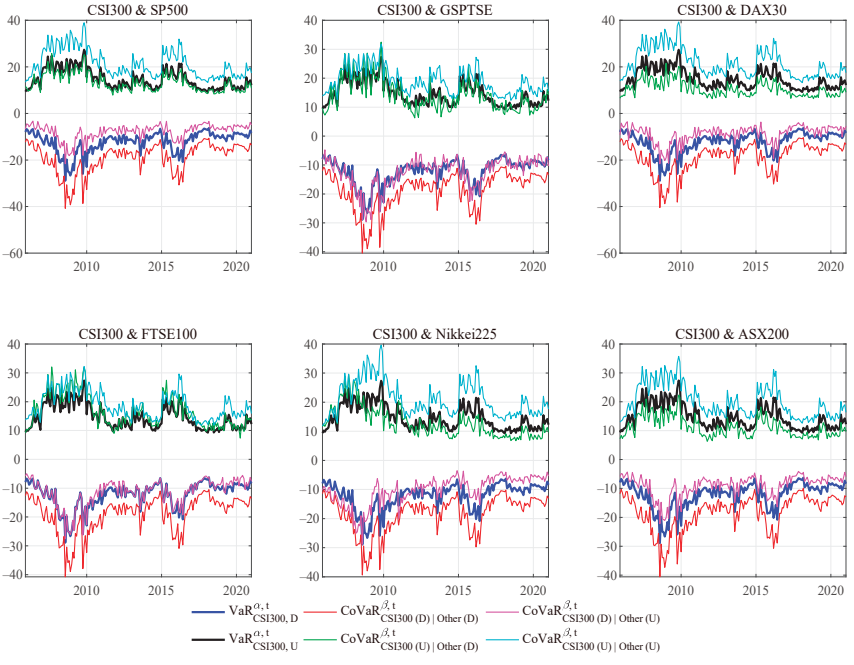


Figure 3. The dynamic trajectories of the VaR and the CoVaR.

Table 8. The hypothesis testing for equalities of CoVaR and VaR.

Null Hypotheses	CSI300-S&P500	CSI300-GSPTSE	CSI300-DAX30	CSI300-FTSE100	CSI300-Nikkei225	CSI300-ASX200
$\frac{CoVaR^{\beta,t}_{CSI300(D)\|Other(D)}}{VaR^{\alpha,t}_{CSI300,D}} =$	0.593 [a] (0.000)	0.577 [a] (0.000)	0.577 [a] (0.000)	0.582 [a] (0.000)	0.550 [a] (0.000)	0.550 [a] (0.000)
$\frac{CoVaR^{\beta,t}_{CSI300(D)\|Other(U)}}{VaR^{\alpha,t}_{CSI300,D}} =$	0.582 [a] (0.000)	0.077 (0.637)	0.456 [a] (0.000)	0.159 [b] (0.017)	0.330 [a] (0.000)	0.445 [a] (0.000)
$\frac{CoVaR^{\beta,t}_{CSI300(U)\|Other(D)}}{VaR^{\alpha,t}_{CSI300,U}} =$	0.181 [a] (0.004)	0.220 [a] (0.000)	0.478 [a] (0.000)	0.132 (0.077)	0.324 [b] (0.047)	0.412 [a] (0.000)
$\frac{CoVaR^{\beta,t}_{CSI300(U)\|Other(U)}}{VaR^{\alpha,t}_{CSI300,U}} =$	0.533 [a] (0.000)	0.456 [a] (0.000)	0.544 [a] (0.000)	0.418 [a] (0.000)	0.517 [a] (0.000)	0.473 [a] (0.000)

Note: superscript a and b represent the significant levels at 1% and 5% respectively.

To further evaluate the intensity of risk spillovers in different routes and analyze its asymmetry, Table 9 reports the summary statistics of the CoVaR normalized by VaR (CoVaR/VaR). It can be observed that the mean values of $\frac{CoVaR^{\beta,t}_{CSI300(D)|Other(D)}}{VaR^{\alpha,t}_{CSI300,D}}$ are greater than those of $\frac{CoVaR^{\beta,t}_{CSI300(D)|Other(U)}}{VaR^{\alpha,t}_{CSI300,D}}$, and the mean values of $\frac{CoVaR^{\beta,t}_{CSI300(U)|Other(U)}}{VaR^{\alpha,t}_{CSI300,U}}$ are greater than those of $\frac{CoVaR^{\beta,t}_{CSI300(U)|Other(D)}}{VaR^{\alpha,t}_{CSI300,U}}$, indicating that the positive and negative risk contagion effects are asymmetric, and the positive effect is stronger than the negative effect. Meanwhile, the mean values of $\frac{CoVaR^{\beta,t}_{CSI300(D)|Other(D)}}{VaR^{\alpha,t}_{CSI300,D}}$ are greater than those of $\frac{CoVaR^{\beta,t}_{CSI300(U)|Other(U)}}{VaR^{\alpha,t}_{CSI300,U}}$, and the mean values of $\frac{CoVaR^{\beta,t}_{CSI300(U)|Other(D)}}{VaR^{\alpha,t}_{CSI300,U}}$ are greater than those of $\frac{CoVaR^{\beta,t}_{CSI300(D)|Other(U)}}{VaR^{\alpha,t}_{CSI300,D}}$ except in CSI300-GSPTSE pairwise returns, implying the asymmetry between upside and downside risk contagion effects, and the downside effect is generally stronger, while the opposite effect is in negative contagion. The analyses are statistically supported by K-S tests (see in Tables A3 and A4 of Appendix A).

Table 9. Summary statistics of the CoVaR/VaR.

	CSI300-SP500	CSI300-GSPTSE	CSI300-DAX30	CSI300-FTSE100	CSI300-Nikkei225	CSI300-ASX200
$\frac{CoVaR^{\beta,t}_{CSI300(D)\|Other(D)}}{VaR^{\alpha,t}_{CSI300,D}}$	1.571 (0.082)	1.551 (0.082)	1.548 (0.079)	1.569 (0.080)	1.504 (0.069)	1.513 (0.080)
$\frac{CoVaR^{\beta,t}_{CSI300(D)\|Other(U)}}{VaR^{\alpha,t}_{CSI300,D}}$	0.625 (0.056)	1.014 (0.161)	0.707 (0.101)	0.926 (0.094)	0.828 (0.264)	0.729 (0.054)
$\frac{CoVaR^{\beta,t}_{CSI300(U)\|Other(D)}}{VaR^{\alpha,t}_{CSI300,U}}$	0.903 (0.056)	0.943 (0.162)	0.721 (0.063)	1.110 (0.113)	0.871 (0.185)	0.762 (0.051)
$\frac{CoVaR^{\beta,t}_{CSI300(U)\|Other(U)}}{VaR^{\alpha,t}_{CSI300,U}}$	1.461 (0.055)	1.317 (0.110)	1.478 (0.059)	1.299 (0.118)	1.441 (0.111)	1.372 (0.061)

Note: this table presents the means and the standard errors (in parentheses) of the CoVaR/VaR.

6. Conclusions

The risk contagion between Chinese and mature markets has attracted more and more attention from both scholars and market participants. In this work, we construct a novel Ms-M-Clayton copula model to identify both positive and negative dependences and revisit the risk contagion between Chinese market and six mature markets in the Americas, Europe, and Asia–Oceania. Four basic Clayton copulas with various rotations are weighted to capture different tail correlations, and a two-state transition mechanism following Markov

chain is introduced to allow the copula depicting dynamic risk correlations. Based on the estimated results, we calculate the CoVaR to measure the risk contagion between markets. The major conclusions are as follows:

Firstly, the financial risk dependence structures are asymmetric, and the correlations are heterogeneous. Overall, the positive dependence is dominant between Chinese and mature markets. Meanwhile, the downside risk correlation is stronger than the upside one between Chinese and American, German, and Japanese markets, while the opposite relevance is observed for Chinese and Canadian, British, and Australian markets. It is noted that compared to the Americas and European markets, the risk correlation between Chinese and Asia–Oceania markets is relatively weak. Moreover, the negative dependence should not be ignored as it may emerge in a volatile market environment and provide market participants with signals to manage their exposure. Then, the financial risk contagion is also asymmetric, which manifests in both positive and negative contagion effects, as well as in both upside and downside contagion effects. Overall, the positive effect is stronger than the negative effect, and the downside effect is stronger than the upside effect in positive structure. Compared with mature markets in Europe and the Americas, the risk spillover from Asia–Oceania markets is relatively weak, indicating that the Japanese and Australian markets can be considered as a potential choice for the investors in the Chinese market to diversify their portfolios.

This work enriches the understanding of financial risk contagion mechanism of Chinese and mature markets, which provides both practical and policy implications for investor and supervisors. With respect to practical aspects, before constructing an international portfolio, it is necessary for investors to use such quantitative models to identify and filter out markets with stronger downside risk correlation in order to better diversify their investment risks. In this study, the Chinese stock market generally has weaker risk relationship and contagion effects with mature markets in Asia–Oceania compared to the Americas and Europe thus, the Japanese and Australian markets can be regarded as feasible choices for Chinese market investors to diversify investment risks. In addition, since the Ms-M-Clayton has the capability to detect negative risk dependence structures, it is possible for investors to leverage it to discover the unusual opportunities to hedge investment risk by constructing cross-market portfolios. In the policy-making perspective, for the emerging markets at a disadvantage in risk contagion, it is essential to improve their financial system and decrease the pressure of capital outflows under extreme conditions. Specifically, according to the findings of this work, the dependence structures between markets are generally stable, which creates the possibility for supervisors to predict future risk scenarios and formulate guiding or regulatory policies using the Ms-M-Clayton copula. Moreover, as the model is sensitive to the transition probability in risk dependence states, and the supervisors are able to perceptively monitor the potential risk changes and implement risk prevention measures on previous experience. Furthermore, the Ms-M-Clayton copula model is also applicable for the series analyses of various engineering fields.

To mention, we focus only on the risk contagion between Chinese and mature stock markets in this work. Several fast-growing economies, such as Brazil, Russia, India, and South Africa, constituting the BRICS group with China, represent over 18% of the population and approximately 8% of the GDP around the world. A comparative analysis of their stock markets may be a promising topic in future research.

Author Contributions: Conceptualization, H.N. and K.X.; methodology, K.X.; software, K.X.; investigation, K.X.; data curation, M.X.; writing—original draft preparation, K.X.; writing—review and editing, H.N.; supervision, H.N.; project administration, H.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: No applicable.

Data Availability Statement: The data is available on request.

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

Appendix A

Table A1. Diagnostic tests of stationarity, autocorrelation, and heteroscedasticity.

	CSI300	S&P500	GSPTSE	DAX30	FTSE100	Nikkei225	ASX200
ADF	−11.633 ***	−12.099 ***	−11.859 ***	−12.234 ***	−13.419 ***	−11.760 ***	−12.253 ***
PP	−12.293 ***	−12.145 ***	−11.920 ***	−12.181 ***	−13.418 ***	−11.763 ***	−12.248 ***
KPSS	0.104	0.179	0.033	0.045	0.059	0.135	0.048
Q (5)	21.335 ***	6.294	7.096	7.109	3.385	4.423	3.236
Q (10)	25.902 ***	15.362	14.193	15.727	9.440	8.564	6.966
Q2(5)	28.419 ***	48.390 ***	16.423 ***	5.253	17.739 ***	5.174	11.914 **
Q2(10)	58.267 ***	52.676 ***	18.408 **	12.539	39.787 ***	10.733	14.600
ARCH (1)	0.704	26.472 ***	15.538 ***	2.126	15.553 ***	3.634 *	11.383 ***
ARCH (5)	14.973 ***	40.282 ***	17.452 ***	4.533	17.119 ***	7.610	11.441 **

Note: ***, **, and * represent the significant levels at 1%, 5%, and 10%, respectively.

Table A2. Parameter estimation results of the marginal distribution models and diagnostic tests.

	Parameters	CSI300	S&P500	GSPTSE	DAX30	FTSE100	Nikkei225	ASX200
Panel A. AR-GARCH model	ϕ_0	1.348 **	1.092 ***	0.282 ***	0.589	0.447	0.659	0.760 ***
		(0.668)	(0.240)	(0.100)	(0.398)	(0.280)	(0.494)	(0.286)
	ϕ_1	0.092 *						
		(0.071)						
	ϕ_4	0.189 ***						
		(0.062)						
	α_0	4.201	1.137	0.642 ***		1.106	4.700	11.320 ***
		(4.241)	(0.937)	(0.211)		(1.058)	(3.963)	(2.144)
	α_1	0.153	0.248 ***	0.535 ***		0.132 **	0.128 **	0.356 ***
		(0.116)	(0.102)	(0.103)		(0.062)	(0.053)	(0.130)
	β_1	0.797 ***	0.723 ***	0.342 ***		0.809 ***	0.730 ***	
		(0.135)	(0.105)	(0.097)		(0.102)	(0.149)	
	GED.	1.193 ***	1.286 ***			1.652 ***		1.494 ***
		(0.135)	(0.213)			(0.262)		(0.242)
Panel B. Diagnostic tests	Log-L	−627.552	−510.743	−343.105	−578.161	−511.389	−579.028	−517.528
	AIC	6.973	5.546	3.732	6.228	5.553	6.314	5.608
	Q (5)	3.820	1.698	3.799	7.109	0.712	0.474	1.594
	Q (10)	6.385	6.374	7.398	15.727	4.119	3.896	5.102
	Q2 (5)	2.549	3.639	3.505	5.971	5.056	2.588	1.393
	Q2 (10)	9.202	9.977	9.374	12.817	15.250	7.690	9.668
	ARCH (1)	0.278	1.998	2.009	2.124	2.289	0.796	0.590
	ARCH (5)	2.491	3.607	3.678	4.533	4.583	3.340	1.273

Note: ***, **, and * represent the significant levels at 1%, 5%, and 10%, respectively.

Table A3. The K-S test for CoVaR/VaR between positive and negative risk spillovers.

Hypotheses	CSI300-S&P500	CSI300-GSPTSE	CSI300-DAX30	CSI300-FTSE100	CSI300-Nikkei225	CSI300-ASX200
$H_0: \frac{CoVaR^{\beta,t}_{CSI300(D)\|Other(D)}}{VaR^{\alpha,t}_{CSI300,D}} = \frac{CoVaR^{\beta,t}_{CSI300(D)\|Other(U)}}{VaR^{\alpha,t}_{CSI300,D}}$; $H_1: \frac{CoVaR^{\beta,t}_{CSI300(D)\|Other(D)}}{VaR^{\alpha,t}_{CSI300,D}} \neq \frac{CoVaR^{\beta,t}_{CSI300(D)\|Other(U)}}{VaR^{\alpha,t}_{CSI300,D}}$.	1.000 *** (0.000)	1.000 *** (0.000)	1.000 *** (0.000)	1.000 *** (0.000)	0.945 *** (0.000)	0.550 *** (0.000)
$H_0: \frac{CoVaR^{\beta,t}_{CSI300(U)\|Other(U)}}{VaR^{\alpha,t}_{CSI300,U}} = \frac{CoVaR^{\beta,t}_{CSI300(U)\|Other(D)}}{VaR^{\alpha,t}_{CSI300,U}}$; $H_1: \frac{CoVaR^{\beta,t}_{CSI300(U)\|Other(U)}}{VaR^{\alpha,t}_{CSI300,U}} \neq \frac{CoVaR^{\beta,t}_{CSI300(U)\|Other(D)}}{VaR^{\alpha,t}_{CSI300,U}}$.	1.000 *** (0.000)	0.824 *** (0.000)	1.000 *** (0.000)	0.577 *** (0.000)	0.896 *** (0.000)	0.445 *** (0.000)

Note: this table summarize the results of the Kolmogorov–Smirnov (KS) tests; *** represent the significant levels at 1%, and the p-values for the KS statistics are reported in parentheses.

Table A4. The K-S test for CoVaRs/VaRs between upside and downside risk spillovers.

Hypotheses	CSI300-S&P500	CSI300-GSPTSE	CSI300-DAX30	CSI300-FTSE100	CSI300-Nikkei225	CSI300-ASX200
$H_0: \frac{CoVaR^{\beta,t}_{CSI300(D)\|Other(D)}}{VaR^{\alpha,t}_{CSI300,D}} = \frac{CoVaR^{\beta,t}_{CSI300(U)\|Other(U)}}{VaR^{\alpha,t}_{CSI300,U}}$; $H_1: \frac{CoVaR^{\beta,t}_{CSI300(D)\|Other(D)}}{VaR^{\alpha,t}_{CSI300,D}} \neq \frac{CoVaR^{\beta,t}_{CSI300(U)\|Other(U)}}{VaR^{\alpha,t}_{CSI300,U}}$.	0.615 *** (0.000)	0.797 *** (0.000)	0.440 *** (0.000)	0.830 *** (0.000)	0.247 *** (0.000)	0.412 *** (0.000)
$H_0: \frac{CoVaR^{\beta,t}_{CSI300(D)\|Other(U)}}{VaR^{\alpha,t}_{CSI300,D}} = \frac{CoVaR^{\beta,t}_{CSI300(U)\|Other(D)}}{VaR^{\alpha,t}_{CSI300,U}}$; $H_1: \frac{CoVaR^{\beta,t}_{CSI300(D)\|Other(U)}}{VaR^{\alpha,t}_{CSI300,D}} \neq \frac{CoVaR^{\beta,t}_{CSI300(U)\|Other(D)}}{VaR^{\alpha,t}_{CSI300,U}}$.	0.989 *** (0.000)	0.346 *** (0.000)	0.253 (0.217)	0.703 *** (0.000)	0.484 *** (0.000)	0.473 *** (0.000)

Note: this table summarize the results of the Kolmogorov–Smirnov (KS) tests; *** represent the significant levels at 1%, and the *p*-values for the KS statistics are reported in parentheses.

References

1. Vogl, M. Chaos Measure Dynamics and a Multifactor Model for Financial Markets. Available at SSRN 4251673. 2022. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4251673 (accessed on 20 October 2022).
2. Vogl, M. Quantitative modelling frontiers: A literature review on the evolution in financial and risk modelling after the financial crisis (2008–2019). *SN Bus. Econ.* **2022**, *2*, 183. [CrossRef] [PubMed]
3. Liu, Y.; Wei, Y.; Wang, Q.; Liu, Y. International stock market risk contagion during the COVID-19 pandemic. *Financ. Res. Lett.* **2022**, *45*, 102145. [CrossRef] [PubMed]
4. Marfatia, H.A. A fresh look at integration of risks in the international stock markets: A wavelet approach. *Rev. Financ. Econ.* **2017**, *34*, 33–49. [CrossRef]
5. Marfatia, H.A. Investors' risk perceptions in the US and global stock market integration. *Res. Int. Bus. Financ.* **2020**, *52*, 101169. [CrossRef]
6. Bhatti, M.I.; Nguyen, C.C. Diversification evidence from international equity markets using extreme values and stochastic copulas. *J. Int. Financ. Mark. Inst. Money* **2012**, *22*, 622–646. [CrossRef]
7. An, S. Dynamic Multiscale Information Spillover among Crude Oil Time Series. *Entropy* **2022**, *24*, 1248. [CrossRef]
8. Lai, Y.H.; Tseng, J.C. The role of Chinese stock market in global stock markets: A safe haven or a hedge? *Int. Rev. Econ. Financ.* **2010**, *19*, 211–218. [CrossRef]
9. Zhong, Y.; Liu, J.P. Correlations and volatility spillovers between China and Southeast Asian stock markets. *Quart. Rev. Econ. Financ.* **2021**, *81*, 57–69. [CrossRef]
10. Kole, E.; Koedijk, K.; Verbeek, M. Selecting copulas for risk management. *J. Bank Financ.* **2007**, *31*, 2405–2423. [CrossRef]
11. Luo, C.Q.; Xie, C.; Yu, C.; Xu, Y. Measuring financial market risk contagion using dynamic MRS-Copula models: The case of Chinese and other international stock markets. *Econ. Model.* **2015**, *51*, 657–671.
12. Di Persio, L.; Vettori, S. Markov Switching Model Analysis of Implied Volatility for Market Indexes with Applications to S&P 500 and DAX. *J. Math.* **2014**, *2014*, 1–17.
13. Fermanian, J.-D. Recent Developments in Copula Models. *Econometrics* **2017**, *5*, 34. [CrossRef]
14. Ji, Q.; Liu, B.-Y.; Cunado, J.; Gupta, R. Risk spillover between the US and the remaining G7 stock markets using time-varying copulas with Markov switching: Evidence from over a century of data. *N. Am. J. Econ. Financ.* **2020**, *51*, 100846. [CrossRef]
15. Di Persio, L.; Frigo, M. Gibbs sampling approach to regime switching analysis of financial time series. *J. Comput. Appl. Math.* **2016**, *300*, 43–55. [CrossRef]
16. Segnon, M.; Trede, M. Forecasting market risk of portfolios: Copula-Markov switching multifractal approach. *Eur. J. Financ.* **2017**, *24*, 1123–1143. [CrossRef]
17. Rajwani, S.; Kumar, D. Measuring dependence between the USA and the Asian economies: A time-varying Copula approach. *Glob. Bus. Rev.* **2019**, *20*, 962–980. [CrossRef]
18. Wang, K.; Chen, Y.-H.; Huang, S.-W. The dynamic dependence between the Chinese market and other international stock markets: A time-varying copula approach. *Int. Rev. Econ. Financ.* **2011**, *20*, 654–664. [CrossRef]
19. Jiang, C.X.; Li, Y.Q.; Xu, Q.F.; Liu, Y. Measuring risk spillovers from multiple developed stock markets to China: A vine-copula-GARCH-MIDAS model. *Int. Rev. Econ. Financ.* **2021**, *75*, 386–398. [CrossRef]
20. Liu, X.-D.; Pan, F.; Cai, W.-L.; Peng, R. Correlation and risk measurement modeling: A Markov-switching mixed Clayton copula approach. *Reliab. Eng. Syst. Saf.* **2020**, *197*, 106808. [CrossRef]
21. Abakah, E.J.A.; Tiwari, A.K.; Alagidede, I.P.; Gil-Alana, L.A. Re-examination of risk-return dynamics in international equity markets and the role of policy uncertainty, geopolitical risk and VIX: Evidence using Markov-switching copulas. *Financ. Res. Lett.* **2022**, *47*, 102535. [CrossRef]
22. Reinhart, C.M.; Calvo, S. *Capital Flows to Latin America: Is There Evidence of Contagion Effects*; Peterson Institute for International Economics: Washington, DC, USA, 1996.

23. Forbes, K.; Rigobon, R. No contagion, only interdependence: Measuring stock market comovements. *J. Financ.* **2010**, *57*, 2223–2261. [CrossRef]
24. Mihai, N.; Maria, M.P. Time-varying dependence in European equity markets: A contagion and investor sentiment driven analysis. *Econ. Model.* **2020**, *86*, 133–147.
25. Ajaya, K.P.; Pradiptarathi, P.; Swagatika, N.; Parad, A. Information bias and its spillover effect on return volatility: A study on stock markets in the Asia-Pacific region. *Pac.-Basin Financ. J.* **2021**, *69*, 101653.
26. Fan, H.C.; Gou, Q.; Peng, Y.C. Spillover effects of capital controls on capital flows and financial risk contagion. *J. Int. Money Financ.* **2020**, *105*, 102189. [CrossRef]
27. Alberto, B.; David, L.T.; Danilo, L.; Marsiglio, S. Financial contagion and economic development: An epidemiological approach. *J. Econ. Behav. Organ.* **2019**, *162*, 211–228.
28. Cheng, H.; Glascock, J.L. Stock market linkages before and after the Asian financial crisis: Evidence from three greater china economic area stock markets and the us. *Rev. Pac. Basin Financ.* **2006**, *9*, 297–315. [CrossRef]
29. Li, H. International linkages of the Chinese stock exchanges: A multivariate GARCH analysis. *Appl. Financ. Econ.* **2007**, *17*, 285–297. [CrossRef]
30. Chatziantoniou, I.; Gabauer, D.; Marfatia, H.A. Dynamic connectedness and spillovers across sectors: Evidence from the Indian stock market. *Scott. J. Political Econ.* **2021**, *69*, 283–300. [CrossRef]
31. Sklar, A. Fonctions de repartition à n dimensions et leurs marges. *Publ. De L'institut De Stat. De L'université De Paris* **1959**, *8*, 229–231.
32. Chang, K.L. Does REIT index hedge inflation risk? new evidence from the tail quantile dependences of the Markov-switching GRG copula. *N. Am. J. Econ. Financ.* **2017**, *39*, 56–67. [CrossRef]
33. Huang, J.J.; Lee, K.J.; Liang, H.M.; Lin, W.F. Estimating value at risk of portfolio by conditional copula-GARCH method. *Insur. Math. Econ.* **2009**, *45*, 315–324. [CrossRef]
34. Hussain, S.I.; Li, S. The dependence structure between Chinese and other major stock markets using extreme values and copulas. *Int. Rev. Econ. Financ.* **2018**, *56*, 421–437. [CrossRef]
35. Luo, C.Q.; Liu, L.; Wang, D. Multiscale financial risk contagion between international stock markets: Evidence from EMD-Copula-CoVaR analysis. *N. Am. J. Econ. Financ.* **2021**, *58*, 101512. [CrossRef]
36. Patton, A.J. Modelling asymmetric exchange rate dependence. *Int. Econ. Rev.* **2006**, *47*, 527–556. [CrossRef]
37. Zhang, X.; Zhang, T.; Lee, C.C. The path of financial risk spillover in the stock market based on the R-vine-Copula model. *Physics A* **2022**, *600*, 127470. [CrossRef]
38. Andrieu, C.; Thoms, J. A tutorial on adaptive MCMC. *Stat. Comput.* **2008**, *18*, 343–373. [CrossRef]
39. Huang, C.W.; Hsu, C.P.; Chiou, W.J.P. *Can Time-Varying Copulas Improve the Mean-Variance Portfolio*; Springer: New York, NY, USA, 2014.
40. Wang, Y.C.; Wu, J.L.; Lai, Y.H. A revisit to the dependence structure between the stock and foreign exchange markets: A dependence-switching copula approach. *J. Bank. Financ.* **2013**, *37*, 1706–1719. [CrossRef]
41. Ji, Q.; Bouri, E.; Roubaud, D.; Shahzad, S.J.H. Risk spillover between energy and agricultural commodity markets: A dependence-switching CoVaR-copula model. *Energy Econ.* **2018**, *75*, 14–27. [CrossRef]
42. Reboredo, J.C.; Ugolini, A. Systemic risk in European sovereign debt markets: A CoVaR-copula approach. *J. Int. Money Financ.* **2015**, *51*, 214–244. [CrossRef]
43. Reboredo, J.C.; Ugolini, A. Downside/upside price spillovers between precious metals: A vine copula approach. *N. Am. J. Econ. Financ.* **2015**, *34*, 84–102. [CrossRef]
44. Xiao, Y. The risk spillovers from the Chinese stock market to major East Asian stock markets: A MSGARCH-EVT-copula approach. *Int. Rev. Econ. Financ.* **2020**, *65*, 173–186. [CrossRef]
45. Sun, X.; Liu, C.; Wang, J.; Li, J. Assessing the extreme risk spillovers of international commodities on maritime markets: A GARCH-Copula-CoVaR approach. *Int. Rev. Financ. Anal.* **2020**, *68*, 101453. [CrossRef]
46. Bai, X.; Lam, J.S.L. A copula-GARCH approach for analyzing dynamic conditional dependence structure between liquefied petroleum gas freight rate, product price arbitrage and crude oil price. *Energy Econ.* **2019**, *78*, 412–427. [CrossRef]
47. Nguyen, Q.N.; Bedoui, R.; Majdoub, N.; Guesmi, K.; Chevallier, J. Hedging and safe-haven characteristics of gold against currencies: An investigation based on multivariate dynamic copula theory. *Resour. Policy* **2020**, *68*, 101766. [CrossRef]
48. Liu, X.D.; Pan, F.; Yuan, L.; Chen, Y. The dependence structure between crude oil futures prices and Chinese agricultural commodity futures prices: Measurement based on Markov-switching GRG copula. *Energy* **2019**, *182*, 999–1012. [CrossRef]
49. Joe, H. *Multivariate Models and Dependence Concepts*; Chapman & Hall: London, UK, 1997.
50. Dempster, A.P. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **1977**, *39*, 1–38.
51. Li, M.; Lu, Y. Genetic Algorithm Based Maximum Likelihood DOA Estimation. In Proceedings of the 2002 International Radar Conference, Edinburgh, UK, 15–17 October 2002; pp. 502–506, IET Digital Library.
52. Gray, S.F. Modeling the conditional distribution of interest rates as a regime-switching process. *J. Financ. Econ.* **1996**, *42*, 27–62. [CrossRef]

53. Li, X.F.; Wei, Y. The dependence and risk spillover between crude oil market and China stock market: New evidence from a variational mode decomposition-based copula method. *Energy Econ.* **2018**, *74*, 565–581. [CrossRef]
54. Huang, Q.; Wang, X.; Zhang, S. The effects of exchange rate fluctuations on the stock market and the affecting mechanisms: Evidence from BRICS countries. *N. Am. J. Econ. Financ.* **2021**, *56*, 101340. [CrossRef]

entropy

MDPI

Article

A Network Model Approach to International Aid

Joe Scattergood and Steven Bishop *

Department of Mathematics, University College London, London WC1E 6BT, UK
* Correspondence: s.bishop@ucl.ac.uk

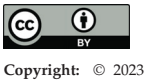
Abstract: Decisions made by international aid donors regarding the allocation of their aid budgets to recipients can be mathematically modelled using network theory. The many countries and multilateral organisations providing developmental aid, mostly to developing countries, have numerous competing or conflicting interests, biases and motivations, often obscured by a lack of transparency and confused messaging. Using network theory, combined with other mathematical methods, these inter-connecting and inter-dependent variables are identified, revealing the complicated properties and dynamics of the international aid system. Statistical techniques are applied to the vast amount of available, open data to first understand the complexities and then identify the key variables, focusing principally on bilateral aid flows. These results are used to create a weighted network model which is subsequently adapted for use by a hypothetical aid recipient. By incorporating modern portfolio theory into this weighted network model and taking advantage of a donor's reasons for allocating their aid budgets to that recipient, a simulation is carried out treating the problem as an optimal investment portfolio of aid determinant 'assets' which illustrates how a recipient can maximise their aid receipts. Suggestions are also made for further uses and adaptations of this weighted network model.

Keywords: international aid; foreign aid; complex systems; network science; network theory; econometrics; financial mathematics; portfolio theory

Citation: Scattergood, J.; Bishop, S. A Network Model Approach to International Aid. *Entropy* **2023**, *25*, 641. https://doi.org/10.3390/e25040641

Academic Editor: Panos Argyrakis

Received: 1 March 2023
Revised: 2 April 2023
Accepted: 4 April 2023
Published: 11 April 2023



1. Introduction

US$162bn of foreign aid was donated by developed countries ('donors') to developing countries ('recipients') in 2020 [1]. Democratic governments of donor countries are faced with decisions regarding how and where to allocate their foreign aid budgets, not solely for poverty alleviation but also to achieve a diverse set of specific goals and unique strategies.

Significant drivers of how donors allocate their foreign aid budgets are based upon achieving certain political and strategic objectives, both domestic and global. Global objectives include the projection of soft power, control over foreign resources, biases towards allies or ex-colonies and gaining global influence. These motivations, behaviours and determinants are complicated and difficult to capture in a mathematical model. Nevertheless, in democratic societies at least, justifications for aid allocation decisions and transparency are often demanded, and mathematical methods and models can help provide these, even if they are not used to determine forward action.

Foreign aid dynamics and interactions are particularly complicated, as detailed in Ben Ramalingam's book, *Aid on the Edge of Chaos* [2], with the complexity of the determinants of aid flows well documented by [3]. There are many interacting variables and dynamics of foreign aid networks, and the relative importance which donors place on their specific and numerous aid determinants is often not known. This causes difficulties when analysing and concluding on many aspects of overseas aid, which further makes it problematic to design and create a useful mathematical model that can capture the dynamics of foreign aid networks and successfully incorporate the interacting and inter-dependent variables.

To attempt this, the many variables and determinants which create the complicated foreign aid dynamics firstly need to be identified and understood. By studying other

research conducted on this topic, and sourcing and analysing additional data using statistical techniques, the results from this article will inform the adaptation and use of a weighted network model, first proposed by [4], that captures the properties, interactions and dynamics of the international aid system.

Much of the literature and research into foreign aid dynamics focuses on the donor. Econometric techniques, primarily regression and ordinary least squares ([5–7]), are commonly employed in an attempt to reveal the relative importance of donor motivations and potential biases behind their aid allocation decisions. More recently, there has been research conducted on the growing field of network theory and the utilisation of related mathematical methods to model the allocation of aid that goes beyond regression ([4,8]).

However, it is rarer to find research and analysis focusing on aid recipients. By understanding donor motivations and biases, aid recipients could exploit these 'assets' and potentially increase their aid receipts if they are viewed as a portfolio of investments.

By identifying and quantifying significant donor motivations for allocating their aid budgets, inputting these variables into a general weighted network model [4] and then adapting it using financial mathematics (modern portfolio theory), aid recipients could use the model to optimise their aid income portfolio, treating donor variables similarly to assets in an investment portfolio. This is illustrated in this article using a simulation.

The principal aim, then, of this article is to illustrate the power of network science and mathematical modelling when applied to the complex and dynamical system of international aid. The potential impact is an increase in transparency of the often-opaque motivations and biases of aid donors, which subsequently could be employed by recipients to increase their aid income.

2. Methods

2.1. Data and Data Analysis

The first step to evaluating, and then adapting, the general weighted network model [4] is to identify the significant motivations and preferences shown by selected donors regarding the allocation of their aid budgets. These will be used as the model's variables. Subsequently, the accuracy of the model's mechanics and outputs can be tested against actual historical data for selected donors and recipients.

Data from the OECD and World Bank were sourced and analysed using various statistical techniques to identify and understand the inter-connecting and inter-dependent variables that drive the data. The pertinent results are summarised here.

2.1.1. Economic and Foreign Aid Data

To compare the economic fortunes of one country versus another, gross national income (GNI), a key measure of economic well-being and a superior metric for assessing the overall economic condition of a country, especially for countries that have large foreign receivables or outlays, will be used for identifying the level of need of an aid recipient ('recipient need'). Furthermore, to assist with country comparisons, GNI per capita will be used rather than absolute GNI.

Table 1 lists the top 10 aid recipients in 2019 by net official development assistance (ODA) receipts, classified as total net ODA flows from Development Assistance Committee (DAC) countries, multilateral organisations and non-DAC countries.

When identifying the top donor countries, rather than looking at absolute aid donated, the affordability of a donor country to provide aid is assessed using aid donated as a percentage of country GNI. This is summarised in Figure 1, which lists the members of the DAC, a development committee of the OECD.

Table 1. Top 10 ODA recipients, including significant regional aid donations, and figures for all developing counties for comparison [9]. All figures in US$m unless otherwise stated.

	Net ODA Receipts					GNI/CAP (US$)	GNI	ODA/GNI (%)
Country/Region	**2015**	**2016**	**2017**	**2018**	**2019**	**2019**	**2019**	**2019**
Syrian Arab Republic	4920	8900	10,428	9997	10,252	-	-	-
Ethiopia	3239	4084	4125	4941	4810	850	95,641	5.03
Bangladesh	2593	2533	3782	3045	4518	1940	316,907	1.43
Yemen	1778	2301	3234	7985	4397	-	-	-
Afghanistan	4274	4069	3812	3792	4285	540	19,402	22.08
Nigeria	2432	2498	3359	3305	3531	2030	433,449	0.81
Kenya	2464	2188	2480	2491	3251	1750	93,578	3.47
Democratic Republic of the Congo	2599	2102	2293	2514	3026	520	45,879	6.59
Jordan	2141	2728	2980	2526	2797	4300	43,429	6.44
India	3174	2679	3198	2462	2611	2130	2,843,902	0.09
Regional (not specific to any country)								
South of Sahara	2435	2635	2759	3137	3410			
Africa region	2184	2777	3017	3241	3201			
All developing countries	146,742	158,811	165,090	166,540	168,588	511,750	292,854,611	0.58

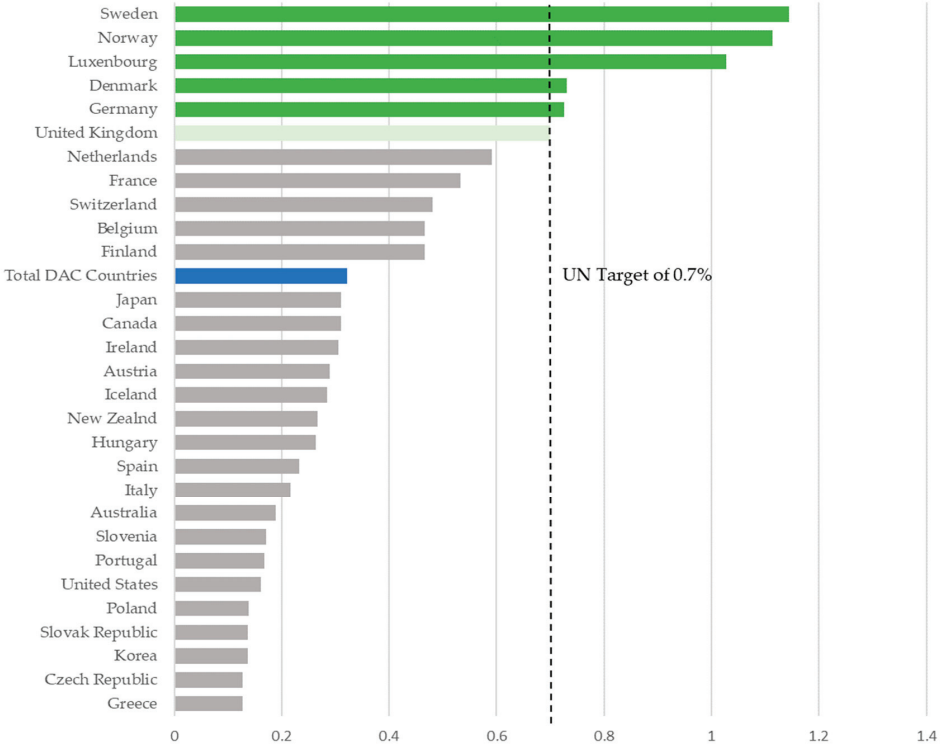


Figure 1. ODA grant equivalent as percentage of GNI in 2020 for DAC donors in the OECD. The grey bars identify those countries that contribute less than the United Nations (UN) target of 0.7%, the blue bar shows total of the DAC countries as a percentage of their GNI and the green bars highlight those countries who contribute over the UN target of 0.7%.

Figure 1 is sourced from the OECD website [10] and arranged in descending order based on the percentage of DAC-country GNI donated in 2020, with Sweden donating the highest percentage of its GNI at 1.15%.

Donor affordability is epitomised by the 0.7% target agreed by the United Nations (UN) in 1970 for aid contributions by DAC countries to developing countries. It is reasonable then to assume that since the UN members agreed to 0.7%, they can therefore afford to

donate 0.7% of their GNI. However, as shown in Figure 1, this target is not being met by most UN countries, including the USA.

Bilateral aid flows—aid given directly from a country donor to a recipient donor—comprised circa. 67% of total ODA donated in 2019, with the remaining third being flows from multilateral institutions and international financial institutions (for example, the World Bank). However, the proportion of bilateral aid reduced significantly in 2020 by 36% on 2019 levels to 42% of the total aid donated, with multilateral institutions taking up the slack, due mainly to the impact of COVID [11].

2.1.2. Aid Flows from Donors to Former Colonies

There are robust conclusions in the research performed by [3,5,6], among many others, that a strong motivation behind aid allocation decisions by donors lies in whether the recipient is an ex-colony or not. Former colonies receive proportionately more aid from their former colonial masters than other recipients.

Indeed, according to [5], between 1970 and 1994, France gave 57% of its total bilateral aid to its former colonies, the UK gave 78% and Portugal 99.6%. Moreover, according to the OECD [12], in 2009, the largest recipient of UK aid was India and, by 2019, this was Pakistan, both former UK colonies. Thus, colonial history is positively correlated with aid, as identified by [5] and confirmed by own analysis performed.

2.1.3. Trade Activity

Before correlation techniques were applied to detect any interdependencies between trade activity and aid donations, the raw data were analysed. Trade data are sourced from the World Integrated Trade Solution (WITS) website, a sister site of the World Bank specifically focused on trade [13], for the period 1993 to 2019. By charting this trade data with aid data sourced from the World Bank [9], a pattern of aid versus trade can be viewed over time. This suggested a positive correlation, confirmed by calculating correlations between the two data sets over many periods. This result is also backed by research performed by [5,6,14].

2.1.4. Recipient Need

The literature is mixed regarding the relative importance of recipient need as a variable in a donor's aid allocation decisions. In [5], the authors are clear on donor motivations being based mainly on self-interest and political and strategic considerations over aid recipient needs. However, later studies, such as [14], dispute this conclusion stating that self-interest, while still a significant input into aid allocation decisions, is not as important as recipient need. Moreover, [6] conclude that the USA behaves very differently from all other aid donors, except Japan, by putting much less emphasis on recipient need and much more emphasis on donor self-interest.

2.1.5. The Herding Phenomenon (the Bandwagon Effect)

Another variable to consider for inclusion in a mathematical model of foreign aid is herding behaviour often exhibited by donors, also termed the 'bandwagon effect'. This refers to the actions and impulses of a group of agents, countries, politicians, or financial traders to follow the actions of the 'crowd' rather than trust their own individual judgment. The phenomenon has similar attributes to 'groupthink'. It is an emergent behaviour of a dynamical system due to the many interactions taking place within that system.

Herding is commonly associated with financial market behaviour, for example asset bubbles [15]. Grounded in behavioural finance, herd mentality refers to investors' bias to follow what other investors are doing, being largely influenced by emotion and intuition, rather than by their own evaluations of potential investments.

In terms of aid allocation, the bandwagon effect manifests itself when a recipient receives more aid from one donor, leading to an increase in aid from many more donors. In

other words, the more aid a recipient receives, the more it attracts. It likely depends on the relative influence of the lead donor rather than characteristics of the recipient.

Research conducted by [6] attempted to measure the effect using regression and incorporating aid from other sources, not only ODA. They find that there is some support for the herding argument, but it is far from conclusive. Ref. [16] gives the phenomenon a more thorough review, concluding that there is around an 11% impact on aid donations through donor herding, which is relatively significant.

2.2. A Network Model for Foreign Aid

The principal outcome of the research conducted, using data analysis and statistical techniques, is the identification of the following significant aid determinants:

- Past colonial relationships;
- Trade activity and commercial interests;
- Poverty alleviation (recipient need); and
- Bandwagon impacts ('herding').

These variables will now be incorporated into a weighted network model to demonstrate how such a model can be modified and used by an aid recipient to treat their various donor-sourced aid receipts as an investment portfolio and maximise their aid income using modern portfolio theory.

The weighted network model introduced allows for additional variables to be incorporated and, indeed, further variables were considered for inclusion. Variables such as the occurrence of war, migration and recipient corruption could be reflected in the model; however, the focus here is on long-term and relatively stable determinants of aid. Furthermore, the bandwagon impact may partially and indirectly incorporate these variables; for example, the war in Afghanistan in the early 2000s led to significant amounts of aid donated by the USA to Afghanistan, swiftly followed by aid donated by other donors.

2.2.1. The General Weighted Network Model

The general model proposed by [4] follows a weighted network model approach utilising donor-specific preference functions to measure donor motivations and biases when deciding aid allocations. The preference functions quantify the *relative* contributions of aid determinants used in aid allocation input decisions, such as poverty, trade activity, past colonial relationships and bandwagon impacts ('herding'), into 'weights' which are applied to a network model, revealing donor behaviours and the relative importance placed on these aid determinants.

Figure 2 is an archetypal bipartite network model which has nation donors on the left representing the set of nodes D and the recipients on the right representing the set of nodes R. D and R are disjoint sets of nodes in which links can exist only between the two sets and not within each set, thus illustrating the flow of aid which is directed only from elements of D to elements of R. In total, there are six nodes split into two disjoint sets of three donors, $d_i \in D$, and three recipients, $r_j \in R$, where $i, j = 1, 2, 3$ represent donors and recipients, respectively.

The n-vector node-specific information u^k_β represents the n quantities (or vector elements) associated with each country, $k = i, j$, in the network, where $\beta = 1, \ldots, n$ is used to denote the numbered element of the vector u^k. For example, recipient Ethiopia's node in Figure 2 is labelled r_3. In the case that this node's specific information contains poverty levels, u^3_1, colonial history, u^3_2, and trade activity, u^3_3, then the vector u^3 has 3 elements $n = 3$, denoted as u^3_β, where $\beta = 1, 2, 3$.

Further, the links between each set holds m-vector link weights l^{ij}_α, representing the m relationships between donors i and recipients j, where $\alpha = 1, \ldots, m$ denotes the numbered element of the vector l^{ij}. For example, the vector representing trade activity and colonial history between the UK, d_2, and Bangladesh, r_1, are denoted as l^{21}_1 and l^{21}_2, respectively, where $m = 2$ in this example, and is denoted as l^{21}_α, where $\alpha = 1, 2$.

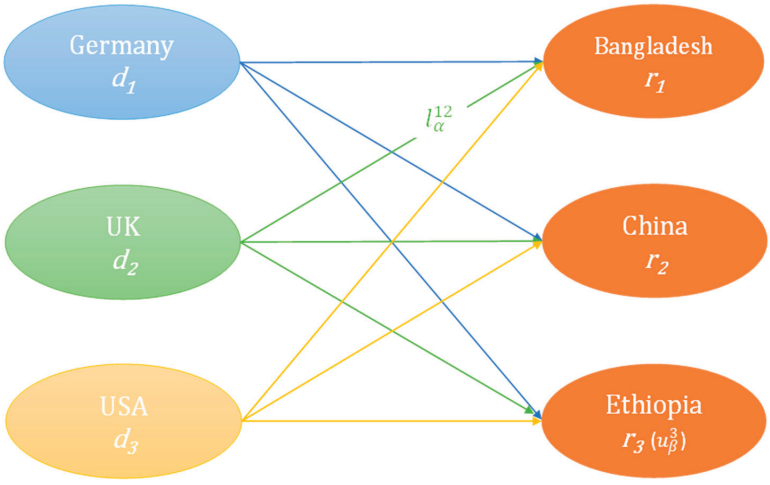


Figure 2. A complete bipartite graph of a weighted network containing notation which underpins the model.

The node and link vector information, as defined, can be quantified into weights and input into a preference Function (1), supplemented by input Functions (2) and (3), outputting a percentage of aid allocated by a donor, d_i, to a recipient, r_j.

$$P^{ij}\left(l^{ij}, u^j\right) := \prod_\alpha f_\alpha^i\left(l_\alpha^{ij}, l_\alpha^{i\bullet}\right) \cdot \prod_\beta g_\beta^i\left(u_\beta^j, u_\beta^\bullet\right) \tag{1}$$

The superscript $\bullet$ in preference Function (1) denotes all recipients in the set of nodes R, and its role is seen in the denominators of (2) and (3). The input functions, f_α^i and g_β^i, quantify donor preferences towards specific determinants of aid, such as trade activity and recipient need, into proportioned weights before input into (1):

$$f_\alpha^i\left(l_\alpha^{ij}, l_\alpha^{i\bullet}\right) = \left[\frac{l_\alpha^{ij}}{\sum_{k\in V_2} l_\alpha^{ik}}\right]^{\mu_\alpha^i} \tag{2}$$

$$g_\beta^i\left(u_\beta^j, u_\beta^\bullet\right) = \left[\frac{u_\beta^j}{\sum_{k\in V_2} u_\beta^k}\right]^{\eta_\beta^i} \tag{3}$$

where the exponent parameters, μ_α^i and η_β^i, hold only non-negative real values. These are referred to as 'power' parameters.

The terms in the square brackets in (2) and (3) are functional inputs holding information on chosen aid determinants positively correlated with aid allocation and expressed as a proportion. The greater the proportion, the higher the value generated by the function and therefore the greater the weight due to a particular aid determinant for input into (1).

There is a difference in usage between the input Functions (2) and (3). The function $f_\alpha^i\left(l_\alpha^{ij}, l_\alpha^{i\bullet}\right)$ in (2) is used for link-specific weights, l_α^{ij}, and quantifies behaviours and relationships that exist between a donor d_i and a recipient r_j, for example the levels of trade activity. The function $g_\beta^i\left(u_\beta^j, u_\beta^\bullet\right)$ in (3) is used for node-specific weights, u_β^j, and quantifies a specific recipient metric, such as the poverty ratio among recipients, which is a determinant of aid that bears no direct relationship to a particular donor. Note u_β^j in

(3) is specific to a recipient r_j; however, it can also be specific to a donor, u^i_β, to quantify determinants specific to a donor, d_i, such as donor affordability.

The form of (2) and (3) assumes that a positive correlation exists between the determinant in question and aid allocated. For negative correlations, these terms are modified by subtracting the functions from (1), resulting in a recipient with a lesser proportion receiving a higher weight of preference and therefore aid allocated relative to the other recipients in the model.

As an example, assume 'recipient need' was chosen to be an aid determinant by a particular donor. This variable can be measured in several ways. First, say it is measured by poverty levels per capita. This measure is assumed positively correlated with aid: the more people in poverty, then the more aid the recipient should attract, resulting in a relatively higher proportional output by Equation (3)—the recipient-specific weight—assuming the power parameter η^i_β is unity. The output from (3) is then an input into the preference Function (1), resulting in a higher percentage of aid to that recipient. On the other hand, if recipient need was instead measured using recipient GNI, the output of Function (3) would need to be subtracted from (1) because recipient GNI is assumed negatively correlated with aid received. The result of these two approaches in terms of the output by preference Function (1) should be roughly equal.

To allow for biases when allocating aid to certain recipients based on aid determinants, the power parameters μ^i_α and η^i_β on Functions (2) and (3), respectively, allow for a choice to be made by a donor with regards to the relative contribution and, thus, importance of particular determinants on the final aid allocations output by (1), which is in the form of percentages of the total aid budget. Donors can dial up or dial down the level of influence that their selected determinants have on the outcome by changing the values of the power parameters.

For example, if a donor wanted to allocate more aid to recipients with which it experiences large amounts of trade activity over those recipients with higher poverty levels, then the donor will choose a higher value for the power parameter applicable to the relevant functional equation that is quantifying trade activity. These parameters then, also provide a means for deducing historical donor behaviours and biases in simulations.

If a mathematical model is to be used by politicians, countries and organisations, then it needs to be simple, effective and able to be communicated. An important feature of this weighted network model is indeed its simplicity and transparency with the inputs into Equations (2) and (3) and, in turn, into preference Function (1), determined by verifiable, properly sourced, factual data.

The weighted network model's initial purpose was to reflect the decisions made by donors with regards to their motivations towards aid allocation based on certain factors, such as trade and recipient poverty. Donors can decide how much emphasis these factors have on the final allocation of their aid budgets. With historical data, sourced from the World Bank and OECD, input into Functions (2) and (3), and with the historical aid allocation figures which are the outputs from preference Function (1) also known, this leaves the power parameter values as the only unknowns. These values can be determined by playing the role of balancing figures and, from these estimated values, donor motivations and the relative importance placed on certain aid determinants can be studied.

2.2.2. Adaptation of the General Model

Before adaptation of the general weighted network model for use by an aid recipient, the model Functions (1)–(3) need to first reflect the analysis conducted in Section 2.1 and the network model in Figure 2. Therefore, three donors, three recipients and the four identified significant aid determinants are to be incorporated into the general model.

The set of donors, D, are Germany, the UK and the USA, respectively; $d_1, d_2, d_3 \in D$, each having its own preference function. The set, R, contains the three recipients: Bangladesh, Afghanistan and Ethiopia, respectively; $r_1, r_2, r_3 \in R$.

From analysis performed, trade relationships were found to be positively correlated with aid and bilateral trade activity was identified as one of the four significant aid determinants that a donor considers when allocating their aid budgets. For incorporation into the general model, trade activity is to be represented by the variable t^{ij}. For example, the superscript $i = 1$ identifies the donor as Germany, d_1, and superscript $j = 1$ represents the recipient Bangladesh, r_1. The trade ratio between a donor and its aid recipients, measured in terms of exports to the recipient in US$, was calculated from [13]. For example, for Germany, $t^{11} : t^{12} : t^{13} \equiv 13{:}1{:}5$ for the year 2019.

Recipient poverty was another significant aid determinant identified and is to be represented in the adapted model by p^j, which denotes the poverty levels in recipient r_j. The ratio of poverty levels is determined by using GNI per capita to quantify recipient need [9]. Correlation analysis indicated that this metric is negatively correlated with aid; therefore, there will be a subtraction from 1 in the poverty-specific functional input equation.

Another of the significant aid determinants identified was colonial relationships, represented by the variable c^{ij} between donor d_i and recipient r_j. This variable will be quantified using a binary zero-one integer programming variable defined as

$$c^{ij} = 1 + \begin{cases} 1, & \text{colonial relationship existed between } d_i \text{ and } r_j \\ 0, & \text{no colonial relationship} \end{cases} \tag{4}$$

The only colonial relationships of relevance to the simulation are Afghanistan and Bangladesh, both ex-UK colonies and protectorates, and thus $c^{21} = 2$ and $c^{22} = 2$; with $c^{ij} = 1$ for all other combinations of i and j, where $i, j = 1, 2, 3$.

The final significant aid determinant identified was the bandwagon effect, or herding, discussed in Section 2.1.5. Simply, it refers to the tendency of aid donors to follow other donors in allocating aid to certain recipients, who then gain 'star' status in the network. This can reveal itself when a recipient attracts a larger proportion of the total aid donated for no discernible reason, controlling for other factors; see [16]. The weighted network model framework can quantify the bandwagon effect, to be denoted b^j, by capturing the phenomenon using aid received by a recipient, r_j, in the previous period as a proportion of total aid donated by all donors in the entire network. This captures the herding effect by measuring recipients' previous success in receiving aid relative to other recipients, thus becoming a 'star' node in the model network.

The four determinants have now been allocated specific variables and associated data to be input into an adapted model. The aid allocated by the three donors to the three recipients in Figure 2 is to be used as the output values of the adapted model's preference function, sourced from [9]. Thus, the only remaining unknowns are the values of the four power parameters in the four input equations, each representing one of the four aid determinants. These parameter values can be estimated by running the model using the known inputs (the aid determinants) and known outputs (actual historical data) to provide important insights into donors' individual and relative motivations and behaviours with regards to allocating their aid budgets. The higher the power parameter value, the more bias has been baked into the aid allocation output from that aid determinant.

To use the model over multiple consecutive time periods, it needs to be made temporal. Starting at time $t = 1$, the weighted network model can be iterated forward in time with the outputs of the equations changing at each t due to the recipients' economic response to aid receipts, which feed back into donors' decisions on the allocation of aid at $t + 1$, acting as a feedback mechanism. For example, assume aid donated at time t led to a reduction in poverty in a recipient. By rolling the model forward to the next time-period, $t + 1$, this reduced level of poverty will be fed into the model at $t + 1$, producing a different aid allocation output percentage for the donor at $t + 1$ compared to t.

By denoting recipient poverty as p_t^j, trade relationships as t_t^{ij}, and the bandwagon effect as b_t^j where the subscript t represents the time-period, the model becomes dynamic

with respect to time. Note that colonial history, c^{ij}, quantified in (4) is a static measure: it does not change with time and, consequently, has no subscript.

After the alterations discussed, the preference Function (1) and input Functions (2) and (3) in the general model have been adapted to create Equations (5)–(9), with the four determinant functions in the preference Function (5) and four input Equations (6)–(9).

$$P^{ij}\left(c^{ij}, t_t^{ij}, p_t^j, b_t^j\right) := f_1^i\left(c^{ij}, c^{i\bullet}\right) \cdot f_2^i\left(t_t^{ij}, t_t^{i\bullet}\right) \cdot g_1^i\left(p_t^j, p_t^\bullet\right) \cdot g_2^i\left(b_t^j, b_t^\bullet\right) \tag{5}$$

$$f_1^i\left(c^{ij}, c^{i\bullet}\right) = \left[\frac{c^{ij}}{\sum_{k\in V_2} c^{ik}}\right]^{\mu_1^i} \tag{6}$$

$$f_2^i\left(t_t^{ij}, t_t^{i\bullet}\right) = \left[\frac{t_t^{ij}}{\sum_{k\in V_2} t_t^{ik}}\right]^{\mu_2^i} \tag{7}$$

$$g_1^i\left(p_t^j, p_t^\bullet\right) = \left[1 - \frac{p_t^j}{\sum_{k\in V_2} p_t^k}\right]^{\eta_1^i} \tag{8}$$

$$g_2^i\left(b_t^j, b_t^\bullet\right) = \left[\frac{b_t^j}{\sum_{k\in V_2} b_t^k}\right]^{\eta_2^i} = \left[\frac{\sum_{m=1}^3 A_{t-1}^{mj}}{\sum_{k=1}^3 \sum_{n=1}^3 A_{t-1}^{nk}}\right]^{\eta_2^i} \tag{9}$$

where A_{t-1}^{mj} in (9) is the amount of aid donated by donor m to recipient j in the previous period $t-1$, and the denominator of (9) quantifies the total aid donated within the network at period $t-1$.

The model is iterated forward starting from $t = 1$ (year 2015) to $t = 5$ (year 2019). Note that $t-1$ is the year 2014, for which actual aid data is input into Function (9). By iterating forward, the power parameters can be backward calculated for each year from 2015 to 2019. The values are shown in Table 2. Entries marked N/A for not applicable are included where the relationship was not relevant for the year in question.

Table 2. Power parameters required by the model to recreate the actual aid allocation results for each year 2015 to 2019 for donors Germany, UK and USA and recipients Afghanistan, Bangladesh and Ethiopia. Data that is not relevant is labelled N/A for not applicable.

Donor	Aid Determinant	2015	2016	2017	2018	2019
Germany	Colonial history	N/A	N/A	N/A	N/A	N/A
	Trade relationship	1.5	1.4	2.0	0.8	-
	Poverty	1.1	1.6	1.1	2.0	1.0
	Bandwagon	4.0	6.0	9.0	5.0	2.2
UK	Colonial history	-	-	0.5	-	-
	Trade relationship	0.4	0.4	0.4	0.1	-
	Poverty	-	-	0.6	0.2	0.1
	Bandwagon	1.0	1.0	0.6	0.3	0.1
USA	Colonial history	N/A	N/A	N/A	N/A	N/A
	Trade relationship	-	-	-	1.1	-
	Poverty	1.4	1.6	1.5	1.6	0.5
	Bandwagon	1.0	0.6	0.4	0.3	0.9

The values in Table 2 can be put into matrix form. For each year the model was iterated, the matrix of power parameter values was input into (6)–(9) and fed into (5) to create the next period's aid allocations. For example, for the year 2015, the matrix was the following:

$$\Phi_{\alpha}^{i} = \begin{pmatrix} \mu_1^1 & \mu_2^1 & \eta_1^1 & \eta_2^1 \\ \mu_1^2 & \mu_2^2 & \eta_1^2 & \eta_2^2 \\ \mu_1^3 & \mu_2^3 & \eta_1^3 & \eta_2^3 \end{pmatrix} = \begin{pmatrix} 0.0 & 1.5 & 1.1 & 4.0 \\ 0.0 & 0.4 & 0.0 & 1.0 \\ 0.0 & 0.0 & 0.4 & 1.0 \end{pmatrix} \tag{10}$$

The general weighted network model represented by Equations (1) to (3) has been tested, modified to Equations (5) to (9), and the relevant power parameters now calculated. All significant inputs and outputs for the years 2015 to 2019 are now known. Next, the model is adapted for use by an aid recipient by incorporating modern portfolio theory before performing a simulation to illustrate how that recipient could optimise their aid receipts.

2.2.3. An Aid Recipient's Investment Strategy Using Network Theory

If a recipient invests in increasing its trade activity with donors, then that recipient should expect an increase in aid receipts from those donors who place a relatively high value on commercial trade in their aid allocation decisions. This increase would then be compounded by the herding effect, leading to additional receipts that can be re-invested back into trade activity or other similar investment 'assets', creating a virtuous cycle of investment and increasing returns.

Paradoxically, recipients may not have an incentive to reduce poverty since it may lead to a fall in aid receipts. Instead, if recipients focus primarily on increasing trade activities, then their GNI should naturally increase and poverty should be reduced. This argument is limited however as it depends on other limiting factors such as the quality of governance and institutions in the recipient country. The fruits of increased trade activity may also fuel corruption rather than being devoted to alleviating poverty. Often, increased trade activity is performed by state-owned companies with the recipient's President as the main shareholder. Donors may wish to accommodate this in their aid decisions, which the weighted network model can do.

Despite these complications, the main interest here is regarding the ability of the weighted network model, illustrated by Figure 2 and Equations (5)–(9), to be used by a recipient as an investment tool to maximise their aid receipts.

By creating a foreign aid network model, a recipient would initially discover how influential it is in the network using centrality measures, the links it holds with donors and those that it does not. Specific weights can be added to links and nodes containing proportions of aid received, trade activity and other recipient–donor dyad information. This network model could also indicate if the recipient should seek out new donors, invest in current donors or a combination of the two.

Recipients can treat their aid network model much like a company seeking to attract funding. They could view the aid determinants used by donors as an 'asset portfolio', safeguarding and maximising the value of those assets by treating them as investments. Recipients can invest their aid income into the asset portfolio, for example by investing in trade relationships with donors. The recipient may also need to invest in other sub-activities such as governance quality and public relations activities, which the network model and portfolio can identify.

An investment plan for a typical aid recipient is outlined as follows:

Step (1): Create a weighted network model, providing insights into links, level of influence and current donors in the recipient's foreign aid network. Analyse each donor's aid determinant preferences, motivations and biases.

Step (2): Produce an asset portfolio representing the donor preferences identified, e.g., trade activity and poverty alleviation, with the USA being highlighted as a highly influential donor.

Step (3): Identify those assets in the portfolio that provide the highest returns, then invest in these. For example, the recipient could invest to increase trade activity with the USA, and in the related governance quality and infrastructure.

Step (4): The investment should lead to higher returns in the form of increased aid income, which is re-invested into the asset investment portfolio; e.g., increased trade activity with the USA should lead to further aid receipts donated by the USA, which then feeds back into the donor's aid allocation model for the following years. The herding phenomenon then compounds the effect.

Treating aid determinants like assets in a portfolio implies the existence of an optimal mix of such variables which provides maximum return for minimal risk. There are in fact two main models that can be used for asset portfolio analysis: Modern Portfolio Theory (MPT) and the Capital Asset Pricing Model (CAPM). The CAPM model is more robust with fewer inputs; whereas the MPT model, though elegant, loses some practicality from the attempt to find asset returns, volatilities and correlations.

Unfortunately, the CAPM model's principal purpose is for modelling and pricing equity market assets and their equivalents, where risk and returns are measured against some trade index such as the FTSE100. There is no equivalent transparently priced market for aid determinants and, therefore, the CAPM model cannot be used here. Instead, MPT is used to illustrate the concept using a simulation.

Let us assume that the portfolio for aid recipient r_j contains two controllable assets, $N = 2$, 'owned' by the recipient: trade activity, t_t^j, and poverty, p_t^j, at time subscript t, denoted in a set by

$$P_t^j = \left\{ t_t^j, p_t^j \right\} \tag{11}$$

These assets are 'investable' with varying risk-reward ratios and could be correlated or uncorrelated since increasing trade volumes do not always translate into reducing poverty, dependent on the recipient country and its regime as discussed earlier. For simplicity in this simulation, it is assumed that the assets are uncorrelated ($\rho = 0$); however, equations can be adapted for the case when the assets are correlated and the correlation coefficient $\rho \neq 0$, discussed in Section 3.1.

The mean and variance of the two-asset portfolio (11) can be written as

$$\mu_{P_t^j} = W\mu_{t^j} + (1 - W)\mu_{p^j} \tag{12}$$

$$\sigma^2_{P_t^j} = W^2\sigma^2_{t^j} + 2W(1 - W)\rho_{t_t^j, p_t^j}\sigma_{t^j}\sigma_{p^j} + (1 - W)^2\sigma^2_{p^j} \tag{13}$$

with the correlation between the assets subject to the constraint $-1 \leq \rho_{t_t^j, p_t^j} \leq 1$.

In (12) and (13), $W \in [0,1]$ is a parameter that determines the proportion of aid receipts invested in trade activity, i.e., W is the weight of the trade activity 'asset', t_t^j, in the portfolio. The weight on the poverty alleviation 'asset', p_t^j, must be $1 - W$, because

$$\sum_{i=1}^{N} W_i = 1 \tag{14}$$

Further, if the two assets are uncorrelated, then $\rho_{t_t^j, p_t^j} = 0$ and the variance (13) becomes

$$\sigma^2_{P_t^j} = W^2\sigma^2_{t^j} + (1 - W)^2\sigma^2_{p^j} \tag{15}$$

The value of parameter W is important since the mean and standard deviation of returns of each asset should technically be known. As W is varied, the risk and reward dynamics of the portfolio change in response.

The minimum variance of the portfolio ('risk-minimising portfolio') is calculated by setting the first derivative of (15) to zero:

$$\frac{\partial \sigma^2_{P^j_t}}{\partial W} = 2W\sigma^2_{t^j} + 2(1-2W)\rho_{t^j_t,p^j_t}\sigma_{t^j}\sigma_{p^j} - 2(1-W)\sigma^2_{p^j} := 0 \tag{16}$$

from which the value of W at the minimum can be found by (17):

$$W = \frac{\sigma^2_{p^j} - \rho_{t^j_t,p^j_t}\sigma_{t^j}\sigma_{p^j}}{\sigma^2_{t^j} + \sigma^2_{p^j} - 2\rho_{t^j_t,p^j_t}\sigma_{t^j}\sigma_{p^j}} \tag{17}$$

For uncorrelated assets, (17) can be simplified to

$$W = \frac{\sigma^2_{p^j}}{\sigma^2_{t^j} + \sigma^2_{p^j}} \tag{18}$$

3. Results

Let us start with an initial amount to invest; for simplicity, assume recipient r_j invests 90% of total aid it receives with the remaining 10% lost to errors, corruption and aid spillage. Then, investment of these aid receipts in portfolio P^j is defined by $0.9W$ in asset t^j and $0.1(1-W)$ in asset p^j.

It is assumed that recipient r_j is a rational investor and aims to maximise portfolio returns for the minimum risk (variance). It is further assumed that investment in its trade activity asset results in an increase in trade levels with donors, and an investment in its poverty alleviation asset at least maintains the current poverty level due to rising populations. Additionally, an investment in trade is assumed to be a risky investment, since money may be lost in the process, and poverty alleviation is deemed relatively risk-free since if it does not work, aid receipts should continue at the current level.

The values of the means, μ, and standard deviations, σ, for both portfolio assets to be input into the MPT Equations (12)–(18) can be calculated from the asset returns for each period, t. Returns on assets are usually calculated by taking the difference between the current asset value and the previous period's asset value, the periodic asset income, and dividing by the previous period's asset value, i.e., for the trade activity asset:

$$R_{t^j_t} = \frac{t^j_t - t^j_{t-1}}{t^j_{t-1}} \tag{19}$$

Such returns can be calculated for each time-period, t, starting at $t = 1$. The mean and standard deviations of these returns can then be calculated.

However, using (19) is too simplistic for this simulation since trade volumes and poverty levels change substantially each period for many reasons, not solely due to any 'return' on an investment in these assets. The return arising purely from investment in these assets needs to be isolated from any additional 'noise' which may be causing their values to change. A recipient could isolate this 'pure' return using knowledge of the dynamics of their trade activity and poverty levels, controlling for the impact from any other variables. For the purposes of this simulation, it is assumed that this has been done by the chosen recipient, Ethiopia, resulting in the figures in Table 3.

Table 3. Parameters relating to Ethiopia's investment in its portfolio of assets.

μ_{t^j}: Return on Trade	σ_{t^j}: Risk of Trade	μ_{p^j}: Return on Poverty	σ_{p^j}: Risk on Poverty
2.0	0.5	1.0	0.2

In the MPT model Formulas (11)–(18), set recipient r_j to be Ethiopia, $j = E$. It is also to be assumed that the portfolio asset returns in (11) are uncorrelated, $\rho_{t_t^E, p_t^E} = 0$, and therefore Equations (12), (15) and (18) are applicable to this simulation. Inputting the values from Table 3 into these equations creates a line in risk-return space, parameterised by W, which can be plotted, thus sketching out a hyperbola as W is varied. See Figure 3.

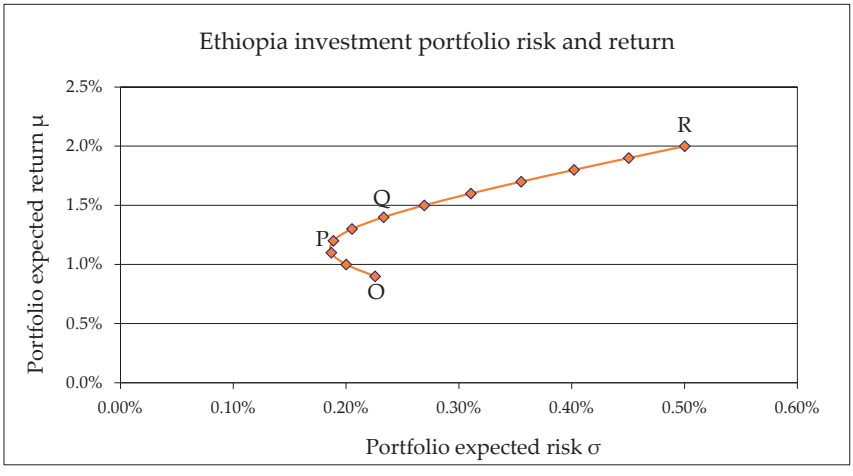


Figure 3. Ethiopia's investment portfolio risk and return profile for two uncorrelated assets. A hyperbola is created from varying the relative weights of the portfolio assets resulting in different risk and reward profiles. The top of the hairpin, PQR, is the efficient market frontier containing optimised portfolio asset weights. The point O represents the portfolio with the lowest return.

PQR of the concave function in Figure 3 is termed the efficient frontier. Choosing a portfolio mix that fits this line results in an optimum portfolio from a risk vs. reward perspective. Ethiopia could choose a portfolio mix anywhere in the risk-reward space in Figure 3. The efficient frontier identifies the possible portfolios that have the highest return for the least possible risk for that return in this risk-reward space.

In general terms, the efficient frontier contains portfolios which mathematically can be defined as solving

$$\sigma_{\mathrm{P}} = min \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} W_i W_j \rho_{ij} \sigma_i \sigma_{\mathrm{j}}} \tag{20}$$

subject to the constraints

$$\mu_{\mathrm{P}} = \sum_{i=1}^{N} W_i \mu_i \tag{21}$$

$$\sum_{i=1}^{N} W_i = 1 \tag{22}$$

Ethiopia's risk preferences will dictate where it wants to be on the curve in Figure 3. Clearly, Q is a better portfolio mix than O since they share similar levels of risk, for which Q offers the higher reward. A rational investor will always choose portfolio Q over O.

The values in Table 3 imply that investing in trade activity is higher risk, but provides a higher return, than poverty. If Ethiopia wanted to maximise return, it would choose portfolio R; or to minimise risk, portfolio P would be the best option. It depends on the recipient's preferences and risk appetite.

If Ethiopia wanted to minimise risk, the minimum variance portfolio (MVP) should be targeted, which is calculated using Equation (18) as $W = 13.8\%$. This means that Ethiopia,

with the risk and return characteristics in Table 3, should invest 13.8% of its 90% aid income into trade activities, with the remaining invested in poverty alleviation, producing an expected portfolio return of $\mu_{P_t^j} = 1.14\%$ and a portfolio risk of $\sigma_{P_t^j} = 0.19\%$,

Or, Ethiopia can choose portfolio R, maximizing both return and risk, for which it would invest all its available aid receipts into trade activity and nothing into poverty alleviation. There is no correct answer as to which portfolio mix Ethiopia should invest in, except that it should be one lying on the efficient frontier, PQR, in Figure 3.

For the purposes of simulating this model, assume Ethiopia decides to invest 50:50, so $W = 0.5$. This means the portfolio chosen is one to the right of portfolio Q in Figure 3 on the efficient frontier. This portfolio has a return of 1.5%, with 1% originating from the investment in trade activity, $W\mu_{tj} = 0.5 \times 2$ and 0.5% from poverty reduction $(1 - W)\mu_{pj} = 0.5 \times 1$, leading to an overall portfolio risk of 0.27%.

Next, the weighted network model (Equations (5)–(9)) is adapted to incorporate the described investment portfolio, with the year 2015 being $t = 1$, and assuming the aid receipts for Ethiopia in 2014 were invested in trade and poverty assets for the year 2015 in accordance with the optimal portfolio mix. Then, using the inputs and data as described, including the values of the power parameters from Table 2, the model is iterated forward.

Throughout the simulation, the amount of aid in US$ donated in the years 2015 to 2019 by the three donors was fixed, as were the data fed into the functional input equations relating to Bangladesh and Afghanistan. Simply, the simulation adapts the model by using the MPT approach applied to Ethiopia only, which will result in different aid allocation percentages for all recipients for each period, compared to actual historical receipts. These updated allocations are fed back into the model at each annual iteration. By keeping all else fixed, the impact of an investment portfolio approach by Ethiopia, which will affect all recipients' aid receipts, can be isolated.

Results of the simulation are presented in Figure 4, demonstrating the impact of Ethiopia investing its aid donations into trade activity and poverty alleviation in an optimum portfolio.

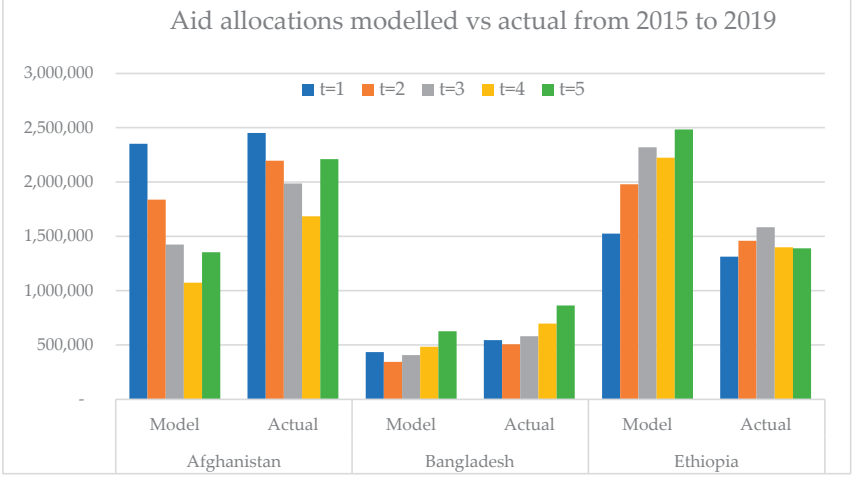


Figure 4. Aid allocations from 2015 $(t = 1)$ to 2019 $(t = 5)$, modelled by simulating the adapted model using Equations (5)–(9), adjusted for MPT, and comparing to actual aid donated. The model results incorporate Ethiopia's aid investments in an optimum portfolio.

Figure 4 demonstrates that if Ethiopia had taken the MPT approach in 2015, as detailed above, then its aid receipts over the period 2015 to 2019 would be 48% higher than they were as a result of increasing trade activity further compounded by the bandwagon effect. This

would come at the loss of aid receipts experienced by Bangladesh of 28% and Afghanistan of 24%.

Figure 5 shows the aid donated by each donor to each recipient in 2019 ($t = 5$), including the total, demonstrating that Ethiopia is the 'winner' in terms of aid donations from all donors, although the UK allocations have not altered much because its parameters in Table 2 for 2019 indicate close to a uniform allocation per aid determinant, i.e., low biases for these aid determinants were shown by the UK in 2019 bordering on ambivalence.

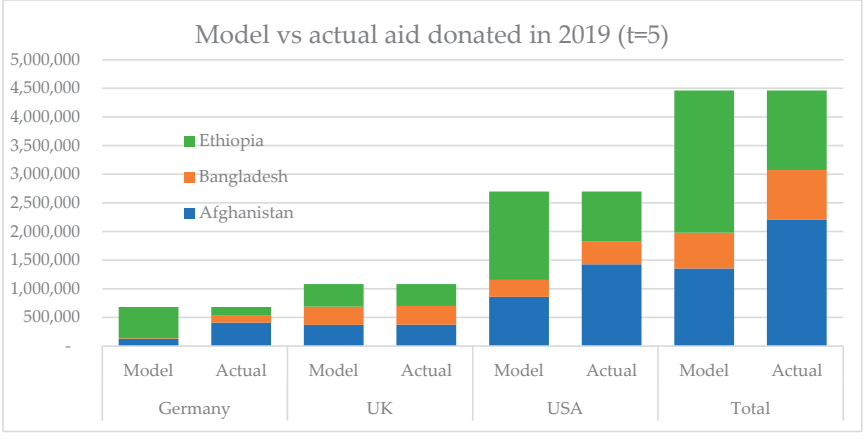


Figure 5. Aid allocations modelled vs. actual in 2019 ($t = 5$) by donor to each recipient.

Figure 6 below shows the total share of aid allocations from 2015 ($t = 1$) to 2019 ($t = 5$) using pie charts set side by side, with modelled aid donations (ODA) on the left and actual donations on the right. Ethiopia's share would have increased from 34% actual to 50% as modelled, if it had invested in trade activity and poverty alleviation in 2015 using MPT, creating an optimum portfolio of these assets, further magnified by the herding phenomenon captured by the model.

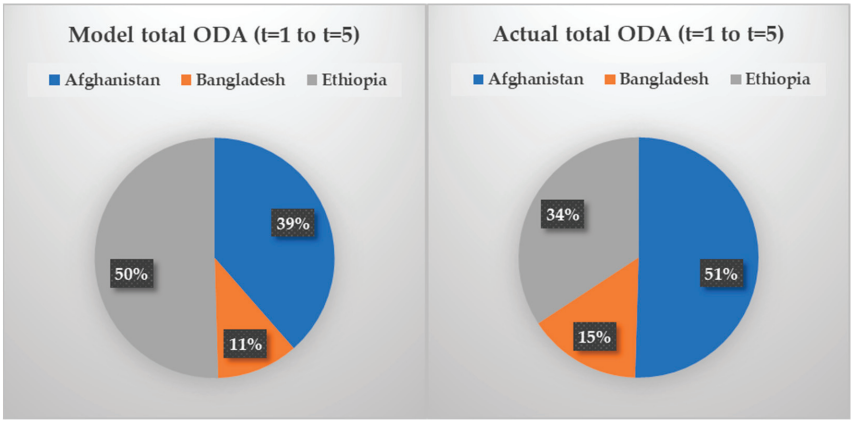


Figure 6. Pie charts of the total aid allocation percentages modelled and actual for the period from 2015 ($t = 1$) to 2019 ($t = 5$) per aid recipient.

This simulation has clearly demonstrated how the weighted network model can be adapted and incorporate MPT to be used as an investment tool by recipients to maximise their aid income.

3.1. Correlated Portfolio Assets

In the MPT simulation, it was assumed for simplicity of demonstration that the assets in Ethiopia's portfolio were uncorrelated. However, what if they were in fact correlated? Assuming high governance quality and no corruption, increasing trade activity experienced by Ethiopia would be expected to reduce its poverty levels. Hence the returns from two assets, trade and poverty, assumed in the simulation should be negatively correlated.

Using MPT, the creation of an optimum portfolio follows the same process as detailed earlier, except now with a value for the correlation function, ρ, included in the Equations (12), (13) and (17). Assuming $\rho_{t^E,p^E} = -0.8$, defining strong negative correlation between the trade and poverty assets in the portfolio, and using the variables as defined in Table 3, the same process performed in the simulation should be followed.

The presence of correlation produces a different efficient frontier curve than when there is no correlation, as shown in Figure 7 and compared to Figure 3. The shape of the curve has a large influence on the optimal portfolio.

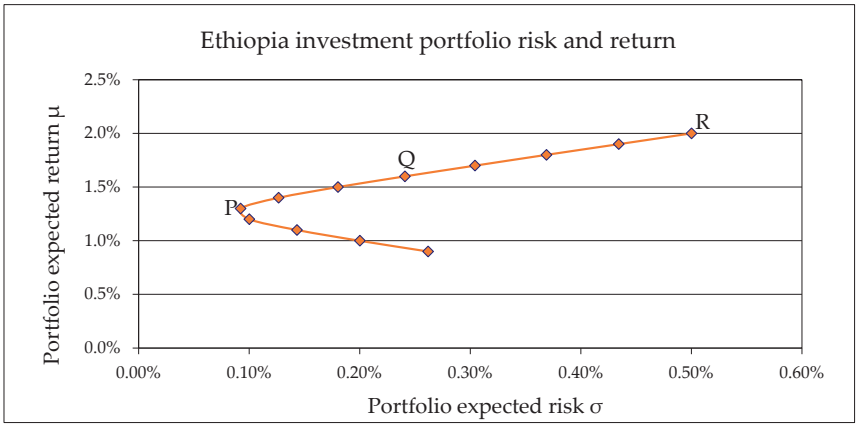


Figure 7. Ethiopia investment portfolio risk and return. Compared to Figure 3, the presence of asset correlation produces a different risk and reward profile, providing different optimum portfolio combinations.

The portfolio at point P, in Figure 7, is the minimal variance portfolio (MVP) for which the weight W is calculated using Equation (17), with values from Table 3 and $\rho_{t^E,p^E} = -0.8$:

$$W = \frac{\sigma_{p^E}^2 - \rho_{t^E,p^E}\sigma_{t^E}\sigma_{p^E}}{\sigma_{t^E}^2 + \sigma_{p^E}^2 - 2\rho_{t^E,p^E}\sigma_{t^E}\sigma_{p^E}} = 27\% \tag{23}$$

Using $W = 27\%$, the MVP has an expected return of 1.27% and portfolio risk of 0.09%.

The effect of correlation on portfolio volatility (which can be measured either by the standard deviation of portfolio returns, as here, or the variance) for this asset portfolio can be seen in Figure 8 showing a monotonic increasing function with an upper and lower bound of volatility when the asset correlation is 1 and -1, respectively. The function plotted is (24) using the inputs from Table 3, Equation (23), and $\rho_{t^E,p^E} = -0.8$ (c.f. Equation (13)).

$$\sigma_{pE} = \sqrt{W^2\sigma_{t^E}^2 + 2W(1-W)\rho_{t^E,p^E}\sigma_{t^E}\sigma_{p^E} + (1-W)^2\sigma_{p^E}^2} \tag{24}$$

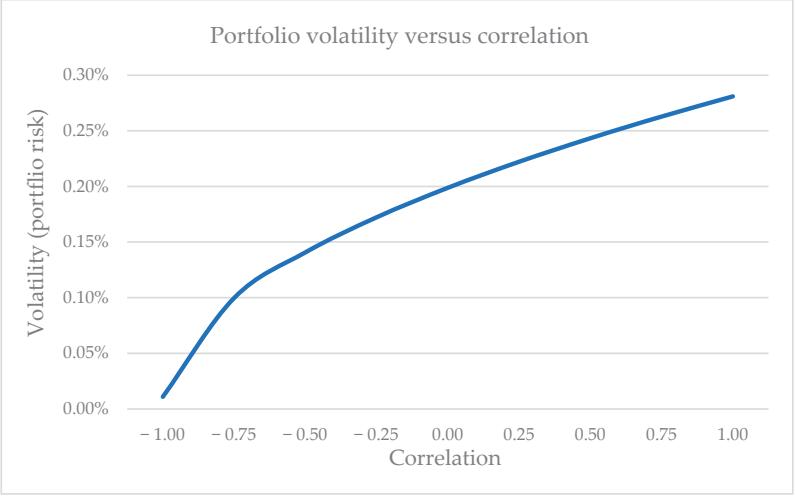


Figure 8. Volatility of the portfolio against correlation of the assets in the portfolio.

4. Conclusions and Discussion

This paper explores the ability and effectiveness of a mathematical model, grounded in network theory, to capture the properties, dynamics and inter-dependencies inherent in foreign aid networks, and replicate a variety of donor and recipient behaviours. By doing so, this progresses the narrative around the analysis of foreign aid, illustrating how mathematics can be used to reveal useful features and intricate properties of real-world foreign aid networks.

Until now, regression analysis has been the dominant method used to analyse donor behaviour ([5–7]). However, here, data analysis of real-world foreign aid donations identifies the key parties involved and reveals their complex interactions. The network model developed can then also be used to investigate historical donor behaviour, akin to regression analysis, informing on the relative values of aid determinants used and their contribution to the donor's final aid allocation decision. Nation donors do not publicise many of the determinants used in their allocation decisions; hence the data analysis described and performed was vital to identify the relevant variables and their interdependencies to provide input into this model. Furthermore, the parameter values μ_α^i and η_β^i contain information that reflect past motives and biases of a donor, and the relative importance the donor places on certain aid determinants.

The model was demonstrated to be flexible and adaptable enough to be used by aid recipients and donors. As an example of the model's use, it was shown how Ethiopia could create a portfolio of assets based on this recipient's determinants and apply modern portfolio theory to maximise aid income.

The approach can potentially be used by donors to replicate the properties of their foreign aid network and apply weights which control the allocation of their aid budget according to their own motivations and biases. Significantly, the weights that a donor would use is then explainable to the public, providing a transparent means of communication for politicians to justify their motivations behind their aid allocation decisions, and the model can also be iterated forward in time enabling a feedback mechanism to occur, in which donors (and recipients) can see the impact of their decisions on future aid allocations. Moreover, a donor could also treat the model like an investment tool, requiring a certain level of 'return' on their aid donations that can be quantified. To do so, donors would first create their specific foreign aid network model using the network science tools described earlier. The next step would be to adapt the model Equations (1)–(3) to include chosen determinants and their parameter values which reflect their motivations and the

investments they want to make. An obvious one is trade activity: donors can use the model to ensure that more aid is allocated to countries that provide higher levels of trade activity with the donor, such as that seen between Germany and China [17].

Finally, by providing a framework to explore the properties of foreign aid networks and the impact that decision variables will have on those properties, including the final aid allocations, the weighted network model can help donors and recipients, and potentially multilateral organisations, with one of the issues associated with foreign aid: aid spillage, by reducing the costs arising from inappropriate use of foreign aid budgets.

In conclusion, the weighted network model, underpinned by network theory, has been demonstrated to successfully model the international aid system and is able to shed new light on the complexity and interactions inherent in foreign aid networks.

Author Contributions: Writing—review and editing, S.B. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data regarding aid donations are obtained from reference [1].

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

References

1. OECD. Development Finance Data. 10 June 2021. Available online: https://www.oecd.org/dac/financing-sustainable-development/development-finance-data/ (accessed on 27 February 2023).
2. Ramalingam, B. *Aid on the Edge of Chaos*; Oxford University Press: Oxford, UK, 2011.
3. Schraeder, P.J.; Hook, S.W.; Taylor, B. Clarifying the Foreign Aid Puzzle: A Comparison of American, Japanese, French, and Swedish Aid Flows. *World Politics* **1998**, *50*, 294–323. [CrossRef]
4. Downes, R.J.; Bishop, S.R. Aid Allocation: A Complex Perspective. In *Global Dynamics: Approaches from Complexity Science*; Wiley: New York, NY, USA, 2016; pp. 271–290.
5. Alesina, A.; Dollar, D. Who gives foreign aid to whom and why? *J. Econ. Growth* **2000**, *5*, 33–63. [CrossRef]
6. Harrigan, J.; Wang, C. A New Approach to the Allocation of Aid Among Developing Countries: Is the USA different from the Rest? *World Dev.* **2011**, *39*, 1281–1293. [CrossRef]
7. McGillivray, M. *Modelling Aid Allocation: Issues, Approaches and Results*; WIDER Discussion Papers/World Institute for Development Economics: Helsinki, Finland, 2003.
8. Swiss, L. Foreign Aid Allocation from a Network Perspective: The Effect of Global Ties. *Soc. Sci. Res.* **2016**, *63*, 111–123. [CrossRef] [PubMed]
9. World Bank. June 2021. Available online: https://data.worldbank.org/indicator (accessed on 27 February 2023).
10. OECD. Data Visualisations. 6 June 2021. Available online: https://www.oecd.org/dac/financing-sustainable-development/datavisualisations/ (accessed on 27 February 2023).
11. Development Initiatives. Aid Data 2019–2020: Analysis of Trends before and during COVID. 8 February 2021. Available online: https://devinit.org/resources/aid-data-2019-2020-analysis-trends-before-during-covid/#section-1-3 (accessed on 27 February 2023).
12. OECD. DAC List of ODA Recipients. 5 June 2021. Available online: https://www.oecd.org/dac/financing-sustainable-development/development-finance-standards/daclist.htm (accessed on 27 February 2023).
13. WITS. World Integrated Trade Solution. 2021. Available online: https://wits.worldbank.org/ (accessed on 27 February 2023).
14. Hoeffler, A.; Outram, V. Need, Merit or Self-Interest-What Determines the Allocation of Aid? University of Oxford: Oxford, UK, 2008.
15. Bikhchandani, S.; Sharma, S. IMF. 1 March 2000. Available online: https://www.imf.org/external/pubs/ft/wp/2000/wp0048.pdf (accessed on 27 February 2023).
16. Frot, E.; Santiso, J. Herding in Aid Allocation. *KYKLOS* **2011**, *64*, 54–74. [CrossRef]
17. Karnitschnig, M. Facing China. 10 September 2020. Available online: https://www.politico.eu/article/germany-china-economy-business-technology-industry-trade-security/ (accessed on 27 February 2023).

entropy

MDPI

Article

Multi-Criteria Analysis of Startup Investment Alternatives Using the Hierarchy Method

Tamara Kyrylych * and Yuriy Povstenko

Department of Mathematics and Computer Sciences, Faculty of Science and Technology, Jan Dlugosz University in Czestochowa, al. Armii Krajowej 13/15, 42-200 Czestochowa, Poland
* Correspondence: t.kyrylych@ujd.edu.pl

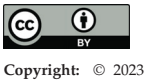
Abstract: In this paper, we discuss the use of multi-criteria analysis for investment alternatives as a rational, transparent, and systematic approach that reveals the decision-making process during a study of influences and relationships in complex organizational systems. It is shown that this approach considers not only quantitative but also qualitative influences, statistical and individual properties of the object, and expert objective evaluation. We define the criteria for evaluating startup investment prerogatives, which are organized in thematic clusters (types of potential). To compare the investment alternatives, Saaty's hierarchy method is used. As an example, the analysis of three startups is carried out based on the phase mechanism and Saaty's analytic hierarchy process to identify investment appeal of startups according to their specific features. As a result, it is possible to diversify the risks of an investor through the allocation of resources between several projects, in accordance with the received vector of global priorities.

Keywords: multi-criteria analysis; criteria composition; investment; startup; Saaty's method; global priority vector; choosing alternatives

Citation: Kyrylych, T.; Povstenko, Y. Multi-Criteria Analysis of Startup Investment Alternatives Using the Hierarchy Method. *Entropy* **2023**, *25*, 723. https://doi.org/10.3390/e25050723

Academic Editors: Stanisław Drożdż and Panos Argyrakis

Received: 24 February 2023
Revised: 7 April 2023
Accepted: 24 April 2023
Published: 27 April 2023



1. Introduction

The positive tendencies towards economic development require updated business entities according to the current market conditions and the emergence of new structural units, all of which form a competitive economic system. The active development of any economy is not possible without the constant emergence of new economic enterprises. This process stimulates the formation of the market environment with healthy competition and ensures scientific and reproducible functioning. Currently, we observe the positive tendency towards building potential for realizing business ideas through the creation of startups, whose business concepts have been dictated by the needs of the modern society and industries. A startup is a strategic economic unit with innovative concepts with the potential to enter the market. First, we outline the essential features of startups:

(1) the innovation of an idea;
(2) the necessity of capital investment;
(3) reproducibility (possibility to sell the inventive solution multiple times);
(4) business expansion;
(5) the existence of a detailed and structured business plan;
(6) generally, a startup is a project in initial stages of implementation;
(7) the possibility of significant growth of the project;
(8) often, startups propose new technologies;
(9) uniqueness;
(10) the potential team of professionals;
(11) the riskiness of the investments;
(12) the concentration of management decisions by the startup founders;
(13) the flexibility as well as quick and efficient adaptation to changes in the environment;

(14) the possibility to individualize the products, according to the demands of consumers;
(15) the dependence on credit resources;
(16) the close relations between the founder and the employees, etc.

Currently, one of the biggest problems is finding investors for startups, the qualified and objective evaluation of the concepts in terms of costs and benefits for future investment, and the successful presentation of the project to investors. Often, this work is entrusted to consulting agencies that professionally evaluate innovative ideas. For the investor, it is important to have a final estimation containing not only a list of factors justifying the appropriateness of investments in the suggested startups but also the method used for comparing several investment alternatives. For an objective and comprehensive assessment of a startup, a large number of criteria should be taken into account; however, this complicates the evaluation process and prolongs its execution. To assess investment alternatives, many methods, mechanisms, techniques, and tools enable the investigation of investments from different points of view. Research has been concentrated in several directions: the economic basis of startups, the mechanisms of their initiation, and the behaviors of investors. The most substantiated and successful in practice are mathematical models that predict the best investment alternatives. Based on the startup founder's viewpoint, a comprehensive analysis of investment alternatives should involve the requirements from the idea to launch, from the gathering and successful use of information to the potential of the startup's innovation in a functioning market. This step-by-step mechanism for building a business was precisely outlined in [1].

The behaviors of investors (especially, business "angels") towards newly created enterprises in the early stages of their development, the ways of evaluating those enterprises, and the interactions of investors and entrepreneurs were described in [2,3]. The basics of practical venture capital management and the details of the cooperation of venture capitalists and entrepreneurs were presented in [4]. Practical advice and the confirmation of the importance of a correct, accurate assessment of the business opportunities of startups were given in [5]. An analysis of venture capital from the viewpoint of current and future investing in an uncertain environment and the high level of competition confirms complexity of the investment choice [6].

An important step towards identifying the most attractive startup for investment involves not only formulating the list of criteria but also establishing their importance (weights). Today, many consulting companies use expert assignment methods to identify the weights of the criteria, but sometimes, the methods are too subjective and dependent on the composition of the expert team, the expert engagement, and lobbying interests. In this area, special attention is paid to the decision-making theory and the Saaty hierarchy method. The multi-criteria decision-making analysis, known as the analytic hierarchy process, was elaborated by Saaty [7–13]. This approach has been applied to many areas, such as economics, management, engineering, mathematics, information systems, cybernetics, mechanics, design, chemistry, health service, etc. The literature on this subject is considerable, including the following books [14–19] and review articles [20–33], where additional references can be found.

The choice and the comparison of the criteria are important parts of decision-making. As the criteria and their weights can significantly influence decision-making, several approaches to solve this problem have been elaborated. In the analytic hierarchy process (AHP), several prioritization methods have been used for deriving weights, such as the eigenvalue (EV) method [8,10,34,35], the logarithmic least squares (LLS) method [36,37], the weighted least squares (WLS) method [38,39], the fuzzy preference programming (FPP) method [40–43], and the cosine maximization method (CMM) developed in [44]. A good description of several of the most-used methods was given by Srdjevic [45]. The main feature of the step-wise weight assessment ratio analysis (SWARA) [46,47] is the possibility to estimate the opinions of experts and interested groups according to the significance ratio of the criteria in the process of their weight determination. In the best–worst method (BMW) [48–50], two vectors of pair-wise comparison were used to determine the weights

of the criteria. The full consistency method (FUCOM) [51–53] is based on the pairwise comparison of the criteria and the satisfaction of the mathematical transitivity conditions. The level based weight assessment (LBWA) model [54,55] is suitable for use in complex multi-criteria models with a large number of criteria, and it allows for the additional corrections of the values of the weight coefficients, depending on the preferences of the decision-makers.

The main purpose of this article is to provide a comparison of several startups from the investor viewpoint. In this paper, we discuss the use of a multi-criteria analysis for investment alternatives as a rational, transparent, and systematic approach that reveals the decision-making process during the study of influences and relationships in complex organizational systems. The proposed methods can be useful for consulting agencies, investors, and also for startups founders, who can then assess their competitive position against offers from other competitors in the selected economic branch or industrial sector. The procedure of the startup assessment, especially during the initial stages of implementation (development, operation, execution phases, etc.) is often subjective and challenging, as it requires the determination and account of many indexes as well as extended expert consultation, the formation of criteria, and so on. We propose new criteria and a new criteria composition for evaluating the investment appeal of startups. As an example, we consider three alternative investments in startups: the production of LED traffic lights, the manufacture of information–reference electronic terminals, and the manufacture of rotor-reactive turbo-rotational heaters of liquids. The analysis of the three startups is carried out based on the phase mechanisms and Saaty's analytic hierarchy process to identify the investment appeal of the startups accounting for their specific features. The consistency index, the consistency ratio, and the global priority vector are calculated. As a result, it is possible to diversify the risks of an investor through the allocation of resources among several projects, in accordance with the calculated vector of global priorities.

2. Criteria Composition for Evaluating Investment Attractiveness of Startups

Based upon the review of the literature, the study of the practice of founding and launching startups, successful experiences of investing in startup enterprises, and the results of our previous research, we suggest the following criteria composition, which are consolidated into 12 blocks (Table 1). Similar grouping of sub-criteria into blocks was considered, for example, in [41]. We used several of the block-criteria discussed in [56–59], and then we supplemented and extended these according to our own criteria.

This criteria could be adjusted according to the economic branch or industrial sector, according to the special features of the business plans presented to the investor. The criteria allow us to analyze the characteristics of startups in a variety of ways, and grouping the proposed criteria could enable potential investors to predetermine the priority groups of the criteria and use the proposed "sketch" of the influential factors to focus attention on the current trends. This criteria-composition model aims to draw the attention of the researcher (investor, consultant) not only on the "classical" list of basic investment indicators (such as payback period and the value of investments) but also to the governmental support of the industry, the innovation and autonomy of startups, time, and resources, as well as the social, scientific, technical, informational, and environmental characteristics.

Table 1. Criteria composition for evaluating investment appeal of a startup.

No	Type of Potential	Criteria
1.	Financial strength	1. Value of investment 2. Payback period (PBP) 3. Expected profitability 4. Risk level 5. Full or partial investor control of the startup 6. Possibility of reverse repurchase (RRP) 7. Possibility of tranche-funding, depending on the stage of the project
2.	Product/service potential	8. Availability of samples or models of the product
3.	Marketing potential	9. Startup position in the market 10. Forecasted level of demand for the product/service 11. Level of competition in the economic branch or industrial sector 12. Evaluation of startup competitiveness 13. Significant target audience 14. Availability of marketing strategy 15. Requirements to attract and interact with customers within the startup initial stage
4.	Organizational potential	16. Availability of organizational plan
5.	Scientific and technical potential	17. Innovation of idea 18. Innovation of technology 19. Availability of project plan for technical realization 20. Availability of intellectual property rights
6.	Staff potential	21. Availability of potential specialists 22. Uniqueness of specialists
7.	Potential of the governmental, international, economic, and political situation	23. The level of development of economic branch or sector in which the startup will operate 24. The level of governmental support of industry branch
8.	Time potential	25. Period of project completion 26. Stage of project development 27. Duration of product introductory period/start of retail service
9.	Autonomy potential	28. Dependence of the startup on other economic branches or industrial sectors 29. Dependence of the startup on other similar projects
10.	Ecological potential	30. Level of negative impact on the environment
11.	Social potential	31. Accessibility of project's social utility
12.	Information potential	32. Availability, reliability, and quality of information in economic branch or industrial sector in which the startup will operate

Figure 1 presents the structure of the Saaty method as the operational algorithm, indicating the priority of investments in startups.

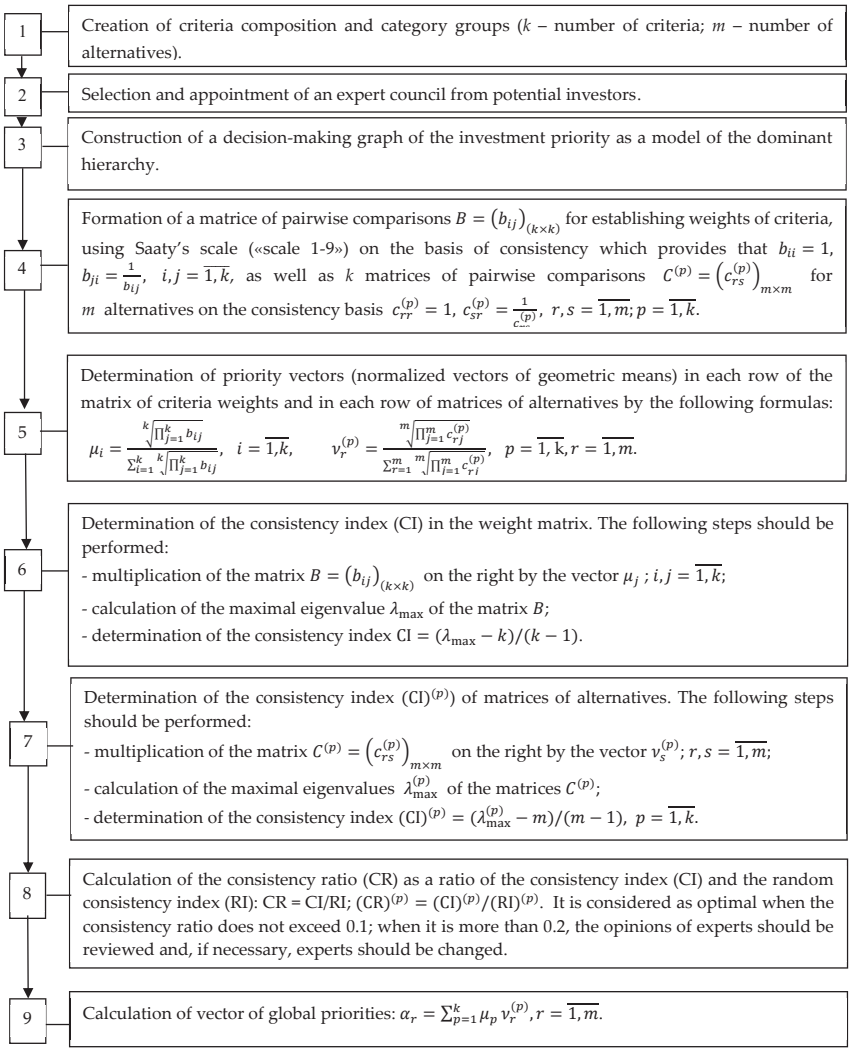


Figure 1. Saaty's analytic hierarchy process for the identification of the investment appeal of startups based on their specific features.

3. Implementation of the Saaty Method for Identified Criteria Composition

To illustrate practically the Saaty method, we analyze three investment alternatives of startups: the production of LED traffic lights, the manufacture of information–reference electronic terminals, and the manufacture of rotor-reactive turbo-rotational heaters of liquids. The structure of the method is first presented as the dominant hierarchy model in an oriented graph (Figure 2).

After considering the business plans of three investment alternatives and establishing the criteria for assessing the prerogatives of investing in the compared startups, we identified the investment priorities. First, we determined the weights of the criteria according to the sequence of the algorithm; this was the fourth step of the hierarchical procedure, as shown in Figure 1. Table 2 presents the results of the criteria comparison for evaluating the startups using Saaty's scale ("scale 1–9") [60,61]. Therefore, we obtained the matrix of pairwise comparisons for establishing the weights of the criteria. The numbers 1–12 in

the top row and the first column correspond to the name of the criteria in Table 1. The priority vector (μ_i) is calculated as the normalized geometric means in accordance with step 5 (see Figure 1). The column RM presents the results of the multiplication of the paired comparison matrix B_{ij} on the right by the vector μ_j. The column DV is obtained by dividing the component of the vector in the column RM by the corresponding component of the vector μ_j. The approximation of the maximal eigenvalue is calculated as the arithmetic mean of the components of the vector in the column DV and equals $\lambda_{\max} = 12.72$. The consistency index $\mathrm{CI} = (12.72 - 12)/11 = 0.06545$. According to [7], for $k = 12$, the random consistency index $\mathrm{RI} = 1.48$; therefore, the consistency ratio is $\mathrm{CR} = \mathrm{CI}/\mathrm{RI} = 0.06545$ and does not exceed 0.1.

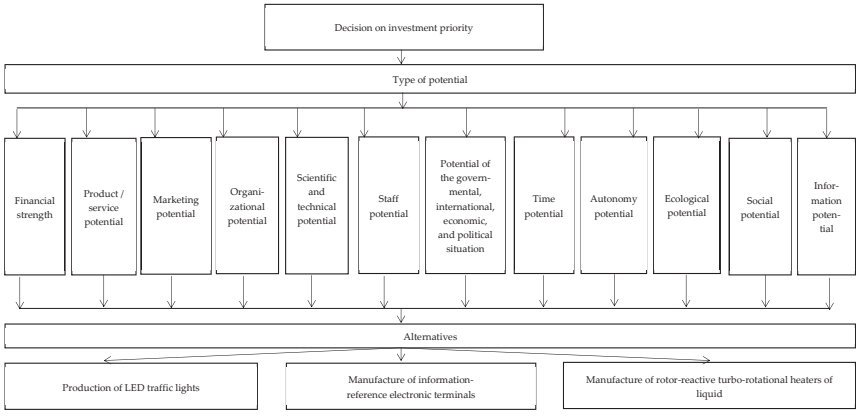


Figure 2. The dominant hierarchical representation of the problem of choosing investment alternatives in startups.

Table 2. The matrix of pairwise comparisons to determine the validity of 12 groups of criteria.

Groups of Criteria	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	μ_i	**RM**	**DV**
1	**1**	3	3	4	2	1	2	2	4	5	4	3	0.1893	2.3312	12.31
2	1/3	**1**	1/2	1	1/2	1/2	1	1/2	3	3	3	4	0.0800	1.1099	13.87
3	1/3	2	**1**	2	1	1/2	1	1	2	2	2	2	0.0911	1.1345	12.45
4	1/4	1	1/2	**1**	1/2	1/5	1	1/2	1	1	1	1	0.0490	0.6099	12.45
5	1/2	2	1	2	**1**	1	2	1	2	2	2	2	0.1058	1.2968	12.26
6	1	2	2	5	1	**1**	1	2	2	2	2	2	0.1282	1.6571	12.93
7	1/2	1	1	1	1/2	1	**1**	1	1	1	1	1	0.0666	0.8525	12.80
8	1/2	2	1	2	1	1/2	1	**1**	2	2	2	2	0.0942	1.1661	12.38
9	1/4	1/3	1/2	1	1/2	1/2	1	1/2	**1**	1	1	1	0.0483	0.5950	12.32
10	1/5	1/3	1/2	1	1/2	1/2	1	1/2	1	**1**	1/2	1/2	0.0422	0.5329	12.63
11	1/4	1/3	1/2	1	1/2	1	1	1/2	1	2	**1**	1/2	0.0511	0.6742	13.19
12	1/3	1/4	1/2	1	1/2	1/2	1	1/2	1	2	2	**1**	0.0542	0.6975	12.87

A similar analysis was performed for the 12 matrices with 3 alternatives. The results for the group of criteria "Financial strength" are shown in Table 3. In this case, we obtain $\lambda_{\max}^{(1)} = 3.0183$. The consistency index $(\mathrm{CI})^{(1)} = (3.0183 - 3)/2 = 0.0092$. The random consistency index $(\mathrm{RI})^{(1)} = 0.52$ for $m = 3$ [7]. The consistency ratio $(\mathrm{CR})^{(1)} = 0.0158$ and does not exceed 0.1. Taking into account the 12 criteria groups, the final results are shown

in Table 4. The conducted research allows us to assert that the startup for manufacturing information–reference electronic terminals is most attractive for investment, as its global priority of 0.3855 is the highest among the analyzed investment proposals. At the same time, the values of the global priorities for the startups producing LED traffic lights and manufacturing rotor-reactive turbo-rotational heaters of liquids are equal to 0.2547 and 0.3599, respectively.

Table 3. The matrix of pairwise comparisons for the group of criteria "Financial strength".

Startup	Production of LED Traffic Lights	Manufacture of Information–Reference Electronic Terminals	Manufacture of Rotor-Reactive Turbo-Rotational Heaters of Liquids	Priority Vector (The Normalized Vector of Geometric Means) $v_r^{(1)}$	RM	DV
Production of LED traffic lights	**1**	1/3	1	0.20984	0.63337	3.01835
Manufacture of information–reference electronic terminals	3	**1**	2	0.54994	1.65990	3.01833
Manufacture of rotor-reactive turbo-rotational heaters of liquids	1	1/2	**1**	0.24021	0.72503	3.01832

Table 4. The optimal choice of startups according to the investment alternatives, based on the groups of criteria.

	Investing Alternatives in Startups	Production of LED Traffic Lights	Manufacture of Information–Reference Electronic Terminals	Manufacture of Rotor-Reactive Turbo-Rotational Heaters of Liquids
No	**Groups of Criteria**		**Priority Vectors**	
1.	Financial strength	0.2098	0.5499	0.2402
2.	Product/service potential	0.2000	0.4000	0.4000
3.	Marketing potential	0.2000	0.4000	0.4000
4.	Organizational potential	0.2500	0.5000	0.2500
5.	Scientific and technical potential	0.1634	0.5396	0.2970
6.	Staff potential	0.1958	0.3108	0.4934
7.	Potential of governmental, international, economic, and political situation	0.2500	0.2500	0.5000
8.	Time potential	0.5936	0.1571	0.2493
9.	Autonomy potential	0.1634	0.2970	0.5396
10.	Ecological potential	0.2500	0.5000	0.2500
11.	Social potential	0.3333	0.3333	0.3333
12.	Information potential	0.3325	0.1396	0.5278
13.	Vector of global priorities	**0.2547**	**0.3855**	**0.3599**

4. Concluding Remarks

New criteria and new criteria composition for the comparison of investment alternatives were proposed. Considering the sub-criteria could aid establishing weights of groups of criteria. The criteria and alternatives are mutually independent. A multi-criteria approach based on the analytic hierarchy method was used providing a gradual, clear, and logically structured assessment of the parameters of the given alternatives to ensure a successful solution. The proposed approach also has some limitations. For a large number of criteria and alternatives, Saaty's scale 1–9 could not be enough. The decision-making process could also be time consuming for a large number of criteria and alternatives. For

example, in the case of 32 sub-criteria, there appears a large matrix of pairwise comparisons, and for $k > 15$ in the literature there is no value of the random index (RI) and only an approximate estimation of the consistency ratio (CR) can be obtained. Therefore, we grouped the 32 new sub-criteria proposed in this study into 12 blocks (potentials). Despite these limitations, the AHP approach is one of the most popular and objective methods for multi-criteria decision-making. The proposed use of the Saaty method for optimal decision-making has a number of advantages, as well. It does not require the unification of the units of measurement for different criteria. It ensures the accuracy of the evaluation by increasing the possibility of intra-matching within the selected criteria. In addition, the presence of a numeric scale allows the relations between the factors to be clearly identified. Finally, this method is adaptable, enabling the criteria composition to be modified by adding or eliminating factors. We compared the maximal eigenvalues $\lambda_{\max}$ obtained as the arithmetic mean of the vector in column DV and the value of $\lambda_{\max}$ obtained using the available mathematical package. With a precision of four digits, the results were the same. It should be emphasized that the consistency ratio (CR) of the pairwise comparison of the 12 groups of criteria, as well as all the 12 consistency ratios $\mathrm{CR}^{(p)}$, $p = 1, 2, \ldots, 12$, did not exceed 0.1; therefore, the evaluation was consistent.

In the future, we are planning to extend our research to compare our results with results obtained by other techniques, in particular, using the Bellman–Zadeh fuzzy set approach.

Author Contributions: Conceptualization, T.K. and Y.P.; methodology, T.K.; validation, Y.P.; formal analysis, T.K.; investigation, T.K. and Y.P.; writing—original draft preparation, T.K. and Y.P; writing—review and editing, T.K. and Y.P.; supervision, Y.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the reviewers for the helpful comments that allowed us to improve the final version of the paper.

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

References

1. Blank, S.; Dorf, B. *The Startup Owner's Manual: The Step-by-Step Guide for Building a Great Company*; K&S Ranch Press: Pescadero, CA, USA, 2012.
2. Benjamin, G.A.; Margulis, J.B. *The Angel Investor's Handbook: How to Profit from Early-Stage Investing*; Bloomberg Press: Princeton, NJ, USA, 2001.
3. Belton, V.; Stewart, T. *Multiple Criteria Decision Analysis: An Integrated Approach*; Springer: New York, NY, USA, 2002.
4. Campbell, K. *Smarter Ventures: A Survivor's Guide to Venture Capital through the New Cycle*; Prentice Hall: Harlow, UK, 2003.
5. Kessler, A. *Eat People: Furthermore, Other Unapologetic Rules for Game-Changing Entrepreneurs*; Penguin Group: New York, NY, USA, 2011.
6. Li, Y. Duration analysis of venture capital staging: A real options perspective. *J. Bus. Ventur.* **2008**, *23*, 497–512. [CrossRef]
7. Saaty, T.L. *The Analytical Hierarchy Process: Planning, Priority Setting, Resource Allocation*; McGraw-Hill: New York, NY, USA, 1980.
8. Saaty, T.L. *Multicriteria Decision Making: The Analytical Hierarchy Process*; RWS Publications: Pittsburgh, PA, USA, 1988.
9. Saaty, T.L. *Decision Making with Dependence and Feedback: The Analytical Network Process*, 2nd ed.; RWS Publications: Pittsburgh, PA, USA, 2001.
10. Saaty, T.L. *Fundamentals of Decision Making and Priority Theory with the Analytical Hierarchy Process*, 2nd ed.; RWS Publications: Pittsburgh, PA, USA, 2006.
11. Saaty, T.L.; Vargas, L.G. *Decision Making with the Analytical Network Process: Economic, Political, Social and Technological Applications with Benefits, Opportunities, Costs and Risks*; Springer: New York, NY, USA, 2006.
12. Saaty, T.L. *Decision Making for Leaders: The Analytical Hierarchy Process for Decisions in a Complex World*, 3rd ed.; RWS Publications: Pittsburgh, PA, USA, 2012.
13. Saaty, T.L.; Vargas, L.G. *Models, Methods, Concepts & Applications of the Analytical Hierarchy Process*, 2nd ed.; Springer: New York, NY, USA, 2012.
14. Brunelli, M. *Introduction to the Analytical Hierarchy Process*; Springer: Cham, Switzerland, 2015.

15. Roy, U.; Majumder, M. *Vulnerability of Watersheds to Climate Change Assessed by Neural Network and Analytical Hierarchy Process*; Springer: Singapore, 2016.
16. De Felice, F.; Saaty, T.L.; Petrillo, A. (Eds.) *Applications and Theory of Analytical Hierarchy Process—Decision Making for Strategic Decisions*; IntechOpen: London, UK, 2016.
17. Ozsahin, D.U.; Hüseyin Gökçekuş, H.; Uzun, B.; LaMoreaux, J. (Eds.) *Application of Multi-Criteria Decision Analysis in Environmental and Civil Engineering*; Springer: Cham, Switzerland, 2021.
18. Thakkar, J.J. *Multi-Criteria Decision Making*; Springer: Singapore, 2021.
19. Kulakowski, K. *Understanding the Analytical Hierarchy Process*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2022.
20. Pohekar, S.D.; Ramachandran, M. Application of multi-criteria decision making to sustainable energy planning—A review. *Renew. Sustain. Energy Rev.* **2004**, *8*, 365–381. [CrossRef]
21. Vaidya, O.S.; Kumar, S. Analytical hierarchy process: An overview of applications. *Eur. J. Oper. Res.* **2006**, *169*, 1–29. [CrossRef]
22. Ho, W. Integrated analytical hierarchy process and its applications—A literature review. *Eur. J. Oper. Res.* **2008**, *186*, 211–228. [CrossRef]
23. Liberatore, M.J.; Nydick, R.L. The analytical hierarchy process in medical and health care decision making: A literature review. *Eur. J. Oper. Res.* **2008**, *189*, 294–307. [CrossRef]
24. Ishizaka, A.; Labib, A. Review of the main developments in the Analytical Hierarchy Process. *Expert Syst. Appl.* **2011**, *38*, 14336–14345.
25. Subramanian, N.; Ramanathan, R. A review of applications of Analytical Hierarchy Process in operations management. *Int. J. Prod. Econ.* **2012**, *138*, 215–241. [CrossRef]
26. Schmidt, K.; Aumann, I.; Hollander, I.; Damm, K.; von der Schulenburg, J.M.G. Applying the Analytical Hierarchy Process in healthcare research: A systematic literature review and evaluation of reporting. *BMC Med. Inform. Decis. Mak.* **2015**, *15*, 112. [CrossRef]
27. Russo, R.F.S.M.; Camanho, R. Criteria in AHP: A systematic review of literature. *Procedia Comput. Sci.* **2015**, *55*, 1123–1132. [CrossRef]
28. Nisel, S.; Özdemir, M. AHP/ANP in sports: A comprehensive literature review. *Int. J. Anal. Hierarchy Process* **2016**, *8*, 405–429.
29. Emrouznejad, A.; Marra, M. The state of the art development of AHP (1979–2017): A literature review with a social network analysis. *Int. J. Prod. Res.* **2017**, *55*, 6653–6675. [CrossRef]
30. Rajput, V.; Kumar, D.; Sharma, A.; Singh, S.; Rambhagat. A literature review on AHP (Analytical Hierarchy Process). *J. Adv. Res. Appl. Sci.* **2018**, *5*, 349–355.
31. Darko, A.; Chan, A.P.C.; Ameyaw, E.E.; Owusu, E.K.; Pärn, E.; Edwards, D.J. Review of application of analytical hierarchy process (AHP) in construction. *Int. J. Constr. Manag.* **2019**, *19*, 436–452.
32. Goyal, P.; Kumar, D.; Kumar, V. Application of multi-criteria decision analysis in the area of sustainability: A literature review. *Int. J. Anal. Hierarchy Process* **2020**, *12*, 512–545.
33. Madzík, P.; Falát, L. State-of-the-art on analytical hierarchy process in the last 40 years: Literature review based on Latent Dirichlet Allocation topic modelling. *PLoS ONE* **2022**, *17*, e0268777. [CrossRef]
34. Saaty, T.L.; Hu, G. Ranking by eigenvector versus other methods in the Analytical Hierarchy Process. *Appl. Math. Lett.* **1998**, *11*, 121–125. [CrossRef]
35. Saaty, T.L. Decision-making with the AHP: Why is the principal eigenvector necessary. *Eur. J. Oper. Res.* **2003**, *145*, 85–91. [CrossRef]
36. Crawford, G.B. The geometric mean procedure for estimating the scale of a judgment matrix. *Math. Model.* **1987**, *9*, 327–334. [CrossRef]
37. Csató, L. A characterization of the Logarithmic Least Squares Method. *Eur. J. Oper. Res.* **2019**, *276*, 212–216. [CrossRef]
38. Wang, L.; Xu, L.; Feng, S.; Meng, M.Q.-H.; Wang, K. Multi–Gaussian fitting for pulse waveform using Weighted Least Squares and multi-criteria decision making method. *Comput. Biol. Med.* **2013**, *43*, 1661–1672. [CrossRef]
39. Wu, S.; Fu, Y.; Lai, K.K.; Leung, W.K.J. A Weighted Least-Square Dissimilarity Approach for multiple criteria ABC inventory classification. *Asia-Pac. J. Oper. Res.* **2018**, *35*, 1850025. [CrossRef]
40. Mikhailov, L. Fuzzy programming method for deriving priorities in the Analytical Hierarchy Process. *J. Oper. Res. Soc.* **2000**, *51*, 341–349. [CrossRef]
41. Wang, J.; Fan, K.; Wang, W. Integration of fuzzy AHP and FPP with TOPSIS methodology for aeroengine health assessment. *Expert Syst. Appl.* **2010**, *37*, 8516–8526. [CrossRef]
42. Almulhim, T.; Mikhailov, L.; Xu, D.-L. A fuzzy group prioritization method for deriving weights and its software implementation. *Int. J. Artif. Intell. Interact. Multimed.* **2013**, *2*, 7–14. [CrossRef]
43. Fallahpour, A.; Wong, K.Y.; Rajoo, S.; Olugu, E.U.; Nilashi, M.; Turskis, Z. A fuzzy decision support system for sustainable construction project selection: An integrated FPP-FIS model. *J. Civ. Eng. Manag.* **2020**, *26*, 247–258. [CrossRef]
44. Kou, G.; Lin, C. A cosine maximization method for the priority vector derivation in AHP. *Eur. J. Oper. Res.* **2014**, *235*, 225–235. [CrossRef]
45. Srdjevic, B. Combining different prioritization methods in the analytical hierarchy process synthesis. *Comput. Oper. Res.* **2005**, *32*, 1897–1919. [CrossRef]

46. Keršulienė, V.; Zavadskas, E.K.; Turskis, Z. Selection of rational dispute resolution method by applying new stepwise weight assessment ratio analysis (SWARA). *J. Bus. Econ. Manag.* **2010**, *11*, 243–258. [CrossRef]
47. Stanujkic, D.; Karabasevic, D.; Zavadskas, E.K. A framework for the selection of a packaging design based on the SWARA method. *Eng. Econ.* **2015**, *26*, 181–187. [CrossRef]
48. Rezaei, J. Best-worst multi-criteria decision-making method. *Omega* **2015**, *53*, 49–57. [CrossRef]
49. Rezaei, J. Best-worst multi-criteria decision-making method: Some properties and a linear model. *Omega* **2016**, *64*, 126–130. [CrossRef]
50. Pamučar, D.; Ecer, F.; Cirovic, G.; Arlasheedi, M.A. Application of improved best worst method (BWM) in real-world problems. *Mathematics* **2020**, *8*, 1342. [CrossRef]
51. Pamučar, D.; Stević, Ž.; Sremac, S. A new model for determining weight coefficients of criteria in MCDM models: Full Consistency Method (FUCOM). *Symmetry* **2018**, *10*, 393. [CrossRef]
52. Fazlollahtabar, H.; Smailbašić, A.; Stević, Ž. FUCOM method in group decision-making: Selection of forklift in a warehouse. *Decis. Mak. Appl. Manag. Eng.* **2019**, *2*, 49–65. [CrossRef]
53. Stević, Ž.; Brković, N. A novel integrated FUCOM-MARCOS model for evaluation of human resources in a transport company. *Logistics* **2020**, *4*, 4. [CrossRef]
54. Žižović, M.; Pamučar, D. New model for determining criteria weights: Level Based Weight Assessment (LBWA) model. *Decis. Mak. Appl. Manag. Eng.* **2019**, *2*, 126–137. [CrossRef]
55. Božanić, D.; Ranđelović, A.; Radovanović, M.; Tešić, D. A hybrid LBWA - IR-MAIRCA multi-criteria decision-making model for determination of constructive elements of weapons. *Facta Univ. Ser. Mech. Eng.* **2020**, *18*, 399–418. [CrossRef]
56. Greblikaitė, J.; Daugėlienė, R. Cluster analysis of expression of entrepreneurship characteristics in the EU innovative projects for SME's and KTU regional science park. *Eur. Integr. Stud.* **2009**, *3*, 184–189.
57. Yüksel, I. Developing a multi-criteria decision making model for PESTEL analysis. *Int. J. Bus. Manag.* **2012**, *7*, 52–66. [CrossRef]
58. Leyva Vázquez, M.; Hechavarría Hernández, J.; Batista Hernández, N.; Alarcón Salvatierra, J.A.; Gómez Baryolo, O. A framework for PEST analysis based on fuzzy decision maps. *Rev. Espac.* **2018**, *39*. Available online: https://www.revistaespacios.com/a18v39n16/18391603.html (accessed on 10 January 2018).
59. Greblikaitė, J.; Astrovienė, J.; Montvydaitė, D. Value-added agricultural bio-business development in European countries. *Manag. Theory Stud. Rural Bus. Infrastruct. Dev.* **2020**, *42*, 235–247. [CrossRef]
60. Saaty, R.W. The analytical hierarchy process – what it is and how it is used. *Math. Model.* **1987**, *9*, 161–176. [CrossRef]
61. Saaty, T.L. Decision making with the analytical hierarchy process. *Int. J. Serv. Sci.* **2008**, *1*, 83–98.

entropy

MDPI

Article

Organizational Labor Flow Networks and Career Forecasting

Frank Webb [1,†], Daniel Stimpson [2], Miesha Purcell [2] and Eduardo López [1,*,†]

[1] Department of Computational and Data Sciences, George Mason University, Fairfax, VA 22030, USA
[2] United States Army Acquisition Support Center (USAASC), 9900 Belvoir Road, Fort Belvoir, VA 22060, USA
* Correspondence: elopez22@gmu.edu
† These authors contributed equally to this work.

Abstract: The movement of employees within an organization is a research area of great relevance in a variety of fields such as economics, management science, and operations research, among others. In econophysics, however, only a few initial incursions have been made into this problem. In this paper, based on an approach inspired by the concept of labor flow networks which capture the movement of workers among firms of entire national economies, we construct empirically calibrated high-resolution networks of internal labor markets with nodes and links defined on the basis of different descriptions of job positions, such as operating units or occupational codes. The model is constructed and tested for a dataset from a large U.S. government organization. Using two versions of Markov processes, one without and another with limited memory, we show that our network descriptions of internal labor markets have strong predictive power. Among the most relevant findings, we observe that the *organizational labor flow networks* created by our method based on operational units possess a power law feature consistent with the distribution of firm sizes in an economy. This signals the surprising and important result that this regularity is pervasive across the landscape of economic entities. We expect our work to provide a novel approach to study careers and help connect the different disciplines that currently study them.

Keywords: labor flow networks; firm-size distribution; career studies; career sequences; manpower analysis

Citation: Webb, F.; Stimpson, D.; Purcell, M.; López, E. Organizational Labor Flow Networks and Career Forecasting. *Entropy* **2023**, *25*, 784. https://doi.org/10.3390/e25050784

Academic Editor: Panos Argyrakis

Received: 28 February 2023
Revised: 21 April 2023
Accepted: 1 May 2023
Published: 11 May 2023

1. Introduction

The study of job change is of great practical and academic relevance, as it is one of the fundamental components of the employment process of any economic system. Several disciplines study versions of this problem, including economics [1,2], management science [3], and operations research [4–6]. Although a great deal of progress has been made in elucidating this critical process, numerous questions remain outstanding. In particular, there is yet to be an integrated interdisciplinary picture that explains both the micro and macro aspects of the problem while maintaining the true system heterogeneity.

In recent years, a new way to approach the problem of job change has started to develop based on the observation that once a person makes a job transition between two firms (i.e., two employers), the probability to observe other subsequent job transitions between the same two firms is significantly larger than what would be expected by random chance [7]. This result has provided empirical support for the development of a new class of large-scale, high-resolution job change, Labor Flow Networks (LFNs) [8–10], which conceptualize the system as a set of nodes representing firms and links representing pairs of nodes between which a job transition is relatively likely to occur (in economics terms, such job changes have low friction). Based on data from two different countries, Finland and Mexico [8–10], the first examples of LFNs were created, encoding large cross-sections of the employers and employees in the workforce in their respective countries (for Finland, the data are comprehensive for about a decade). From the physical standpoint, LFN models are constituted by complex random environments that harbor non-equilibrium transport

processes operating near equilibrium and, as such, can be understood from many of the rules of non-equilibrium statistical mechanics [11]. A number of important observations have emerged from the LFNs literature. First, it has been realized that firms contribute in heterogeneous ways to unemployment [9,10] with some firms being responsible for more unemployed people. Second, the firm-size distribution in an economy [12,13] is related to both network and temporal features displayed by the LFNs [10]. Third, that socio-economic status and race play important roles in occupational mobility [14]. Fourth, that the relationship between vacancies and jobs in an economy (the so-called Beveridge curve) cycles in a counter-clockwise manner [15] tracing a hysteretic curve that does not retrace its steps as it returns to a previous state through a business cycle. This continues to be an active area of research, with new directions being explored [16].

As an empirical model, the key ingredients of LFNs are (i) data that relate employers and employees and (ii) a statistical test that confirms that job changes among those employers are not random. This provides a flexible, data-driven framework that makes it possible to model a multitude of systems with considerable accuracy, especially if the dynamics of the system are sufficiently slow. This opens the possibility to study a related, and yet-to-be explored context of employment, internal labor markets. The study of such job markets has the potential to bring about greater conceptual understanding of the job change process because, when information about organizations is available, it can offer greater depth than national level data (organizations usually record personnel details such as academic accomplishments, years of work experience, job responsibilities, etc.). Because the system sizes of organizations are limited and thus cannot achieve the regularity of the thermodynamic limit, and because people's careers inside organizations are not long enough for the phase space to be effectively explored, the dynamics of careers in organizations lives in the space between mesoscopic and macroscopic systems.

In this article, aided by the availability of data from a large US governmental organization, we apply the network approach for the study of so-called internal labor markets, that is, the jobs internal to an organization among which individual workers transition while pursuing an organizational career. Internal labor markets are not just smaller versions of large economic systems, but instead have different operating rules and are organized differently than a national economy and thus cannot be assumed to display the same regularities as national employment landscapes. For example, there are no independent firms inside an organization that can be viewed as the employers (nodes) of the network. Another distinction is that job changes inside an organization can occur from mechanisms different than job search (such as organizational reorganization or promotions based on seniority).

Here, to test the application of LFNs to internal labor markets, we first study ways to identify the relevant network nodes based on one of several possible job descriptors that we refer to as *"location"*. We work with three different descriptors, operating units, occupational codes, and geographic locations. Thus, when the network is constructed of, say, operating units, job transitions by individuals connect the operating units that individuals exit and join immediately after. We show that the different networks produced by using these different choices of nodes all display predictive power (i.e., their links are better predictors of future job changes than random chance), although some networks perform much better. Another important finding is that the choice of network node leads to networks that may have interesting topological properties. Most notably, we find a reproduction of the Zipf-law of firm-size distribution when nodes correspond to operating units of the organization [10,12,13,17]. Furthermore, because the approach microscopically tracks the movements of each person, forecasting of individual work trajectories (that is, so-called organizational careers) inside internal labor markets becomes possible. Careers can be forecast with memoryless Markov chains [10], the most similar model to a physical system, or with memory of prior jobs by using the method in [18]. We evaluate the quality of the predictions through a variety of methods including Jensen-Shannon divergence [19] and Jaccard indices, and find strong agreement between observation and prediction with

both methods, although the use of memory leads to even stronger agreement with empirical observation.

The ability to track so-called organizational careers through the labor flow network method should not be understated. While the study of labor markets at national levels can yield limited information about careers, in general the data sources are not capable of providing enough information to perform accurate studies. Some data lose visibility of individuals, others only track workers for limited periods of time, and there is almost no information about the nature of the jobs undertaken by individuals (no concrete data on job responsibilities or tasks). In contrast, within an organizational setting, where personnel data are available, the ability to use LFNs to understand the job landscape within the organization becomes a tool that clarifies employment dynamics at both the individual and organizational levels over time. This means that the network approach is able to bring together concepts of operations research, management science (specifically career studies), and economics of job search for the first time in such a concrete way, providing an opportunity to develop an integrated view of internal labor markets that is currently missing.

The remainder of the paper is organized as follows. Section 2 discusses the data for the Army Acquisition workforce, defines the methods of analysis, the logic behind those methods, and the relevant notation. Then, in Section 3 we present the results of our analysis. Of particular interest, Section 3.2 addresses the similarity between the organization we analyze and the firm-size distribution. Finally, we discuss our findings and their relevance to the study of oganizations and careers in Section 4.

2. Materials and Methods

2.1. Data

The data we study are for the Army Acquisition Workforce (AAW). This is a civilian organization that is part of the United States Army whose function is to provide logistical support to the military component of the Army by purchasing equipment, training, and a number of other logistical needs. The AAW has both uniformed military and civilian components. Job changes within the civilian component (which are the only ones we study here) are not based on military orders, but function as a typical job market, where employees apply for jobs as openings emerge. Thus, the organization has freedom of career mobility based on qualifications and individuals can join or leave as with any other job in the private sector. Information of the AAW is public, and can be found online [20]. The data have two parts, one associated with individuals and the other with the structure of the AAW. The datasets cover the period between 2012 and 2020. All employee records are anonymized by associating to each individual a hashed key. Each employee record contains the position occupied each month the employee is part of the AAW. This information includes the individual's operational unit as well as his/her occupational series [21], a code assigned by the US government to positions that imply certain responsibilities and other requirements. Over the period, the AAW has ranged in size from under 35,000 to close to 42,000 individuals. There are around 1000 operational units in the AAW, and employees span close to 100 occupational codes.

2.2. Methods

2.2.1. Basic Network Elements

To describe the job landscape inside an organization, we distinguish between two types of entities, employed individuals ($\alpha, \beta, \ldots$) and the "location" of the employment ($i, j, \ldots$) (the term location is not ideal, but other choices such as *class*, used in the manpower literature [22] are also problematic and thus we choose location because it better fits our analysis). Our data identify individuals as well as several possible choices of locations such as operating units within an organization, geographic locations (such as US States) where some part of the organization operates, or types of occupations that the organization requires (say, data analysts or accountants, recorded with a standardized code system [21]).

Here, we use the term location as a descriptor to indicate where the individual can be found within the organization. For example, if we are interested in knowing the movements of the workforce by geography, i would represent a particular US state where some of the organization has facilities. On the other hand, if we want to know the distribution of the workforce by occupation, i would be an occupational code.

Our characterization of the system is based on the structure of labor markets, which are studied by looking at the interrelated dynamics of individuals employed or looking for employment and the jobs those individuals occupy or the vacancies they may aspire to fill. In previous studies of LFNs, the choice of location was not discussed in itself, perhaps determined by the data available (e.g., in [8,10] only firms and employees are recorded, making locations represent firms). However, in our case, not only does the data provide an opportunity to explore several possibilities, but there is a genuine question about which choice of location to use in terms of better accuracy of the models, something we address below and return to in Section 4.

Given a choice for i, an Organizational Labor Flow Network (OLFN) is generated in the following way. Consider a set of individuals $\mathcal{E} = \{\alpha, \beta, \dots\}$ and job locations $\mathcal{N} = \{i, j, \dots, \}$. We denote the sizes of the sets as $e = |\mathcal{E}|$ and $n = |\mathcal{N}|$. The work histories of individuals, usually called *sequences* in career studies, are typically recorded at discrete and uniformly distributed time points $t_o, t_o + 1, \dots, t_o + T$, where t_o is the initial time of observation, T the number of time units of observation (equal to the duration of the data) and, in our case, the units are in months. Thus, we define the *employment sequence of agent* α by

$$c_\alpha(t) = i \quad [i \in \mathcal{E}, t \in \{t_{\alpha,o}, t_{\alpha,o} + 1, \dots, t_{\alpha,o} + \tau_\alpha\}], \tag{1}$$

where i is a job location, $t_{\alpha,o}$ the first time α is observed to be in the organization, and τ_α is the so-called job tenure (the number of time units an employee spends in the organization). Note that $t_o \leq t_{\alpha,o} \leq t_o + T$ and $0 \leq \tau_\alpha \leq T - (t_{\alpha,o} - t_o)$ and that information about individual starting and ending times is necessary to know given that many employees join and/or leave over a period of time.

The nodes of an OLFN are constituted by the job locations $\mathcal{E}$. A link (i, j) between two nodes in the OLFN is possible only if there are job transitions from i to j, but this may not guarantee a link. Instead, (i, j) would be included as a link in the OLFN if statistically significant job transitions are observed between the nodes [10]. The statistical test is explained in Section 2.2.2.

2.2.2. Statistical Significance of Organizational Labor Flow Networks

An OLFN can be defined in several ways beyond the choice of locations, just as long as it leads to reliable networks in terms of forecasting future job change. This means that, to construct an OLFN, we must check that information gathered at some period of time can be used to forecast a subsequent time period. This requires that we statistically test the reliability of past information in terms of providing information about the future. But, how to design this test?

In this system, job changes past or future appear as job transitions. Thus, we must find a way to take information about transitions and convert this to links in a network. In other words, links may only be introduced between pairs of locations (nodes) that have had job transitions between them, although the final decision can depend on additional criteria (see below). Following on, we must further consider whether linking a node pair should be done independently or related to linking other node pairs. We can quickly realize that to choose links between node pairs independently of each other runs the risk of ignoring correlations. For example, some locations are characterized by many people (large operating units or popular occupation codes), while others by a few. For the case of a large location, it is likely that it sends and receives many workers, an effect that is felt across many of the node pairs that involve that node. This acts as a correlation between the large node and the transitions involving other nodes, effectively coupling its possible links. Therefore, it is generally more appropriate to decide on adding links by taking into account

their correlated structure. The simplest way to do this would be to correlate links that connect to the same node, independent of other nodes. This approach, however, is likely to ignore higher order correlations that trickle through from node to node. Therefore, an even better strategy would be to decide on links on the basis of the whole network structure.

We must also choose a time frame with job transitions that help us predict future transitions. In this case, we pick a simple strategy that works well in that it provides proof of principle. Thus, we divide the data into two equal-size time periods, $t_o, \ldots, t_o + \lfloor T/2 \rfloor$ and $t_o + \lfloor T/2 \rfloor, \ldots, t_o + T$, that we refer to respectively as $\mathcal{T}_<$ and $\mathcal{T}_>$. The first time period acts as the baseline, whereas the second corresponds to the forecasting (or test) period. From the baseline period, we take transitions and consider them as candidates for links. The testing period is then used to determine whether our choices of links have been appropriate in terms of making our OLFN useful for prediction.

Having established time windows, we must decide how to introduce links. While an approach that would explore the entire space of possible combinations of linking in some designated period of time could be imagined, in practice this is very challenging due to the combinatorial explosion of possibilities. Instead, we take an approach related to [7] but also addresses the fact that their method ignores the link correlations we identified above. In [7], each pair of firms in the city of Stockholm is considered independently and a single transition between firms is used to gauge subsequent likelihood of transitions. Here, as in [7], we adopt the notion that observing a transition between a node pair suggests they should be linked, but introduce a numerical threshold $\mathcal{W}$ representing the minimum number of transitions (in either direction) between two locations i and j in order to make that pair of nodes a *candidate* to have a link . This generates a *candidate network* where node pairs have tentative links if they satisfy the threshold.

The final step of the statistical test is to check if the candidate network is indeed predictive. To do this, we construct possible random future networks (meant to be in time period $\mathcal{T}_>$) and compare them with information from the candidate network in the past (from time period $\mathcal{T}_<$). Two pieces of information have been used in [10] for this purpose. First, let us define $\kappa_i^{\text{in}}(\mathcal{T}_>)$ and $\kappa_i^{\text{out}}(\mathcal{T}_>)$ as, respectively, the number distinct nodes from which workers transition into node i and the number distinct nodes to which workers transition to from i, both within time period $\mathcal{T}_>$. Similarly, we define $\sigma_i^{(\text{in})}(\mathcal{T}_>)$ as the number of workers that transition from other nodes into i over the period $\mathcal{T}_>$, and $\sigma_i^{(\text{out})}(\mathcal{T}_>)$ the number of workers that transition from i to other nodes in the same period. These quantities are versions of the concepts of node degree and node strength [23].

It was found in [10] that the most demanding version of test was the one that preserved $\sigma_i^{(\text{in})}(\mathcal{T}_>)$ and $\sigma_i^{(\text{out})}(\mathcal{T}_>)$ because the statistic that measures the amount of deviation from random transitions produced the smallest (yet highly significant) results. To perform this test, we generate a large number of distinct realizations of random networks using Monte Carlo. Each such network is created by randomly assigning transitions between nodes in the period $\mathcal{T}_>$ while requiring that $\sigma_i^{(\text{in})}(\mathcal{T}_>)$ and $\sigma_i^{(\text{out})}(\mathcal{T}_>)$ remain true in each and every realization for all nodes. To generate a statistic, the random model lets us estimate an expectation value for how many transitions can randomly occur between a pair of nodes that is a candidate link, based on a given threshold $\mathcal{W}$, from period $\mathcal{T}_<$. Introducing the notation $\mathcal{C}_<$ for the set of candidate links during $\mathcal{T}_<$ and $\mathcal{C}_>^{(s)}$ for set of stochastic transitions predicted by each realization of one of the random models during $\mathcal{T}_>$, all the null models generate an expected density of overlaps

$$\wp^{(s)}(\mathcal{W}) = \frac{\langle |\mathcal{C}_< \cap \mathcal{C}_>^{(s)}| \rangle}{|\mathcal{C}_<|} \tag{2}$$

which measures the expected fraction of candidate links from $\mathcal{T}_<$ that would also have a transition during $\mathcal{T}_>$ simply by random chance. In this expression, the customary notation of size of a set $|\,|$ and expectation $\langle\,\rangle$ have been used. Clearly, the choice of stochastic model alters the resulting set $\mathcal{C}_>^{(s)}$.

From the standpoint of observation, we want to know among the $\mathcal{C}_<$, what fraction of them were observed to have transitions during $\mathcal{T}_>$. Labeling the observed transitions as $\mathcal{C}_{>T}^{(o)}$, the fraction of transitions matching candidate links is given by

$$\wp^{(o)} = \frac{|\mathcal{C}_< \cap \mathcal{C}_>^{(o)}|}{|\mathcal{C}_<|} \tag{3}$$

The statistic of interest is finally defined as the ratio between the two quantities, which we call the excess probability $x_{\mathcal{W}}$, given by

$$x_{\mathcal{W}} = \frac{\wp^{(o)}}{\wp^{(s)}}. \tag{4}$$

If $x_{\mathcal{W}}$ is above 1 with a large degree of certainty (has a very small p-value), we conclude that the threshold $\mathcal{W}$ leads to an OLFN that is useful for prediction. To provide intuition for this statement, note that $x_{\mathcal{W}}$ measures the network averaged increase in probability with respect to the random model that a transition during $\mathcal{T}_>$ occurs along a pair of nodes with $\mathcal{W}$ or more transitions during $\mathcal{T}_<$. To illustrate, if $x_{\mathcal{W}}$ is 2, the transitions actually observed during $\mathcal{T}_>$ are twice as likely to occur along pairs of nodes that had $\mathcal{W}$ transitions during $\mathcal{T}_<$ than what would be expected from the random model. Therefore, a value of $x_{\mathcal{W}} > 1$ (and the greater the better) means that transitions during $\mathcal{T}_>$ are predictable on the basis of transitions during $\mathcal{T}_<$ because they prefer to occur along candidate links by a factor of $x_{\mathcal{W}}$ than along random node pairs. Finally, as a technical point, the p-values can be determined semi-analytically (or analytically in the case of the uncorrelated random model, where only the total number of system transitions is preserved) by using the methodology in [10].

2.2.3. Career Sequences and Their Probability Distributions

To study careers, we are interested in the *non-degenerate* version of the sequences encoded in Equation (1). To illustrate what this means, consider an employment sequence c_α in which α spends from t to $t + \Delta t$ working at location i, or $c_\alpha(t) = \cdots = c_\alpha(t + \Delta t) = i$ but $c_\alpha(t - 1) \neq c_\alpha(t)$ and $c_\alpha(t + \Delta t) \neq c_\alpha(t + \Delta t + 1)$. We will refer to such a time period of uninterrupted work at a given location as a *spell*.

Since our primary interest is in the locations (or career steps) individuals take, we create a non-degenerate version of c_α called u_α such that only the location of a spell is recorded but not the number of time steps spent in a location. Thus, for example, if α's career is spent only in two locations, i and j and $c_\alpha = \{i, i, i, \ldots, i, j, j, \ldots, j\}$, the corresponding career sequence is $u_\alpha = \{i, j\}$. We should note that u_α preserves temporal ordering so that if α first worked in location i and then in j, these appear in that same order in u_α. Our sequences also possess the feature that if an individual were to *return* to a previous location, this would be captured in the sequence. Thus, an individual with an employment sequence of the form $\{i, i, j, j, i, i\}$ would have career sequence $\{i, j, i\}$.

The frequencies with which career sequences occur is very useful information because they offer insights on the sorts of choices individuals make under the constraints of the opportunities that become available within the organization (an individual cannot change into a job that is not offered, an important observation from the perspective of modeling made by the seminal work of White [24]). In order to understand how common or rare specific career sequences are, we define the distribution of observed sequences $\hat{\phi}(\mathbf{u})$, where $\mathbf{u}$ is the random variable of career sequences. For a given time period of observation,

$$\hat{\phi}(\mathbf{u} = u) = \frac{\sum_\alpha \delta_{u_\alpha, u}}{\sum_\alpha u_\alpha} \tag{5}$$

where u_α corresponds to the career sequence of α, $\delta_{u_\alpha, u}$ is the Kronecker delta equal to 1 when α's career matches the desired sequence u and 0 otherwise, and the denominator is

the total number of distinct careers observed. Described intuitively, Equation (5) precisely defines how we count careers to determine their probability of occurring.

Because careers can be sensitive to the initial location, we further specialize our analysis to distinguish careers on the basis of their initial location. Let us label the first location of career u as u_o (or $u_{\alpha,o}$ when the career refers to that of individual α). Then, we are interested in the set of conditional distributions

$$\hat{\phi}(\mathbf{u} = u | u_o = i) = \frac{\sum_\alpha \delta_{u_\alpha,u} \delta_{u_{\alpha,o},i}}{\sum_\alpha \delta_{u_{\alpha,o},i}}. \quad (6)$$

2.2.4. Temporal Statistics: Length of Service

Research on manpower identified early on some important features about the study of workforces inside organizations. When the emphasis is not on specific individuals, manpower studies are very similar to population studies with one critical difference: in the latter, survival times of segments of the population can be known quite well and vary slowly over time (the number of people of a certain ethnicity of a given age) whereas in the former the population of employees is much more changeable [22]. Thus, the concept of the *completed length of service* emerged [6,25].

The key conceptual point still carries over in terms of career forecasting: as an individual enters an organization, it is important to anticipate how long that individual is likely to stay in the organization. For simplicity, we approach this question here in a similar way to the manpower literature. In fact, we hinted at this point already in our definition of employment sequences (Equation (1)), where we introduced the quantity τ_α to represent α's job tenure in the organization. This quantity corresponds to the length of service random variable $\boldsymbol{\tau}$. Given the e individuals in the data, to determine the length of service distribution $\psi(\boldsymbol{\tau})$ we exclude from $\mathcal{E}$ all those employment sequences for which the last location recorded occurs in the last time unit in the data. This is because at this point, we are not capable to tell if any of those individuals exit the organization in that very last time unit, or if they continue in the organization.

Due to the sensitive nature of the data, we do not report the specific distribution of length of service of individuals in the organization, but use it in order to model careers in the ways we explain next (Section 2.2.5).

2.2.5. Markov Models of Career Sequences

To test the usefulness of OLFNs in modeling the movement of personnel across an organization, we construct two Markov chains, one which relies solely on the network structure (based on [9,10]) and another that uses the network structure plus memory (when applicable) about the prior transition [18]. At the most basic level, Markov chains require that one defines states of the system and probabilities to go between states. Our method based solely on network structure uses as states the current job (node) held by an employee, and the probability to transition between jobs is estimated on the basis of the transitions made by all workers over some selected period of time of the data (for example, the first half of the years in the data). On the other hand, our method to include memory generally defines as a state the tuple made of the current and previous job a worker has held (with exceptions needed to handle the first job of the worker), and the transition probabilities are estimated from other workers and the last two jobs they held. We now describe these details.

Let us start by clarifying that both models simply lead to the creation of simulated employment sequences and their associated career sequences. Since we mostly focus on career sequences, we introduce $\mathbf{r}^{(o)}$ and $\mathbf{r}^{(1)}$ to represent random career sequences respectively created from the Markov network model or the Markov model with one-step memory. These random variables are characterized by the distributions $\phi^{(o)}(\mathbf{r}^{(o)})$ and $\phi^{(1)}(\mathbf{r}^{(1)})$. These distributions are created from a large number of model realizations. There are two kinds of such realizations. On the one hand, a single random walker can only

generate a single career, not enough to generate useful distributions $\phi^{(o)}(\mathbf{r}^{(o)})$ or $\phi^{(1)}(\mathbf{r}^{(1)})$. Therefore, to generate these distributions, we use M_w walkers which correspondingly generate M_w careers from which to create the distributions. A second way to introduce multiple realizations is to generate M_d distributions $\phi^{(o)}(\mathbf{r}^{(o)})$ and $\phi^{(1)}(\mathbf{r}^{(1)})$ so that no one single realization of M_w walkers dominates the results. Our ultimate goal is to determine the quality of the models, which we do by defining below a set of metrics that compare each of these distributions to $\hat{\phi}(\mathbf{u})$.

A common feature to both models is the fact that individuals can begin to work at the organization at any time in any one of its locations. For the purposes of modelling their career sequences, one could ignore the specific point in time unless there were reasons to assume that temporal interactions play an important role. The initial location, on the other hand, is always relevant in terms of the number of either employment of career sequences generated. Thus, we should keep in mind that all the distributions we study are in reference to careers that start at each specific location (node) in the network.

One last feature shared by both models is that the number of time steps an individual travels is drawn from the length of service distribution $\psi(\boldsymbol{\tau})$. The effect is that each individual has a randomly drawn, fixed lifetime in the organization so that after time τ, the individual's career sequence (either $\mathbf{r}^{(o)}$ or $\mathbf{r}^{(1)}$) is completed and counted toward the appropriate distribution.

The model based on [10] makes use of the network structure but, deviating from that article, also includes weights to construct the transition rates between nodes. In the model, a simulated individual located at i at time step t has a probability p_{ij} to choose j as their next location, and this probability is constant in time. To determine p_{ij}, we make use of all the employment sequences in Equation (1). Such sequences can be used from the entire data (all the time points) or limited to parts of the time (e.g., $\mathcal{T}_<$ which would require some small adjustments like redefining work spells). Assuming we are using the entire data, we first count the number of moves f_{ij} from node i to j on the basis of the number of sequences (and the number of times in that sequence) where a transition occurs from i to j. Concretely,

$$f_{ij} = \sum_{\alpha} \sum_{t=t_{\alpha,o}}^{t_{\alpha,o}+\tau_\alpha-1} \delta_{i,c_\alpha(t)} \delta_{j,c_\alpha(t+1)}. \tag{7}$$

This equation states that f_{ij} is given by the number of times any individual makes a transition from i to j. For the Markov process, the probability of the transition i to j is then given by the proportion of all transition out of i that go to j with respect to all transitions out of i, or

$$p_{ij} = \frac{f_{ij}}{\sum_j f_{ij}} \quad [i, j \in \mathcal{N}]. \tag{8}$$

Note that the definitions of f_{ij} and p_{ij} include diagonal terms. Thus, the diagonal of the transition matrix of the Markov chain accounts for the very frequent occurrence of individuals remaining in their locations.

In contrast to the pure network model, the model that keeps track of the previous step (if the career has visited at least one other node) makes use of a slightly more complicated transition matrix. Note that when an individual enters the network at a node and has not yet made transitions to other nodes, the model is applied as if it was the pure network model described above; only after one transition can memory begin to play a role. To make use of memory, let us focus on a node j. The probability that an individual transitions from j to h given that it had previously transitioned from i to j is based on the number of careers that have previously made the same sequence of moves. Therefore, if $f_{(i,j),(j,h)}$ is given by

$$f_{(i,j),(j,h)} = \sum_{\alpha} \sum_{t=t_{\alpha,o}}^{t_{\alpha,o}+\tau_\alpha-2} \delta_{i,c_\alpha(t)} \delta_{j,c_\alpha(t+1)} \delta_{h,c_\alpha(t+2)}, \tag{9}$$

the probability for an individual to go from j to h given that they came from i is given by

$$p_{(i,j),(j,h)} = \frac{f_{(i,j),(j,h)}}{\sum_h f_{(i,j),(j,h)}} \quad [i, j, h \in \mathcal{N}]. \tag{10}$$

In both types of models, it is possible that the probabilities are 0 for an individual to move beyond their current location. If that is the case, the individual merely remains in the node until either the simulation finishes or the number of time units τ assigned to the individual are complete. We should note that a single realization for a walker can last up to the length of time we choose to model.

2.2.6. Evaluating Predicted Career Sequences

Next, we describe the metrics we use to assess the quality of the models. Essentially, we are interested in knowing whether the models tend to produce with high probability the careers actually observed, along with their observed frequencies. Symbolically, this is equivalent to testing for the similarity of the numerical values between $\hat{\phi}(\mathbf{u} = u|u_o = i)$ and $\phi^{(m)}(\mathbf{r}^{(m)} = r|r_o = i)$ when $u = r$ over the space of possibilities of u (the sample space), where $m = 0, 1$ for the memoryless Markov model or the one-step memory model, respectively. As a practical matter, we note that because all careers are distinguished by their initial location i, all the quantities we define are computed according to their initial location. Stated in plain English, the data show certain career paths and the models try to imitate these. Therefore, evaluating the models is done by checking how "similar" the imitation created by the models is to the observed careers.

In an ideal scenario, two distributions are similar if their sample spaces are similar and the probabilities of events (the elements of the sample space) are also similar. To be precise about what similar means, we now proceed to introduce several different quantitative measures of that similarity and highlight how each focuses on a particular aspect of that similarity.

Let us first concentrate on the similarity between probabilities $\hat{\phi}(\mathbf{u} = u|u_o = i)$ and $\phi^{(m)}(\mathbf{r}^{(m)} = r|r_o = i)$. In this case, similarity means that the observed and modeled probabilities of the same career u starting at node i have similar values, i.e., $\phi^{(m)}(\mathbf{r}^{(m)} = u|r_o = i) \approx \hat{\phi}(\mathbf{u} = u|u_o = i)$. But this comparison has to be done carefully because for any given initial node i, u is not independent of other careers starting from i. Let us denote all the observed careers starting from i as $\mathcal{U}(i) = \{u_\alpha\}_{\alpha\in\mathcal{E};u_o=i}$. Then, they are related by the fact that $\sum_{u\in\mathcal{U}(i)} \hat{\phi}(\mathbf{u} = u|u_o = i) = 1$ which is the normalization condition for $\hat{\phi}$. Modeled careers also satisfy a similar relation; calling the set of these careers $\mathcal{R}^{(m)}(i) = \{r_\theta^{(m)}\}_{\{\theta\};r_o=i}$ for model m, they satisfy $\sum_{u\in\mathcal{R}^{(m)}(i)} \phi^{(m)}(\mathbf{r}^{(m)} = u|r_o = i) = 1$. Note that $\mathcal{U}(i) = \{u_\alpha\}_{\alpha\in\mathcal{E};u_o=i}$ and $\mathcal{R}^{(m)}(i) = \{r_\theta^{(m)}\}_{\{\theta\};r_o=i}$ are, respectively, the sample spaces of the observed and modeled careers starting at i. The relation between the probabilities of all careers starting at a single node means that it is not enough to know that one particular career u is such that $\phi^{(m)}(\mathbf{r}^{(m)} = u|r_o = i) \approx \hat{\phi}(\mathbf{u} = u|u_o = i)$. Instead, we need to know that the entire collection of careers starting from i have approximately equal values of probability between observation and model. An effective way to study this is through information theoretic methods. Here we apply the Jensen-Shannon divergence (JSD) for this purpose [19]. This quantity measures information divergence between distributions in such a way that, unlike the Kullback-Liebler divergence, is efficient in handling possible mismatches in the sample spaces of the distributions. Defining the entropy of a random

variable X with distribution $P(X)$ as $H(P) = -\sum_X P(X)\log P(X)$, the JSD applied to $\hat{\phi}(\mathbf{u} = u|u_o = i)$ and $\phi^{(m)}(\mathbf{r}^{(m)} = r|r_o = i)$ takes the form

$$JSD^{(m)}(i) = H\left[\frac{1}{2}\hat{\phi}(\mathbf{u}|u_o = i) + \frac{1}{2}\phi^{(m)}(\mathbf{r}^{(m)}|r_o = i)\right] - \frac{1}{2}\Big[H(\hat{\phi}(\mathbf{u}|u_o = i)) + H(\phi^{(m)}(\mathbf{r}^{(m)}|r_o = i))\Big]. \quad (11)$$

Intuitively, the Jensen-Shannon divergence measures how much information two distributions share, with a value of 0 if they share all information (the distributions are identical), and a maximum possible value of $\log(2)$ when one distribution has no information about the other.

Since the distributions $\phi^{(m)}(\mathbf{r}^{(m)} = r|r_o = i)$ are generally different between different Monte Carlo realizations, we generate one $JSD^{(m)}(i)$ for each of the M_d realizations. To perform a complete test in terms of JSD, we create two versions of it, one that computes the JSD between pairs of distributions $\phi^{(m)}(\mathbf{r}^{(m)} = r|r_o = i)$ emerging from the Monte Carlo realizations (providing $M_d(M_d - 1)/2$ distinct values of JSD) and another comparing the real distribution $\hat{\phi}(\mathbf{u}|u_o = i)$ of careers against the simulated distributions (providing M_d values of JSD).

To explain this strategy further (using M_d realizations), note that the random distribution $\phi^{(m)}(\mathbf{r}^{(m)} = r|r_o = i)$ and $\hat{\phi}(\mathbf{u}|u_o = i)$ are both sample distributions. First, the modeled distribution $\phi^{(m)}(\mathbf{r}^{(m)} = r|r_o = i)$ emerges from generating M_w walks that begin at i and generate a set of walks $\mathcal{R}^{(m)}(i)$. Second, the distribution $\hat{\phi}(\mathbf{u}|u_o = i)$ is formed by all the observed careers beginning at i. Because both distributions emerge from a finite number of samples, even if either of the models $m = 0$ or 1 was perfectly correct, one cannot expect the two distributions to overlap perfectly. Thus, a more realistic evaluation of their similarity comes from observing how much $\hat{\phi}(\mathbf{u}|u_o = i)$ typically differs from $\phi^{(m)}(\mathbf{r}^{(m)} = r|r_o = i)$. This leads us to the need for creating M_d versions of $\phi^{(m)}(\mathbf{r}^{(m)} = r|r_o = i)$ to compare against $\hat{\phi}(\mathbf{u}|u_o = i)$. When needed, we label each such realization by the index $q = 1, \ldots, M_d$. Finally, note that the comparison between simulated career distributions allows us to develop a baseline for how well the observed career distribution is expected to match simulations. As a practical matter regarding numerical estimation of entropy, our situation is dominated by careers out of virtually all starting nodes where the most common career is to stay at that node; this means that we are able to estimate entropy via simple naive methods as in our case these are not particularly affected by problems such as those highlighted in the literature on entropy estimation [26–28].

Shifting to sample space testing, we introduce the Jaccard index which determines how similar two sets are by checking for the proportion of elements that are common between the sets; when both sets have the same elements the Jaccard index is 1, and when they share no elements it is 0. Thus, for a given location i, we define the Jaccard index $J^{(m)}(i)$ of node i due to model m as

$$J^{(m)}(i) = \frac{|\mathcal{U}(i) \cap \mathcal{R}^{(m)}(i)|}{|\mathcal{U}(i) \cup \mathcal{R}^{(m)}(i)|} \quad (12)$$

which quantifies how much the sets $\mathcal{U}(i)$ and $\mathcal{R}^{(m)}(i)$ resemble each other. Since $\mathcal{R}^{(m)}(i)$ is a product of simulations, one does not expect $J^{(m)}(i)$ to be the same for every realization. One simple approach (that we adopt here) to deal with this is to create a union of the simulated careers, $\prod_q^{M_d} \cup \mathcal{R}_q^{(m)}(i)$ and compare this set with $\mathcal{U}(i)$. Note that the choice to check against the union over $\mathcal{R}_q^{(m)}(i)$ is well justified on the basis that we are not after a test of probability, only sample space.

As a final check, we introduce a ratio test for careers. This check is useful for several purposes. For one, it can identify particular career sequences that are especially rare compared to random expectation. Another advantage is that it can be put to use in

generating career profiles for each starting node that provide a sense for how well the collection of modeled careers match the collection of observed careers. A final use comes as an alternative to the measurements from JSD and can be readily applied to obtaining full descriptions of a model over the entire network. All these depend on the definition

$$d^{(m)}(u,i) = \log\left[\frac{\hat{\phi}(\mathbf{u} = u|u_o = i)}{\phi^{(m)}(\mathbf{r}^{(m)} = u|r_o = i)}\right], \tag{13}$$

which compares the observed probability of career u with initial location i against its simulated probability. The quantity approaches 0 as the simulated and observed probabilities of a career become more similar (i.e., $\phi^{(m)}(\mathbf{r}^{(m)} = u|r_o = i) \approx \hat{\phi}(\mathbf{u} = u|u_o = i)$). On the other hand, if a model overestimates the frequency of u, $d^{(m)}(u,i) > 0$; if it is underestimated, $d^{(m)}(u,i) < 0$.

Using $d^{(m)}(u,i)$ over all observed careers beginning at i provides another way to test the models. This can be done, for a given node, by measuring the average $d^{(m)}(u,i)$ over observed career paths, or

$$\langle d^{(m)}(i)\rangle = \frac{\sum_{u\in\mathcal{U}(i)} d^{(m)}(u,i)}{|\mathcal{U}(i)|}. \tag{14}$$

As indicated, this quantity can also serve as a measure of the quality of a model at the level of each individual starting point for careers. A related quantity that can be derived is the variance of $d^{(m)}(u,i)$, defined as

$$\mathrm{var}(d^{(m)}(i)) = \frac{\sum_{u\in\mathcal{U}(i)}\left[d^{(m)}(u,i) - \langle d^{(m)}(i)\rangle\right]^2}{|\mathcal{U}(i)|} \tag{15}$$

which provides a measure of how well models capture the totality of the careers predicted to start at i.

A final use for $d^{(m)}(u,i)$ is introduce is the creation of a profile for the effectiveness of each model to recover individual observed careers. Let us create a rank-ordered list of careers $u \in \mathcal{U}(i)$ so that u_0 is the most probable career departing i (that is $\hat{\phi}(\mathbf{u} = u_0|u_o = i) > \hat{\phi}(\mathbf{u} = u|u_o = i)$ for $u \neq u_0$). Similarly, u_1 is the second most probable career from i, which means that $\hat{\phi}(\mathbf{u} = u_0|u_o = i) > \hat{\phi}(\mathbf{u} = u_1|u_o = i) > \hat{\phi}(\mathbf{u} = u|u_o = i)$ for $u \neq u_0, u_1$. After ordering all careers, we can construct the curve $\pi_i^{(m)}(c) = (c, 10^{d^{(m)}(u_c,i)})$ where $c = 0, 1, \ldots, |\mathcal{U}(i)| - 1$. This profile for node i shows in decreasing order of importance how closely model m is able to reproduce careers in i. A perfect model will tend to produce a flat curve of the form $(c, 1)$. On the other hand, if some careers deviate strongly, there will be noticeable jumps.

3. Results

3.1. The Validity of Organizational Labor Flow Networks

To verify that OLFNs are in fact informative, we apply the method in Section 2.2.2 where the time steps are monthly periods and the $\mathcal{T}_<$ and $\mathcal{T}_<$ are quarterly periods (3 months). To test that the information of previous transitions is strong enough, we simply impose $\mathcal{W} = 1$ and measure the time series of x_1 over the years of data we possess.

Given our ability to choose the definition of locations, we explore the three versions mentioned above, operating units, occupational series code, and geographic location (in this case, at the state level). The results are shown in Figure 1. The model used corresponds to fixed strength of nodes based on candidate links, the most demanding test based on results from [10]. For all choices of the definition of location, the excess probabilities $x_{\mathcal{W}}$ are considerably above 1 which means that defining and OLFN on the basis of any of these locations produces networks on which a walker (representing an employee) can travel along careers that are likely to be found in the real data. However, the value of $x_{\mathcal{W}}$ is larger

for units than other definitions of location (solid blue line). This result in interesting in that it reinforces the value of work done in [8–10] where nodes are defined on the basis of firms in the economy. The similarity is that, just like firms, operating units are the actual administrative units within which people work.

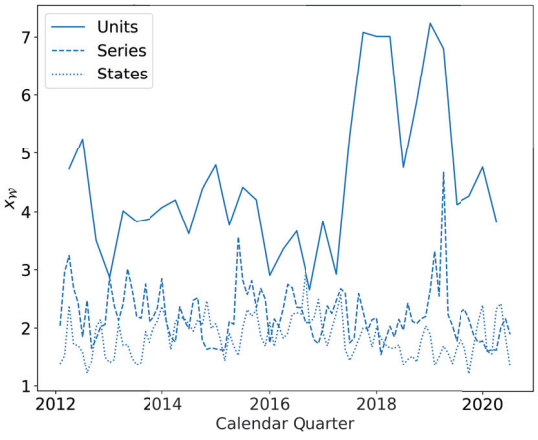


Figure 1. Quarterly excess probabilities $x_{\mathcal{W}}$ over the time frame of the data. We make the measurements with three different definitions of locations, operating units (dotted line), occupational series codes (circle), and US state where the employee is located (dashed line). The model corresponds to fixed strength of nodes based on candidate links, the most demanding test based on results from [10]. Even in this case, it is clear that $x_{\mathcal{W}}$ is markedly above 1.

Given the effectiveness of using operating units for predicting job change, we further explore this definition of network. In Figure 2, we study the effect of the threshold $\mathcal{W}$ on the excess probability $x_{\mathcal{W}}$. The temporal tracking is the same as in Figure 1. In this case, we see that increasing $\mathcal{W}$ leads to modest gains in predictive ability of the network, yet remaining within the same order of magnitude as $\mathcal{W} = 1$.

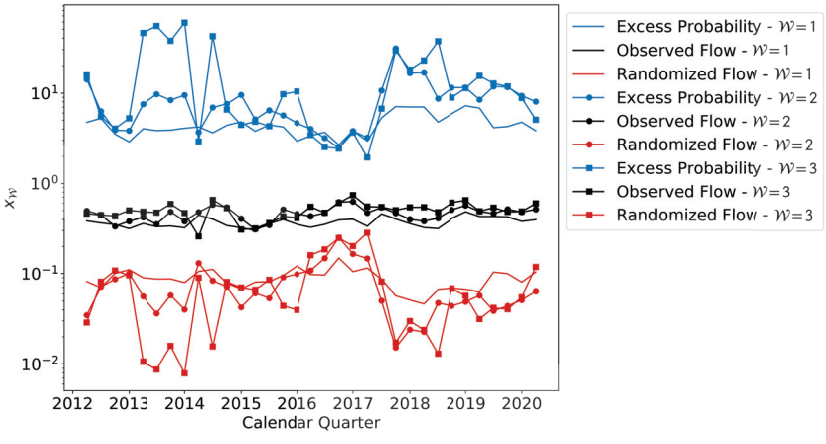


Figure 2. Quarterly excess probabilities $x_{\mathcal{W}}$ and the values of $\wp^{(o)}$ and $\wp^{(r)}$ across the time frame of the data for units in the AAW, tested across increasing $\mathcal{W}$. The dotted lines correspond to $\mathcal{W} = 1$, the circles to $\mathcal{W} = 2$, and dashed lines to $\mathcal{W} = 3$. The bundle of curves in the middle of the plot correspond to $\wp^{(o)}$. The lower bundle of curves represent $\wp^{(s)}(\mathcal{W})$ due to random models. Finally, the excess probabilities $x_{\mathcal{W}}$ are represented by the upper bundle of curves.

Based on this analysis, we conclude that even a single observed transition ($\mathcal{W} = 1$) between a node pair has considerable predictive power regarding future transitions and therefore, in the absence of some pre-established tolerance level, we adopt even a single job transition to be an acceptable link in an OLFN. Clearly, our result confirms that the idea of OLFNs is not just theoretical, but one that actually captures real employment affinity and can help predict future job changes. The results of this analysis also dictate how we define the probabilities of transitions in our Markov models (see Section 2.2.5).

3.2. Structure of Organizational Labor Flow Networks

Once networks are generated, we check their general topological characteristics. As indicated above, three possible definitions of nodes can be used, operational unit, occupational series code, or geography. However, given that geographic location appears to provide the smallest values of $x_{\mathcal{W}}$ above 1, we concentrate on the topological features of the OLFNs generated with locations defined as operational units and occupational series.

In Figure 3, we present the degree distributions $\Pr(k)$ of the OLFNs defined with $\mathcal{W} = 1$, where k represented the degree of a node. Given the small number of nodes present in the network built on occupational codes, the degree distribution (left) does not seem to provide a clear structure. The effect of the number of nodes on this lack of structure is another reason why our expectations for obtaining systematic results based on state locations as nodes are low, further justifying our obviating this analysis (while there are about 100 distinct occupations, there are only 50 states in the US; this small number of nodes is unlikely to show much connectivity structure).

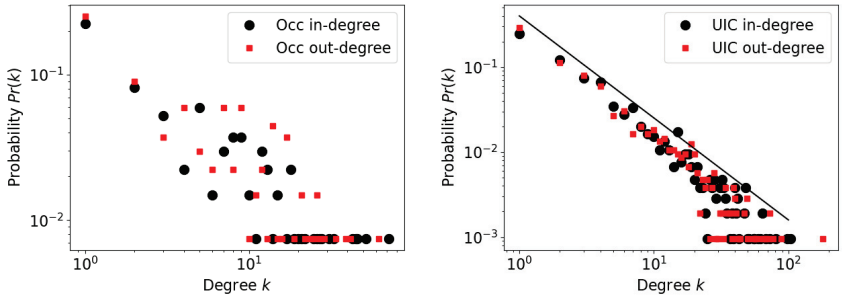


Figure 3. Degree distributions for OLFNs defined by occupational series (**left**) or operational units (**right**). The plots are shown in log-log scale. For units, we add a reference solid line that decays as $k^{-1.2}$.

On the other hand, when the network is defined in terms of operational units, much more topological information can be seen. First, the right panel of Figure 3 exhibits a long tail distribution of degree, with close to two decades of steady, near-linear decay in double logarithmic scale which is consistent with a power-law. Assuming this shape of the degree distribution (a power-law), we find by inspection a decaying slope of a value of ≈ -1.2, or $\Pr(k) \sim k^{-1.2}$. This slope is close to the value observed in much on the literature on the firm-size distribution, known to show exponents in a range near -1 but with considerable variation that includes the value -1.2 (see [29,30]). Although here we are reporting the probability of a node to have degree k, the degree is a consequence of job transitions which are proportional to the size of units. Consequently, the exponent we measure can be directly compared to that of the firm size distribution. In addition, no prior empirical work has addressed the internal structure of firms, and the simulation studies that have been performed [31] have predicted that the distribution of unit sizes inside a firm should *grow*, which is the opposite prediction to our observations.

The agreement between the exponent value found here and exponents in the literature on the firm-size distribution suggests that large organizations, even if they have highly

controlled structures, somehow organize themselves in a way that mimics the organization of entire economies. After the seminal paper by Simon and Bonani recognizing this phenomenon [12], and given the abundant literature on this topic (see e.g. [13,17]), we do not attempt to explain this phenomenon here. However, we do note the importance of this finding in the context of this debate because it suggest that the phenomenon is a truly emergent feature of the functioning of economic entities.

3.3. Jensen-Shannon Divergence

Moving beyond the macro-structure of the system, we now focus on the probabilistic structure of careers. For this purpose, we apply the JSD explained in Section 2.2.6. Given the limited value shown in defining careers in terms of geographic locations, we narrow our focus to operating units and occupational series only.

Evaluating the numerical values of JSD requires establishing a baseline, as explained above, that compares careers among the random distributions versus the comparison of careers between a random and the observed distribution. In Figure 4, we illustrate the nature of the results of our analysis. The left panel contains JSD distributions for one illustrative occupational series code. The model without memory is represented by the red and green distributions. The red distribution corresponds to M_d distinct values of the JSD between the distribution of observed careers $\hat{\phi}$ commencing in the occupational code of interest and the M_d modeled distributions $\phi^{(o)}$ of careers starting at the same occupation code. In contrast, the green distribution is constructed from the $M_d(M_d-1)/2$ distinct JSD values that emerge from comparing all the pairs of distributions M_d distributions $\phi^{(o)}$ with each other. From the figure, we see that the green distribution among the random career realizations is characterized by lower values of JSD. This should be expected from the fact that the careers generated by the model are fundamentally similar to each other. The red distribution, in contrast, has larger values of JSD because observed and random careers need not be as similar. It is notable, however, that the JSD values are small indicating that both random models perform well.

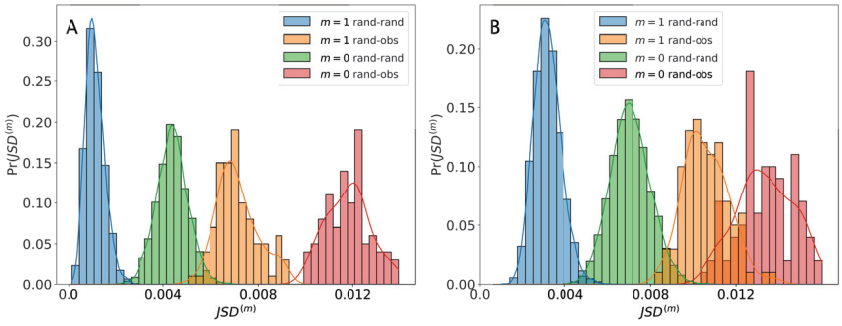


Figure 4. Distributions of values of JSD when comparing careers generated by random modeling and observed careers. Locations defined by operating unit on panel (**A**) and by occupational series are shown on panel (**B**). The model without memory can be seen on both panels, with the green distributions showing the pairwise comparisons between the random distributions of careers and the red showing the comparisons between the observed distribution against each of the random distributions. Similarly, the one-step memory model can also be seen on both panels (in blue), with the distributions showing the pairwise comparisons between the random distributions of careers and the orange showing the comparisons between the observed distribution against each of the random distributions.

When memory is introduced, generating random career distributions $\phi^{(1)}$, the orange and blue JSD distributions emerge. Once again, the JSD values that emerge from comparing $M_d(M_d-1)/2$ random distributions in pairs are lower (blue) than the M_d distinct JSD

values that compare $\hat{\phi}$ with $\phi^{(1)}$. We can also observe that memory lowers the JSD values of these distributions in comparison to the ones from the model without memory.

Similar results can be gleaned when locations are redefined to operating units (Figure 4, right panel). While the examples presented in Figure 4 correspond to a particular unit and occupational code, the qualitative characteristics observed are consistent for the remaining nodes and definitions of locations.

3.4. Jaccard Index

Having tested the similarity of the distributions, we are now in a position to determine if the structure of careers predicted by the models is similar to real observed careers. As explained above in Section 2.2.6, the Jaccard index eliminates the advantage that comes to popular careers when evaluated through the JSD. Instead, all careers are compared on equal footing, providing much more clarity about the difference between the models and the real-world.

Although it would be perfectly informative to generate distributions of values of the Jaccard index, it is very useful to compare the two models we use directly on the basis of their ability to achieve large values of Jaccard index approaching 1. Each point in Figure 5 corresponds to a starting career location i (left are occupational series, right are units) where the horizontal coordinate represents the Jaccard index of $\mathcal{U}(i) \cap \left[\prod^{M_d} \cup \mathcal{R}^{(1)}(i)\right]$ and the vertical coordinate to the Jaccard index of $\mathcal{U}(i) \cap \left[\prod^{M_d} \cup \mathcal{R}^{(0)}(i)\right]$.

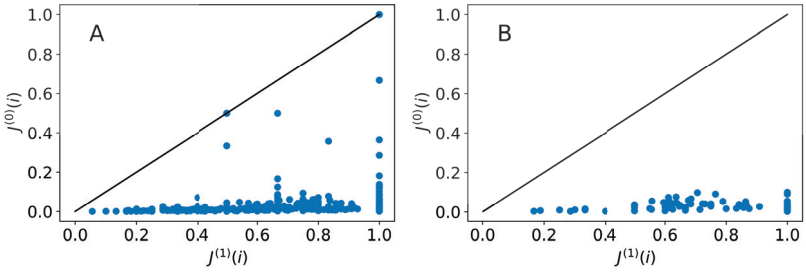


Figure 5. Comparison of Jaccard indices calculated from models with memory and without memory for units (**B**) and occupational codes (**A**) for locations in the OLFN . Each point is an initial node for careers. The horizontal coordinate captures the Jaccard index of the collected M_d careers created in the one-step memory model, and the vertical the Jaccard index of the collected M_d careers created with the memoryless model. The solid line highlights the diagonal of the plot.

The results clearly illustrate the situation. The memoryless model is hardly ever able to approach the value 1, generating values that are almost exclusively confined in the range between 0 and 0.1 (with some exceptions). On the other hand, the model with one-step memory is partially successful at achieving Jaccard indices of 1, as well as generating other less optimal, yet better performing values between 0 and 1 in comparison to the memoryless model.

3.5. Career Profiles and Overall Evaluation of Career Forecasts

In order to develop better intuition about the ability of our models to replicate observation, we also study the career profiles generated.

In Figure 6 we present the profiles $\pi_i^{(m)}(c)$ for the same operating unit and occupational code as those in Figure 4 with both the memoryless and one-step memory models. In both panels, it is clear that generally the one-step memory model performs better than the memoryless model. Deviations tend to be more attenuated. In both examples, the quality of the forecast of the most likely careers starting from each of the nodes (the points to the

left of the plot representing $u_0, u_1, \ldots$) is high, represented by the fact that the symbols can be located near the reference line at height 1.

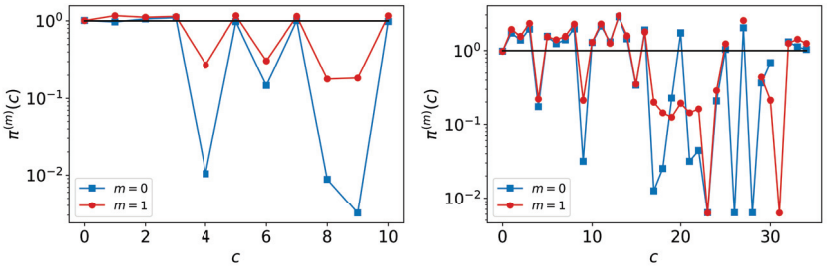


Figure 6. Career profiles $\pi_i^{(m)}(c)$ starting from location i. The left panel uses the same unit that is displayed in the left panel of Figure 4; the right panel uses the same occupation as in the right panel of Figure 4. The memoryless model career profiles are shown in blue and the one-step memory models in red. Since the most important careers from the standpoint of probability distribution are found on the left of the plot (smallest values of c), it is clear that the models perform well. However, one-step memory is more effective. In addition, forecasting on the basis of occupations proves to be less reliable than using units, as can be seen from the larger number of large fluctuations in the profiles.

To assess the models globally in terms of their performances, we show the distribution of Equations (13) measured across starting career nodes (units and occupations) and models. For $d^{(m)}(u, i)$, we present Figure 7 covering units and occupations with memoryless and one-step memory models. From the figures, we see that the majority of careers are forecasted correctly (with appropriate values of probability), seen by the concentration of the peaks around 0. Note, however, that the efficiency of the OLFN based on units is superior.

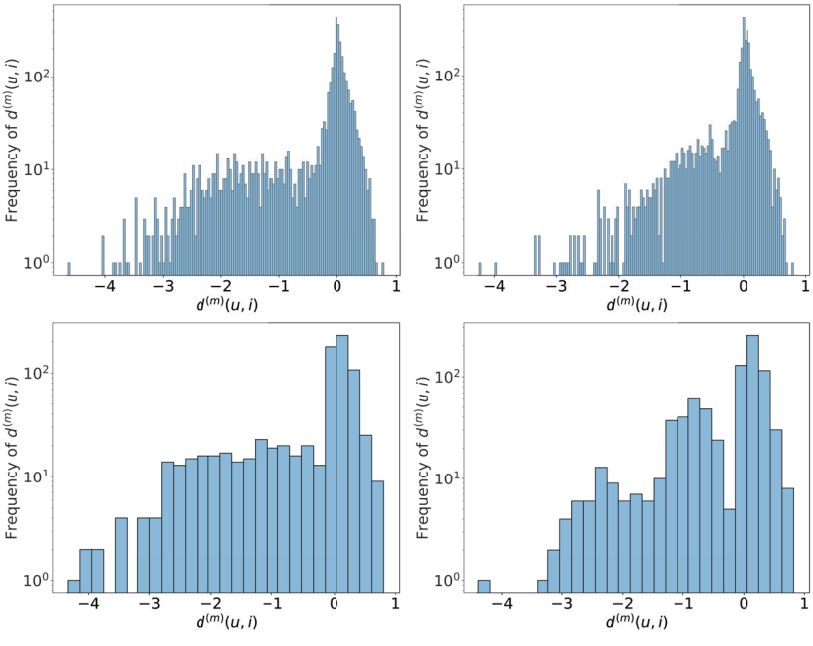


Figure 7. Histograms of $d^{(m)}(u, i)$ for the memoryless (**left column**) and one-step memory (**right column**) models for OLFNs of units (**top row**) and occupations (**bottom row**) across all starting nodes

i and all observed careers. Despite the tails on the left of the histograms, the large frequencies are highly concentrated around 0, indicating the general effectiveness of the models in reproducing real careers. Memory indeed helps arrive at better predictions of careers, as can be seen from the reduction of the amount of the mass of the histograms for negative values of $d^{(m)}(u, i)$. The plots have different numbers of bins because OLFNs based on units (top row) have many more nodes than those based on occupations (bottom row); see Section 2.1.

As a final assessment, we present in Figure 8 histograms of the values of $\mathrm{var}(d^{(m)}(i))$ from Equation (15), for OLFNs generated from units and occupations. In this case, we see once again the that models perform well given the large frequency of 0. In addition, OLFNs with unit-defined nodes continue to perform best.

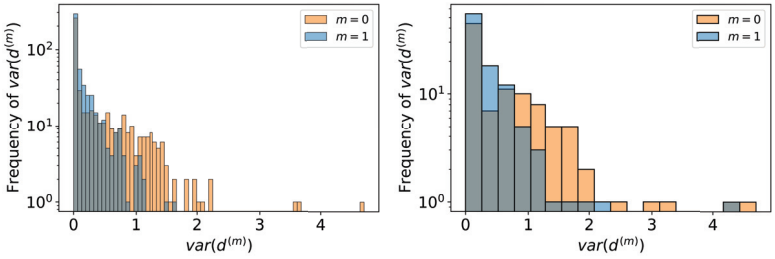


Figure 8. Histograms of the values of $\mathrm{var}(d^{(m)}(i))$ over all starting nodes for units (**left**) and occupations (**right**). The memoryless model is captured by the orange histogram and the one-step memory by the blue histogram. The variances are concentrated around 0 and the tails of the distributions decay very fast, with units generally displaying smaller values of variance and hence better forecasting capabilities.

4. Some Final Discussions and Conclusions

The results we have obtained in the manuscript show that the LFN network approach can be successfully applied to internal labor markets, leading to OLFNs. Further, we find that even a single transition is effective at providing evidence that two nodes in a labor network should be connected, a result that is in agreement with the finding in [7–10]. With this in mind, we are able to generated OLFNs that provide the substrate on which to model careers (either for memoryless or one-step memory models) that allows us to forecasts the workforce job changes at a microscopic level, i.e., for any career sequence.

The finding that OLFNs of units behave in the same way as the firm-size distribution is new and of critical importance. The lack of availability of data such as the one presented here has made the reporting of this regularity impossible. However, the relevance of this observation is that it opens a new window into our understanding of this interesting yet not-fully explained phenomenon. In particular, given the supervised nature of the structure of a single organization such as this one, it suggests that the firm-size distribution may be a consequence of an optimization process that seeks to make the functioning of the interacting units as efficient as possible.

Another advancement of our paper is the introduction of a notion of career sequences occurring on a network of operational units, new in the study of careers. We expect that, as we focus more on its details, numerous relevant features of the system will start to emerge such as the value of work or friendship ties in people's careers.

We find that the introduction of memory in the modeling of careers is an essential component that has been missing from the approaches that have so far been deployed for this problem. A variety of techniques have been used in order to understand movements of individuals across an organization, including pattern clustering of sequences [32,33], Markov modeling performed at several levels of sophistication [6,24,34], and manpower analysis [4]. However, the use of memory has so far been neglected as an effective approach to model careers.

A limitation of our current methodology is that it is calibrated against observed job transitions rather than possible job transitions. This is an important issue because the finite nature of the system does not provide enough observation of rare transitions to be extracted from the data. In order to overcome this, study of the characteristics of each job (say, occupational series, location, career field) offers a new direction to pursue in order create a more flexible model that may be able to predict what could happen even if it has never been observed.

We believe that the analysis performed here, including the application of new ideas and techniques, will spark interest in pushing this topic forward, and attempting to bring together the related but generally non-intersecting approaches that have so far been deployed in career studies.

Author Contributions: Conceptualization, F.W. and E.L.; methodology, F.W., E.L., D.S. and M.P.; software, F.W., E.L. and M.P.; data analysis F.W., D.S. and E.L., writing—original draft preparation, F.W. and E.L., supervision, E.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been funded by the Army Research Office under contract W911NF2020117 and the Army Research Institute under contract W911NF2210250.

Institutional Review Board Statement: The current project was found to be exempt under US federal law from institutional review by the George Mason Review Board due to being secondary research for which consent is not required.

Informed Consent Statement: Not applicable.

Data Availability Statement: Due to the proprietary nature of the data, these cannot be shared without the express authorization from the US Department of the Army.

Acknowledgments: We acknowledge helpful discussions with Marko Nikituk.

Conflicts of Interest: The funders of the work (Army Research Office and Army Research Institute) have played no role in the the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. The Office of the Director of Acquisition Career Management of the US Army has collaborated in this research but has not influenced the analyses or interpretation of the data. The study design is such that no elements of the data and analyses could compromise personal privacy or US national security.

Abbreviations

The following abbreviations are used in this manuscript:

LFN	Labor Flow Network
OLFN	Organizational Labor Flow Network
JSD	Jensen-Shannon Divergence

References

1. Petrongolo, B.; Pissarides, C.A. Looking into the Black Box: A Survey of the Matching Function. *J. Econ. Lit.* **2001**, *39*, 390–431. [CrossRef]
2. Rogerson, R.; Shimer, R.; Wright, R. Search-Theoretic Models of the Labor Market: A Survey. *J. Econ. Lit.* **2005**, *43*, 959–988. [CrossRef]
3. Gunz, H.; Lazarova, M.; Mayrhofer, W. *The Routledge Companion to Career Studies*; Routledge: Abingdon, UK, 2019.
4. De Feyter, T.; Guerry, M. Markov models in manpower planning: A review. In *Handbook of Optimization Theory: Decision Analysis and Applications*; Varela, J., Acuña, S., Eds.; Nova Science Publishers: Hauppauge, NY, USA, 2011; pp. 67–88.
5. Young, A.; Almond, G. Predicting Distributions of Staff. *Comput. J.* **1961**, *3*, 246–250. [CrossRef]
6. Bartholomew, D. *Stochastic Models for Social Processes*, 3rd ed.; Wiley Series in Probability and Mathematical Statistics Series; Wiley: Hoboken, NJ, USA, 1982.
7. Collet, F.; Hedström, P. Old friends and new acquaintances: Tie formation mechanisms in an interorganizational network generated by employee mobility. *Soc. Netw.* **2013**, *35*, 288–299. [CrossRef]
8. Guerrero, O.A.; Axtell, R.L. Employment Growth through Labor Flow Networks. *PLoS ONE* **2013**, *8*, e60808. [CrossRef]

9. Axtell, R.L.; Guerrero, O.A.; López, E. Frictional unemployment on labor flow networks. *J. Econ. Behav. Organ.* **2019**, *160*, 184–201. [CrossRef]
10. López, E.; Guerrero, O.A.; Axtell, R.L. A network theory of inter-firm labor flows. *EPJ Data Sci.* **2020**, *9*, 33. [CrossRef]
11. Krapivsky, P.L.; Redner, S.; Ben-Naim, E. *A Kinetic View of Statistical Physics*; Cambridge University Press: Cambridge, UK, 2010. [CrossRef]
12. Simon, H.A.; Bonini, C.P. The Size Distribution of Business Firms. *Am. Econ. Rev.* **1958**, *48*, 607–617.
13. Axtell, R.L. Zipf Distribution of U.S. Firm Sizes. *Science* **2001**, *293*, 1818–1820. [CrossRef]
14. Villarreal, A. The U.S. Occupational Structure: A Social Network Approach. *Sociol. Sci.* **2020**, *7*, 187–221. [CrossRef]
15. del Rio-Chanona, R.M.; Mealy, P.; Beguerisse-Díaz, M.; Lafond, F.; Farmer, J.D. Occupational mobility and automation: A data-driven network model. *J. R. Soc. Interface* **2021**, *18*, 20200898. [CrossRef]
16. Wang, F.; Smolyak, A.; Dong, G.; Tian, L.; Havlin, S.; Sela, A. Group Homophily as a Measure of Self-Liking of Communities: Application in Vocational Networks. *Res. Sq.* **2022**. [CrossRef]
17. Stanley, M.H.R.; Amaral, L.A.N.; Buldyrev, S.V.; Havlin, S.; Leschhorn, H.; Maass, P.; Salinger, M.A.; Stanley, H.E. Scaling behaviour in the growth of companies. *Nature* **1996**, *379*, 804–806. [CrossRef]
18. Rosvall, M.; Esquivel, A.V.; Lancichinetti, A.; West, J.D.; Lambiotte, R. Memory in network flows and its effects on spreading dynamics and community detection. *Nat. Commun.* **2014**, *5*, 4630. [CrossRef]
19. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
20. Department of the Army. Army Acquisition Workforce. 2023. Available online: https://asc.army.mil/web/career-development/about-aaw/ (accessed on 19 April 2023).
21. Office of Personnel Management, Classifying General Schedule Positions. Available online: https://www.opm.gov/policy-data-oversight/classification-qualifications/classifying-general-schedule-positions/ (accessed on 27 February 2023).
22. Bartholomew, D.J. The Statistical Approach to Manpower Planning. *J. R. Stat. Soc. Ser. D* **1971**, *20*, 3–26. [CrossRef]
23. Newman, M. *Networks*; Oxford University Press: Oxford, UK, 2018.
24. White, H.C. *Chains of Opportunity, System Models of Mobility in Organizations*; Harvard University Press: Cambridge, MA, USA; London, UK, 1970. [CrossRef]
25. Bryant, D.T. A Survey of the Development of Manpower Planning Policies. *Br. J. Ind. Relations* **1965**, *3*, 279–290. [CrossRef]
26. Paninski, L. Estimation of Entropy and Mutual Information. *Neural Comput.* **2003**, *15*, 1191–1253. [CrossRef]
27. Grassberger, P. Entropy Estimates from Insufficient Samplings. *arXiv* **2008**, arXiv:0307138.
28. DeDeo, S.; Hawkins, R.X.D.; Klingenstein, S.; Hitchcock, T. Bootstrap Methods for the Empirical Study of Decision-Making and Information Flows in Social Systems. *Entropy* **2013**, *15*, 2246–2276. [CrossRef]
29. Fujiwara, Y.; Di Guilmi, C.; Aoyama, H.; Gallegati, M.; Souma, W. Do Pareto–Zipf and Gibrat laws hold true? An analysis with European firms. *Phys. A Stat. Mech. Its Appl.* **2004**, *335*, 197–216. [CrossRef]
30. Guerrero, O.A. Coupled Dynamics of Labor and Firms through Complex Networks. Ph.D. Thesis, George Mason University, Fairfax, VA, USA, 2013.
31. Amaral, L.A.N.; Buldyrev, S.V.; Havlin, S.; Salinger, M.A.; Stanley, H.E. Power Law Scaling for a System of Interacting Units with Complex Internal Structure. *Phys. Rev. Lett.* **1998**, *80*, 1385–1388. [CrossRef]
32. Abbott, A.; Hrycak, A. Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers. *Am. J. Sociol.* **1990**, *96*, 144–185. [CrossRef]
33. Aisenbrey, S.; Fasang, A.E. New life for old ideas: The "second wave" of sequence analysis bringing the "course" back into the life course. *Sociol. Methods Res.* **2010**, *38*, 420–462. [CrossRef]
34. Stewman, S. Demographic Models of Internal Labor Markets. *Adm. Sci. Q.* **1986**, *31*, 212–247. [CrossRef]

entropy

MDPI

Article

Entropy of Financial Time Series Due to the Shock of War

Ewa A. Drzazga-Szczęśniak [1], Piotr Szczepanik [2], Adam Z. Kaczmarek [3] and Dominik Szczęśniak [3,*]

[1] Department of Physics, Faculty of Production Engineering and Materials Technology, Częstochowa University of Technology, 19 Armii Krajowej Ave., 42200 Częstochowa, Poland; ewa.drzazga@pcz.pl
[2] Institute of Pricing and Market Analysis, Analitico, 49/8 Królewska Str., 47400 Racibórz, Poland; pszczepanik@instytut-analitico.pl
[3] Department of Theoretical Physics, Faculty of Science and Technology, Jan Długosz University in Częstochowa, 13/15 Armii Krajowej Ave., 42200 Częstochowa, Poland; adam.kaczmarek@doktorant.ujd.edu.pl
* Correspondence: d.szczesniak@ujd.edu.pl

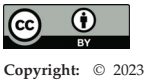
Abstract: The concept of entropy is not uniquely relevant to the statistical mechanics but, among others, it can play pivotal role in the analysis of a time series, particularly the stock market data. In this area, sudden events are especially interesting as they describe abrupt data changes with potentially long-lasting effects. Here, we investigate the impact of such events on the entropy of financial time series. As a case study, we assume data of the Polish stock market, in the context of its main cumulative index, and discuss it for the finite time periods before and after outbreak of the 2022 Russian invasion of Ukraine. This analysis allows us to validate the entropy-based methodology in assessing changes in the market volatility, as driven by the extreme external factors. We show that some qualitative features of such market variations can be well captured in terms of the entropy. In particular, the discussed measure appears to highlight differences between data of the two considered timeframes in agreement with the character of their empirical distributions, which is not always the case in terms of the conventional standard deviation. Moreover, the entropy of cumulative index averages, qualitatively, the entropies of composing assets, suggesting capability for describing interdependencies between them. The entropy is also found to exhibit signatures of the upcoming extreme events. To this end, the role of recent war in shaping the current economic situation is briefly discussed.

Keywords: entropy; volatility; information theory; econophysics; sudden events; war; time series; data science

Citation: Drzazga-Szczęśniak, E.A.; Szczepanik, P.; Kaczmarek, A.Z.; Szczęśniak, D. Entropy of Financial Time Series Due to the Shock of War. *Entropy* **2023**, *25*, 823. https://doi.org/10.3390/e25050823

Academic Editor: Panos Argyrakis

Received: 5 April 2023
Revised: 8 May 2023
Accepted: 18 May 2023
Published: 21 May 2023



1. Introduction

In general, sudden or extreme events translate to the atypical patterns and deviations from the expected observations. As such, the ability to detect and address accordingly aforesaid anomalies is of great importance in various areas of science, technology, or even social studies [1–5]. This is to say, timing and occurrence of sudden events is essential when considering reliability of a system under extreme external conditions. A special attention to these aspects is given in the field of economy, where sudden events correspond to a notable incline/decline in economic activity or may even mark a breakdown of some economic models, e.g., by exposing their limitations in terms of efficiency and rationality of the market [6]. In what follows, it is crucial to account for such events during economic modeling when considering processes such as the forecasting, decision making, or anomaly detection [7]. This is conventionally carried out on the grounds of the time series analysis, a vital part of data science [8]. The main reason for that is related to character of the time series itself, which are derived from the financial data and intrinsically encode information about economic events [9]. Thus, to allow discussion of the extreme changes in economy, appropriate tools in the time series domain are required.

In the context of the above, entropy appears as an intriguing analytical concept, which spans beyond its original field of thermodynamics. While in terms of the statistical mechanics, this property relates to the discrete probabilities of microstates, in the area of time series entropy, it is considered as an extension of the information theory [10–13], in accordance with the groundbreaking works by Shannon and Kolmogorov [14,15]. In particular, entropy can quantify the uncertainty, disorder, or simply randomness of the time series, without adding constraints on the corresponding probability distribution [13,16–18]. Hence, it constitutes an attractive alternative to the standard deviation for measuring market volatility [19,20]. However, entropy allows for discussing not only the magnitude of such fluctuations but also their distributions and patterns [21,22]. It can account for the nonlinearities and correlations in the datasets, simultaneously capturing interdependence between assets [23–25]. Moreover, since volatility relates to the degree of an asset movement over time, the corresponding entropy should be inherently sensitive to the sudden events or the economic shocks of interest. As a result, entropy constitutes potentially highly relevant framework for discussing impact of sudden events on the market and a pivotal tool in econophysics [10,12,13,18,26].

So far, the studies on the economic sudden events in terms of entropy have been limited mainly to a few instances, such as the investigations related to the 2008 economic crisis [27] or to the outbreak of the COVID-19 pandemic [11]. However, recent Russian invasion of Ukraine resulted in a yet another prominent economic shock, which is well defined in terms of the timeframes, and influences multiple market branches [28]. The economic consequences of this event constitute not only a perfect platform to investigate the impact of the shock of war on the modern economy but also to validate the entropic methodology in assessing market changes due to the extreme external factors. These arguments, along with the aforementioned general characteristics of entropy in the field of econophysics, constitute intriguing motivation to analyze this new measure in terms of the market volatility description and the resulting potential for the detection of sudden events. Herein, we provide our contribution to this still not fully explored area. In detail, we concentrate our study on the behavior of the main cumulative index of the Polish stock market (WIG20) and conduct our calculations with respect to the conventional Shannon entropy. The WIG20 index was chosen due to the direct proximity of the corresponding market to the theater of war as well as the relatively high development of the Polish economy. For convenience, the obtained results are compared with the predictions of the standard deviation. This analysis allows us to verify efficiency and predictive capabilities of the entropy-based formalism and to outline pertinent perspectives for the future research. It also provides the possibility to give preliminary insights into the other factors potentially influencing the WIG20 index, besides the pivotal shock of war.

2. Methodology

The present analysis is conducted for the time series of the daily log-returns, as calculated based on the financial data of interest. In particular, the daily log-returns (R_i) are derived by following the relation:

$$R_i = \ln \frac{P_i}{P_{i-1}} \approx \frac{P_i - P_{i-1}}{P_{i-1}}, \tag{1}$$

where P_i (P_{i-1}) is the closing price of an asset on day i ($i-1$). In this manner, we obtain convenient time series data which are additive and symmetric in accordance with the scope of the present analysis. While the former property simply means that the log-returns are additive over time, the second one is much less self-explanatory. In brief, the symmetry of log-returns relates to the fact that positive and negative log-returns of equal magnitude are equidistant from zero on the logarithmic scale, yielding no net change when compared.

The volatility of the above time series is explored based on the two measures, namely, the standard deviation and the entropy. The former parameter is given by:

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(R_i - \mu)^2}, \tag{2}$$

for the N data points and μ being the arithmetic mean of all the returns. On the other hand, the latter measure is calculated based on the Shannon entropy [14]:

$$H = -\sum_{i=1}^{M} p_i \ln p_i. \tag{3}$$

In Equation (3), M stands for the number of bins (known also as the *intervals* or *classes* [29]) in the discrete probability density function of the returns and p_i is the probability related to a given bin. Herein, p_i is calculated by employing the Riemann approximation as follows:

$$p_i = (x_{i+1} - x_i) f(x_{i+1}), \tag{4}$$

where $x_i(x_{i+1})$ is the left (right) width endpoint of a bin and $f(x_{i+1})$ denotes the corresponding height. Note that, in Equation (3), when the logarithm base is e, the entropy is measured in *nats*. One can also use base equal to 2 or 10, resulting in the units of *shannons* or *hartleys*, respectively. Obviously, the change of units does not influence the qualitative behavior of entropy.

In the present study, the above theoretical model is fed with the financial data of the WIG20 cumulative index and its composing stocks, as divided into two one-year-long datasets. The first set corresponds to the one-year timeframe before the invasion (24 February 2021–23 February 2022), whereas the second considers a similar period but after the beginning of the invasion (24 February 2022–23 February 2023). In the following, we arrive with the total of $N = 251$ data points for each set, providing sufficient economic perspective for our calculations. Note that the WIG20 index serves here as a pivotal parameter for comparison between the two approaches in modeling volatility. However, due to its cumulative character, this index measures only the total fluctuations, and to gain better insight into the underlying correlations of the market, the composing stocks are discussed. All of these stocks, including the WIG20 index, are listed in Table A1 along with their full names, market symbols, and the basic summary statistics in Appendix A. This list is valid for the assumed-here time period but it is obviously subject to changes in the future. For the sake of completeness, it is also crucial to note that the component company Pepco was introduced to the stock market on 26 May 2021, i.e., the corresponding records do not cover the entire one-year period before the invasion. In addition, the composition of the WIG20 index changed four times over the analyzed timeframe of two years. In detail, on 18 March 2022, the already-mentioned Pepco and other company named mBank replaced previously indexed stocks of Tauron and Mercator, respectively. Similarly, on 16 September 2022, the company Kęty replaced Lotos, and on 16 December 2022, Kurk switched with the PGING. All the described changes are appropriately marked in the Section 3 and in Appendix A.

To this end, for the purpose of the present study, both the datasets of interest are divided into the finite number of bins, which compose the discrete probability density function of the returns. There is no general and valid rule that determines the number and character of such bins [29]. The final choice is always strongly related to the population of data points and their variability. In general, one should never stay with the empty bins or decrease their number to the point that resolution of the probability distribution is too low. In reference to the multiple models for the bin number, we observe that $M = 20$ is optimal for our case. In the first place, the chosen M value provides relatively high resolution of

the probability distribution on equal footing across all considered time series, allowing us to not hinder information about the tails in some of the instances. Secondly, the assumed number of bins does not simultaneously exceed the upper theoretical limits for $N \sim 250$, as set by the Velleman formula [29].

3. Results

In Figure 1, we depict standard deviation as calculated for the WIG20 index and its composing stocks. According to the initial assumptions, the results are presented here for the one-year timeframe before (orange) and after (blue) the beginning of the invasion. Note that Figure 1 is divided into three panels: the first for the constant component companies, whereas the second (third) panel corresponds to the stocks, which at some point were introduced to (removed from) the WIG20 index. For convenience, the corresponding numerical results and the percentage difference between estimates obtained for the two considered timeframes are given in Table A2 in Appendix A.

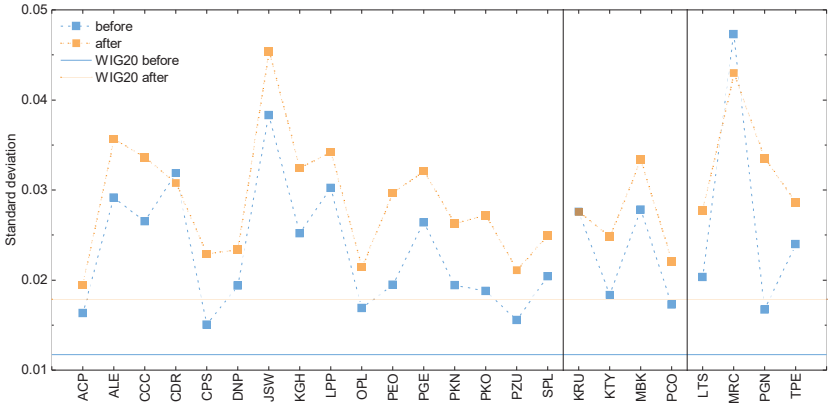


Figure 1. The standard deviation for the WIG20 index and its composing stocks. The first panel is for the constant component companies and the second (third) for the stocks introduced to (removed from) the index at some point. The results are given for the one-year time period before the beginning of the Russian invasion of Ukraine (blue) and after this event (orange). The solid lines correspond to the WIG20 index, whereas closed symbols represent estimates for the component stocks. Dashed lines are the guide for an eye.

Upon the analysis of Figure 1, the total standard deviation appears to be higher after the beginning of the invasion than for the time period before it. This means that the volatility of the market visibly increases for the former dataset. In other words, this indicates higher degree of stock price variations in the second considered period, which can be caused by not only the decline but also incline of the asset value. Similar behavior can be observed for most of the composing stocks. In detail, one can notice that only companies such as Kruk (debt management and purchase), Mercator (medical devices), and CD Projekt (video game developer and publisher) do not comply with this trend. The first company shows practically indistinguishable values for the two considered datasets, while the two latter ones present inverse behavior in comparison to the total standard deviation. The observed standard deviation for the first two companies is potentially related to the fact that their stocks were not included in the WIG20 index for the entire time, meaning their impact on the total index was limited. Moreover, Mercator capitalization, as a producer of medical gloves, was heavily reduced by the end of the COVID-19 pandemic. Finally, the value of CD Projekt was subject to turbulence due to the mixed reviews of their flagship video game product Cyberpunk 2077. Thus, the standard deviation for each of the three

companies is the results of not only the wartime market changes but also other, external factors. Despite these deviations, it can be stated that most of the composing assets as well as the cumulative results highlight the impact of the shock of war.

Nonetheless, the results for the individual stocks still allow us to observe that the largest volatility increase is present for the bank sector, with other notable examples in petroleum and telecommunication sectors (see Table A2 in Appendix A for details). Interestingly, by comparing the component estimates with the results for the cumulative index, we can note that the total standard deviation measure does not average values obtained for the individual stocks. In fact, this measure is always lower than any of the corresponding component values. This is true for both considered sets of data and can originate from the way that the cumulative index is calculated or potentially from the shortcomings of the standard deviation approach.

To investigate more in detail the already observed trends, in Figures 2 and 3 we present the discrete probability density function of the returns for the total index and its composing stocks, respectively. Note that these are the empirical distributions of the pooled returns. All the distributions are given for the time period before and after the beginning of the invasion, with the same color scheme as before. Based on Figure 2, it can be observed that the probability distributions for the WIG20 index resemble normal distribution. However, the wartime dataset is characterized by the fatter tails and lower central maxima than the distribution corresponding to the index values before the conflict outbreak. This observation is in qualitative agreement with the results obtained within the standard deviation approach, which suggest higher volatility of the market after the beginning of the invasion. The situation is once again similar when inspecting return distributions for the component stocks, i.e., volatility for most of the stocks is higher after the beginning of the Russian invasion. Still, there is some visible exception from this trend in terms of Pepco data. This is potentially due to the fact that, as mentioned earlier, data for Pepco do not cover the entire year before the invasion because of its relatively late introduction to the market on 26 May 2021.

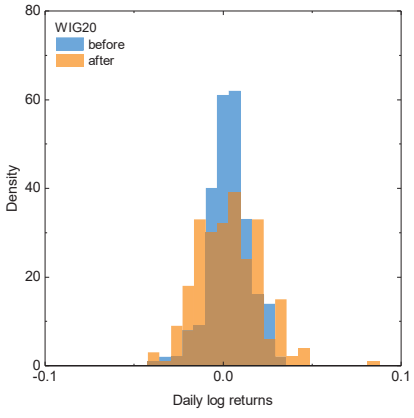


Figure 2. The discrete probability density function for the WIG20 index, for the one-year period before (blue) and after (orange) the beginning of the Russian invasion of Ukraine.

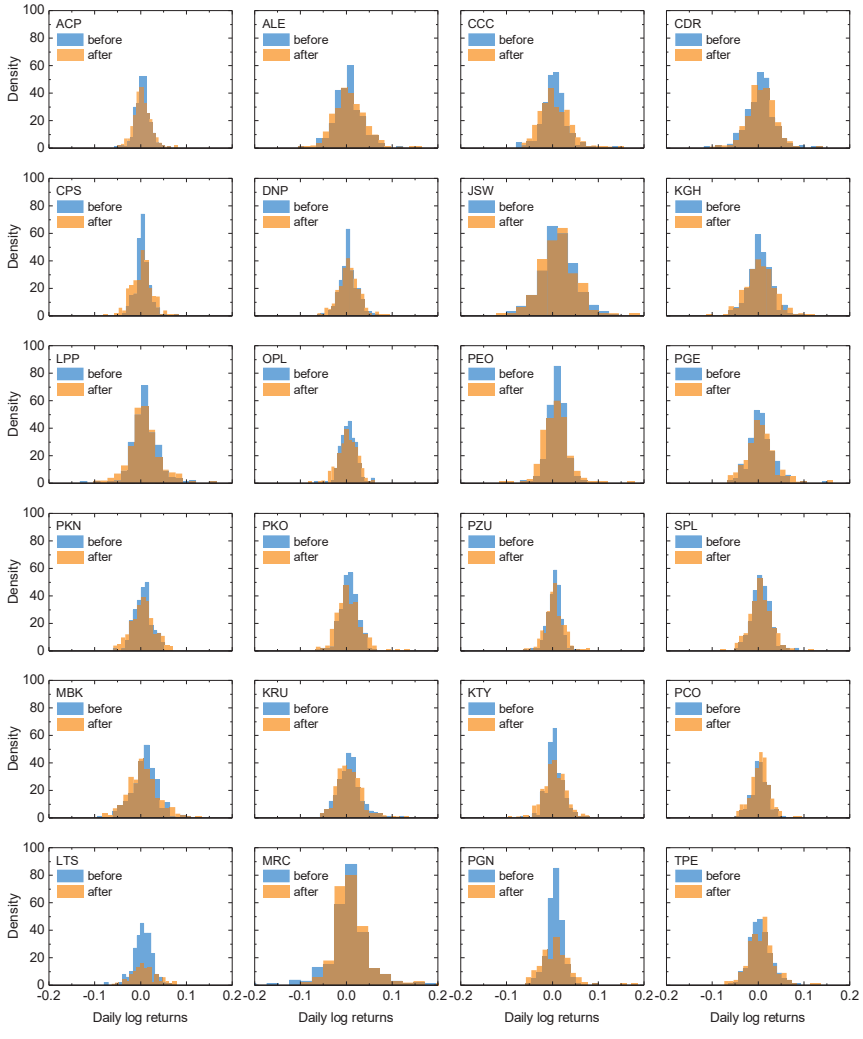


Figure 3. The discrete probability density function for the component stocks of the WIG20 index. The first four rows are for the constant component companies, and the fifth (sixth) row is for the stocks introduced to (removed from) the index at some point. The results are presented for the one-year time period before the beginning of the Russian invasion of Ukraine (blue) and after this event (orange).

It is next instructive to compare all the above results with the predictions of the entropic model. These are presented in Figure 4, in the form of the entropy estimates for the WIG20 index and its component stocks, based on the two types of the datasets of interest. In general, the total entropy, as well as the relative behavior, between composing entropies is similar to the standard deviation predictions. However, closer inspection of the results allows us to observe that, contrary to the previous case, here, all the component stocks exhibit higher entropy after the war outbreak. The only exception is Mercator, relatively late in the WIG20 index and experiencing the COVID-19-related problems during the entire analyzed period, as described before. Moreover, this time, the results for the total index qualitatively average results for the component companies. The mentioned observation is particularly visible for the data corresponding to the timeframe after the beginning of

the invasion. The results allow us also to note that the percentage difference between the results before and after invasion is smaller for each of the calculated entropies than in the case of the standard deviation results (see Table A2 in Appendix A for details). Finally, the obtained entropies appear to follow the character of the discrete probability density function in Figures 2 and 3, in terms of the differences between results obtained for the two considered timeframes.

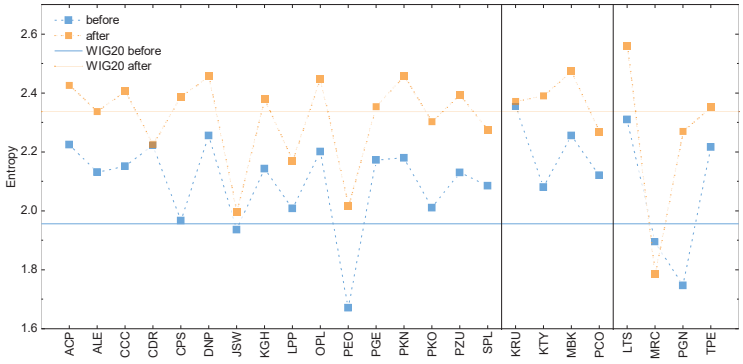


Figure 4. The Shannon entropy for the WIG20 index and its composing stocks. The first panel is for the constant component companies and the second (third) for the stocks introduced to (removed from) the index at some point. The results are given for the one-year time period before the beginning of the Russian invasion of Ukraine (blue) and after this event (orange). The solid lines correspond to the WIG20 index, whereas closed symbols represent estimates for the component stocks. Dashed lines are the guide for an eye.

To supplement our analysis, we additionally plot the entropic index of WIG20 index for various time periods within the here-assumed timeframes. In Figure 5, we present the obtained results for the datasets before (left panel) and after (right panel) the beginning of the considered conflict. Both panels depict different behavior, namely, before the invasion, the entropic index clearly increases when the assumed time distance from the invasion date becomes smaller. On the other hand, the entropy is relatively stable throughout the entire period after the invasion data, independent of the number of considered days.

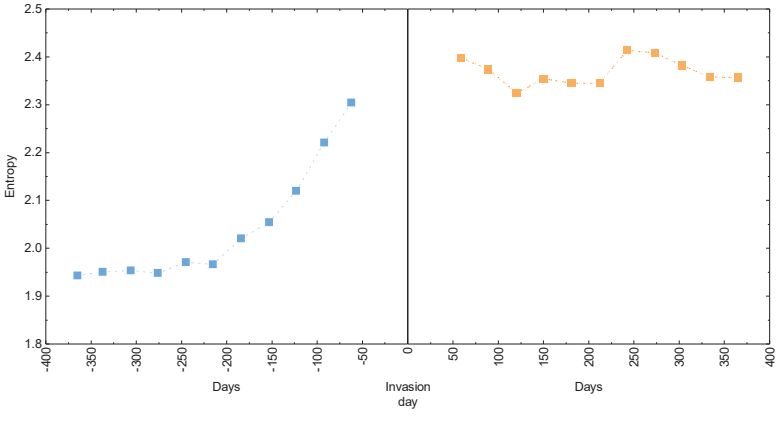


Figure 5. The Shannon entropy for the WIG20 index as calculated for different periods of time before (blue) and after (orange) the beginning of the Russian invasion of Ukraine. Dashed lines are the guide for an eye.

4. Conclusions

In the present study, we validated the entropy-based theoretical framework in describing behavior of financial time series under the influence of sudden and extreme external events. This was carried out in the context of the WIG20 main cumulative index of the Polish stock market for a one-year-long data samples before and after the Russian invasion of Ukraine, respectively. In particular, it was shown that entropy reproduces some of the features of the standard deviation when describing the effects of the shock of war. The obtained results confirmed that entropy can indeed be used as an alternative measure of volatility. These findings not only agree with the previous studies on applications of entropy in finances [10,11,18,24,27], but also supplement them by considering the wartime-driven changes in the stock market. For convenience, all the numerical results are summarized in Table A2 in Appendix A.

In addition to the above, the present study reveled several noteworthy differences between entropy and standard deviation measures. First, the entropy was found to capture the character of empirical data in qualitative agreement with the discrete probability distribution function, which was not always the case when considering the standard deviation measure. As a result, it is concluded that the entropy was better in highlighting differences between results obtained for the two timeframes of interest. This was particularly visible in the case of CD Projekt data, where standard deviation predicted inverse behavior to the probability distribution function and entropy. Finally, it was also revealed that the entropy of cumulative index qualitatively averages entropies of the composing stocks, again in contrast to the standard deviation estimates. This finding is particularly interesting since it shows that entropy holds potential in encompassing interdependencies between assets.

The last part of the analysis revealed that the entropy measure can be used to quantify anomalies in time series toward their better detection. In particular, entropy exhibits different functional character when considering it for various time periods, before and after the beginning of the invasion. In other words, it can be argued that entropy shows signatures of the upcoming economic shock. That means that the impact of a potential sudden event can be visible in the entropy behavior when the time range is sufficiently small and the context data are available for a long time range. In the future, entropy may constitute a building block for future tools aimed at sudden (extreme) event prediction. Interestingly, these results also clearly indicate that the shock of war has a long-lasting effect of increased volatility of the market, at least within the one-year time perspective. To further verify the presented observations, we note that the analysis can be extended toward other more complex or larger datasets and be conducted via more sophisticated entropic models based not only on the Shannon entropy but also on other formulations, e.g., by Rényi [30] or Tsallis [31].

To this end, all the obtained results allow us to make some preliminary statements on the role of invasion in shaping the current economic situation in Poland. The calculated standard deviation and entropy measures clearly point out that the volatility of the Polish market is higher after the crisis outbreak than before. That is to say, the presented study allows us to conclude that the shock of war visibly impacts the Polish economy, according to the fact that the entire analysis was conducted with respect to this well-defined point in time. However, it is difficult to judge how big this impact is in comparison to other factors, such as the still-persisting effects of the COVID-19 pandemic or the internal economic decisions of the Polish government and related financial institutions e.g., in terms of changes in the interest rates of the National Bank of Poland [32]. To address the impact of an additional factors, besides the considered shock of war, extended investigations spanning beyond the scope of the present analysis are required. This can be carried out by identifying the aspects of influence and then by analyzing them separately, but on the same footing, within the economic model of choice. However, due to the potential complexity of the problem, it is argued here that the proposed analysis should incorporate a more sophisticated approach based, for example, on the network and behavioral modeling, in agreement with the recent insights from the field of complex systems [33]. Since similar observations can be made at

the level of the European or even global market, it is expected that our entropic approach may provide interesting results with the dependence on the proximity to the conflict zone or the bond strength with the Ukrainian or Russian market. Still, such analysis will be limited in terms of influencing factors, and the above complex approach is expected to be also valuable for such large-scale simulations. In summary, the shock of war appears to be an important factor of recent economic turmoil, but its magnitude in the context of other factors is yet to be determined.

Author Contributions: Conceptualization, E.A.D.-S. and D.S.; methodology, E.A.D.-S. and D.S.; software, E.A.D.-S.; validation, E.A.D.-S., P.S. and A.Z.K.; formal analysis, E.A.D.-S. and P.S.; investigation, E.A.D.-S., P.S. and A.Z.K.; data curation, E.A.D.-S., P.S. and A.Z.K.; writing—original draft preparation, E.A.D.-S. and A.Z.K.; writing—review and editing, E.A.D.-S. and D.S.; visualization, E.A.D.-S. and P.S.; supervision, D.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Appendix A. Summary Statistics

The current Appendix section contains supplementary data to the discussion presented in the main text.

In Figure A1, the daily log-returns for the WIG20 cumulative index are depicted. The left panel presents data before the conflict outbreak (blue), whereas the right panel depicts data after the beginning of the invasion (orange). The inset presents a more detailed view in the vicinity of the initial invasion day. Qualitatively similar behavior can be observed for each of the WIG20 composing assets.

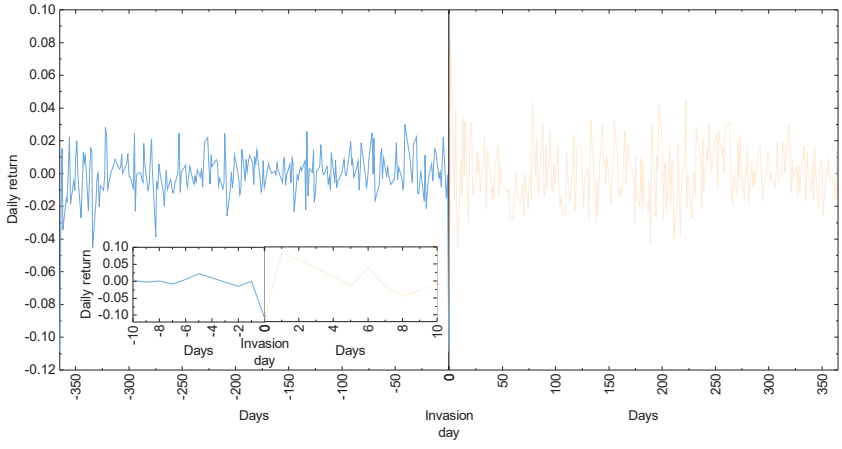


Figure A1. The daily log-returns for the WIG20 cumulative index before (blue) and after (orange) the beginning of the 2022 Russian invasion of Ukraine. For convenience, the inset presents data in the vicinity of the initial invasion day.

In Table A1, the basic summary statistics of the daily log-returns are given for the WIG20 cumulative index and its composing assets. The data are provided for two considered timeframes, i.e., one year before the invasion day and one year after this date. In Table A2, the numerical values of the calculated standard deviation and entropy, as

obtained within the presented analysis, are collected. The data are given in a similar fashion as in Table A1.

Table A1. The summary statistics of the daily log-returns before (outside the brackets) and after (inside the brackets) the outbreak of the 2022 Russian invasion of Ukraine. The cumulative index data are followed by the first group for the constant component companies and then by the second (third) group for the stocks introduced to (removed from) the index at some point. The dates of the introduction/removal of the composing companies are given next to the name of the corresponding company.

Name	Symbol	Mean	Minimum	Maximum	Skewness	Kurtosis
Cumulative Index	WIG20	0.00027 (0.00017)	−0.0455 (−0.0452)	0.0299 (0.0844)	−0.2908 (0.5006)	1.1820 (1.3099)
Asseco	ACP	0.00074 (0.00056)	−0.0595 (−0.0488)	0.0608 (0.0765)	−0.0201 (0.4742)	1.7600 (1.3305)
Allegro	ALE	−0.0026 (0.0010)	−0.1123 (−0.1027)	0.1067 (0.1578)	0.1747 (0.5547)	1.1735 (1.9310)
CCC	CCC	−0.00196 (−0.00050)	−0.0851 (−0.068)	0.1327 (0.1490)	0.4102 (0.7795)	3.4547 (1.5699)
CD Projekt	CDR	−0.00118 (−0.00022)	−0.1257 (−0.1024)	0.1307 (0.1347)	0.1356 (0.0001)	2.3658 (1.6826)
Cyfrowy Polsat	CPS	0.00029 (−0.00150)	−0.0395 (−0.0847)	0.0729 (0.0782)	0.5066 (0.0800)	2.2896 (0.8452)
Dino	DNP	0.00076 (0.00157)	−0.0670 (−0.0673)	0.0659 (0.0909)	0.1096 (0.2339)	0.9815 (1.3584)
JSW	JSW	0.00144 (0.00251)	−0.1117 (−0.1345)	0.1170 (0.3130)	−0.0557 (1.5369)	0.5071 (9.5325)
KGHM	KGH	−0.00088 (0.00011)	−0.0673 (−0.1178)	0.0853 (0.1168)	0.0967 (0.2907)	0.6175 (1.2485)
LPP	LPP	0.00241 (0.00052)	−0.1393 (−0.1090)	0.1468 (0.1581)	0.3934 (0.3732)	4.5029 (2.1642)
Orange	OPL	0.00128 (−0.00021)	−0.0653 (−0.0870)	0.0578 (0.0572)	0.2096 (−0.3807)	1.3278 (0.6936)
Bank Pekao	PEO	0.00256 (−0.00016)	−0.0711 (−0.1221)	0.0625 (0.1714)	−0.3068 (0.7317)	1.5768 (5.6249)
PGE	PGE	0.00076 (0.00062)	−0.0721 (−0.0691)	0.1358 (0.1567)	0.6716 (0.9674)	2.6880 (3.5208)
Orlen	PKN	0.00070 (0.00042)	−0.0640 (−0.0620)	0.0457 (0.1096)	−0.0736 (0.2118)	0.2344 (0.6775)
Bank PKO	PKO	0.00167 (−0.00029)	−0.0718 (−0.0717)	0.0017 (0.1327)	−0.2323 (0.8032)	1.0216 (2.7165)
PZU	PZU	0.00053 (0.00084)	−0.0644 (−0.0660)	0.0523 (0.0784)	−0.3524 (0.2138)	1.7112 (1.4921)
Santander Bank	SPL	0.00190 (0.00035)	−0.0536 (−0.0867)	0.0823 (0.1143)	0.4418 (0.3425)	1.3168 (1.7196)
Pepco (18 March 2022)	PCO	−0.00050 (0.00074)	−0.0470 (−0.0523)	0.0558 (0.1143)	0.1075 (0.6496)	0.4942 (3.1461)
mBank (18 March 2022)	MBK	0.00275 (0)	−0.1002 (−0.0851)	0.0943 (0.1264)	−0.0907 (0.4069)	0.9970 (0.7190)
Kęty (16 September 2022)	KTY	0.00084 (−0.00001)	−0.0684 (−0.1017)	0.0744 (0.0757)	0.3034 (−0.1884)	1.8525 (1.2969)
Kruk (16 December 2022)	KRU	0.00219 (0.00136)	−0.0622 (−0.0584)	0.1313 (0.1220)	0.9346 (0.9295)	2.7356 (2.2608)
Tauron (18 March 2022)	TPE	−0.00014 (0.00053)	−0.0617 (−0.0796)	0.0853 (0.1301)	0.4780 (0.5036)	0.6674 (2.1356)
Mercator (18 March 2022)	MRC	−0.00614 (0.00050)	−0.2348 (−0.0980)	0.2080 (0.2571)	0.0931 (1.9378)	5.0186 (7.6932)
LOTOS (16 September 2022)	LTS	0.00113 (0.00481)	−0.0826 (−0.0611)	0.0614 (0.0740)	−0.5117 (0.3371)	1.8753 (0.0411)
PGNiG (16 December 2022)	PGN	−0.00016 (0.00044)	−0.0603 (−0.0644)	0.0568 (0.1793)	−0.1985 (1.4098)	1.3151 (5.8434)

Table A2. The numerical values of the standard deviation and entropy, as calculated for the WIG20 index and its composing stocks. Similarly to Table A1, the cumulative index data are followed by the first group for the constant component companies and then by the second (third) group for the stocks introduced to (removed from) the index at some point. The dates of the introduction/removal of the composing companies are given next to the name of the corresponding company. The results are presented for the one-year time period before the beginning of the Russian invasion of Ukraine and after this event. For convenience, the percentage difference between estimates obtained for the two timeframes of interest is given for the standard deviation and entropy.

Name	Symbol	Standard Deviation Before	Standard Deviation After	Percentage Difference	Entropy Before	Entropy After	Percentage Difference
Cumulative Index	WIG20	0.012	0.018	40%	1.956	2.336	17.71%
Asseco	ACP	0.016	0.019	17.63%	2.224	2.426	8.68%
Allegro	ALE	0.029	0.036	21.54%	2.131	2.336	9.18%
CCC	CCC	0.027	0.034	22.95%	2.151	2.405	11.15%
CD Projekt	CDR	0.032	0.031	3.180%	2.223	2.226	0.14%
Cyfrowy Polsat	CPS	0.015	0.023	42.11%	2.387	1.967	19.29%
Dino	DNP	0.019	0.023	19.05%	2.256	2.456	8.49%
JSW	JSW	0.038	0.045	16.87%	1.936	1.998	3.15%
KGHM	KGH	0.025	0.032	24.56%	2.143	2.378	10.39%
LPP	LPP	0.030	0.034	12.50%	2.008	2.170	7.76%
Orange	OPL	0.017	0.021	21.05%	2.201	2.449	10.66%
Bank Pekao	PEO	0.019	0.030	44.90%	1.671	2.016	18.71%
PGE	PGE	0.026	0.032	20.69%	2.173	2.353	7.95%

Table A2. *Cont.*

Name	Symbol	Standard Deviation Before	Standard Deviation After	Percentage Difference	Entropy Before	Entropy After	Percentage Difference
Orlen	PKN	0.019	0.026	31.11%	2.180	2.460	12.07%
Bank PKO	PKO	0.019	0.027	34.78%	2.010	2.302	13.54%
PZU	PZU	0.016	0.021	27.03%	2.130	2.393	11.63%
Santander Bank	SPL	0.020	0.025	22.22%	2.085	2.274	8.67%
Pepco (18 March 2022)	PCO	0.017	0.022	25.64%	2.121	2.267	6.66%
mBank (18 March 2022)	MBK	0.028	0.033	16.39%	2.262	2.473	8.912%
Kęty (16 September 2022)	KTY	0.018	0.025	32.56%	2.080	2.390	13.87%
Kruk (16 December 2022)	KRU	0.026	0.028	3.640%	2.353	2.371	0.76%
Tauron (18 March 2022)	TPE	0.024	0.029	18.87%	2.216	2.352	5.95%
Mercator (18 March 2022)	MRC	0.047	0.043	8.89%	1.894	1.785	5.93%
LOTOS (16 September 2022)	LTS	0.020	0.028	33.33%	2.310	2.558	10.19%
PGNiG (16 December 2022)	PGN	0.017	0.033	64%	1.747	2.269	26%

References

1. He, C.; Wen, Z.; Huang, K.; Ji, X. Sudden shock and stock market network structure characteristics: A comparison of past crisis events. *Technol. Forecast. Soc. Chang.* **2022**, *180*, 121732. [CrossRef]
2. Weinberg, D.H.; Andrews, B.H.; Freudenburg, J. Equilibrium and sudden events in chemical evolution. *Astrophys. J.* **2017**, *837*, 183. [CrossRef]
3. Aminikhanghahi, S.; Cook, D. A survey of methods for time series change point detection. *Knowl. Inf. Syst.* **2017**, *51*, 339–367. [CrossRef] [PubMed]
4. Suriani, N.S.; Hussain, A.; Zulkifley, M.A. Sudden event recognition: A survey. *Sensors* **2013**, *13*, 9966–9998. [CrossRef] [PubMed]
5. Ramage, C. Sudden events. *Futures* **1980**, *12*, 268–274. [CrossRef]
6. Evangelos, V. Efficient markets hypothesis in the time of COVID-19. *Rev. Econ. Anal.* **2021**, *13*, 45–62.
7. Musmeci, N.; Aste, T.; Matteo, T.D. Interplay between past market correlation structure changes and future volatility outbursts. *Sci. Rep.* **2016**, *6*, 36320. [CrossRef]
8. Montgomery, D.; Jennings, C.; Kulahci, M. *Introduction to Time Series Analysis and Forecasting*; Wiley: Hoboken, NJ, USA, 2015.
9. Plerou, V.; Gopikrishnan, P.; Rosenow, B.; Amaral, L.A.; Stanley, H.E. Econophysics: Financial time series from a statistical physics point of view. *Phys. A Stat. Mech. Its Appl.* **2000**, *279*, 443–456. [CrossRef]
10. Rodriguez-Rodriguez, N.; Miramontes, O. Shannon Entropy: An econophysical approach to cryptocurrency portfolios. *Entropy* **2022**, *24*, 1583. [CrossRef]
11. Sheraz, M.; Nasir, I. Information-theoretic measures and modeling stock market Volatility: A Comparative Approach. *Risks* **2021**, *9*, 89. [CrossRef]
12. Velichko, A.; Heidari, H. A method for estimating the entropy of time series using artificial neural networks. *Entropy* **2021**, *23*, 1432. [CrossRef] [PubMed]
13. Yin, Y.; Shang, P. Weighted permutation entropy based on different symbolic approaches for financial time series. *Phys. A Stat. Mech. Its Appl.* **2016**, *443*, 137–148. [CrossRef]
14. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 623–656. [CrossRef]
15. Kolmogorov, A. On tables of random numbers. *Theor. Comput. Sci.* **1998**, *207*, 2. [CrossRef]
16. Tenreiro Machado, J.A. Entropy analysis of integer and fractional dynamical systems. *Nonlinear Dyn.* **2010**, *62*, 371–378. [CrossRef]
17. Shi, W.; Shang, P. Cross-sample entropy statistic as a measure of synchronism and cross-correlation of stock markets. *Nonlinear Dyn.* **2013**, *71*, 539–554. [CrossRef]
18. Dionisio, A.; Menezes, R.; Mendes, D. An econophysics approach to analyse uncertainty in financial markets: An application to the Portuguese stock market. *Eur. Phys. J. Condens. Matter Complex Syst.* **2006**, *50*, 161–164. [CrossRef]
19. Bentes, S.R.; Menezes, R.; Mendes, D.A. Long memory and volatility clustering: Is the empirical evidence consistent across stock markets? *Phys. A Stat. Mech. Its Appl.* **2008**, *387*, 3826–3830. [CrossRef]
20. Bentes, S.R.; Menezes, R. Entropy: A new measure of stock market volatility? *J. Phys. Conf. Ser.* **2012**, *394*, 012033. [CrossRef]
21. Delgado-Bonal, A. Quantifying the randomness of the stock markets. *Sci. Rep.* **2019**, *9*, 12761. [CrossRef]
22. Sheraz, M.; Dedu, S.; Preda, V. Entropy measures for assessing volatile markets. *Procedia Econ. Financ.* **2015**, *22*, 655–662. [CrossRef]
23. Darbellay, G.A.; Wuertz, D. The entropy as a tool for analysing statistical dependences in financial time series. *Phys. A Stat. Mech. Its Appl.* **2000**, *287*, 429–439. [CrossRef]

24. Almog, A.; Shmueli, E. Structural entropy: Monitoring correlation-based networks over Time with Application to Financial Markets. *Sci. Rep.* **2019**, *9*, 10832. [CrossRef] [PubMed]
25. Lahmiri, S.; Bekiros, S. Randomness, informational Entropy, and volatility interdependencies among the major world markets: The role of the COVID-19 pandemic. *Entropy* **2020**, *22*, 833. [CrossRef]
26. Jakimowicz, A. The role of entropy in the development of economics. *Entropy* **2020**, *22*, 452. [CrossRef]
27. Bose, R.; Hamacher, K. Alternate entropy measure for assessing volatility in financial markets. *Phys. Rev. E* **2012**, *86*, 056112. [CrossRef]
28. Fiszeder, P.; Małecka, M. Forecasting volatility during the outbreak of Russian invasion of Ukraine: Application to commodities, stock indices, currencies, and cryptocurrencies. *Equilib. Q. J. Econ. Econ. Policy* **2022**, *17*, 939–967. [CrossRef]
29. Doğan, N.; Doğan, I. Determination of the number of bins/classes used in histograms and frequency tables: A short bibliography. *J. Stat. Res.* **2010**, *7*, 77–86.
30. Rényi, A. On measures of information and entropy. In Proceedings of the 4th Berkeley Symposium on Mathe- matical Statistics and Probability, Berkeley, CA, USA, 20–30 July 1960 ; pp. 547–561.
31. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487. [CrossRef]
32. Brzeszczyński, J.; Gajdka, J.; Kutan, A.M. Evolution of the impact of the interest rates changes announced by Narodowy Bank Polski (NBP) on the financial markets in the high, medium and low level of interest rates environments in Poland. *NBP Work. Pap.* **2019**, *303*, 1–101.
33. Battiston, S.; Farmer, J.D.; Flache, A.; Garlaschelli, D.; Haldane, A.G.; Heesterbeek, H.; Hommes, C.; Jaeger, C.; May, R.; Scheffer, M. Complexity theory and financial regulation. *Science* **2016**, *351*, 818–819. [CrossRef] [PubMed]

entropy

MDPI

Article

Local Phase Transitions in a Model of Multiplex Networks with Heterogeneous Degrees and Inter-Layer Coupling

Nedim Bayrakdar [1], Valerio Gemmetto [1] and Diego Garlaschelli [1,2,3,*]

[1] Lorentz Institute for Theoretical Physics, University of Leiden, 2333 CA Leiden, The Netherlands
[2] IMT School of Advanced Studies Lucca, 55100 Lucca, Italy
[3] INdAM-GNAMPA Istituto Nazionale di Alta Matematica, 00185 Rome, Italy
* Correspondence: garlaschelli@lorentz.leidenuniv.nl

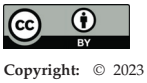
Abstract: Multilayer networks represent multiple types of connections between the same set of nodes. Clearly, a multilayer description of a system adds value only if the multiplex does not merely consist of independent layers. In real-world multiplexes, it is expected that the observed inter-layer overlap may result partly from spurious correlations arising from the heterogeneity of nodes, and partly from true inter-layer dependencies. It is therefore important to consider rigorous ways to disentangle these two effects. In this paper, we introduce an unbiased maximum entropy model of multiplexes with controllable intra-layer node degrees and controllable inter-layer overlap. The model can be mapped to a generalized Ising model, where the combination of node heterogeneity and inter-layer coupling leads to the possibility of local phase transitions. In particular, we find that node heterogeneity favors the splitting of critical points characterizing different pairs of nodes, leading to link-specific phase transitions that may, in turn, increase the overlap. By quantifying how the overlap can be increased by increasing either the intra-layer node heterogeneity (spurious correlation) or the strength of the inter-layer coupling (true correlation), the model allows us to disentangle the two effects. As an application, we show that the empirical overlap observed in the International Trade Multiplex genuinely requires a nonzero inter-layer coupling in its modeling, as it is not merely a spurious result of the correlation between node degrees across different layers.

Keywords: multiplex networks; maximum entropy models; World Trade Multiplex; mean-field Ising model

Citation: Bayrakdar, N.; Gemmetto, V.; Garlaschelli, D. Local Phase Transitions in a Model of Multiplex Networks with Heterogeneous Degrees and Inter-Layer Coupling. *Entropy* **2023**, *25*, 828. https://doi.org/10.3390/e25050828

Academic Editor: Panos Argyrakis

Received: 1 March 2023
Revised: 6 May 2023
Accepted: 9 May 2023
Published: 22 May 2023



1. Introduction

The wide variety of different phenomena that occur around us are often the result of systems that emerge and (self-)organize dynamically. These systems consist of a multitude of basic constituents interacting with each other in complicated ways and forming complex patterns. Many of these systems can be represented as networks sustaining various processes. Examples of such systems include social networks, transportation networks, biological networks, financial networks, and technological networks. In particular, social, financial, and economic networks are an important class of systems that, in the wake of recent global crises (such as the 2007–2008 financial crisis, the COVID-19 pandemic, and the ongoing Ukraine crisis), have been attracting attention given the possibility of studying the propagation of shocks among their constituents. Generally, individuals, banks, firms, or countries can be represented as nodes, and the relationships among them can be represented as links [1–3]. Other types of economic and financial networks are obtained as some form of projection from time series data [3–7]. The study of these networks may increase our understanding of a variety of processes that take place through them, such as the spreading of diseases, the diffusion of (mis)information, the stability of financial markets, and the resilience of the economy.

The simplest approach is to map each constituent within a system onto a single node and to map each interaction between pairs of constituents onto a link of a single type,

regardless of the nature of the interaction. In this approach, all the links in a network are treated on an equal footing, making it a single-layer network representation, which might, however, lead to an oversimplification that fails to capture the details of a multirelational system. For instance, production and trade networks are the result of the functioning of global supply chains, involving the exchange of multiple products between firms and countries, which determines nontrivial dependencies between product-specific layers of the network. In order to realistically follow the propagation of shocks in the economy, knowledge of the nature of the links is essential. The inability to properly represent multirelational systems using single-layer networks has lead to the introduction of so-called *multilayer networks* [8–12]. Multilayer networks allow us to describe multirelational systems by representing each type of relationship in a separate layer of the network, where each node is present in all layers, and the different types of connections are reported in the corresponding layers. Returning to the example of social networks, the different types of relationships between people, such as kinship, friendship, coworkership, etc., would each be represented by links in a different layer [13], and could be analyzed in their mutual dependencies.

However, in order to assess true dependencies across layers, one should use proper null models. In recent years, there has been an increase in attention towards null models of networks constructed as random graph ensembles [14–20]. A class of such models is the so-called Exponential Random Graph Models (ERGMs) [17–27]. ERGMs are used commonly within the social network analysis community, and have been more recently re-derived within a statistical physics maximum entropy framework [19,20,27]. This has allowed researchers to utilize techniques that are common in statistical physics. In the ERGM framework, one chooses the probability distribution on graphs such that it maximizes the entropy. This maximization is performed while the expected values of certain chosen graph properties are constrained to be equal to desired values.

Real-world multilayer networks have been compared against null ERGMs with independent layers [28,29]. This comparison has highlighted various properties of real multilayer networks that result from the interdependence of layers. Two such properties are the *overlap* and the *multiplexity* [9,28]. The overlap and the multiplexity essentially contain similar information and capture the correlation of a node's connectivity across two or more layers. For example, in a social network, people may communicate with their friends through multiple means of communication, such as talking on the phone, sending emails, or sending instant text messages. In this example, the layer that represents communication through email has a significant overlap with the layer of communication through text messages. A more specific example is a study of the so-called World Trade Multiplex (representing international trade in different commodities among countries [30]), which showed that, despite the fact that each layer of the multiplex is separately well described by a maximum entropy model with given node degrees [31–33], the observed trade overlap across different commodity-specific layers is significantly different from the overlap predicted by a null model with independent layers [28]. This result is not unexpected, since one can imagine that the trade of a certain product between two countries may increase/decrease the possibility of the trade of a *different* product between the same two countries. Other examples of networks displaying a significant overlap are airport networks, on-line social games, collaboration networks, and citation networks [34–36].

An important conclusion that has been reached after comparing real-world multiplexes against null models with independent layers is that a significant part of the observed overlap in many real networks could actually be spuriously created by the correlations among node degrees across different layers, even if the latter are conditionally independent of each other, instead of resulting from genuine inter-layer dependencies [28,29]. Indeed, if node degrees are correlated among layers, then there will be an increased probability of a link between two nodes being present in multiple layers, while the probability of a link occurring in one layer will not necessarily influence the presence of a link occurring in another layer. The measured overlap of the network therefore consists of a part resulting

from 'spurious' coupling between the layers and of a part resulting from genuine coupling between the layers. This spurious coupling increases as the density and/or heterogeneity of the degrees of the network increases. Real-world networks are often dense and have strongly heterogeneous degrees; therefore, the assessment of inter-layer coupling in these real-world networks will be severely affected.

The focus of this paper is the introduction of interdependencies between the layers of a multilayer network in the ERGM through the explicit inclusion of the overlap as an extra constraint. This inclusion of the overlap in the ERGM will aid us in understanding which (higher-order) properties of the network structure may be (highly) dependent on the overlap. Additionally, it will help us distinguish between the overlap in the network due to the correlation of single-node properties across layers and the overlap due to a genuine coupling between the layers. Finally, it will allow us to generate null models with the desired amount of spurious overlap and genuine overlap. It turns out that this problem is mathematically identical to solving the Ising model on a complete graph (which is also known as the mean-field Curie–Weiss model) and leads to a phase transition between a 'multiplexed' (magnetized) and a 'non-multiplexed' (non-magnetized) phase. However, the problem is more general because the locality of the constraints on the degrees of nodes will imply different parameter values, and hence different properties for the phase transitions relative to different pairs of nodes. For instance, it will, in general, not be possible to enforce a 'zero-field' spontaneous symmetry breaking condition for all pairs of nodes simultaneously. Therefore, for a given specification of the constraints, different pairs of nodes may realize different symmetry-broken values of their contribution to the overall inter-layer overlap. Crucially, this property arises only from the simultaneous presence of the two constraints (on the global overlap and on the heterogeneous local degrees), and would not be realized in the absence of one of them.

The rest of the paper is organized as follows: In Section 2, we mathematically define quantities and models that are relevant to this paper. This includes the derivation of a benchmark model, where the layers of the multiplex network are independent. In Section 3, we introduce, and solve analytically, our new model, where the layers of the multiplex are interdependent due to the inclusion of the overlap. Section 4 contains a discussion regarding the possible local phase transitions of the model. In Section 5, we explore our model by using various numerical methods. In Section 6, we briefly analyze the World Trade Multiplex, and show that the empirical overlap in this real-world network is not merely the result of the heterogeneity of the network, but requires a nonzero coupling between the layers in its modeling. Finally, we provide some concluding remarks in Section 7, and some technical details in Appendices A and B.

2. Background Theory

This section contains some background notions, definitions, and models.

2.1. Single-Layer Network Definitions

We will limit our discussion to the case of *binary and undirected* networks. A binary undirected network can be defined as a graph that is an ordered pair $G = (V, E)$, where $V = \{v_1, v_2, ..., v_N\}$ is a set of N *vertices* or *nodes*, and E is a set of *unordered* pairs of different vertices called *edges* or *links*. Note that the definition of E depends on the relevant class of relations between the constituents of the system. The vertex $v_i \in V$ will be referred to simply as i throughout the rest of the paper. If $(i, j) \in E$, the vertices i and j are said to be connected, and may be referred to as *neighbors* of each other. The number of links L of the graph is given by the cardinality of E: $L = |E|$.

Matrix Representation

A graph G is represented by its *adjacency matrix* $G = \{g_{ij}\}$. This is an $N \times N$ matrix where

$$g_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

We define E as containing pairs of *distinct* vertices, which means that a vertex cannot have a connection to itself (self-loop). It is then natural to define the diagonal elements as $g_{ii} \equiv 0$. Since we limit our discussion to undirected graphs, the adjacency matrix is always symmetric, $g_{ij} = g_{ji}$, and it therefore contains $N(N-1)/2$ independent elements that fully specify the matrix and ultimately the graph.

Degrees and Degree Distribution

One of the main topics in the analysis of complex networks is the identification of the different roles that nodes play [37]. For instance, there are a variety of measures that characterize the structural importance of a node in a network. The degree $k_i(G)$ of the graph G is defined as the number of connections node i has to other nodes in the network.

$$k_i(G) = \sum_{j=1}^{N} g_{ij} \tag{2}$$

The list $\{k_i(G)\}_{i=1}^{N}$ of degrees is called the *degree sequence* of the graph G. The degree distribution $P(k)$ is defined as the fraction of nodes in the network with degree k. Real-world networks systematically show a degree distribution with heavy tails, where the degrees vary over a broad range, often spanning several orders of magnitude [38,39]. The majority of the vertices of these real-world networks have a small number of links to other vertices, while a few vertices have a relatively high number of links to other vertices, which are also referred to as 'hubs'. An example is the World Wide Web, where some pages are incredibly popular and are pointed to by thousands of other pages, while generally, most pages are almost unknown. The heavy tails of real-world degree distributions can often be, but not necessarily, approximated by power laws of the form $P(k) \sim k^{-\gamma}$. In any case, vertices with a degree much larger than the average degree $\langle k \rangle$ occur with a non-negligible probability. This is a signature of a high level of statistical heterogeneity in real-world networks. Encoding this heterogeneity will be a crucial ingredient of our models.

2.2. Multiplex Network Definitions

A binary undirected multiplex network can be defined in terms of the previously defined single-layer networks. A multiplex network is a set $\vec{G} = \{G^\alpha\}_{\alpha=1}^{M}$ of M undirected binary graphs $G^\alpha = (V, E^\alpha)$ that share the same set of N nodes. In the context of multilayer networks, G^α is called a layer of $\mathcal{M}$, and will be referred to simply as α throughout the rest of the paper. Note that a multiplex network is a type of multilayer network that does not allow inter-layer connections between two layers α and β where $\alpha \neq \beta$.

Matrix Representation

The layer G^α and its intra-layer links can then be represented by the adjacency matrix $G^\alpha = \{g_{ij}^\alpha\}$. This is an $N \times N$ matrix where

$$g_{ij}^\alpha = \begin{cases} 1 & \text{if } (i,j) \in E^\alpha, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Multilinks in Multiplex Networks

In order to capture the information regarding the presence of the links between the pair of nodes (i, j) in any of the M layers, we define the object

$$m_{ij} \equiv (g_{ij}^1, g_{ij}^2, \ldots, g_{ij}^M) \tag{4}$$

which is also known as the *multilink* of (i, j). Additionally, we define the set $\mathcal{M}_{ij}$ as the set that contains all the 2^M possible configurations of m_{ij}.

Multidegrees

The multidegree of a node $i \in V$ of a multiplex network $\vec{G}$ is the object

$$\vec{k}_i(\vec{G}) \equiv \left(k_i^1(\vec{G}), k_i^2(\vec{G}), \ldots, k_i^M(\vec{G})\right) \tag{5}$$

where

$$k_i^\alpha(\vec{G}) = \sum_{j \neq i}^{N} g_{ij}^\alpha \tag{6}$$

is the degree of the node i in the layer α [9,40]. From the vector definition of the multidegree, one can obtain a scalar quantity defined as the *layer-averaged degree*:

$$\bar{k}_i(\vec{G}) = \frac{1}{M} \sum_{\alpha=1}^{M} k_i^\alpha(\vec{G}), \tag{7}$$

which is the degree of node i averaged over all the M layers. Note that, in each layer α, the total layer-specific degree of all nodes equals twice the number of links in that layer, which we denote as L^α:

$$\sum_{i=1}^{N} k_i^\alpha(\vec{G}) = \sum_{i<j} g_{ij}^\alpha = 2L^\alpha(\vec{G}). \tag{8}$$

Summing the above relationship for the M layers, we get

$$M \sum_{i=1}^{N} \bar{k}_i(\vec{G}) = \sum_{\alpha=1}^{M} \sum_{i<j} g_{ij}^\alpha = 2 \sum_{\alpha=1}^{M} L^\alpha(\vec{G}) = 2L(\vec{G}), \tag{9}$$

where $L(\vec{G})$ denotes the total number of links over the entire multiplex:

$$L(\vec{G}) = \sum_{\alpha=1}^{M} \sum_{i<j} g_{ij}^\alpha. \tag{10}$$

Overlap

There are many properties that encode the interdependence between the layers of a multilayer network, but we will limit our discussion to one such property: the overlap. The overlap $O^{\alpha\beta}(\vec{G})$ between two layers α and β of the multiplex $\vec{G}$ is defined as the number of links that appear in both layers α and β [34,41]:

$$O^{\alpha\beta}(\vec{G}) = \sum_{i<j} g_{ij}^\alpha g_{ij}^\beta \tag{11}$$

where, throughout the paper, using $\sum_{a<b}$ and $\prod_{a<b}$, we denote a *double* sum and a *double* product for all possible (unrepeated) pairs of values of the two indices, a and b (with $a \neq b$), respectively. The *global overlap* $O(\vec{G})$ is defined as the sum of $O^{\alpha\beta}(\vec{G})$ for all pairs of layers:

$$O(\vec{G}) = \sum_{\alpha<\beta} \sum_{i<j} g_{ij}^\alpha g_{ij}^\beta. \tag{12}$$

As the names of these properties suggest, they are a measure of how overlapping the layers of the multiplex network are.

2.3. Exponential Random Graph Models for Multiplexes

ERGMs are ensemble models, which means that they are defined as probability distributions over many possible (multiplex) networks. Given the observed (or desired) value $C_i^* \equiv C_i(\vec{G}^*)$ for K graph properties $\{C_i(\vec{G})\}_{i=1}^K$ defined on each possible multiplex $\vec{G}$ (where $\vec{G}^*$ represents a particular, e.g., real-world, multiplex of interest), an ERGM generates a probability distribution $P(\vec{G})$ over multiplex networks that maximizes the entropy, under the constraint that the expected value of $C_i(\vec{G})$ equals C_i^*, for all $i = 1, K$. This method provides us with a general framework for modeling maximally random (maximum entropy) multiplex networks, to be used as null models that can be compared against the empirical multiplex $\vec{G}^*$ to detect higher-order patterns that are irreducible to the K enforced constraints. Maximizing the entropy subject to a set of constraints is also widely used in problems with incomplete information [42,43].

Let $\mathcal{G}_N^M$ be the set of (binary undirected) multiplex networks consisting of N vertices and M layers (note that this set includes single-layer networks for $M = 1$), let $\vec{G} = \{G_1, G_2, ..., G_M\} \in \mathcal{G}_N^M$ be a multiplex network in that set, and let $P(\vec{G})$ be the sought-for probability of $\vec{G}$ within the ensemble. We want $P(\vec{G})$ to be such that the expectation value of each graph observable $C_i(\vec{G})$ (in the chosen set of K observables) is equal to the corresponding observed or desired value C_i^*. This type of probability distribution is also referred to as a *canonical ensemble*. The ideal probability distribution is the one that maximizes the Gibbs–Shannon entropy

$$S = -\sum_{\vec{G}\in\mathcal{G}_N^M} P(\vec{G}) \ln P(\vec{G}) \tag{13}$$

under the normalization condition

$$\sum_{\vec{G}\in\mathcal{G}_N^M} P(\vec{G}) = 1 \tag{14}$$

and the other K constraints

$$C_i^* = \langle C_i \rangle, \qquad i = 1, \ldots, K, \tag{15}$$

where

$$\langle C_i \rangle \equiv \sum_{\vec{G}\in\mathcal{G}_N^M} P(\vec{G}) C_i(\vec{G}). \tag{16}$$

The maximization of the entropy is achieved by introducing a global Lagrange multiplier η for the normalization condition and a specific multiplier θ_i for each constraint $\langle C_i \rangle = C_i^*$, $i = 1, \ldots, K$. This leads to the parametric solution

$$P(\vec{G}, \vec{\theta}) = \frac{e^{-H(\vec{G},\vec{\theta})}}{Z(\vec{\theta})} \tag{17}$$

where $H(\vec{G}, \vec{\theta})$ is the graph Hamiltonian

$$H(\vec{G}, \vec{\theta}) \equiv \sum_{i=1}^{K} \theta_i C_i(\vec{G}) = \vec{\theta} \cdot \vec{C}(\vec{G}) \tag{18}$$

and $Z(\vec{\theta})$ is the partition function determined by the normalization condition

$$Z(\vec{\theta}) \equiv e^{\eta+1} = \sum_{\vec{G}\in\mathcal{G}_N^M} e^{-H(\vec{G},\vec{\theta})}. \tag{19}$$

The parametric form of $P(\vec{G},\vec{\theta})$, if inserted back into Equation (13), leads to the explicit expression for the entropy:

$$S(\vec{\theta}) = -\sum_{\vec{G}\in\mathcal{G}_N^M} P(\vec{G},\vec{\theta}) \ln P(\vec{G},\vec{\theta}) = \vec{\theta}\cdot\langle\vec{C}\rangle + \ln Z(\vec{\theta}). \tag{20}$$

2.4. Maximum Likelihood Parameter Estimation

Equations (17)–(19) fully define the ERGM, apart from the specification of the parameters $\vec{\theta}$. In principle, by treating these Lagrange multipliers as free parameters, one can study the effects that the specification of certain graph observables $\{C_i\}$ has on other aspects of network structure [27,44–47]. This approach, however, does not allow one to consider ERGMs as null models of a particular real network [17,19]. In the latter case, maximum likelihood parameter estimation leads to the unique (given the choice of constraints) ERGM representing a null model for a particular real (multiplex) network $\vec{G}^*$, and hence, enforcing Equation (15) exactly, as we briefly recall below. This null model can then be used to detect statistically significant deviations of empirical structural properties of $\vec{G}^*$ from the ensemble.

The log-likelihood of the particular multiplex $\vec{G}^*$ is

$$\mathcal{L}(\vec{G}^*,\vec{\theta}) = \ln P(\vec{G}^*,\vec{\theta}) = -\sum_{i=1}^{K} \theta_i C_i^* - \ln Z(\vec{\theta}). \tag{21}$$

This function has the following properties [19]:

$$\frac{\partial\mathcal{L}(\vec{G}^*,\vec{\theta})}{\partial\theta_i} = \langle C_i\rangle - C_i^* \tag{22}$$

$$\frac{\partial^2\mathcal{L}(\vec{G}^*,\vec{\theta})}{\partial\theta_i\,\partial\theta_j} = -\langle C_iC_j\rangle + \langle C_i\rangle\langle C_j\rangle. \tag{23}$$

Equation (22) means that the stationary points $\vec{\theta} = \vec{\theta}^*$ of $\mathcal{L}$ are precisely those that satisfy the constraints (15), i.e.,

$$\langle C_i\rangle_{\vec{\theta}^*} = \sum_{\vec{G}\in\mathcal{G}_N^M} C_i(\vec{G})P(\vec{G},\vec{\theta}^*) = \sum_{\vec{G}\in\mathcal{G}_N^M} C_i(\vec{G})\frac{e^{-\sum_{j=1}^{K}\theta_j^* C_j(\vec{G})}}{Z(\vec{\theta})} = C_i(\vec{G}^*), \quad i = 1,\ldots,K \tag{24}$$

where $\langle C_i\rangle_{\vec{\theta}^*}$ indicates that the ensemble average is evaluated at the values $\vec{\theta}^*$. Equation (23) indicates that $\mathcal{L}$ is concave, since the matrix with entries $\partial^2\mathcal{L}/\partial\theta_i\,\partial\theta_j$ has the form of a negative covariance matrix, and must therefore be non-positive definite [48]. The solutions $\vec{\theta}^*$ of the coupled equations $\langle C_i\rangle_{\vec{\theta}^*} = C_i^*$ in Equation (15) can therefore be found by maximizing the log-likelihood $\mathcal{L}$. If $\partial^2\mathcal{L}/\partial\theta_i\,\partial\theta_j$ is negative definite, which will be true if the functions $C_i(\vec{G})$ are linearly independent [48] (i.e., the chosen constraints are non-redundant), then there will be, at most, one solution, and it will be the unique maximum of $\mathcal{L}$. Maximizing a concave function is generally easier than solving the system of coupled nonlinear equations in Equation (24). Once the solution $\vec{\theta} = \vec{\theta}^*$ is found, it can be used to generate a null model

of $\vec{G}^*$. Moreover, inserting the value θ^* back into Equation (21) and using Equation (20), we obtain the important relation

$$\begin{aligned}\mathcal{L}(\vec{G}^*,\vec{\theta}^*) &= \ln P(\vec{G}^*,\vec{\theta}^*)\\ &= -\sum_{i=1}^{K}\theta_i^* C_i^* - \ln Z(\vec{\theta}^*)\\ &= -\sum_{i=1}^{K}\theta_i^* \langle C_i\rangle_{\vec{\theta}^*} - \ln Z(\vec{\theta}^*)\\ &= -S(\vec{\theta}^*),\end{aligned} \tag{25}$$

i.e., the maximized log-likelihood equals minus the entropy for the particular value $\vec{\theta}^*$, which in turn represents the 'entropy of the data' given the chosen constraints. This result allows one to easily calculate the entropy of the data $S(\vec{\theta}^*) = -\mathcal{L}(\vec{G}^*,\vec{\theta}^*)$ automatically as part of the likelihood maximization procedure, rather than as a much more complicated formal sum of all configurations, as in the general definition (13).

2.5. Benchmark: Independent Layers Model

As anticipated in the Introduction, our goal is that of considering how the empirical overlap between links in different layers of a multiplex is jointly determined by both a 'genuine' coupling between the M layers and a 'spurious' correlation resulting from the heterogeneous (and correlated across layers) degrees of the N nodes. As a null benchmark before inserting both components in an ERGM of a multiplex, we first consider only the layer-averaged degrees of all vertices as constraints, as defined in Equation (7). We can therefore create a null model of a real multiplex $\vec{G}^*$ using the ERGM in combination with the maximum likelihood method. This model will be referred to as the *Average Configuration Model (ACM)*, and will allow us to study the sole effects of correlated heterogeneous degrees on the inter-layer overlap. The Hamiltonian of this model, denoted as H_0, since it represents a benchmark for a more complicated model to be defined later, is

$$H_0(\vec{G},\vec{\theta}) = M\sum_{i=1}^{N}\theta_i\bar{k}_i(\vec{G}) = \sum_{\alpha=1}^{M}\sum_{i<j}(\theta_i+\theta_j)g_{ij}^{\alpha} \tag{26}$$

where we have reparametrized by exposing M for convenience. The partition function is

$$\begin{aligned}Z_0(\vec{\theta}) &= \sum_{\vec{G}\in\mathcal{G}_N^M} e^{-\sum_{\alpha=1}^{M}\sum_{i<j}(\theta_i+\theta_j)g_{ij}^{\alpha}}\\ &= \sum_{\vec{G}\in\mathcal{G}_N^M}\prod_{\alpha=1}^{M}\prod_{i<j} e^{-(\theta_i+\theta_j)g_{ij}^{\alpha}}\\ &= \prod_{\alpha=1}^{M}\prod_{i<j}\sum_{g_{ij}^{\alpha}=0}^{1} e^{-(\theta_i+\theta_j)g_{ij}^{\alpha}}\\ &= \prod_{\alpha=1}^{M}\prod_{i<j}\left[1+e^{-(\theta_i+\theta_j)}\right]\\ &= \prod_{i<j}\left[1+e^{-(\theta_i+\theta_j)}\right]^{M}.\end{aligned} \tag{27}$$

The probability distribution over the ensemble is then given by

$$P_0(\vec{G},\vec{\theta}) = \prod_{\alpha=1}^{M}\prod_{i<j}\frac{e^{-(\theta_i+\theta_j)g_{ij}^{\alpha}}}{1+e^{-(\theta_i+\theta_j)}}, \tag{28}$$

from which we see that pairs of nodes and pairs of layers are all independent of each other, each entry g_{ij}^{α} being an independent Bernoulli random variable with success probability $p_{ij}^{\alpha}(\vec{\theta})$ and expected value $\langle g_{ij}^{\alpha}\rangle_{\vec{\theta}}$ given by

$$p_{ij}^{\alpha}(\vec{\theta}) = \langle g_{ij}^{\alpha}\rangle_{\vec{\theta}} = \frac{e^{-(\theta_i+\theta_j)}}{1+e^{-(\theta_i+\theta_j)}} \equiv p_{ij}(\vec{\theta}). \tag{29}$$

Clearly, $p_{ij}^{\alpha}(\vec{\theta}) = p_{ij}(\vec{\theta})$ is the probability that a link occurs between node i and j in layer α, which turns out to be independent of α given our choice of the layer-averaged (not layer-specific) degree as a constraint.

The log-likelihood of the multiplex $\vec{G}^*$ is

$$\mathcal{L}_0(\vec{G}^*, \vec{\theta}) = -M\sum_{i=1}^{N}\theta_i\overline{k}_i^* - M\sum_{i<j}\ln\left[1+e^{-(\theta_i+\theta_j)}\right], \tag{30}$$

where $\overline{k}_i^* = \overline{k}_i(\vec{G}^*)$. The parameter value θ_m^* maximizing the log-likelihood must satisfy

$$\left.\frac{\partial\mathcal{L}_0(\vec{G}^*,\vec{\theta})}{\partial\theta_m}\right|_{\vec{\theta}=\vec{\theta}^*} = -M\overline{k}_m^* + M\sum_{j\neq m}\frac{e^{-(\theta_m^*+\theta_j^*)}}{1+e^{-(\theta_m^*+\theta_j^*)}} = 0 \qquad \forall m \tag{31}$$

or equivalently,

$$\overline{k}_i^* = \sum_{j\neq i}\frac{e^{-(\theta_i^*+\theta_j^*)}}{1+e^{-(\theta_i^*+\theta_j^*)}} \qquad \forall i. \tag{32}$$

The above results show that, as expected from the general result reported in Equation (24), according to the maximum likelihood principle, the empirical layer-averaged degree $\overline{k}_i^* = \overline{k}_i(\vec{G}^*)$ of the real multiplex $\vec{G}^*$ is equal to the ensemble average $\langle\overline{k}_i\rangle_{\vec{\theta}^*}$:

$$\begin{aligned}\overline{k}_i^* &= \sum_{j\neq i}p_{ij}(\vec{\theta}^*)\\ &= \frac{1}{M}\sum_{\alpha=1}^{M}\sum_{j\neq i}p_{ij}^{\alpha}(\vec{\theta}^*)\\ &= \frac{1}{M}\sum_{\alpha=1}^{M}\sum_{j\neq i}\langle g_{ij}^{\alpha}\rangle_{\vec{\theta}^*}\\ &= \langle\overline{k}_i\rangle_{\vec{\theta}^*}.\end{aligned} \tag{33}$$

The probability distribution $P_0(\vec{G},\vec{\theta}^*)$ can then be written as a product of the layers:

$$P_0(\vec{G},\vec{\theta}^*) = \prod_{\alpha=1}^{M}P_0^{\alpha}(G^{\alpha},\vec{\theta}^*) \tag{34}$$

where P_0^{α} is the probability distribution over a single layer, i.e.,

$$P_0^{\alpha}(G^{\alpha},\vec{\theta}^*) = \prod_{i<j}[p_{ij}(\vec{\theta}^*)]^{g_{ij}^{\alpha}}[1-p_{ij}(\vec{\theta}^*)]^{1-g_{ij}^{\alpha}}. \tag{35}$$

This means that each layer α can be generated by using the link probability $p_{ij}(\vec{\theta}^*)$ that is equal throughout the layers. This is again a consequence of exclusively constraining properties defined as the overall averages of the layers. This null model can be used as a benchmark to determine the expected value of the inter-layer overlap $O(\vec{G})$ defined in Equation (12), which is due solely to the correlation between the degree of the same node i across the M layers, and not to any genuine inter-layer dependency. This expected value is

$$\langle O\rangle_{\vec{\theta}^*} = \sum_{\alpha<\beta}\sum_{i<j}\langle g_{ij}^{\alpha}g_{ij}^{\beta}\rangle_{\vec{\theta}^*} = \sum_{\alpha<\beta}\sum_{i<j}\langle g_{ij}^{\alpha}\rangle_{\vec{\theta}^*}\langle g_{ij}^{\beta}\rangle_{\vec{\theta}^*} = \sum_{\alpha<\beta}\sum_{i<j}p_{ij}^2(\vec{\theta}^*), \tag{36}$$

where we have used the independence $\langle g_{ij}^{\alpha}g_{ij}^{\beta}\rangle_{\vec{\theta}^*} = \langle g_{ij}^{\alpha}\rangle_{\vec{\theta}^*}\langle g_{ij}^{\beta}\rangle$ between layers $\alpha \neq \beta$. Deliberately, we have chosen the layer-averaged degree as the only constraint so that the expected degree of a node is the same across all layers, thereby creating a strong correlation between degrees in different layers, while keeping the layers themselves independent. Using Equations (25) and (30), we can calculate the entropy of the data, given the model, as

$$S_0(\vec{\theta}^*) = -\mathcal{L}_0(\vec{\theta}^*) = -\ln P_0(\vec{G}^*,\vec{\theta}^*) = M\sum_{i=1}^{N}\theta_i^*\overline{k}_i^* + M\sum_{i<j}\ln\left[1+e^{-\left(\theta_i^*+\theta_j^*\right)}\right], \tag{37}$$

which only requires the knowledge of $\vec{\theta}^*$ and of the layer-averaged degrees $\overline{k}_i(\vec{G}^*)$, $i = 1, N$.

3. The Overlapping Average Configuration Model

Having illustrated all the ingredients that are necessary to define and model basic properties of multiplex networks within a maximum entropy framework, in this section, we introduce a model of multiplex networks with genuinely interdependent layers. To this end, we incorporate the overlap as an extra constraint in the ERGM, and study the model in combination with the maximum likelihood method. This model is a generalization of the previous ACM benchmark, and will therefore be referred to as the *Overlapping Average Configuration Model (OACM)*, as it includes not only the intra-layer degrees, but also the inter-layer coupling, as building blocks.

3.1. Constructing the Hamiltonian

We want to define a model of a multiplex with M layers, N vertices, and given expected layer-averaged degrees (as defined in Equation (7)) and global inter-layer overlap (as defined in Equation (12)). The Hamiltonian of our ERGM is, in this case,

$$H(\vec{G},\vec{\theta},J) = M\sum_{i=1}^{N}\theta_i\bar{k}_i(\vec{G}) - \frac{4J}{M}O(\vec{G}) = \sum_{i<j}\sum_{\alpha=1}^{M}(\theta_i+\theta_j)g_{ij}^{\alpha} - \frac{4J}{M}\sum_{i<j}\sum_{\alpha<\beta}g_{ij}^{\alpha}g_{ij}^{\beta} \tag{38}$$

where $(\vec{\theta},J)$ are the Lagrange multipliers coupled to the $N+1$ constraints. We have defined the Lagrange multiplier for the overlap as $-4J/M$ for later convenience. Clearly, $H(\vec{G},\vec{\theta},J) = H_0(\vec{G},\vec{\theta})$ where H_0 is the benchmark Hamiltonian of the ACM without overlap defined in Equation (26). Using the multilink m_{ij} defined in Equation (4) and defining

$$\theta_{ij} \equiv \theta_i + \theta_j, \tag{39}$$

the Hamiltonian in Equation (38), this can be written as a sum of the pairs of vertices:

$$H(\vec{G},\vec{\theta},J) = \sum_{i<j}h_{ij}(m_{ij},\theta_{ij},J) \tag{40}$$

where

$$h_{ij}(m_{ij},\theta_{ij},J) \equiv (\theta_i+\theta_j)\sum_{\alpha=1}^{M}g_{ij}^{\alpha} - \frac{4J}{M}\sum_{\alpha<\beta}g_{ij}^{\alpha}g_{ij}^{\beta} \tag{41}$$

will be referred to as the *pair Hamiltonian*. As we shall see in a moment, the pair Hamiltonian can be mapped exactly to a mean-field Ising model coupling the M layers homogeneously. To arrive at this mapping, we transform the Boolean variables $g_{ij}^{\alpha} \in \{0,1\}$ to new 'spin' variables $\sigma_{ij}^{\alpha} \in \{-1,1\}$, as follows:

$$g_{ij}^{\alpha} = \frac{1}{2}(\sigma_{ij}^{\alpha}+1). \tag{42}$$

From now on, we assume that M is large (multiplex with several layers) and expand expressions accordingly. By defining

$$s_{ij} \equiv \{\sigma_{ij}^1, \sigma_{ij}^2, \ldots, \sigma_{ij}^M\} \tag{43}$$

as the multilink for the node pair (i,j) in terms of the $\sigma_{ij}^\alpha = \pm 1$ variables, we see that Equation (42) can be used to transform Equation (41) into

$$h_{ij}(s_{ij}, \theta_{ij}, J) = \left(\frac{\theta_{ij}}{2} - J\right) \sum_{\alpha=1}^{M} \sigma_{ij}^\alpha - \frac{J}{M} \sum_{\alpha<\beta} \sigma_{ij}^\alpha \sigma_{ij}^\beta - \frac{JM}{2} + \frac{M\theta_{ij}}{2}. \tag{44}$$

If we define

$$B_{ij} \equiv J - \frac{\theta_{ij}}{2}, \tag{45}$$

$$v_{ij} \equiv -MB_{ij} + \frac{JM}{2}, \tag{46}$$

then the pair Hamiltonian finally reduces to

$$h_{ij}(s_{ij}, B_{ij}, J) = -B_{ij} \sum_{\alpha=1}^{M} \sigma_{ij}^\alpha - \frac{J}{M} \sum_{\alpha<\beta} \sigma_{ij}^\alpha \sigma_{ij}^\beta + v_{ij}. \tag{47}$$

From the above expression, we see that, for every specific pair of nodes (i,j), the variables σ_{ij}^α can be thought of as Ising spins residing in the M nodes of a fully connected graph, where every Ising spin interacts with every other $M-1$ spins and is coupled to a 'field' B_{ij}. In terms of the multiplex networks being modeled, this means that for every specific pair of nodes (i,j), the edges connecting i and j throughout the M layers are all coupled to a common 'external' field B_{ij}, and are also coupled to each other with a homogeneous interaction strength J/M. A positive coupling $J > 0$ favors more overlap (i.e., more alignment between links in different layers), while $J < 0$ disfavors the overlap. The term v_{ij} is an inessential overall shift in energy independent of the spin configuration. This model is identical to the mean-field Ising or Curie–Weiss model. This exact mapping is what we use in Appendix in order to solve the model analytically, and in particular, to show the existence, for each pair of nodes, of a phase transition separating a 'magnetized' phase and a 'non-magnetized' phase, which here represent a 'multiplexed' phase (where links in different layers tend to 'align' to each other) and a 'non-multiplexed' phase, respectively.

The full Hamiltonian (40) is a summation of the Hamiltonians of *non-interacting* Ising systems, each for a distinct pairs of nodes. Note, however, that despite the independence of different pairs of nodes, the pair Hamiltonians $h_{ij}(s_{ij}, B_{ij}, J)$ share some parameters: J is common to all such Hamiltonians, and $h_{ij}(s_{ij}, B_{ij}, J)$ and (say) $h_{ik}(s_{ik}, B_{ik}, J)$ also share the parameter θ_i, because the latter appears in both B_{ij} and B_{ik}. This is the result of the original constraint on the degree of each node, which results in the same Lagrange multiplier θ_i appearing in all pair Hamiltonians involving the same node i. These common parameters imply that, even if all pairs of nodes are independent, the control parameters of all pair Hamiltonians cannot be chosen independently, resulting in a correlated phenomenology for the various pairs of nodes. In particular, as we shall see, each pair of nodes can undergo *locally* the typical phase transition of the mean-field Ising model, but the features of these pair-specific phase transitions are all nontrivially related to each other.

We also note, from Equations (44) and (47), that if $J = \theta_{ij}/2$ (or equivalently, $B_{ij} = 0$), then the pair Hamiltonian (hence the graph probability) becomes invariant upon a global 'spin flip' ($\sigma_{ij}^\alpha \to -\sigma_{ij}^\alpha \ \forall \alpha$), which here corresponds to the replacement of each existing link with a missing link ($g_{ij}^\alpha = 1 \to g_{ij}^\alpha = 0 \ \forall \alpha$) and, vice versa, of each missing link with an existing link ($g_{ij}^\alpha = 0 \to g_{ij}^\alpha = 1 \ \forall \alpha$). This is due to the vanishing of the 'external field' B_{ij} that, when present, selects a preferred 'spin direction' (up versus down), which here means a preferred density (high versus low). We expect that with the parameter

choice $J = \theta_{ij}/2$, the pair of nodes (i, j) gains an expected $1/2$ density of links across the M layers, i.e., an expected number of links equal to $M/2$, corresponding to half the maximum number of links for that node pair. Additionally, if J is smaller than the critical value, this expected number of links is also the typical value, and basically, the model is not fundamentally different from a model without constraints, where the intermediate density is produced as a result of a completely uniform probability distribution for the multilink. However, if J exceeds the critical value, the intermediate average density is no longer the typical one realized by individual graphs sampled from the model: rather, it is the ensemble average of two typical (high and low) values of the realized density, just like in the equivalent spin system, below the Curie temperature, and without an external field one would typically observe, with the same probability, overall positive and negative magnetization with a zero ensemble average. The numerical simulations access the typical realized values, while the equations still govern the expected value. This situation corresponds to a 'symmetry-broken' phase, where the typical realizations are less symmetric than the Hamiltonian that generates them. However, here, the heterogeneity of the degrees implies different values of the external field $B_{ij} = J - \theta_{ij}/2$, which means that the zero-field spontaneous symmetry breaking condition cannot, in general, be realized for all pairs of nodes simultaneously, leading to a phenomenology governed by the interplay between the values of J and $\{\theta_i\}_{i=1}^{N}$, and ultimately between the values of the inter-layer overlap and the node degrees.

3.2. Calculating the Partition Function

The partition function defined in (19) can be written as the product

$$Z(\vec{\theta}, J) = \sum_{\vec{G} \in \mathcal{G}_N^M} e^{-H(\vec{G}, \vec{\theta}, J)} = \sum_{\vec{G} \in \mathcal{G}_N^M} \prod_{i<j} e^{-h_{ij}(s_{ij}, \theta_{ij}, J)} = \prod_{i<j} z_{ij}(\theta_{ij}, J), \tag{48}$$

where $z_{ij}(\theta_{ij}, J)$ is the *pair partition function*, which is a sum of the set $\mathcal{S}_{ij}$ of all 2^M possible multilinks for (i, j):

$$z_{ij}(\theta_{ij}, J) \equiv \sum_{s_{ij} \in \mathcal{S}_{ij}} e^{-h_{ij}(s_{ij}, \theta_{ij}, J)}. \tag{49}$$

The multiplex probability can be written in terms of the multilink probabilities $P_{ij}(s_{ij}, \theta_{ij}, J)$:

$$P(\vec{G}, \vec{\theta}, J) = \prod_{i<j} P_{ij}(s_{ij}, \theta_{ij}, J) \tag{50}$$

where

$$P_{ij}(s_{ij}, \theta_{ij}, J) \equiv \frac{e^{-h_{ij}(s_{ij}, \theta_{ij}, J)}}{z_{ij}(\theta_{ij}, J)}. \tag{51}$$

The complete partition function and multiplex probability can therefore be obtained as products of pair-specific quantities, where each multilink can be regarded as a configuration of a Curie–Weiss system. To obtain an explicit expression for $z_{ij}(\theta_{ij}, J)$, we use a Hubbard–Stratonovich transformation and the Laplace theorem [49] in the limit $M \to \infty$. The details are provided in Appendix A and are a generalization of the approach used in [50]. The final result is

$$z_{ij}(\theta_{ij}, J) = 2^M e^{-\frac{M}{2}\theta_{ij} - 2JMu_{ij}(u_{ij}-1)} \cosh^M\left(2Ju_{ij} - \frac{\theta_{ij}}{2}\right), \tag{52}$$

where u_{ij} is the solution to the equation

$$u_{ij} = \frac{1}{2} + \frac{1}{2}\tanh\left(2Ju_{ij} - \frac{\theta_{ij}}{2}\right). \tag{53}$$

The solutions to the above equation will be discussed in the next section.

Now, given a particular real multiplex network $\vec{G}^*$, the log-likelihood, as defined, in general, in Equation (21), is

$$\mathcal{L}(\vec{\theta}, J) = \ln P(\vec{G}^*, \vec{\theta}, J) = \sum_{i<j} \Big[-h_{ij}(s^*_{ij}, \theta_{ij}, J) - \ln z_{ij}(\theta_{ij}, J) \Big]. \tag{54}$$

At a stationary point of $\mathcal{L}$, the derivatives of $\mathcal{L}$ with respect to every Lagrange multiplier must equal zero. As we show in Appendix B, this leads to the maximum likelihood equations

$$\sum_{j\neq i}^{N} \sum_{\alpha=1}^{M} g_{ij}^{*\alpha} = M \sum_{j\neq i}^{N} u^*_{ij} \quad \forall i \tag{55}$$

$$\frac{4}{M} \sum_{i<j} \sum_{\alpha<\beta} g_{ij}^{*\alpha} g_{ij}^{*\beta} = 2M \sum_{i<j} \Big(u^*_{ij} \Big)^2 \tag{56}$$

where u^*_{ij}, being the solution to Equation (53) with $(\vec{\theta}, J)$ replaced by $(\vec{\theta}^*, J^*)$, is implicitly related to the maximum likelihood parameters $(\vec{\theta}^*, J^*)$. Note that the quantities on the LHS of Equations (55) and (56) are precisely the quantities that we constrained from the start, namely, $M\overline{k}^*_i$ and $4O^*/M$, respectively. According to the maximum likelihood principle, these empirical quantities must equal their respective ensemble averages, $M\langle \overline{k}_i \rangle_{\theta^*, J^*}$ and $4\langle O \rangle_{\theta^*, J^*}/M$, which appear on the RHS. The quantity u^*_{ij} can therefore be considered as an *average* probability of a link occurring between the nodes i and j, which is equal throughout the M layers and is, therefore, a measure of the density of links in the multilink m_{ij}. This is similar to how we identified p_{ij} to be the connection probability in the ACM, which was based solely on the constraints $\overline{k}_i$. In support of this idea, we see that, in the case $J^* = 0$, the Lagrange multipliers $\vec{\theta}^*$ reduce it to the value $\vec{\theta}^0 \equiv \vec{\theta}^*|_{J^*=0}$, such that

$$u^*_{ij}\Big|_{J^*=0} = \frac{1}{2}\left[1 + \tanh\left(-\frac{\theta^0_i + \theta^0_j}{2} \right) \right] = \frac{e^{-(\theta^0_i + \theta^0_j)}}{1 + e^{-(\theta^0_i + \theta^0_j)}} = p_{ij}(\vec{\theta}^0) \tag{57}$$

which is identical to the expression in Equation (29), providing the link probability p_{ij} obtained in Section 2.5 in the absence of the constraint for the overlap. The quantity u^*_{ij} can therefore possibly be interpreted as a *mean-field* quantity that *globally* incorporates the layer interdependence that was introduced through the overlap O^*, but *locally* treats the layers as if they were independent. A characteristic of mean-field theories is that the effects of all elements of a system on a given element are approximated by a single, average effect.

Formally, we can calculate the entropy of the data, given the model, as the maximized likelihood using Equations (25) and (54):

$$\begin{aligned} S(\vec{\theta}^*, J^*) &= -\mathcal{L}(\vec{\theta}^*, J^*) \\ &= -\ln P(\vec{G}^*, \vec{\theta}^*, J^*) \\ &= H(\vec{G}^*, \vec{\theta}^*, J^*) + \sum_{i<j} \ln z_{ij}(\theta^*_{ij}, J^*) \\ &= M \sum_{i=1}^{N} \theta^*_i \overline{k}^*_i - \frac{4J^*}{M} O^* + \sum_{i<j} \ln z_{ij}(\theta^*_{ij}, J^*), \end{aligned} \tag{58}$$

which requires the knowledge of the parameters $\vec{\theta}^*$ and J^* (which are, however, defined only implicitly through u^*_{ij}). Comparing the above expression with Equation (37), we see that $S(\vec{\theta}^0, 0) = S_0(\vec{\theta}^0)$, as expected, i.e., the model with $J^* = 0$ has the same entropy as the equivalent ACM with no overlap, for the same value of $\vec{\theta}^0$. Similarly, $\mathcal{L}(\vec{\theta}^0, 0) = \mathcal{L}_0(\vec{\theta}^0)$ for the maximized likelihood in the two models. In order to understand the relationship between the entropies of the two models when $J^* \neq 0$, let us first note that a positive (resp. negative) coupling strength J^* means that the empirical overlap O^* is larger (resp. smaller) than the expected overlap under the null model with $J^* = 0$, i.e.,

$$O^* \lessgtr \langle O \rangle_{\vec{\theta}^0} \quad \Leftrightarrow \quad J^* \lessgtr 0 \tag{59}$$

where we have used the notation in Equation (36). However, one should not naively conclude from the combination of Equations (58) and (59) that the entropy of the model with $J^* < 0$ is larger than the entropy of the model with $J^* = 0$, because the two partition functions are different, and also because the two entropies are calculated for different Lagrange multipliers, i.e., $\vec{\theta}^* \neq \vec{\theta}^0$ when $J^* \neq 0$. In fact, we can actually show that the entropy of the model with $J^* \neq 0$ is always smaller than the one for the model with $J^* = 0$. To see this, we introduce the relative entropy (or Kullback–Leibler divergence) between the two models, as follows:

$$R(\vec{\theta}^0, \vec{\theta}^*, J^*) \equiv \sum_{\vec{G} \in \mathcal{G}_N^M} P(\vec{G}, \vec{\theta}^*, J^*) \ln \frac{P(\vec{G}, \vec{\theta}^*, J^*)}{P_0(\vec{G}, \vec{\theta}^0)} \geq 0, \tag{60}$$

where the last inequality is a well-known property of the relative entropy, and the equality is realized if, and only if, $P_0(\vec{G}, \vec{\theta}^0)$ and $P(\vec{G}, \vec{\theta}^*, J^*)$ are identical, which, in turn, requires $J^* = 0$, yielding $\vec{\theta}^0 = \vec{\theta}^*$ and $R(\vec{\theta}^0, \vec{\theta}^0, 0) = 0$. For $J^* \neq 0$, we can write

$$\begin{aligned}
R(\vec{\theta}^0, \vec{\theta}^*, J^*) &= \sum_{\vec{G} \in \mathcal{G}_N^M} P(\vec{G}, \vec{\theta}^*, J^*) \ln P(\vec{G}, \vec{\theta}^*, J^*) - \sum_{\vec{G} \in \mathcal{G}_N^M} P(\vec{G}, \vec{\theta}^*, J^*) \ln P_0(\vec{G}, \vec{\theta}^0) \\
&= -S(\vec{\theta}^*, J^*) + \sum_{\vec{G} \in \mathcal{G}_N^M} P(\vec{G}, \vec{\theta}^*, J^*) \Big[H_0(\vec{G}, \vec{\theta}^0) + \ln Z_0(\vec{\theta}^0) \Big] \\
&= -S(\vec{\theta}^*, J^*) + \sum_{\vec{G} \in \mathcal{G}_N^M} P_0(\vec{G}, \vec{\theta}^0) \Big[H_0(\vec{G}, \vec{\theta}^0) + \ln Z_0(\vec{\theta}^0) \Big] \\
&= -S(\vec{\theta}^*, J^*) + \sum_{\vec{G} \in \mathcal{G}_N^M} P_0(\vec{G}, \vec{\theta}^0) \ln P_0(\vec{G}, \vec{\theta}^0) \\
&= -S(\vec{\theta}^*, J^*) + S_0(\vec{\theta}^0),
\end{aligned} \tag{61}$$

where we have used the fact that $H_0(\vec{G}, \vec{\theta}^0) = M \sum_{i=1}^N \theta_i^0 \, \overline{k}_i(\vec{G})$ has the same expectation value, equal to $M \sum_{i=1}^N \theta_i^0 \, \overline{k}_i(\vec{G}^*)$, under both $P(\vec{G}, \vec{\theta}^*, J^*)$ and $P_0(\vec{G}, \vec{\theta}^0)$:

$$\begin{aligned}
\sum_{\vec{G} \in \mathcal{G}_N^M} P(\vec{G}, \vec{\theta}^*, J^*) H_0(\vec{G}, \vec{\theta}^0) &= M \sum_{i=1}^N \theta_i^0 \left[\sum_{\vec{G} \in \mathcal{G}_N^M} P(\vec{G}, \vec{\theta}^*, J^*) \overline{k}_i(\vec{G}) \right] \\
&= M \sum_{i=1}^N \theta_i^0 \, \overline{k}_i(\vec{G}^*) \\
&= M \sum_{i=1}^N \theta_i^0 \left[\sum_{\vec{G} \in \mathcal{G}_N^M} P_0(\vec{G}, \vec{\theta}^0) \overline{k}_i(\vec{G}) \right] \\
&= \sum_{\vec{G} \in \mathcal{G}_N^M} P_0(\vec{G}, \vec{\theta}^0) H_0(\vec{G}, \vec{\theta}^0).
\end{aligned} \tag{62}$$

Now, applying the inequality $R(\vec{\theta}^0, \vec{\theta}^*, J^*) \geq 0$ in Equation (60) to Equation (62), we get

$$0 \leq S(\vec{\theta}^*, J^*) \leq S_0(\vec{\theta}^0), \tag{63}$$

confirming that the entropy of the model with $J^* \neq 0$ is always smaller than the one for the model with $J^* = 0$, consistent with the fact that the former is more constrained than the latter.

4. Local Phase Transitions in the Model

The number of solutions of Equation (53) depends on the values of the parameters $\theta_{ij} = \theta_i + \theta_j$ and J. We illustrate this fact in Figure 1, where both the LHS and the RHS of

Equation (53) are plotted as a function of u_{ij} for various values of θ_{ij} and J. The appearance of multiple solutions signals the existence of *phase transitions* in the limit when the number M of layers diverges, which determine abrupt changes in the value of u_{ij} and, therefore, also in the properties of the multilink m_{ij} and the structure of the multiplex as a whole. The configurations for m_{ij} that are separated by a phase transition are the *phases* of the multilink. The point where multiple solutions appear or vanish is the *bifurcation point*.

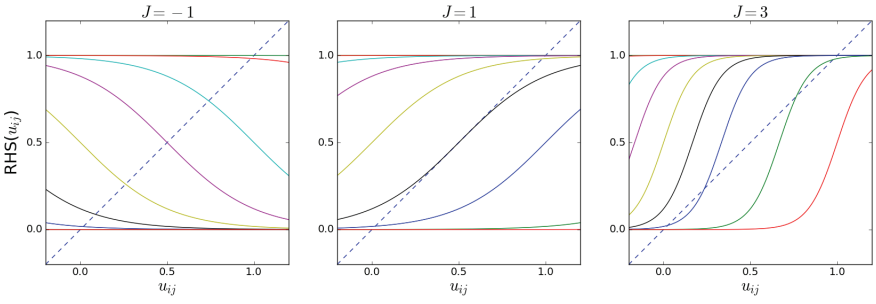


Figure 1. A graphical illustration of the solution(s) of Equation (53). The solid lines show the RHS of Equation (53) as a function of u_{ij} for the different parameters $\theta_{ij} \in \{-12, -8, -4, -2, 0, 2, 4, 8, 12\}$, while the dashed line shows the LHS, which equals u_{ij} itself. For a given parameter value, the solutions of Equation (53) are the intersection between the dashed and the corresponding solid line. Each panel corresponds to a different value of J (in the rest of the paper, we will consider only $J \geq 0$).

Figure 1 shows that, at the interval $0 \leq u_{ij} \leq 1$, there can be either one, two, or three solutions, and that for $\theta_{ij} \to +\infty$ or $\theta_{ij} \to -\infty$ there is always one solution, namely, $u_{ij} = 0$ or $u_{ij} = 1$, respectively. The number of solutions depends on whether the slope (derivative) of the RHS (which depends on the parameters) exceeds the slope of the LHS (which is always equal to 1) of Equation (53) at their intersection. From now on, we will consider only the case $J \geq 0$, which corresponds to a tendency to create an increased inter-layer overlap compared with the model with $J = 0$. The case $J < 0$ corresponds to the opposite case where the overlap is suppressed, which we do not discuss here. New solutions appear or vanish at the point where Equation (53) is satisfied *and* the derivatives of the LHS and RHS of Equation (53) are equal:

$$1 = J\left[1 - \tanh^2\left(2Ju_{ij} - \frac{\theta_{ij}}{2}\right)\right]. \tag{64}$$

Equation (64) cannot be satisfied if $0 \leq J \leq 1$, since $0 \leq \tanh^2(x) < 1$ for $x \in \mathbb{R}$, and, therefore, if $J \leq 1$, a phase transition is impossible, and there is a unique solution for u_{ij}. When $J > 1$, Equation (64) gives us two potential solution branches, $u_{ij}^{\pm} = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - 1/J}$, where we have used $2u_{ij} - 1 = \tanh\left(2Ju_{ij} - \theta_{ij}/2\right)$. Equation (53) can be written as $\theta_{ij} = 4Ju_{ij} - \ln\left[u_{ij}/(1 - u_{ij})\right]$ using the identity $\tanh^{-1} x = \frac{1}{2}\ln\left[(1 + x)/(1 - x)\right]$. By then substituting $u_{ij}^{\pm}$ into this expression for θ_{ij}, we obtain the equations for the two curves in the (J, θ_{ij}) plane that mark the points where additional solutions appear or vanish:

$$\theta_{ij}^{+}(J) = \frac{2\sqrt{J}}{\sqrt{J} - \sqrt{J-1}} - \ln\left(\frac{\sqrt{J} + \sqrt{J-1}}{\sqrt{J} - \sqrt{J-1}}\right), \tag{65}$$

$$\theta_{ij}^{-}(J) = \frac{2\sqrt{J}}{\sqrt{J} + \sqrt{J-1}} - \ln\left(\frac{\sqrt{J} - \sqrt{J-1}}{\sqrt{J} + \sqrt{J-1}}\right), \tag{66}$$

as shown in Figure 2. In the region between the two curves, there are three solutions to Equation (53). Note that the 'zero-field' condition $\theta_{ij} = 2J$ is always in that region when $J > 1$. This means that the condition $J > 1$ is sufficient to ensure that the system is in the magnetized (symmetry-broken) phase when in the absence of the external field. However,

when $\theta_{ij} \neq 2J$, the condition $J > 1$ is necessary but not sufficient. In particular, generally, it may happen that, for a given value of $J > 1$, different pairs of nodes will be in different (magnetized or non-magnetized) phases depending on the value of θ_{ij}. This shows that the system can undergo a multitude of separate phase transitions if the parameters $\{\theta_{ij}\}$ remain fixed and J is varied.

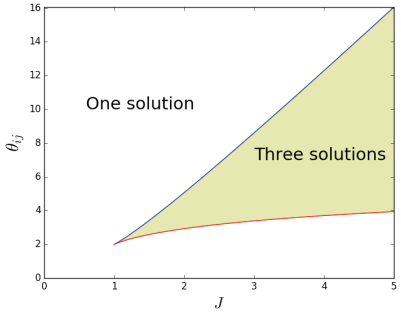


Figure 2. The upper (blue) and lower (red) curves correspond to Equations (65) and (66), respectively, which delimit the region of phase space (yellow area), for which Equation (53) has three solutions. Note that the 'zero-field' condition $\theta_{ij} = 2J$ is always in the yellow area when $J > 1$, so the condition $J > 1$ is sufficient to ensure that the system in zero field is in the magnetized (symmetry-broken) phase.

In the magnetized phase, the phenomenon of symmetry breaking will occur: the typical realized values of the 'magnetization' will not coincide with the corresponding ensemble average. In the zero-field case ($\theta_{ij} = 2J$), the symmetry breaking is 'spontaneous', i.e., not induced by any field pointing in a preferred direction, while in the nonzero-field case, the symmetry is broken by the field itself. This well-known property of the Ising model has specific implications for our problem here. Indeed, while certain values of θ_{ij}, J may solve the maximum likelihood Equations (55) and (56), the corresponding solutions to Equation (53) may not necessarily maximize the likelihood, and are therefore not 'valid' (or *stable*). Once the values θ^*_{ij} and J^* that solve the maximum likelihood equations are found, the graph probability corresponding to this set of values can be written as a function of the configuration of the graph (or the collection of configurations of the multilinks m_{ij}), and one can check which typical configurations (those minimizing the Hamiltonian) arise. As Figure 1 suggests, in the regime where there are three solutions, u_{ij}, one value will be relatively high (which corresponds to a relatively high density of links in m_{ij}), another value will be relatively low (which corresponds to a relatively low density of links in m_{ij}), and the third value will be between the other two, corresponding to an intermediate density of links in m_{ij}. By inspecting the (pair) Hamiltonian in Equation (47) in terms of the $\sigma^\alpha_{ij} = 2g^\alpha_{ij} - 1$ variable, it becomes clear which of the three solutions u^*_{ij} are viable (stable). In the case where $B_{ij} = 0$, or equivalently, when $\theta_{ij} = 2J$, the (pair) Hamiltonian is symmetric with respect to a change in sign, $\sigma^\alpha_{ij} \to -\sigma^\alpha_{ij}$, which means that the high- and low-density solutions are equal. This is the symmetry-broken situation we have discussed in Section 3.1. In this case, the intermediate-density solution will result in a lower value for the Hamiltonian than the high- and low-density solutions. The viable (stable) solutions are therefore the high- and low-density ones. In the case where $B_{ij} \neq 0$, it is clear that the high-density solution minimizes the Hamiltonian when $B_{ij} > 0$ and maximizes it when $B_{ij} < 0$. The low-density solution minimizes the Hamiltonian when $B_{ij} < 0$ and maximizes it when $B_{ij} > 0$. The intermediate solution will, however, never minimize the Hamiltonian when $B \neq 0$, and is therefore never viable (stable). From these considerations, it becomes clear that a phase transition, corresponding to a sudden change in u_{ij}, may only happen when we cross from a negative (positive) B_{ij} to a positive (negative) B_{ij} (when $J > 1$). Figure 3 shows the symmetric stable solutions u_{ij} in the case where $B_{ij} = 0$, with the bifurcation occurring at $J = 1$. In case of the positive field $B_{ij} = +1$, it shows a single stable solution

curve, which is the high-density solution (in the case where $B_{ij} = -1$, this image would be flipped with respect to the $u^*_{ij} = 1/2$ axis). The right panel in Figure 3 shows that the value of the stable solution u_{ij} jumps when B_{ij} crosses from positive to negative, as expected.

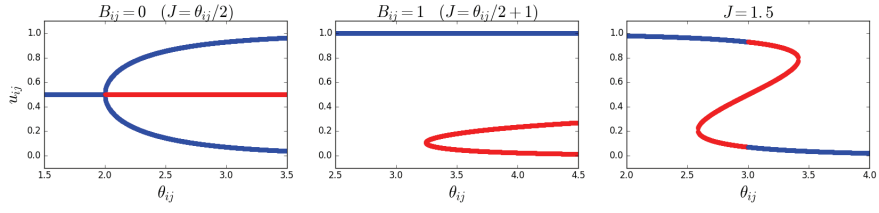


Figure 3. Solutions for u_{ij} as a function of θ_{ij} for different parameter values. The blue and red segments of the curve(s) correspond to the stable and unstable solutions of Equation (53), respectively. **Left** panel: $B_{ij} = 0$ (with J varying accordingly). **Middle** panel: $B_{ij} = 1$ (with J varying accordingly). **Right** panel: constant value of $J = 1.5$, which translates to a non-constant B_{ij}.

Combining the above considerations for all multilinks simultaneously, and adding the other constraint on the layer-averaged degrees, the multiplex will undergo a sequence of phase transitions, determining a hierarchy of increasingly ordered (magnetized, or rather 'multiplexed' in this case) phases where, for an increasing number of pairs of nodes, the links in different layers will tend to 'align' to each other (for $J > 1$). The separations between these phase transitions will depend on the values of the enforced layer-averaged degrees, which determine $\vec{\theta}^*$. The fully ordered phase, where all pairs of nodes are multiplexed, is the one where all the M layers of the multiplex are perfectly aligned, and are, therefore, basically an identical copy of each other. We might say that, in this case, the effective number of independent layers is $M_{\text{eff}} \approx 1$, and the expected overlap is maximal and proportional to the expected number $\langle L \rangle_{\vec{\theta}^*, J^*} = \sum_{\alpha=1}^{M} \sum_{i<j} u^*_{ij}$ of links in the entire multiplex:

$$\langle O \rangle_{\vec{\theta}^*, J^*} \approx \sum_{\alpha<\beta} \sum_{i<j} u^*_{ij} = (M/2) \langle L \rangle_{\vec{\theta}^*, J^*}, \tag{67}$$

since $\langle g^{\alpha}_{ij} g^{\beta}_{ij} \rangle_{\vec{\theta}^*, J^*} \approx \langle g^{\alpha}_{ij} \rangle_{\vec{\theta}^*, J^*} = u^*_{ij}$ for most pairs, i.e., α, β, of layers. In the opposite extreme, we have a fully disordered phase where no pair of nodes is multiplexed (for instance, if $J < 1$), so the effective number of independent layers is maximal ($M_{\text{eff}} \approx M$), and the expected overlap is basically of the order of that given by Equation (36) for the model with $J^* = 0$, i.e.,

$$\langle O \rangle_{\vec{\theta}^*, J^*} \approx \sum_{\alpha<\beta} \sum_{i<j} (u^*_{ij})^2, \tag{68}$$

since $\langle g^{\alpha}_{ij} g^{\beta}_{ij} \rangle_{\vec{\theta}^*, J^*} \approx \langle g^{\alpha}_{ij} \rangle_{\vec{\theta}^*, J^*} \langle g^{\beta}_{ij} \rangle_{\vec{\theta}^*, J^*} = (u^*_{ij})^2$ for most pairs of layers. The relationship between $\langle O \rangle_{\vec{\theta}^*, J^*}$ and $\langle L \rangle_{\vec{\theta}^*, J^*}$ will depend on the specific values of $\{u^*_{ij}\}_{i<j}$, so ultimately, on the enforced degree sequence. Between these two extremes, if the phases are well separated (which here means that the enforced degrees of different nodes have very different values), there will be intermediate regimes where $\langle O \rangle_{\vec{\theta}^*, J^*}$ and $\langle L \rangle_{\vec{\theta}^*, J^*}$ scale in a way that is between the two limiting scalings. All these general considerations will be confirmed in the next sections with numerical, analytical, and empirical analyses.

5. Numerical Analysis

Equations (53), (55), and (56) are the key equations of our OACM model. These equations are generally, however, very difficult to solve. Therefore, before creating a null model for a real-world network by solving the maximum likelihood equations to find the Lagrange multipliers, we shall first treat the Lagrange multipliers as free parameters in order to explore and analyze the properties of the model as a function of these parameters. This analysis shall be performed by utilizing the Metropolis–Hastings algorithm [51]. This algorithm can be used to sample the exponential probability distribution defined by

the Hamiltonian of the model. By sampling the distribution, we numerically obtain various properties of the graph ensemble, which may then be compared to our analytical results in order to test the validity of the latter. Note that the sampling of the exponential distribution defined by a specific Hamiltonian may also be regarded as the simulation of a multiplex that corresponds to that Hamiltonian.

5.1. Exploring the Parameter Space

In order to explore the space of parameters, we are primarily interested in the difference between statistically *homogeneous* networks and statistically *heterogeneous* ones. To this end, we will explore the parameter space $(\theta_1, \ldots, \theta_N, J)$ of the model by specifying a value for J and sampling certain transformed parameters $x_1, \ldots, x_N$ from a distribution for each class, where $x_i \equiv e^{-\theta_i}$. The quantity x_i will be referred to as the 'fitness', or 'hidden variable', of node i. The broader the distribution of the fitness, the more heterogeneous the resulting network structure.

5.1.1. Homogeneous Fitness: Erdős–Rényi Graphs with Overlap

The simplest distribution from which we can sample $x_1, \ldots, x_N$ is the delta distribution centered at x, such that $x_1 = x_2 = \ldots = x_N \equiv x$ and, therefore, $\theta_1 = \theta_2 = \ldots = \theta_N \equiv \theta = -\ln x$, resulting in statistically homogeneous networks. With this choice of parameters, our model is an extension of the Erdős–Rényi model, which is a random graph model that can be derived within the ERGM by solely constraining the total number of links in the network, and where all links occur with the same probability. As we shall see, the extension derives from the fact that the extra constraint on the overlap can lead to a symmetry-breaking phase transition, although the broken symmetry might not manifest at first sight. Indeed, since the parameters are the same for all pairs of nodes, the condition for the existence of multiple solutions is also the same, and, therefore, there is a unique phase transition where, depending on the values of θ and J, pairs of nodes are either all 'magnetized' or all 'non-magnetized'. Similarly, since here $\theta_{ij} = \theta_i + \theta_j = 2\theta \ \forall i, j$, the spontaneous symmetry-breaking condition discussed in Section 3.1 for the vanishing of the external field is the same for all pairs of nodes, and given by $J = \theta$. In the symmetry-broken (magnetized) phase, for all pairs of nodes, the expected value of $\sum_{\alpha=1}^{M} g_{ij}^{\alpha}$ (or equivalently, of the 'magnetization' $\sum_{\alpha=1}^{M} \sigma_{ij}^{\alpha}$) is the same, and is always between the two typical (high-density and low-density) realized values. However, since all pairs are independent, the actual realized values of $\sum_{\alpha=1}^{M} g_{ij}^{\alpha}$ are also independent across pairs, so on average, over the entire network, the magnetization will realize both the low-density and high-density values, with equal probability. In other words, different pairs of nodes are i.i.d. realizations of the same system. This is a peculiar situation where the realized values of L and O (which represent sums of all pairs of nodes) will still coincide with their expected values as if no symmetry breaking was present, even if different pairs of nodes actually realize different symmetry-broken values that are individually different from the expected value. The net result is an expected number of links $\langle L \rangle = MN(N-1)/4$) equal to half the maximum one, or equivalently, an average zero magnetization in the associated spin system. Similar considerations apply to the case $J \neq \theta$, with the difference that, in that case, the symmetry is not broken spontaneously, but by the direction of the external field (value of θ), which implies that the two typical realized values of the magnetization for a given pair of nodes are no longer symmetric around the expected value. Still, both typical values will be realized, independently and with their probabilities, across the entire network, because different pairs of nodes are still independent. So, irrespective of the value of J and θ, we expect to observe realized values of L and O that correspond again to what one would observe without symmetry breaking, using the ensemble averages for each pair, irrespective of the phase of the system. All these considerations are confirmed below.

By looking at Equation (38), we can see that a uniform θ essentially means that instead of constraining the average layer degrees $\overline{k}_i$, we constrain the total number of links L in the

multiplex network. In this case, the combined maximum entropy and maximum likelihood equations become

$$u = \frac{1}{2} + \frac{1}{2}\tanh\left(2J^*u - \theta^*\right) \tag{69}$$

$$\sum_{i<j}^{N}\sum_{\alpha=1}^{M} g_{ij}^{*\alpha} = \frac{MN(N-1)}{2}u^* = \langle L\rangle_{\theta^*,J^*} \tag{70}$$

$$\frac{4}{M}\sum_{i<j}\sum_{\alpha<\beta} g_{ij}^{*\alpha}g_{ij}^{*\beta} = MN(N-1)(u^*)^2 = \frac{4}{M}\langle O\rangle_{\theta^*,J^*} \tag{71}$$

where $u^* = u(\theta^*, J^*)$ is the solution to Equation (69). Note that we now have a single equation for u, confirming the existence of a single *global* phase transition across the multiplex network, rather than separate local phase transitions for every multilink m_{ij}. Additionally, we note that if u^* can be considered as the density (and the link probability) of the network, then the value of u^* is exactly the same as the value of the density p in the Erdős–Rényi model [14,27], which solely constrains the number of links in the network. The difference between our model and the Erdős–Rényi model is that our model contains the possibility of a phase transition. However, since the number of links $\langle L\rangle$ also determines the overlap $\langle O\rangle$, the two quantities cannot be tuned independently of each other.

By using the Metropolis–Hastings algorithm, we have sampled our ERGM for multiplexes with $M = 100$ layers and $N = 100$ nodes for various values of θ and/or J. If we repeat the simulations for $J = 1.5$ and $\theta = 1.4$, $\theta = 1.5$, and $\theta = 1.6$, the system must undergo a phase transition as per Figure 3. We expect an abrupt change in the value of u^*, and according to Equations (70) and (71), we therefore expect an abrupt change in the equilibrium value of both L and O. Figure 4 shows simulations for $\theta \in \{1.4, 1.5, 1.6\}$ confirming the transition from a relatively high to a low density as the value of the field $B = J - \theta$ changes sign. These simulations have been repeated for different combinations of values for J and θ around the point where B changes sign, confirming the results shown here. Note that the middle plot in Figure 4 shows that the algorithm converges to multiplexes with a density of $1/2$, confirming that, when $B = 0$, L is approximately half of the total amount of possible links in the multiplex, as we expected above.

In Figure 5 we test the prediction, given by Equations (70) and (71), of the quadratic relationship $\langle O\rangle = \langle L\rangle^2/N^2$. Note that this quadratic trend is predicted irrespective of the value of $J > 0$, and even coincides with what Equation (68) predicts in the case $J = 0$ for a homogeneous multiplex with constant θ, as considered here. So, in this case, the expected relationship between $\langle O\rangle$ and $\langle L\rangle$ is not informative regarding the phase transition, although the specific values picked up by the system along the curve are. Indeed, we again simulate multiplexes with $M = 100$ layers, $N = 100$ nodes, and a variety of values for θ and J. Each simulation results in a value for $\langle L\rangle$ and a value for $\langle O\rangle$, which we plot against each other. These points are then compared to the theoretical points predicted by Equations (69)–(71) for the chosen parameter values, and added to Figure 5. We see that the relationship between simulated quantities is in agreement with the one predicted by the model. As we had anticipated, this is the result of the fact that different pairs of nodes are i.i.d. realizations of the same system, so that the ensemble average is realized as a sample average of the pairs of nodes across the network, even if in the symmetry-broken phase, the ensemble average of $\sum_{\alpha=1}^{M} g_{ij}^{\alpha}$ is not representative of any of the values realized locally for individual pairs of nodes. Therefore, the only scaling we observe coincides with the one given in Equation (68) for the 'non-magnetized' regime in the case where θ is the same for all nodes. The only, although very important, signature of the phase transition we see in Figure 5 is the fact that, for $J > 1$ and $\theta \neq J$, both the simulated data and the corresponding theoretical predictions 'drift away' from the intermediate values of $\langle L\rangle$ (which are still obtained for $\theta = J$) towards either low ($\theta > J$) or high ($\theta < J$) values of $\langle L\rangle$. This is because the realized multiplex networks are either low-density or high-density,

which is an indication of a phase transition occurring when increasing the value of J, exactly as predicted by Figure 3.

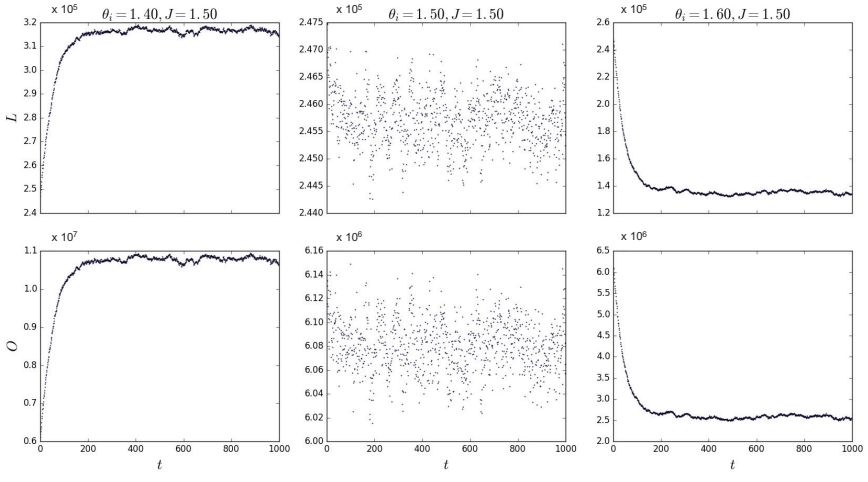


Figure 4. Total number of links L (top panels) and inter-layer overlap O (bottom panels) as a function of simulation time using the Metropolis–Hastings algorithm for $J = 1.5$, $N = 100$, $M = 100$. **Left** panels: $\theta = 1.4$. **Middle** panels: $\theta = 1.5 = J$ (symmetry-broken case). **Right** panels: $\theta = 1.6$. For fixed J, varying θ determines a phase transition from a high-density phase to a low-density phase.

We conclude our discussion of the homogeneous case by noting that, given an empirical multiplex $\vec{G}^*$ of interest, the entropy of the data given, in general, by Equation (58) reduces, in this case, to

$$\begin{aligned} S(\theta^*, J^*) &= M\theta^* \sum_{i=1}^{N} \overline{k}_i^* - \frac{4J^*}{M} O^* + \sum_{i<j} \ln z_{ij}(2\theta^*, J^*) \\ &= 2\theta^* L^* - \frac{4J^*}{M} O^* + \frac{N(N-1)}{2} \ln z(2\theta^*, J^*), \end{aligned} \tag{72}$$

where we have used Equation (9) (denoting, via $L^* = L(\vec{G}^*)$, the total number of links in the multiplex, which also equals the expected value $\langle L \rangle_{\theta^*, J^*}$) and the fact that the pair partition function z_{ij}, given by Equation (52), has the same value $z(2\theta^*, J^*) \equiv z_{ij}(2\theta^*, J^*)$ for all the $N(N-1)/2$ pairs of nodes. From Equation (72), we see that the entropy is determined, as expected, by both L^* and O^*. At the same time, we know that O^* depends uniquely and quadratically on L^* in this homogeneous model. The values achieved by the entropy are, therefore, bound by the relationship between L^* and O^*, which here is the same irrespective of the value of J^*, including when $J^* = 0$. In any case, the entropy also depends on the specific values of (θ^*, J^*), and Equation (63) guarantees that an upper bound for $S(\theta^*, J^*)$ is given by the entropy $S_0(\theta^0)$ of the ACM model with $J^* = 0$ and $\theta^* = \theta^0$ (clearly, the homogeneity implies that $\theta_i^0 = \theta^0$ for all $i = 1, N$ in the ACM model as well).

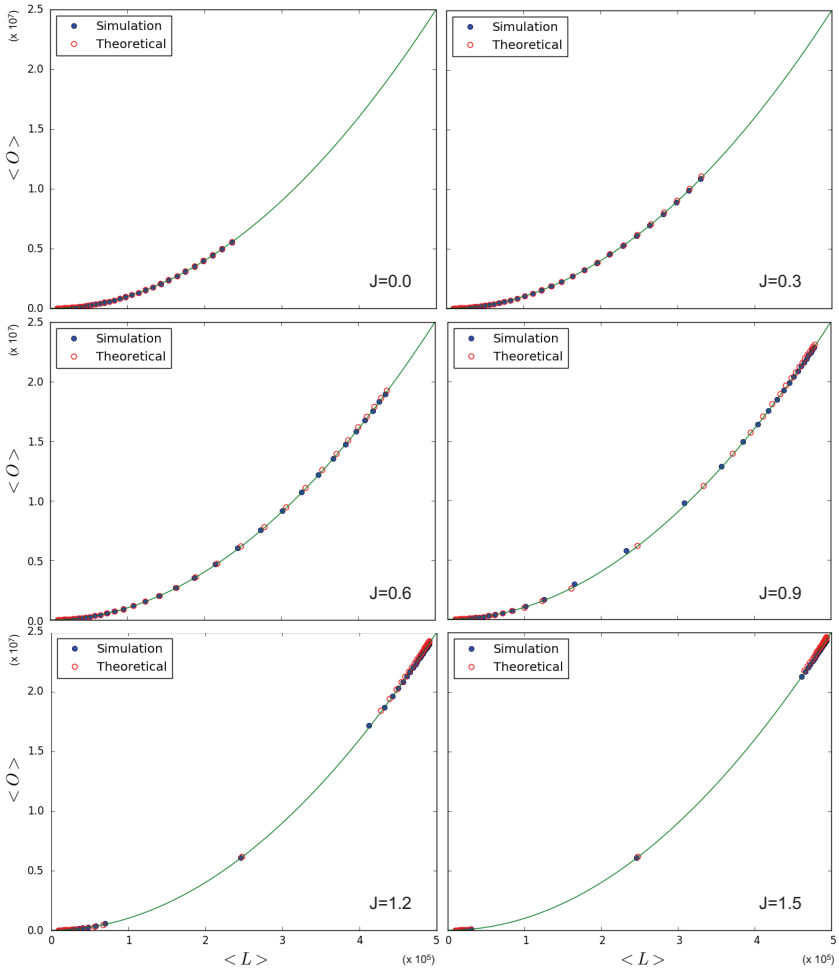


Figure 5. Relationship between the expected inter-layer overlap $\langle O \rangle$ and the total number of links $\langle L \rangle$ in homogeneous multiplexes with $N = 100$, $M = 100$, and $\theta_i = \theta$ for all $i = 1, N$. The blue points correspond to simulations obtained via the Metropolis–Hastings algorithm for $J \in \{0.0, 0.3, 0.6, 0.9, 1.2, 1.5\}$ and $\theta \in [0.05, 2.00]$ in steps of $\Delta\theta = 0.05$. The open red circles are the corresponding theoretically predicted points. The solid curve corresponds to the quadratic trend $\langle O \rangle = \langle L \rangle^2 / N^2$ predicted for all $J \geq 0$. Multiple solutions for u^*_{ij} first appear when $J > 1$, but the system keeps following the quadratic trend, albeit drifting away from the central point obtained for the zero-field case $\theta = J$ (corresponding to a spontaneously broken symmetry).

5.1.2. Power-Law-Distributed Fitness: Scale-Free Networks with Overlap

We now move away from the homogeneous case and consider a situation where the fitness values $\{x_i\}_{i=1}^N$ are drawn from a heavy-tailed distribution, in particular, a power law. This choice will produce a high degree of heterogeneity. In the ACM (see Section 2.5), the expected degree distribution is determined by the Lagrange multipliers θ_i, or equivalently, the transformed hidden variables $x_i = e^{-\theta_i}$. If x is distributed according to a power law, the expected degree distribution shall be distributed according to a power law as well, with the modulo as an upper cut-off. Since our OACM is an extension of the ACM, we will still sample x_i from a power law distribution $P(x) \sim x^{-\gamma}$ for various values

of γ, even though the expected degree distribution is not solely determined by the hidden variables $\{x_i\}$, but depends on J as well. In any case, a higher level of heterogeneity in the hidden variables x_i will lead to a higher level of heterogeneity in the degrees. Since the parameter space is rather large ($N+1$-dimensional), we define

$$x_i = z x_{0,i} \tag{73}$$

where z is a scaling factor. We sample $x_{0,i}$ only *once* from every chosen distribution. The value of x_i is varied by varying the scaling factor z. The parameter space to be explored will then be (z, J), which is 2-dimensional. We deduce that

$$\theta_i = -\ln(z x_{0,i}) \tag{74}$$

which shows that an increasing z leads to a decreasing θ_i. In the ACM, we have shown that the link probability is equal to $p_{ij} = x_i x_j/(1 + x_i x_j)$, which means that larger values of x_i lead to a larger expected degree, so that increasing all the fitness values will increase the density in the network. This qualitative relationship still holds with the addition of the constraint on the expected overlap (for fixed J).

The complexity of Equations (53), (55), and (56) does not allow us to easily derive the expected relationship between the overlap and the number of links in the network, as was the case when θ_i was constant. It is, however, possible to visualize the relationship between the overlap and the number of links by using the Metropolis–Hastings algorithm. Figure 6 shows this relationship, where x_i is sampled from power law distributions with various values of γ, alongside the expected quadratic term previously observed to occur for homogeneous values of the fitness x_i (delta distribution). We see that the overlap for a given number of links is higher in the cases where x is drawn from a power law distribution than when x is drawn from a delta distribution, even though the coupling parameter J is kept constant. The cause of this difference lies in the level of heterogeneity of the fitness distribution: unlike the homogeneous case, now different pairs of nodes have very different values of $\theta_{ij} = \theta_i + \theta_j$, and, therefore, the condition $J = \theta_{ij}/2$ for the vanishing of the 'external field' B_{ij} (spontaneous symmetry-breaking condition) cannot be realized simultaneously by all pairs. The figure also shows the effect of different exponents of the power law distributions of the fitness. A smaller value of γ leads to a higher overlap for a given number of links. By increasing the value of γ, the power law distribution becomes more sharply peaked, and will therefore lead to more homogeneous networks. Note, however, that increasing the value of the coupling parameter J itself also leads to an increase in the overlap for a given number of links for the same distribution.

Importantly, the phase transition now occurs for different pairs of nodes as J is varied. Some pairs of nodes will be in the non-magnetized phase, while others will be in the magnetized phase. The effective number M_{eff} of independent layers will, in general, depend on the choice of parameters. Among the magnetized pairs, the realized values of the overlap are no longer those corresponding to the ensemble average (as in the homogeneous case), but typically to the symmetry-broken solution with lower energy (hence dictated by the value of θ_{ij}), because no other pair of nodes will, in general, exist with the same parameters and such that the two symmetry-broken values are averaged by the resulting value of the realized overlap. In particular, while for $0 < J < 1$ all node pairs are in the non-magnetized phase, as J increases from 1 towards larger values, the pairs of nodes that first undergo the phase transition are the ones with values $\theta_i + \theta_j$ that fall between the limits set by Equations (65) and (66). As those equations and Figure 2 show, there are more and more combinations $\theta_i + \theta_j$ entering the magnetized phase as J increases. When J is sufficiently large, all pairs will be magnetized. Clearly, for any two pairs of nodes, (i, j) and (i, k), that share the same node, i, the values of $\theta_i + \theta_j$ and $\theta_i + \theta_k$ will be correlated, as they share the same term θ_i. This means that the pairs of nodes entering the magnetized phase typically have nodes in common, even if it would be incorrect to say that individual nodes enter the magnetized phase 'one by one', while this is certainly correct for individual node pairs, if the sum $\theta_i + \theta_j$ is different across all of them.

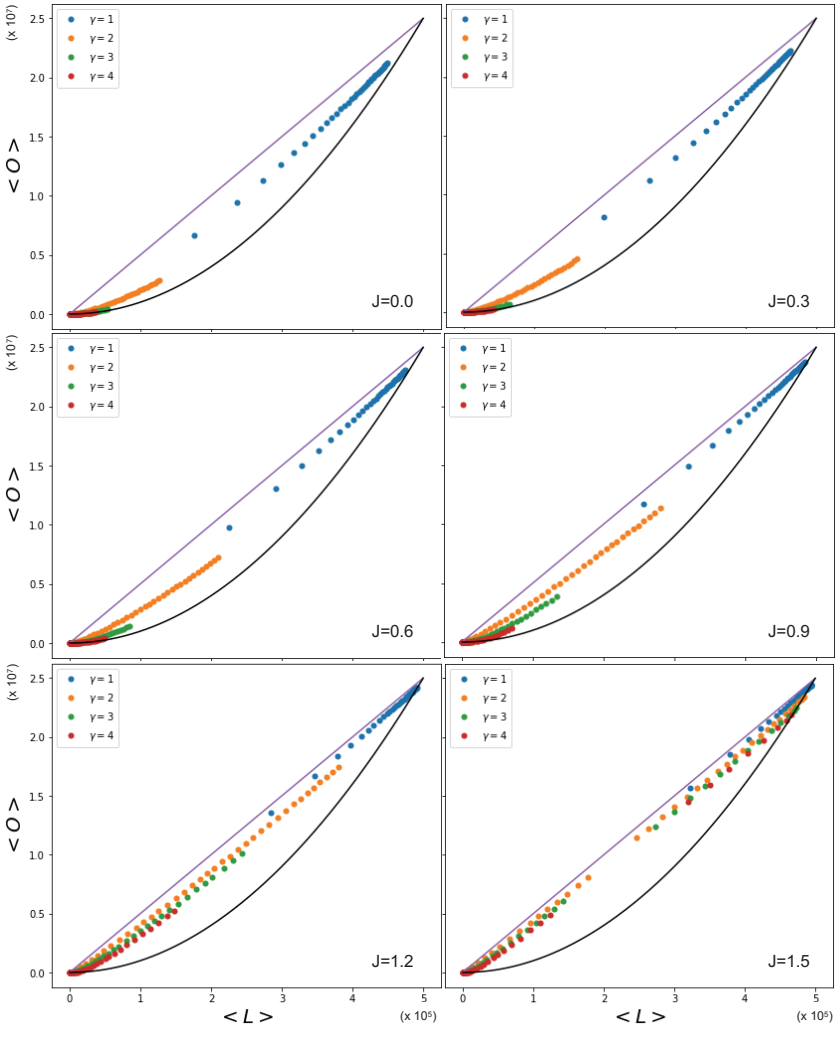


Figure 6. Relationship between the expected inter-layer overlap $\langle O \rangle$ and the total number of links $\langle L \rangle$ in heterogeneous multiplexes with $N = 100$, $M = 100$, and $x_{0,i}$ sampled from a power law distribution with different values for γ. The colored points correspond to simulations obtained via the Metropolis–Hastings algorithm for $J \in \{0.0, 0.3, 0.6, 0.9, 1.2, 1.5\}$ and $z \in [0.05, 2.00]$ in steps of $\Delta z = 0.05$. The straight line corresponds to the upper limit $\langle O \rangle = M\langle L \rangle / 2$ calculated in Equation (67). The solid curve corresponds to the quadratic trend $\langle O \rangle = \langle L \rangle^2 / N^2$ (achieved by homogeneous multiplexes with constant x_i), which here turns out to mark a lower bound. For increasing values of J, and especially as $J > 1$, the system moves closer to the upper bound. For $J = 1.5$, we see that the points are concentrating towards high-density and low-density (symmetry-broken) regimes, drifting away from the intermediate values, like in the homogeneous case. However, this is now the combined result of the behavior of statistically different pairs of nodes, each having a different zero-field condition $\theta_i + \theta_j = 2J$, so the spontaneous symmetry breaking cannot be realized for all node pairs simultaneously.

Figure 6 indeed shows the effect of the changing number of magnetized node pairs as J increases above 1. We note that, for larger and larger J, the relationship between $\langle O \rangle$ and $\langle L \rangle$ tends towards the 'maximally multiplexed' linear extreme (shown as a straight line) given in Equation (67). At the same time, we see that the 'non-multiplexed' case ($J < 1$) described by Equation (68) now realizes values of the overlap that are very different from the quadratic trend achieved by the homogeneous model (also shown as a solid curve in Figure 6), which now turns out to represent a lower bound. We can 'zoom in' to better see this difference by looking at Figure 7, where, by using Equations (53), (55), and (56), we additionally calculate the theoretically predicted values of $\langle O \rangle$ and $\langle L \rangle$ and compare them to the simulation data, where $x_{0,i}$ is sampled from a power law distribution with $\gamma = 1$ (the results for $\gamma \in \{2, 3, 4\}$ are qualitatively similar and are therefore not shown here). The figure confirms a strong deviation from the curve for the homogeneous model, even when $J = 0$ (signaling a much higher but spurious overlap, arising only from the rising correlation among node degrees across different layers), and a close agreement with the maximally overlapping value in Equation (67) already for $J = 1.5$ (corresponding to a further increase in overlap, arising from an additional, genuine coupling between layers).

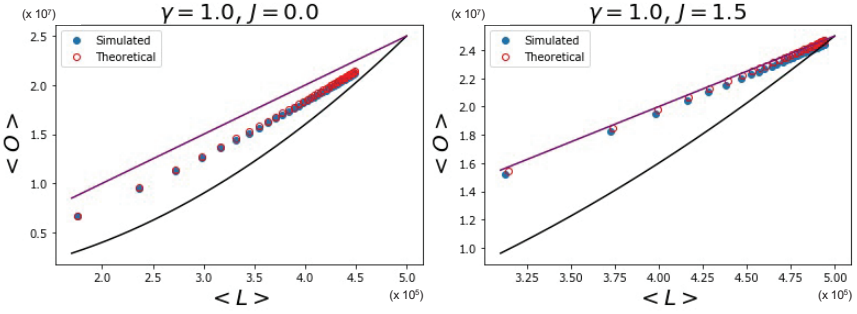


Figure 7. Relationship between the expected inter-layer overlap $\langle O \rangle$ and the total number of links $\langle L \rangle$ in heterogeneous multiplexes with $N = 100$, $M = 100$, and $x_{0,i}$ sampled from a power law distribution with $\gamma = 1$. The blue points correspond to simulations obtained via the Metropolis–Hastings algorithm for $z \in [0.05, 2.00]$ in steps of $\Delta z = 0.05$ with $J = 0$ (**left** panel) and $J = 1.5$ (**right** panel). The red open circles are the theoretically predicted values corresponding to the same parameters used in the simulations. The straight line corresponds to the upper limit $\langle O \rangle = M\langle L \rangle / 2$ calculated in Equation (67). The solid curve corresponds to the quadratic trend $\langle O \rangle = \langle L \rangle^2 / N^2$ (achieved by homogeneous multiplexes with constant x_i), which here turns out to mark a lower bound. We see that, compared with the homogeneous lower bound, the heterogeneity of nodes increases the overlap dramatically, even in the absence of true coupling ($J = 0$). When coupling is present, the overlap is additionally increased and already approaches the upper bound for $J = 1.5$.

5.1.3. Log-Normally Distributed Fitness

The delta and power law distributions we have considered so far represent examples of completely homogeneous and extremely heterogeneous (especially for $\gamma = 1$) distributions, respectively. We now consider the log-normal distribution as a third example between these two extremes. This analysis will indeed lead to results that are in some sense intermediate between what we have observed so far, and useful for interpreting the real-world case that we will present later on. A log-normal distribution is the distribution of a random variable whose logarithm is normally distributed (i.e., if the random variable x is log-normally distributed, then $y = \ln x$ follows a normal distribution). The probability density for a log-normal distribution is

$$P(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2/(2\sigma)}, \tag{75}$$

where μ and σ correspond to the mean and the standard deviation of the normal distribution of $\ln x$. We will vary the value of x_i by again introducing a scaling factor that can be changed such that $x_i = zx_{0,i}$ and $\theta_i = -\ln(zx_{0,i})$, where we sample $x_{0,i}$ *once* from the log-normal distribution for a variety of values for μ and σ.

The log-normal distribution allows us to inspect the transition in the relationship between the overlap and the number of links from the quadratic lower limit to the linear upper limit by varying the value of σ. Indeed, when $0 < \sigma \ll 1$, the normal distribution of $\ln x_{0,i}$ is sharply peaked. By decreasing the value of σ towards 0, $\ln x_{0,i}$ (and, therefore, $x_{0,i}$ as well) shall approach a delta distribution. This is the distribution that led us to the quadratic lower limit for the relationship between the overlap and the number of links in the network. Conversely, when $\sigma \gg 1$, the log-normal distribution approaches a distribution with a power law tail with $\gamma = 1$. This distribution led us to the linear upper limit between the overlap and the number of links in the network (when J was sufficiently large). By increasing the value of σ from 0 to a sufficiently large value (e.g., $\sigma = 10$), we can therefore increase the heterogeneity of the network from a completely homogeneous network achieving the quadratic lower limit to an extremely heterogeneous network close to the linear upper limit relationship in the simulation data.

Figure 8 shows the relationship between the average overlap and the number of links in the network with simulation data that were obtained by using the Metropolis–Hastings algorithm for a variety of values for J and σ. Again, the linear upper limit is illustrated as a straight line and the quadratic lower limit as a solid curve. The figure confirms that in the case where $J = 0$, the data points that correspond to $x_{0,i}$ being sampled from a log-normal distribution with a relatively low value for σ are either on or close to the quadratic lower limit curve. On the other hand, the case where $\sigma = 10$ results in data points where the overlap in the network for a given number of links is almost maximal, and therefore approaches the linear upper limit. This first set of results confirms the strong role of node heterogeneity in determining increased correlations between the degrees of the same node across different layers, which, in turn, increase the inter-layer overlap even without any explicit coupling ($J = 0$), and hence, in a 'spurious' manner. On the other hand, when we increase the value of J, the data points corresponding to relatively low values of σ (e.g., $\sigma = 10^{-5}$ and $\sigma = 10^{-3}$) stay on or close to the quadratic lower limit, a finding similar to the results in Section 5.1.1, showing that the symmetry-broken values realized by different pairs of nodes, when averaged across the network, restore the ensemble average because the node pairs are all independent and (almost) identically distributed. Remarkably this means that, in a certain sense, node homogeneity 'suppresses' the effects of the true inter-layer coupling ($J > 0$) on the realized overlap. For the intermediate value $\sigma = 1.0$, the data are distributed close to the quadratic lower limit curve only for low values of J, while increasing the value of J leads to a more linear trend, eventually approaching the linear upper limit. In this case, the coupling is effective in producing a higher realized overlap. In the case where $\sigma = 10$, the linear trend is instead achieved already for $J = 0.0$ (although the points are aligned below it); hence, increasing the value of J barely influences the value of the overlap for a given number of links.

Therefore the effect of increasing J in networks with a moderate heterogeneity is a transition from multiplex configurations with densities of all levels towards multiplex configurations with either low or high density, which is a result of the phase transition. It also shows that a very high level of heterogeneity leads to an overlap in the network that is already close to maximal for a given number of links, irrespective of the phase transition and the value of J. However, in the case where we have an intermediate level of heterogeneity ($\sigma = 1.0$), we observe that the effect of the coupling can be relatively strong, and we can therefore construct networks with a combination of the overlap and number of links falling between the extreme linear upper limit and the quadratic lower limit in a controlled, systematic manner. Note that Figure 8 also shows that, as J increases above 1, the (symmetry-broken) realized data start to 'drift away' from the intermediate densities, in a way similar to what we observed in Figure 5, but in a more pronounced manner. This is due to the fact that, as J increases, a larger number of multilinks shall be either in the low-density or high-density phase.

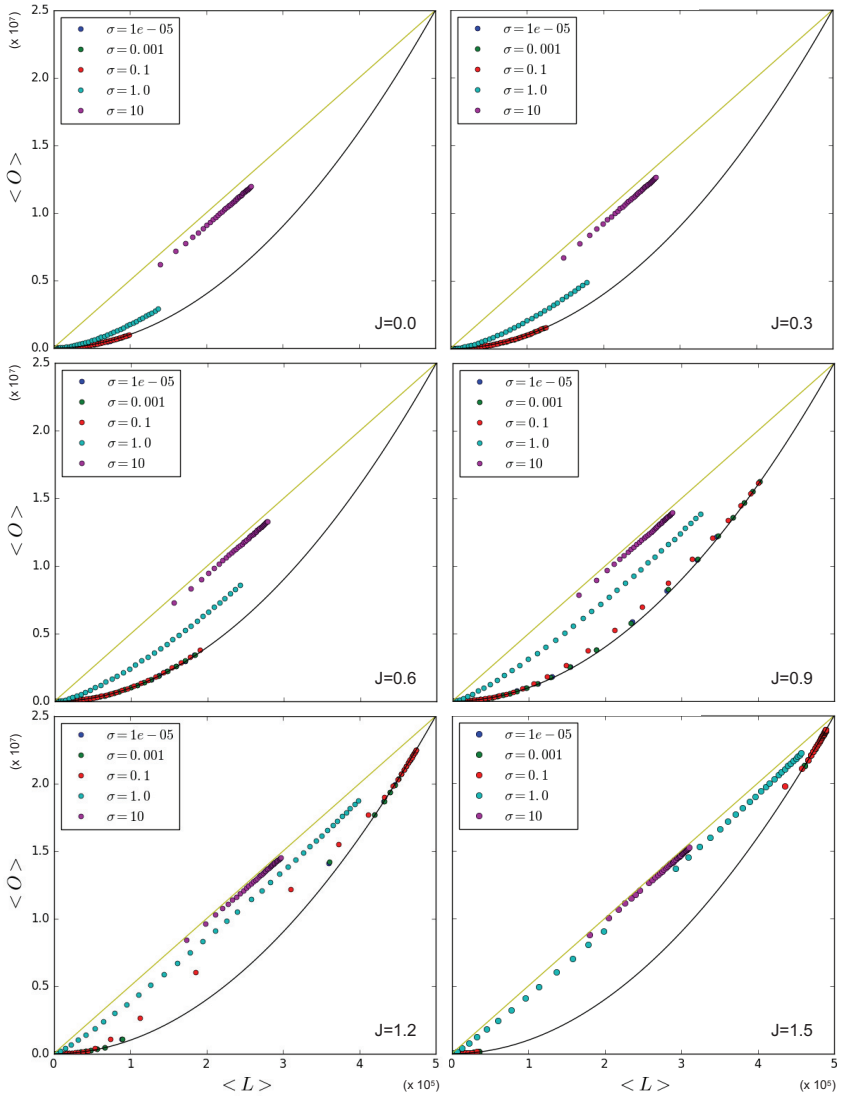


Figure 8. Relationship between the expected inter-layer overlap $\langle O \rangle$ and the total number of links $\langle L \rangle$ in heterogeneous multiplexes with $N = 100$, $M = 100$, and $x_{0,i}$ sampled from a log-normal distribution with different values for σ. The colored points correspond to simulations obtained via the Metropolis–Hastings algorithm for $J \in \{0.0, 0.3, 0.6, 0.9, 1.2, 1.5\}$ and $z \in [0.05, 2.00]$ in steps of $\Delta z = 0.05$. The straight line corresponds to the upper limit $\langle O \rangle = M\langle L \rangle / 2$ calculated in Equation (67). The solid curve corresponds to the quadratic trend $\langle O \rangle = \langle L \rangle^2 / N^2$ (achieved by homogeneous multiplexes with constant x_i), which here marks a lower bound achieved when $\sigma \to 0^+$. For increasing values of J (genuine coupling) and σ (spurious coupling), the system moves closer to the upper bound. For $J > 1$, we see that, starting from the multiplexes with smaller values of σ, the points are concentrating towards high-density and low-density (symmetry-broken) regimes, drifting away from the intermediate values, like in the homogeneous and power law cases. To realize this separation for larger values of σ, a larger value of J is required.

Again, in Figure 9 (which is the counterpart of Figure 7), we 'zoom in', and, using Equations (53), (55), and (56), we show the theoretically predicted values of $\langle O \rangle$ and $\langle L \rangle$ and compare them to the simulation data, where $x_{0,i}$ is sampled from a log-normal distribution with $\sigma = 1$, for $J = 0$ and $J = 1.5$. The results for $\sigma \in \{10^{-5}, 10^{-3}, 10^{-1}, 10^{1}\}$ are not shown here since relatively low and high values for σ lead to results similar to those we have shown in Sections 5.1.1 and 5.1.2, respectively. Figure 9 confirms that the theoretical predictions are in good agreement with the simulation data, apart from the expected 'drifting away' of symmetry-broken values from the corresponding ensemble average.

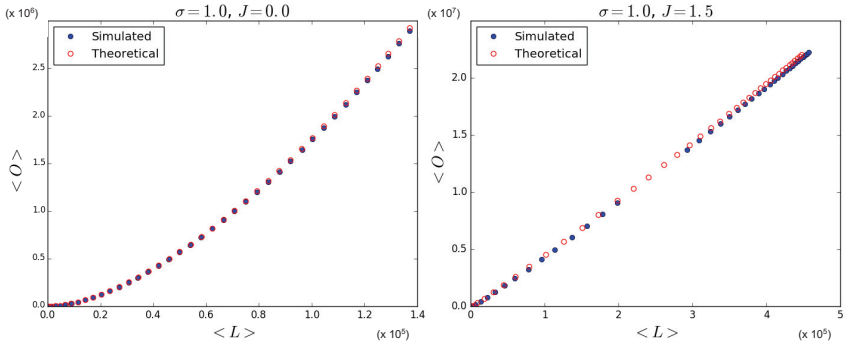


Figure 9. Relationship between the expected inter-layer overlap $\langle O \rangle$ and the total number of links $\langle L \rangle$ in heterogeneous multiplexes with $N = 100$, $M = 100$, and $x_{0,i}$ sampled from a log-normal distribution with $\sigma = 1$. The blue points correspond to simulations obtained via the Metropolis–Hastings algorithm for $z \in [0.05, 2.00]$ in steps of $\Delta z = 0.05$ with $J = 0$ (**left** panel) and $J = 1.5$ (**right** panel). The red open circles are the theoretically predicted values corresponding to the same parameters used in the simulations.

6. Analysis of the World Trade Multiplex

In this section, we finally consider an application of the model to a real-world economic network. Since our models lead to multiplex networks with independent pairs of nodes (i.e., independent multilinks) even when links are correlated across layers, it is important that the real-world network is consistent with this assumption. For instance, networks constructed from time series data [3–5] are not viable, because the known (and strong) correlations between the time series corresponding to different vertices generate dependencies between pairs of nodes (and higher-order patterns) through the triangular inequality [6,7]. For this reason, we select the World Trade Multiplex as an ideal case study for the present analysis, because each separate layer of that network has been successfully modeled in the past via maximum entropy models of networks with given degrees [31–33]. At the same time, it has been shown that certain structural properties of commodity-specific layers are very similar across the different layers of the multiplex [30], and that this similarity (in particular, the correlation among the degrees of the same node in different layers) generates a large spurious component of the inter-layer overlap [28,29], which is not necessarily due to a genuine coupling. In this sense, our analysis here will add a natural novel aspect to the modeling of the network, namely, the explicit comparison with a model with nontrivial coupling among layers, which has not been considered so far. We use the UN-COMTRADE dataset that represents the multiplex network of international trade (https://comtradeplus.un.org, accessed on 2 September 2019). The different layers of this multiplex network represent different commodities. The vertices in this network represent different countries, and a link exists between two countries in a given layer if there is trade between them in that commodity. The data include $N = 206$ countries and $M = 96$ commodities. Some examples of traded commodities are meat, fish, dairy products, coffee, and tobacco [30,33].

Using the international trade data, we wish to identify a possibly nontrivial overlap by creating (L, O) plots similar to the ones depicted in Figures 6 and 7 or 8. We therefore repeatedly filter the network such that each layer α has the same number of links $L^\alpha \equiv L^0$ (where $\alpha = 1, \ldots, M$), and calculate the corresponding overlap O for the specified value of L^0 (note that this means that the total number of links in the entire multiplex is $L = ML^0$). The criterion we follow is choosing the L^0 strongest (highest weight) links in every layer to obtain data with comparable degrees across layers, as in our models. Note that, by using this filtering method, the highest possible density we can achieve is limited by the density of the sparsest layer in the unfiltered network. The results are shown in Figure 10, which indicates that the overlap for a given number of links appears to be around halfway between the quadratic lower limit curve and the linear upper limit curve. This suggests that the degree of heterogeneity of the network is intermediate, similar to that realized by log-normally distributed fitness values, as in our example considered above.

As anticipated, we are currently unable to solve the maximum likelihood equations in order to obtain the joint values of all the Lagrange multipliers in the full OACM model with $J \neq 0$. However, after filtering the original empirical network such that every layer has L^0 links, we can use the values of the hidden variables x_i^* for the null model corresponding to the absence of inter-layer coupling, i.e., to $J^* = 0$. As we have shown in Equations (57), this assumption reduces our model to the ACM discussed in Section 2.5. The maximum likelihood equations in this case are much easier to solve, and can be found using one of the numerical algorithms available at https://meh.imtlucca.it (accessed on 1 May 2023). This procedure is repeated for a range of values for L^0. The cumulative distribution of the hidden variables x_i^* are plotted in Figure 10 for various values of L^0. The figure qualitatively shows that the shape of the cumulative distribution of x is fat-tailed and indeed similar to the one for a log-normal distribution. Moreover, it does not vary with L^0, apart from an overall change of scale.

The null model with $J^* = 0$, when compared to the data for the same choice of L_0, allows us to detect the presence of nontrivial coupling among the layers, when present. Indeed, from Figure 10, we see that the filtered networks have a relatively high overlap, the data points being distributed along a similar trend as the one corresponding to a nonzero J in our previous heterogeneous examples. By using the values of the hidden variables for the model with $J^* = 0$, we can calculate the corresponding expected number of links and the expected overlap under the null hypothesis of no coupling between the layers, but the same average degree sequence in the real network. The results are shown in Figure 10, alongside the curve corresponding to the empirical data. We see that the assumption $J = 0$ leads to an insufficiently overlapping multiplex, demonstrating the necessity of a model that introduces dependencies between the layers of a network. The difference between the two curves can be quantified by fitting both to the curve

$$O = AL^\alpha \tag{76}$$

where A is a proportionality factor and α is an exponent (not to be confused with the label of a layer of the multiplex). For the empirical data, we find a steeper increase characterized by an exponent $\alpha_{\text{empirical}} = 1.19$, while for the predictions from the ACM, we find $\alpha_{\text{CM}} = 1.06$ (see Figure 10). The difference between the two values implies that the difference between the realized and expected overlap increases as L increases, confirming that the observed overlap in the WTM is not only the spurious result of the correlated heterogeneity of the degrees of countries, but reflects genuine ($J^* > 0$) inter-layer dependencies.

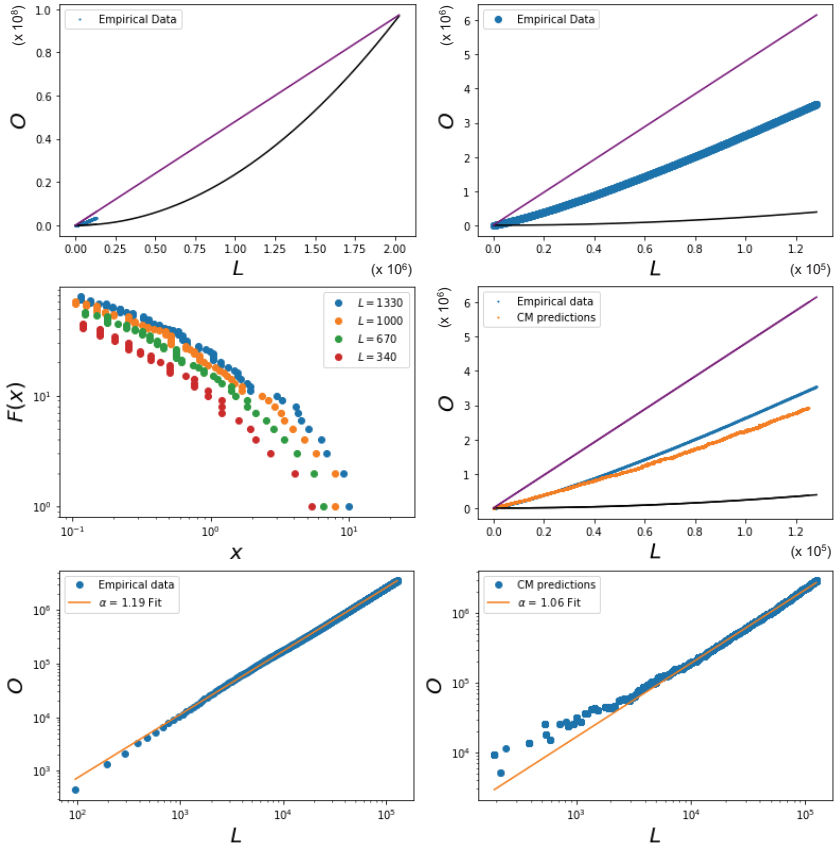


Figure 10. Comparison of the empirical World Trade Multiplex (WTM) with the zero-coupling ($J^* = 0$) benchmark provided by the Average Configuration Model (ACM). The WTM consists of $N = 206$ nodes, each representing a country, and $M = 96$ layers, each representing a commodity group. The filtered data were obtained by retaining the same number L^0 of strongest links in each layer (hence $L = ML^0$ links in the entire multiplex), and varying L^0. **Top left**: relationship between the expected inter-layer overlap $\langle O \rangle$ and the total number of links $\langle L \rangle$ in the WTM (blue), compared with the upper limit $\langle O \rangle = M\langle L \rangle/2$ calculated in Equation (67) (purple straight line) and the quadratic trend $\langle O \rangle = \langle L \rangle^2/N^2$ achieved by homogeneous multiplexes (black solid curve). **Top right**: zoomed-in version of the top left panel, showing that the empirical data follow an intermediate scaling between the two extremes. **Center left**: cumulative distributions reporting the number $F(x)$ of nodes with hidden variable larger than x in the ACM, obtained for different values of L^0 (see legend). **Center right**: same as the top right panel with the addition of the relationship produced by the ACM benchmark, showing that the empirical WTM (blue) has a higher overlap than the corresponding null model having zero inter-layer coupling but the same degree heterogeneity (orange). **Bottom left**: log–log plot of the relationship between the overlap and the number of links in the empirical WTM, along with a power law fit of the form $O = AL^\alpha$, where the fitted exponent is $\alpha = 1.19$. **Bottom right**: log–log plot of the same relationship in the ACM benchmark with no coupling, along with a power law fit of the form $O = AL^\alpha$, where the fitted exponent is $\alpha = 1.06$.

We conclude with a discussion about the entropy in the heterogeneous case, analogous to the one we made in Section 5.1.1 in the homogeneous case. Here we note that, given a

multiplex $\vec{G}^*$ of interest, the entropy $S(\vec{\theta}^*, J^*)$ of the data, given the OACM model, is the one given by Equation (58), which in the heterogeneous case cannot be, in general, reduced to a simpler formula. However, if we define the minimum and maximum values of the hidden variables as

$$\theta^*_{\min} \equiv \min_{i=1,N}\{\theta^*_i\}, \qquad \theta^*_{\max} \equiv \max_{i=1,N}\{\theta^*_i\}, \tag{77}$$

respectively, we can bound the entropy as follows:

$$S_{\min}(\vec{\theta}^*, J^*) \leq S(\vec{\theta}^*, J^*) \leq S_{\max}(\vec{\theta}^*, J^*) \tag{78}$$

where we have defined

$$S_{\min}(\vec{\theta}^*, J^*) \equiv 2\theta^*_{\min}L^* - \frac{4J^*}{M}O^* + \sum_{i<j}\ln z_{ij}(\theta^*_i + \theta^*_j, J^*), \tag{79}$$

$$S_{\max}(\vec{\theta}^*, J^*) \equiv 2\theta^*_{\max}L^* - \frac{4J^*}{M}O^* + \sum_{i<j}\ln z_{ij}(\theta^*_i + \theta^*_j, J^*). \tag{80}$$

The bounds in Equation (78) are alternative to the general ones in Equation (63), and arguably more useful to characterize how the entropy is effectively constrained by, once again, the relationship between L^* and O^*. The latter, unlike the homogeneous case, is not necessarily quadratic, and can follow the diverse trends we have shown in Figures 6, 8 and 10. In particular, the power law relationship captured by Equation (76) for the empirical WTM provides a convenient way of bounding $S(\vec{\theta}^*, J^*)$ via Equations (78)–(80).

7. Conclusions

In this paper we have introduced a maximum entropy model, or ERGM, of multiplex networks with given degrees and inter-layer overlap. The model allowed us to separately control the effects of the correlations between node degrees across different layers (which lead to a spurious overlap) and that of a genuine inter-layer coupling. The nature of the enforced constraints is such that different pairs of nodes are statistically independent, even if the parameters governing them are correlated via those of the nodes they share.

For each pair of nodes, the model can be mapped exactly to a mean-field Ising model featuring a magnetization-like phase transition, which includes the possibility of (spontaneous) symmetry breaking. Given the difficulty of solving the maximum likelihood equations to obtain the values of the Lagrange multipliers corresponding to a particular real network, we first treated the Lagrange multipliers as free parameters in order to explore and analyze the properties of multiplex systems as a function of these parameters using numerical methods. Additionally, the numerical results were compared to our analytical results in order to test the validity of the latter. We have shown that the analytical equations are highly accurate. The combined result, at the level of the entire multiplex, of the properties of all node pairs is nontrivial and crucially depends on the values of the node-specific parameters, which ultimately depend on the enforced degrees.

In the fully homogeneous case, the phase transition occurs at the same critical point for all node pairs simultaneously, because the parameters are identical for all nodes. However, the independence of different node pairs implies that, even in the magnetized phase, the realized values of the inter-layer overlap and total number of links coincide with the ensemble average. This happens because different node pairs realize all the possible symmetry-broken values independently, so that an average of the realized values for a large number of independent node pairs asymptotically equals the ensemble average. The value of J has little effect on the relationship connecting the overlap to the number of links, which remains similar to what we observed for the case $J = 0$, showing that node homogeneity suppresses the effects of a genuine inter-layer coupling.

In the heterogeneous case, the phenomenology is very different since, despite the fact that node pairs are still independent, they are now governed by different parameters, and the ensemble average for a given pair can no longer be realized as an average of the realized values of pairs with the same parameters. This implies that the observed overlap

and number of links will depend on the realized symmetry-broken values, whose typical value does not coincide in general with the ensemble average, and is determined by the node-specific parameters (hence, ultimately by the degrees). Moreover different pairs of nodes are, in general, found in different phases, so the multiplex displays, as a function of the parameters, a hierarchy of phase transitions. We have found that increasing the value of the coupling parameter J generally increases the (genuine) overlap for a given number of links, if there is enough node heterogeneity. However, we have also shown that increasing the heterogeneity of the network increases the (spurious) overlap for a given number of links as well. This is a consequence of the presence of large hubs that appear in a correlated manner across layers, due to the increased heterogeneity of the network. Additionally, every multilink that is connected to these hubs has a relatively low critical threshold for the coupling parameter J. Therefore, these multilinks have a higher probability to be in the high density phase, which leads to a higher overlap as well, which corresponds to increasing the amount of genuine correlation. In general, the overlap for a given number of links can be increased by increasing either the heterogeneity of the network or the value of the coupling parameter, with a subtle interplay between the two. In principle, this can be used in order to create multiplexes with a specific degree of overlap for a given of number of links, provided their combination is within the theoretical limits discussed in Section 5.

Finally, by using a dataset that represents the empirical multiplex network of international trade in several commodity-specific layers, we have used the model to disentangle the spurious overlap arising from the documented strong correlation of node degrees across layers [28,29] from the genuine overlap arising from actual inter-layer coupling. We have found that the assumption that there is no coupling between the layers ($J = 0$), which reduces our model to the ACM, results in a multiplex with insufficient inter-layer overlap. This means that the empirical overlap is not merely the spurious result of the correlated heterogeneity of the network, but requires a true nonzero coupling between layers.

Our results demonstrate the subtleties of the interplay between node heterogeneity and inter-layer dependencies in multiplex networks, highlighting the need for null models that can control these factors separately. In this paper, we have introduced perhaps the simplest, although already very rich, model of this type. Our model can be seen as a minimal one, to be further generalized in the future.

Author Contributions: Conceptualization, D.G. and V.G.; methodology, D.G., V.G. and N.B.; software, N.B.; data curation, V.G. and N.B.; writing—original draft preparation, N.B.; writing—review and editing, V.G. and D.G.; visualization, N.B.; supervision, D.G.; project administration, V.G.; funding acquisition, D.G. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by Stichting Econophysics, Leiden, The Netherlands.

Data Availability Statement: For the empirical analysis of the World Trade Multiplex, the publicly available UN-COMTRADE dataset was analyzed in this study. The data are available at http://comtrade.un.org/ (accessed on 2 September 2019). The codes used for the numerical calculation of the parameters maximizing the likelihood are available at https://meh.imtlucca.it (accessed on 1 May 2023).

Acknowledgments: This work is supported by the European Union—Horizon 2020 Program under the scheme 'INFRAIA-01-2018-2019—Integrating Activities for Advanced Communities', Grant Agreement n.871042, 'SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics' (http://www.sobigdata.eu, accessed on 1 May 2023). This work is also supported by the European Union-NextGenerationEU-National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR), project 'SoBigData.it-Strengthening the Italian RI for Social Mining and Big Data Analytics'-Grant IR0000013 (3264, 28/12/2021). We also acknowledge support from the project NetRes—'Network analysis of economic and financial resilience', Italian DM n. 289, 25-03-2021 (PRO3 Scuole), CUP D67G22000130001 (https://netres.imtlucca.it, accessed on 1 May 2023).

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

Appendix A. Hubbard–Stratonovich Transform

The pair Hamiltonian of our OACM in Equation (47) can be rewritten as

$$h_{ij}(s_{ij}, B_{ij}, J) = -B_{ij} \sum_{\alpha=1}^{M} \sigma_{ij}^{\alpha} - \frac{J}{2M} \left(\sum_{\alpha=1}^{M} \sigma_{ij}^{\alpha} \right)^2 + \frac{J}{2} + v_{ij}. \tag{A1}$$

We want to obtain an expression for the pair partition function:

$$\begin{aligned} z_{ij}(B_{ij}, J) &= \sum_{s_{ij} \in \mathcal{S}_{ij}} e^{-h_{ij}(s_{ij}, B_{ij}, J)} \\ &= \sum_{s_{ij} \in \mathcal{S}_{ij}} \exp \left[\frac{J}{2M} \left(\sum_{\alpha=1}^{M} \sigma_{ij}^{\alpha} \right)^2 + B_{ij} \sum_{\alpha=1}^{M} \sigma_{ij}^{\alpha} - \frac{J}{2} - v_{ij} \right] \\ &= e^{-J/2} e^{-v_{ij}} \sum_{s_{ij} \in \mathcal{S}_{ij}} \exp \left[\left(\sqrt{\frac{J}{2M}} \sum_{\alpha=1}^{M} \sigma_{ij}^{\alpha} \right)^2 + B_{ij} \sum_{\alpha=1}^{M} \sigma_{ij}^{\alpha} \right]. \end{aligned} \tag{A2}$$

The argument of the exponent in the above expression can be linearized by using the Gaussian integral

$$e^{a^2} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} d\xi_{ij} e^{-\xi_{ij}^2/2 + \sqrt{2} a \xi_{ij}}. \tag{A3}$$

In our case, by choosing $a = \sqrt{J/(2M)} \sum_{\alpha=1}^{M} \sigma_{ij}^{\alpha}$ the partition function factorizes with respect to the individual summations of σ_{ij}^{α}:

$$\begin{aligned} z_{ij}(B_{ij}, J) &= \frac{e^{-J/2 - v_{ij}}}{\sqrt{2\pi}} \sum_{s_{ij} \in \mathcal{S}_{ij}} \int_{-\infty}^{\infty} d\xi_{ij} e^{-\xi_{ij}^2/2} \exp \left[\sum_{\alpha=1}^{M} \sigma_{ij}^{\alpha} \left(\sqrt{\frac{J}{M}} \xi_{ij} + B_{ij} \right) \right] \\ &= \frac{e^{-J/2 - v_{ij}}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} d\xi_{ij} e^{-\xi_{ij}^2/2} \sum_{\sigma_{ij}^1 \in \{-1,1\}} \cdots \sum_{\sigma_{ij}^M \in \{-1,1\}} \prod_{\alpha=1}^{M} \exp \left[\sigma_{ij}^{\alpha} \left(\sqrt{\frac{J}{M}} \xi_{ij} + B_{ij} \right) \right] \\ &= \frac{e^{-J/2 - v_{ij}} \, 2^M}{\sqrt{2\pi}} \int_{-\infty}^{\infty} d\xi_{ij} e^{-\xi_{ij}^2/2} \left[\cosh \left(\sqrt{\frac{J}{M}} \xi_{ij} + B_{ij} \right) \right]^M. \end{aligned} \tag{A4}$$

Performing the change of variable $\sqrt{J/M} \xi_{ij} = J y_{ij}$ we obtain

$$z_{ij}(B_{ij}, J) = 2^M \sqrt{\frac{JM}{2\pi}} e^{-J/2} e^{-v_{ij}} \int_{-\infty}^{\infty} d\xi_{ij} \left[\Phi_{J, B_{ij}}(y_{ij}) \right]^M \tag{A5}$$

where

$$\Phi_{J, B_{ij}} \equiv e^{-J y_{ij}^2/2} \cosh \left(J y_{ij} + B_{ij} \right). \tag{A6}$$

We are interested in the large M limit. To proceed in the calculation of $z_{ij}(B_{ij}, J)$, it is useful to define the quantity

$$f_{ij}(B_{ij}, J) \equiv - \lim_{M \to \infty} \frac{1}{M} \ln z_{ij}(B_{ij}, J) = - \lim_{M \to \infty} \ln z_{ij}^{1/M}(B_{ij}, J) \tag{A7}$$

which is the *free energy per layer*. By inserting the result (A5) into (A7), we obtain

$$\begin{aligned} f_{ij}(B_{ij}, J) &= -\ln 2 - \lim_{M \to \infty} \frac{1}{M} \left[\ln \left(e^{-J/2} \sqrt{\frac{JM}{2\pi}} \right) - v_{ij} + \ln \left(\int_{-\infty}^{\infty} dy_{ij} \left[\Phi_{J, B_{ij}}(y) \right]^M \right) \right] \\ &= -\ln 2 + \frac{J}{2} - B_{ij} - \ln \left[\lim_{M \to \infty} \left(\int_{-\infty}^{\infty} dy_{ij} \left[\Phi_{J, B_{ij}}(y) \right]^M \right)^{1/M} \right]. \end{aligned} \tag{A8}$$

In order to obtain a more explicit form of the function $f_{ij(B_{ij},J)}$, we use the Laplace theorem [49]. Let $\phi(y)$ and $\psi(y)$ be continuous and positive functions within a range $c \leq y \leq d$, then

$$\lim_{M\to\infty}\left[\int_c^d \psi(y)(\phi(y))^M\right]^{1/M} = \max_{c\leq y\leq d}\phi(y). \tag{A9}$$

For $\psi(y) = 1$ and $\phi(y) = \Phi_{J,B_{ij}}(y)$, this results in

$$f_{ij}(B_{ij},J) = -\ln 2 + \frac{J}{2} - B_{ij} - \ln\left[\max_{-\infty\leq y_{ij}\leq\infty}\Phi_{J,B_{ij}}(y_{ij})\right]. \tag{A10}$$

The derivative of $\Phi_{J,B_{ij}}(y_{ij})$ with respect to y_{ij} is zero at its maximum:

$$\frac{d\Phi_{J,B_{ij}}(y_{ij})}{dy_{ij}} = Je^{-Jy_{ij}^2/2}\sinh\left(Jy_{ij}+B_{ij}\right) - Jy_{ij}e^{-Jy_{ij}^2/2}\cosh\left(Jy_{ij}+B_{ij}\right) = 0. \tag{A11}$$

The variable y_{ij} therefore obeys the equation

$$y_{ij} = \tanh\left(Jy_{ij}+B_{ij}\right). \tag{A12}$$

Note that this equation is identical to the one obtained for the magnetization in the Ising Model, and, depending on the values of J and B_{ij}, there are either one or three solutions that satisfy Equation (A12). The free energy f_{ij} can now be written as a function of J and B_{ij}:

$$f_{ij}(B_{ij},J) = -\ln 2 + \frac{J}{2} - B_{ij} + \frac{J}{2}(y_{ij})^2 - \ln\left[\cosh\left(Jy_{ij}+B_{ij}\right)\right]. \tag{A13}$$

We then finally arrive at the pair partition function

$$\begin{aligned} z_{ij}(B_{ij},J) &= e^{-Mf_{ij}} \\ &= 2^M e^{-v_{ij}} e^{-JM(y_{ij})^2/2}\cosh^M\left(Jy_{ij}+B_{ij}\right) \end{aligned} \tag{A14}$$

which, returning to the variables θ_{ij}, coincides with Equation (52) in the main text, where u_{ij} is the solution to Equation (53).

Appendix B. Maximum Likelihood

To determine the parameters $(\vec{\theta}^*, J^*)$ that maximize the log-likelihood of the OACM given in Equation (54), we first calculate the derivatives

$$-\frac{\partial\mathcal{L}(\vec{\theta},J)}{\partial\theta_k} = \sum_{i<j}\frac{\partial h_{ij}(m_{ij}^*,\theta_{ij},J)}{\partial\theta_k} + \sum_{i<j}\frac{\partial\ln z_{ij}(\theta_{ij},J)}{\partial\theta_k}, \quad k=1,\ldots,N \tag{A15}$$

$$-\frac{\partial\mathcal{L}(\vec{\theta},J)}{\partial J} = \sum_{i<j}\frac{\partial h_{ij}(m_{ij}^*,\theta_{ij},J)}{\partial J} + \sum_{i<j}\frac{\partial\ln z_{ij}(\theta_{ij},J)}{\partial J}. \tag{A16}$$

We then set the derivatives with respect to θ_k to zero:

$$-\left.\frac{\partial\mathcal{L}(\vec{\theta},J)}{\partial\theta_k}\right|_{\vec{\theta}^*,J^*} = \sum_{\alpha=1}^{M}\sum_{j\neq k} g_{jk}^{*\,\alpha} - M\sum_{j\neq k}u_{jk}^* = 0 \tag{A17}$$

where we utilize the fact that g_{ij}^{α} and u_{ij} are symmetric with respect to the indices i, j, i.e.,

$$\sum_{i<j}g_{ij}^{\alpha}\delta_i^k = \sum_{j=k+1}^{N} g_{jk}^{\alpha}, \qquad \sum_{i<j}g_{ij}^{\alpha}\delta_j^k = \sum_{j=1}^{k-1} g_{jk}^{\alpha}. \tag{A18}$$

Similarly, we set the derivative with respect to J to zero:

$$-\left.\frac{\partial \mathcal{L}(\vec{\theta}, J)}{\partial J}\right|_{\vec{\theta}^*, J^*} = \sum_{i<j}\left(-\frac{4}{M}\sum_{\alpha<\beta} g_{ij}^{*\alpha} g_{ij}^{*\beta} + 2M\left(u_{ij}^*\right)^2\right) = 0. \tag{A19}$$

Taken together, the above calculations lead to the maximum likelihood Equations (55) and (56) in the main text.

References

1. Krackhardt, D. Cognitive social structures. *Soc. Netw.* **1987**, *9*, 109–134. [CrossRef]
2. Padgett, J.F.; Ansell, C.K. Robust Action and the Rise of the Medici, 1400–1434. *Am. J. Sociol.* **1993**, *98*, 1259–1319. [CrossRef]
3. Bardoscia, M.; Barucca, P.; Battiston, S.; Caccioli, F.; Cimini, G.; Garlaschelli, D.; Saracco, F.; Squartini, T.; Caldarelli, G. The physics of financial networks. *Nat. Rev. Phys.* **2021**, *3*, 490–507. [CrossRef]
4. Lacasa, L.; Luque, B.; Ballesteros, F.; Luque, J.; Nuno, J.C. From time series to complex networks: The visibility graph. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 4972–4975. [CrossRef] [PubMed]
5. Tsiotas, D.; Magafas, L.; Argyrakis, P. An electrostatics method for converting a time-series into a weighted complex network. *Sci. Rep.* **2021**, *11*, 11785. [CrossRef]
6. MacMahon, M.; Garlaschelli, D. Community Detection for Correlation Matrices. *Phys. Rev. X* **2015**, *5*, 021006. [CrossRef]
7. Anagnostou, I.; Squartini, T.; Kandhai, D.; Garlaschelli, D. Uncovering the mesoscale structure of the credit default swap market to improve portfolio risk modelling. *Quant. Financ.* **2021**, *21*, 1501–1518. [CrossRef]
8. De Domenico, M.; Solé-Ribalta, A.; Cozzo, E.; Kivelä, M.; Moreno, Y.; Porter, M.A.; Gómez, S.; Arenas, A. Mathematical formulation of multilayer networks. *Phys. Rev. X* **2013**, *3*, 041022. [CrossRef]
9. Battiston, F.; Nicosia, V.; Latora, V. Structural measures for multiplex networks. *Phys. Rev. E* **2014**, *89*, 032804. [CrossRef]
10. Kivelä, M.; Arenas, A.; Barthelemy, M.; Gleeson, J.P.; Moreno, Y.; Porter, M.A. Multilayer networks. *J. Complex Netw.* **2014**, *2*, 203–271. [CrossRef]
11. Battiston, F.; Nicosia, V.; Latora, V. The new challenges of multiplex networks: Measures and models. *Eur. Phys. J. Spec. Top.* **2017**, *226*, 401–416. [CrossRef]
12. Boccaletti, S.; Bianconi, G.; Criado, R.; Del Genio, C.I.; Gómez-Gardenes, J.; Romance, M.; Sendina-Nadal, I.; Wang, Z.; Zanin, M. The structure and dynamics of multilayer networks. *Phys. Rep.* **2014**, *544*, 1–122. [CrossRef] [PubMed]
13. Verbrugge, L.M. Multiplexity in adult friendships. *Soc. Forces* **1979**, *57*, 1286–1309. [CrossRef]
14. Erdos, P.; Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* **1960**, *5*, 17–60.
15. Albert, R.; Barabási, A.L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **2002**, *74*, 47. [CrossRef]
16. Watts, D.J.; Strogatz, S.H. Collective dynamics of'small-world'networks. *Nature* **1998**, *393*, 440. [CrossRef]
17. Squartini, T.; Garlaschelli, D. Analytical maximum-likelihood method to detect patterns in real networks. *New J. Phys.* **2011**, *13*, 083001. [CrossRef]
18. Squartini, T.; Mastrandrea, R.; Garlaschelli, D. Unbiased sampling of network ensembles. *New J. Phys.* **2015**, *17*, 023052. [CrossRef]
19. Squartini, T.; Garlaschelli, D. *Maximum-Entropy Networks: Pattern Detection, Network Reconstruction and Graph Combinatorics*; Springer: Berlin/Heidelberg, Germany, 2017.
20. Cimini, G.; Squartini, T.; Saracco, F.; Garlaschelli, D.; Gabrielli, A.; Caldarelli, G. The statistical physics of real-world networks. *Nat. Rev. Phys.* **2019**, *1*, 58–71. [CrossRef]
21. Holland, P.W.; Leinhardt, S. An exponential family of probability distributions for directed graphs. *J. Am. Stat. Assoc.* **1981**, *76*, 33–50. [CrossRef]
22. Besag, J. Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B (Methodol.)* **1974**, *36*, 192–236. [CrossRef]
23. Frank, O.; Strauss, D. Markov graphs. *J. Am. Stat. Assoc.* **1986**, *81*, 832–842. [CrossRef]
24. Contractor, N.S.; Wasserman, S.; Faust, K. Testing multitheoretical, multilevel hypotheses about organizational networks: An analytic framework and empirical example. *Acad. Manag. Rev.* **2006**, *31*, 681–703. [CrossRef]
25. Wasserman, S.; Faust, K. *Social Network Analysis: Methods and Applications*; Cambridge University Press: Cambridge, UK, 1994; Voluem 8.
26. Carrington, P.J.; Scott, J.; Wasserman, S. *Models and Methods in Social Network Analysis*; Cambridge University Press: Cambridge, UK, 2005; Volume 28.
27. Park, J.; Newman, M.E. Statistical mechanics of networks. *Phys. Rev. E* **2004**, *70*, 066117. [CrossRef]
28. Gemmetto, V.; Garlaschelli, D. Multiplexity versus correlation: The role of local constraints in real multiplexes. *Sci. Rep.* **2015**, *5*, 9120. [CrossRef]
29. Gemmetto, V.; Squartini, T.; Picciolo, F.; Ruzzenenti, F.; Garlaschelli, D. Multiplexity and multireciprocity in directed multiplexes. *Phys. Rev. E* **2016**, *94*, 042316. [CrossRef]

30. Barigozzi, M.; Fagiolo, G.; Garlaschelli, D. Multinetwork of international trade: A commodity-specific analysis. *Phys. Rev. E* **2010**, *81*, 046104. [CrossRef]
31. Squartini, T.; Fagiolo, G.; Garlaschelli, D. Randomizing world trade. I. A binary network analysis. *Phys. Rev. E* **2011**, *84*, 046117. [CrossRef]
32. Fagiolo, G.; Squartini, T.; Garlaschelli, D. Null models of economic networks: The case of the world trade web. *J. Econ. Interact. Coord.* **2013**, *8*, 75–107. [CrossRef]
33. Mastrandrea, R.; Squartini, T.; Fagiolo, G.; Garlaschelli, D. Reconstructing the world trade multiplex: The role of intensive and extensive biases. *Phys. Rev. E* **2014**, *90*, 062804. [CrossRef]
34. Szell, M.; Lambiotte, R.; Thurner, S. Multirelational organization of large-scale social networks in an online world. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 13636–13641. [CrossRef] [PubMed]
35. Cardillo, A.; Gómez-Gardenes, J.; Zanin, M.; Romance, M.; Papo, D.; Del Pozo, F.; Boccaletti, S. Emergence of network features from multiplexity. *Sci. Rep.* **2013**, *3*, 1344. [CrossRef] [PubMed]
36. Menichetti, G.; Remondini, D.; Panzarasa, P.; Mondragón, R.J.; Bianconi, G. Weighted multiplex networks. *PLoS ONE* **2014**, *9*, e97857. [CrossRef] [PubMed]
37. Freeman, L.C. A set of measures of centrality based on betweenness. *Sociometry* **1977**, *40*, 35–41. [CrossRef]
38. Clauset, A.; Shalizi, C.R.; Newman, M.E. Power-law distributions in empirical data. *SIAM Rev.* **2009**, *51*, 661–703. [CrossRef]
39. Barabási, A.L.; Albert, R. Emergence of scaling in random networks. *Science* **1999**, *286*, 509–512. [CrossRef]
40. Berlingerio, M.; Coscia, M.; Giannotti, F.; Monreale, A.; Pedreschi, D. Foundations of multidimensional network analysis. In Proceedings of the Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference, Kaohsiung, Taiwan, 25–27 July 2011; pp. 485–489.
41. Bianconi, G. Statistical mechanics of multiplex networks: Entropy and overlap. *Phys. Rev. E* **2013**, *87*, 062806. [CrossRef]
42. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620. [CrossRef]
43. Jaynes, E.T. On the rationale of maximum-entropy methods. *Proc. IEEE* **1982**, *70*, 939–952. [CrossRef]
44. Garlaschelli, D.; Loffredo, M.I. Multispecies grand-canonical models for networks with reciprocity. *Phys. Rev. E* **2006**, *73*, 015101. [CrossRef]
45. Newman, M.E.; Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **2004**, *69*, 026113. [CrossRef] [PubMed]
46. Anand, K.; Bianconi, G. Entropy measures for networks: Toward an information theory of complex topologies. *Phys. Rev. E* **2009**, *80*, 045102. [CrossRef] [PubMed]
47. Garlaschelli, D.; Loffredo, M.I. Generalized bose-fermi statistics and structural correlations in weighted networks. *Phys. Rev. Lett.* **2009**, *102*, 038701. [CrossRef]
48. Coolen, T.; Annibale, A.; Roberts, E. *Generating Random Networks and Graphs*; Oxford University Press: Oxford, UK, 2017.
49. Pólya, G.; Szegő, G. *Problems and Theorems in Analysis: Series, Integral Calculus, Theory of Functions*; Aeppli, D., Translator; Springer: Berlin/Heidelberg, Germany, 1972.
50. Park, J.; Newman, M.E. Solution of the two-star model of a network. *Phys. Rev. E* **2004**, *70*, 066146. [CrossRef]
51. Hastings, W.K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57*, 97–109. [CrossRef]

entropy

MDPI

Article

Laplacian Spectra of Persistent Structures in Taiwan, Singapore, and US Stock Markets

Peter Tsung-Wen Yen [1], Kelin Xia [2] and Siew Ann Cheong [2,*]

[1] Center for Crystal Researches, National Sun Yat-sen University, 70 Lienhai Rd., Kaohsiung 80424, Taiwan
[2] School of Physical and Mathematical Sciences, Nanyang Technological University, 21 Nanyang Link, Singapore 637371, Singapore
* Correspondence: cheongsa@ntu.edu.sg

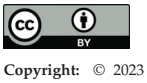
Abstract: An important challenge in the study of complex systems is to identify appropriate effective variables at different times. In this paper, we explain why structures that are persistent with respect to changes in length and time scales are proper effective variables, and illustrate how persistent structures can be identified from the spectra and Fiedler vector of the graph Laplacian at different stages of the topological data analysis (TDA) filtration process for twelve toy models. We then investigated four market crashes, three of which were related to the COVID-19 pandemic. In all four crashes, a persistent gap opens up in the Laplacian spectra when we go from a normal phase to a crash phase. In the crash phase, the persistent structure associated with the gap remains distinguishable up to a characteristic length scale ϵ^* where the first non-zero Laplacian eigenvalue changes most rapidly. Before ϵ^*, the distribution of components in the Fiedler vector is predominantly bi-modal, and this distribution becomes uni-modal after ϵ^*. Our findings hint at the possibility of understanding market crashs in terms of both continuous and discontinuous changes. Beyond the graph Laplacian, we can also employ Hodge Laplacians of higher order for future research.

Keywords: graph laplacian; stock market; complex systems; persistent structure; Fiedler vector

Citation: Yen, P.T.-W.; Xia, K.; Cheong, S.A. Laplacian Spectra of Persistent Structures in Taiwan, Singapore, and US Stock Markets. *Entropy* **2023**, *25*, 846. https://doi.org/10.3390/e25060846

Academic Editors: José F. F. Mendes and Panos Argyrakis

Received: 3 March 2023
Revised: 29 April 2023
Accepted: 17 May 2023
Published: 25 May 2023



1. Introduction

Unlike simple systems, where we can easily identify the few relevant variables and deduce the mathematical equations that they must obey (conservation laws, equations of state, equations of motion), or for thermodynamic systems, where we identify extensive and intensive variables that are statistical sums and averages of the microscopic variables, for complex systems it is difficult to identify a set of simplified (coarse-grained) variables [1,2]. This is especially challenging, since we know that self-organization and emergence is a hallmark of complex systems, implying that the effective variables might change from time to time [3]. One of the directions explored by complex systems scientists is to embed the N variables onto a low-dimensional manifold, using information contained in their time series $X_{i=1,\ldots,N}(t)$ [4,5]. Recently, D'Addese et al. [6] and Villani et al. [7] used information-theoretic methods to identify the *relevant sets of variables* in random Boolean networks, gene-regulatory networks, MAPK signaling pathways in eukaryotes, and other systems, and the manifold they evolve on. Others have turned instead to topological data analysis (TDA) and persistent homology to achieve the same goal [8,9]. Still others have combined information-theoretic methods and simplicial complexes arising from TDA to identify effective variables, and their interactions in the form of higher-order networks [10].

To be useful for describing a complex system, effective variables must change slowly with time, so that we do not need to switch between different sets of effective variables frequently. Of the $N \gg 1$ microscopic variables, we find some combinations that change dramatically over short time scales, as well as other combinations that evolve slowly. We call the former *fast variables*, and the latter *slow variables* [11,12]. Frequently, the slow variables do not evolve independently, but form groups that co-evolve. These are then

persistent structures that are consistent with self-organization (in that their equations of motion are not built into the microscopic dynamics) and emergence (the groups themselves can vary over long times) in the complex system. The first step towards understanding how we should write down the effective variables would be to identify the persistent structures. We attempted to do this in our two previous papers on TDA and Ricci curvature analysis (RCA). In our first paper [8], we applied TDA to identify persistent structures in financial correlation networks during market crashes. This attempt is an extension of our exploration into financial market dynamics using more traditional econophysics methods such as the minimal spanning tree (MST) [13–19], and the planar maximally filtered graph (PMFG) [20–24]. We were attracted to TDA because it can give us more information than graph filtering methods, as illustrated by how the Betti numbers change in toy models where two shells merge through the formation of a bottleneck, or when a shell changes into a torus through intermediate spindle torus and horn torus stages. However, the computation of persistent Betti numbers is tedious and time-consuming, and generally not feasible at large length scales. At smaller length scales, the number of persistent structures is large, making it impossible to identify all of them automatically.

More importantly, in TDA two persistent structures are assumed to have become one, the moment they become connected by a neck. As illustrated in Figure 1, we believe that persistent structures remain distinguishable beyond this first connection, so long as we can tell them apart from the neck region connecting them. Therefore, in our second paper [9], we introduced tools from Ricci curvature to help identify persistent structures with positive Ricci curvatures nearly everywhere, and neck regions with negative Ricci curvatures. By following the evolution of a particular neck over a market crash, we visualized how it was formed (down to the exact component stocks) and destroyed. Nevertheless, challenges remain. First, RCA is not easy to implement and automate. Second, small curvature changes are hard to detect because they involve collective movements of many nodes. To this end, new perspectives and approaches are necessary for the elucidation of the overall dynamical picture.

Drawing upon our experience in studying undergraduate physics, we can solve problems more easily by changing our approach or rephrasing our questions from a different perspective. In solid state physics, we find concepts such as the Brillouin zone, band structure, Fermi level, and band gap emerging naturally when we choose to work in momentum space. Additionally, owing to the band theory of solids so obtained we can predict such emergent phases as conductors, semi-conductors, half-metals, and insulators. In our two TDA papers, we investigated financial correlations in real space by examining simplicial complexes obtained through the filtration process. Here, we make a first attempt at characterizing such correlations in "momentum space". Before we dive into the spectral analysis of financial correlations, we first explain what persistent structures are and how to think of their continuous and discontinuous changes in Section 2, by using a raindrop analogy. Thereafter, in Section 3, we briefly review the filtration procedure in TDA, before arguing for the theoretical connection between symmetries and block-diagonal matrices. In particular, in solid state physics, the symmetries are in real space while the block-diagonal matrices appear in momentum space, whereas for networks or simplicial complexes, diagonal blocks associated with community structure appears in real space, and thus we expect the symmetries to be in momentum space. Communities in networks or simplicial complexes are normally discovered from adjacency matrices A_{ij}, but they can also be discovered from the graph Laplacians L_{ij}, which has interpretations closer to the Hamiltonian matrix H_{ij} in quantum mechanics, and their spectral properties are better understood. In the remainder of Section 3, we illustrate using various toy models of community structures that the existence of persistent clusters separated in space show up as a persistent gap in the spectra of L_{ij}. From the Fiedler eigenvector, associated with the first non-zero eigenvalue λ_1 of L_{ij}, we can identify the neck, in addition to the persistent clusters. We also realize from these studies that the persistent clusters remain distinct even after they become linked, up till the point where λ_1 changes most rapidly with change in

length scale. In Section 4, we apply these insights to analyze the correlations in real-world stock markets, by sliding six-month time windows across four market crashes on three stock exchanges, to see how the topology and geometry of such correlations change with time. We found the existence of two distinct phases in stock markets. In the normal phase, the spectrum of Laplacian eigenvalues has no gaps (consistent with the market being a single giant cluster), whereas in the crash phase, we find a gap emerging at large length scales (consistent with the market breaking into two or more clusters). Finally, we conclude in Section 5.

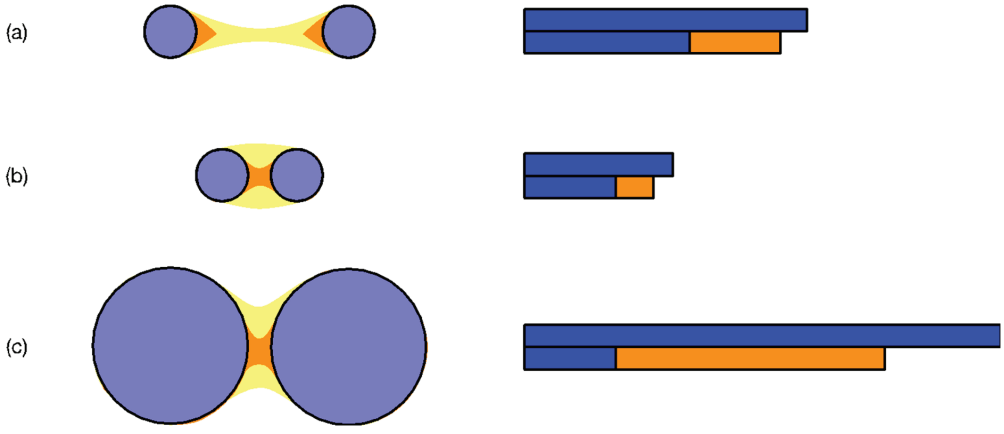


Figure 1. Three pairs of clusters at three increasing filtration parameters ϵ_1 (no necks, communities shown in blue), ϵ_2 (necks shown in orange), and ϵ_3 (necks shown in yellow). For each pair of clusters, we also show the standard TDA barcode (blue bars, from $\epsilon = 0$ to the value of ϵ when the clusters become connected), and an extended barcode shown in orange where the original clusters remain distinguishable. In (**a**), the two small clusters remain distinguishable over a large range of filtration parameters. This is to be contrasted against (**b**), where the two clusters are the same sizes as those in (**a**), but are closer to each other. They are therefore distinguishable only over a small range of filtration parameters. Finally, in (**c**), we have two large clusters whose separation is the same as that in (**b**). However, because of their sizes, the two large clusters are distinguishable over a much larger range of filtration parameters.

2. Intuition on Persistent Structures

Before we formally define persistent structures in Section 3, let us first develop an intuition on these based on a familiar physical phenomenon. In an atmosphere saturated with water vapor, water droplets can nucleate around impurities. When a water droplet first forms, it is small and light, and can be suspended by warm air rising from the earth's surface. The water droplet can then lose mass through evaporation, or it can absorb more water vapor from the atmosphere to become larger and heavier. Eventually, it becomes too heavy to be suspended by the rising warm air and begins to fall toward the earth's surface as a raindrop. As the raindrop falls, it rubs against the air and deforms into the characteristic teardrop shape (Figure 2a). Even though the raindrop now consists of a large number of water molecules (Figure 2b), it continues to lose water molecules through evaporation (Figure 2c), or gain water molecules through absorption (Figure 2d). More importantly, as the raindrop gains speed falling through air, its surface becomes unstable. The trailing end of the raindrop may then breakup into smaller droplets (Figure 2e).

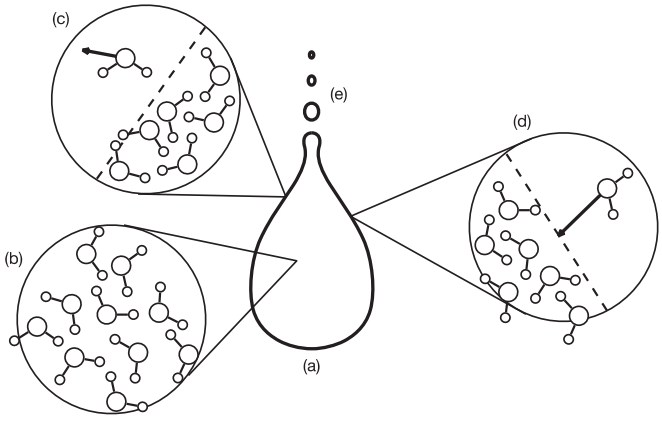


Figure 2. (**a**) A macroscopic raindrop with its characteristic teardrop shape falling through air. (**b**) The raindrop consists of a large number of microscopic water molecules whose relative positions are always changing. (**c**) Every now and then, a water molecule will escape from the surface of the raindrop (shown as dashed line). (**d**) Sometimes, the raindrop (whose surface is shown as a dashed line) can also absorb a water molecule from the air around it. (**e**) If the raindrop falls too fast, its surface will become unstable, and the trailing end of the raindrop may breakup into smaller droplets.

Instead of microscopic water molecules, we prefer to describe the phenomenon in terms of raindrops. This is because many raindrops retain their identities as they descend to the earth's surface. Indeed, if we perform instantaneous hierarchical clustering on the collection of water molecules coming down as rain, each raindrop is a robust cluster at a convenient length scale. However, unlike robust clusters with constant compositions, the compositions of raindrops change across length scale and time. It is thus better to think of a raindrop as a persistent homological structure, from the TDA point of view. Persistent homological structures need not have fixed compositions with respect to changes in length scale and time. They just need to have the same set of defining topological characteristics. For example, when a "sphere" comprising 20 particles grows over time to become one having 1000 particles, we can continue to refer to the structure as a "sphere" ($\beta_0 = 1$), provided it has no holes ($\beta_1 = 0$) and no voids ($\beta_2 = 0$).

Indeed, in this analogy, the raindrop 10 km above ground has a composition different from the raindrop that reaches the ground. Nevertheless, we think of the two as the same raindrop at different times, because it can be tracked continuously from an altitude of 10 km down to the ground. On the other hand, if an old raindrop completely evaporates at a height h_1, and thereafter a new raindrop suddenly forms at height $h_2 < h_1$, we do not consider the new raindrop to be the same persistent structure as the old raindrop. Therefore, a change in composition is admissible for a persistent structure, provided this change is always slow.

Treating the raindrop as a persistent structure and ignoring its compositional changes, we can then describe the time evolution of the raindrop in terms of the position $\vec{R}(t)$ of its centre of mass, and its volume $V(t)$. The former is the mean

$$\vec{R}(t) = \frac{1}{N}\sum_{i=1}^{N} \vec{r}_i(t) \tag{1}$$

of the $N \gg 1$ water molecules making up the raindrop, while the latter is related to the covariances

$$\begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} & \Sigma_{xz} \\ \Sigma_{yx} & \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zx} & \Sigma_{zy} & \Sigma_{zz} \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N}(x_i - X)^2 & \frac{1}{N}\sum_{i=1}^{N}(x_i - X)(y_i - Y) & \frac{1}{N}\sum_{i=1}^{N}(x_i - X)(z_i - Z) \\ \frac{1}{N}\sum_{i=1}^{N}(y_i - Y)(x_i - X) & \frac{1}{N}\sum_{i=1}^{N}(y_i - Y)^2 & \frac{1}{N}\sum_{i=1}^{N}(y_i - Y)(z_i - Z) \\ \frac{1}{N}\sum_{i=1}^{N}(z_i - Z)(x_i - X) & \frac{1}{N}\sum_{i=1}^{N}(z_i - Z)(y_i - Y) & \frac{1}{N}\sum_{i=1}^{N}(z_i - Z)^2 \end{bmatrix}, \quad (2)$$

where $(X, Y, Z) = \vec{R}$. Of course, the shape of the raindrop can also change with time. This is determined by the higher-order statistical moments of $\{(x_i(t), y_i(t), z_i(t))\}_{i=1}^{N}$. However, we can only adopt this hierarchical description in terms of position, size, shape, ... provided the topological characteristics of the raindrop remains unchanged. If the raindrop breaks up into two raindrops, or if an air bubble forms within the raindrop, our description of the first raindrop would have to change discontinuously.

Through this analogy, we hope to convince our readers that persistent structures are the most convenient variables to develop physical theories around. A persistent structure is a collection of microscopic variables that is long-lived (temporal persistence), insensitive to changes in length scales (spatial persistence), and whose statistical moments change continuously with time. The last requirement is guaranteed by topological persistence, i.e., the Betti numbers $\beta_0, \beta_1, \ldots$ remaining constant.

3. Formal Spectral Definition of Persistent Structures

3.1. TDA Definition of Persistence

In Section 2, we saw that a raindrop remains well-defined as a persistent structure over the time it takes to fall to the ground. Therefore, within this time, we can write down equations that govern the continuous changes in its position, velocity, size, and shape. This description is useful because the raindrops are well separated in space. In contrast, the description of a swimming pool in terms of water droplets is not useful, first because there is no natural size to use for such water "droplets", and second because slight "movements" of these "droplets" would make them overlap with each other (and lose their distinctive identities). A discontinuous change occurs when two "droplets" merge, and therefore the structures before and after merging cannot be treated as the same. The structure before does not persist past the merger, while the structure after does not exist until the merger.

It is this spatial persistence that the filtration procedure in TDA identifies. As shown in Figure 3, we draw a link between two data points at filtration parameter ϵ, if their pairwise distance is less than or equal to ϵ. We then write the network obtained in terms of a simplicial complex, which is a set consisting of 0-simplices (nodes), along with 1-simplices (links), 2-simplices (triangles), along with higher-order k-simplices, which are complete graphs with $(k+1)$ nodes. We can also define a *face* of a k-simplex to be a $(k-1)$-simplex making up the k-simplex, and the set of all faces of a k-simplex its *boundary*. In terms of these constructs, a simplicial complex Σ can be precisely defined as a set of simplices satisfying two conditions: (1) any face of a simplex in Σ is also in Σ; and (2) the intersection of any two simplices σ_1 and σ_2 in Σ is either the empty set $\varnothing$, or a face of both σ_1 and σ_2. As ϵ is increased, we find more connected components in Σ. For example, at t_1 and ϵ_1 in Figure 3, the simplicial complex obtained is $\Sigma_1 = \{\langle 1,9\rangle, \langle 5,8\rangle, \langle 1\rangle, \ldots, \langle 9\rangle\}$, which has only two 1-simplices ($\langle 1,9\rangle$ and $\langle 5,8\rangle$) and no 2-simplices, whereas at t_1 and $\epsilon_2 > \epsilon_1$, the simplicial complex $\Sigma_2 = \{\langle 1,7,4\rangle, \langle 1,6,7\rangle, \ldots, \langle 2,10,5\rangle, \langle 1,4\rangle, \langle 1,6\rangle, \ldots, \langle 5,8\rangle, \langle 1\rangle, \ldots, \langle 9\rangle\}$ obtained has six 2-simplices ($\langle 1,7,4\rangle, \langle 1,6,7\rangle, \ldots, \langle 2,10,5\rangle$) and 14 1-simplices. At t_1 and ϵ_3, the simplicial complex obtained is $\Sigma_3 = \{\langle 1,3,6,7\rangle, \langle 2,3,5,10\rangle, \langle 2,5,8,10\rangle, \langle 1,7,4\rangle \ldots, \langle 5,10,8\rangle, \langle 1,4\rangle, \ldots, \langle 5,8\rangle, \langle 1\rangle, \ldots, \langle 9\rangle\}$. In this example, $\langle i\rangle$ is a 0-simplex, $\langle i,j\rangle$ a 1-simplex, $\langle i,j,k\rangle$ a 2-simplex, and $\langle i,j,k,l\rangle$ a 3-simplex.

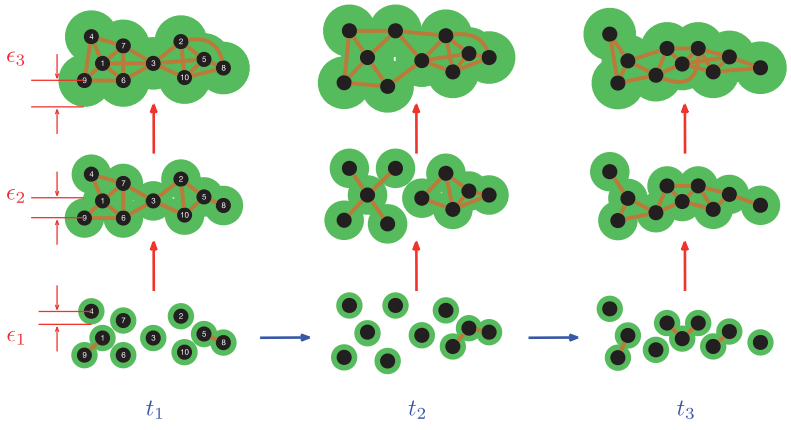


Figure 3. The simplicial complexes obtained by the TDA filtration process for a set of ten data points at three different scales $\epsilon_1 < \epsilon_2 < \epsilon_3$ (the sizes of the green disks), at three different times $t_1 < t_2 < t_3$. In this figure, two data points $\vec{r}_i$ and $\vec{r}_j$ are connected, if $\left|\vec{r}_i - \vec{r}_j\right| \leq \epsilon$.

To follow the dynamics, we start at the smallest scale ϵ_1, to find 6 isolated nodes (0-simplices) and 2 links each connecting 2 nodes (1-simplices) at t_1. At this same scale, we have 7 isolated 0-simplices, 1 1-simplex consisting of 2 links connecting 3 nodes at t_2, as well as 3 0-simplices and 3 1-simplices (2 of them consisting of 1 link connecting 2 nodes, and 1 of them consisting of 2 links connecting 3 nodes) and t_3. In contrast, at the intermediate scale ϵ_2, we find a connected simplicial complex with 10 0-simplices, 14 1-simplices, and 6 2-simplices at time t_1. At the scale ϵ_2, and time t_2, the simplicial complex has two connected components. The first consists of 5 0-simplices and 4 1-simplices. The second consists of 5 0-simplices, 7 1-simplices, and 3 2-simplices. Finally, at t_3, the connected simplicial complex at scale ϵ_2 has 10 0-simplices, 14 1-simplices, and 5 2-simplices.

Not all connected components identified through the filtration process are persistent, because they remain topologically distinct over very small ranges of ϵ. When TDA was first invented, it was applied onto data sets obtained at one point in time or averaged over time. Therefore, the range $(\epsilon_b,\ \epsilon_d)$ between the scale ϵ_b a topologically distinct component first appears (also called the *birth* of the component) and the scale ϵ_d it disappears (also called the *death* of the component) is referred to as its *lifetime*. In TDA, the lifetimes of components are typically shown in the form of a *barcode* or a *persistence* diagram. In a barcode (see Figure 4a), each bar shows the birth (component first appears) and the death (component disappears) of a component in the simplicial complex as ϵ is varied. In a persistence diagram (see Figure 4b), a component is represented as a point whose x coordinate is the birth time, and whose y coordinate is the death time. Persistent components must have long lifetimes, and we can identify these by looking in the barcode in Figure 4a for bars that are significantly longer than the previous ones (the last two bars), or large deviations from the diagonal in the persistence diagram. In the example shown in Figure 4b, there are two 0-dimensional components with lifetimes greater than $\epsilon = 1.0$. These two merged into one at $\epsilon_d = 1.58$, compared to the most recent death at $\epsilon \lesssim 0.5$, and can therefore be thought of as persistent components. In contrast, none of the 1-dimensional components shown in Figure 4b are persistent.

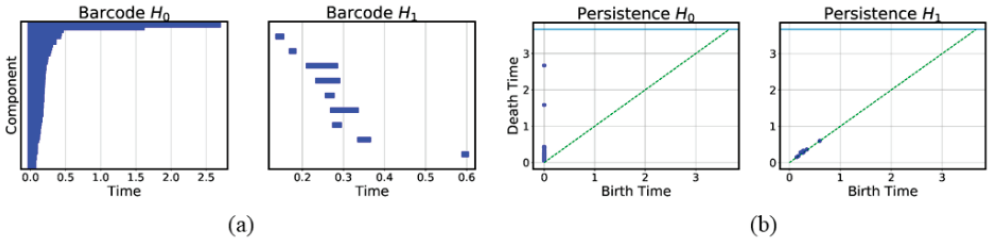


Figure 4. (**a**) The barcodes of the 0-dimensional homology group H_0 and 1-dimensional homology group H_1 for an artificial data set with 50 data points (the same one as in the 1st Figure in Section 3.3.2) undergoing the filtration process. (**b**) The persistence diagrams of the 0-dimensional and 1-dimensional components emerging from the filtration process.

In the example shown in Figure 4, the data set contains two persistent clusters by construction. When there are more persistent structures at different length scales, identifying them from barcodes and persistence diagrams will become challenging. In the rest of this section, we will show that it is easier, and more systematic, to identify persistent structures in spectral space. In fact, this was first demonstrated by Donath and Hoffmann [25], as well as Fiedler [26], who identified communities based on the eigenvectors of the adjacency matrix and the Laplacian matrix respectively. We refer readers to the survey Spielman and Teng [27], and the tutorial on spectral clustering by von Luxburg [28]. To understand why spectral clustering works so well, let us start with what we know about block-diagonal matrices in quantum mechanics.

3.2. Block-Diagonal Matrices in Quantum Mechanics

The barcodes and persistence diagrams described in Section 3.1 are visualizations in real space. It turns out that we can also identify persistent structures in spectral space. To do this, we start from the adjacency matrix representation of the simplicial complex. As we show in Section 3.3.1, there are no persistent structures for a single cluster of data points. Thus, the simplest example that can help us understand how persistent structures are identified would be two well-separated clusters of data points in Section 3.3.2. The adjacency matrix thus has a well-defined community structure, with one diagonal block for the first cluster, and a second diagonal block for the second cluster.

In quantum mechanics, we were first introduced to block-diagonal matrices when we explore the implications of symmetries. For example, we know that the angular momentum operator L^2 and L_z (the z-component of the angular momentum) have the same eigenvectors $|l\ m\rangle$, with eigenvalues $L^2|l\ m\rangle = l(l+1)\hbar^2|l\ m\rangle$ and $L_z|l\ m\rangle = m\hbar|l\ m\rangle$. Since $m = -l, -l+1, \ldots, 0, \ldots, l-1, l$, the matrix representation of L^2 is organized into $(2l+1) \times (2l+1)$ diagonal blocks (see Figure 5a). We were taught that this is the consequence of a symmetry, embodied by the commutation relation $\left[L^2, L_z\right] = 0$, with the diagonal blocks being irreducible representations of this symmetry. In this angular momentum example, the diagonal blocks do not have the same sizes. In contrast, in solid state physics, the diagonal blocks have the same sizes. To see this, consider a crystal made up of $N = N_1N_2N_3$ repeating unit cells. At the boundary of this crystal, we apply the Born-von Karman boundary conditions, to write the wave function as $\psi\left(\vec{r}\right) = \psi\left(\vec{r} + N_1\vec{a}_1 + N_2\vec{a}_2 + N_3\vec{a}_3\right)$, where $\vec{a}_1$, $\vec{a}_2$, $\vec{a}_3$ are the primitive lattice vectors. Furthermore, the periodic crystal has translational symmetry, and thus $\psi(\vec{r} + \vec{R}) = e^{i\vec{k}\cdot\vec{R}}\psi\left(\vec{r}\right)$. Therefore, when we Fourier transform the Hamiltonian matrix in real space, we obtain a Hamiltonian matrix in momentum space that is block-diagonal (see Figure 5b). Each $N \times N$ diagonal block is associated with a distinct wave vector $\vec{k}$. Diagonalizing the block for $\vec{k}$, we would obtain widely separated energy eigenvalues $E_1(\vec{k})$, $E_2(\vec{k})$, $\ldots$, $E_n(\vec{k})$, $\ldots$. Similarly, from the diagonal

blocks of $\vec{k}'$ and $\vec{k}''$, we obtain the energy eigenvalues $\{E_n(\vec{k}')\}$ and $\{E_n(\vec{k}'')\}$. As shown in Figure 5c, $E_n(\vec{k})$, $E_n(\vec{k}')$, and $E_n(\vec{k}'')$ have comparable values, and thus when we combine $E_n(\vec{k})$ for all values of $\vec{k}$, we obtain the nth energy band of the crystal. Between the nth energy band and the $(n+1)$th energy band of the crystal, we find the nth band gap for the band structure of the crystal.

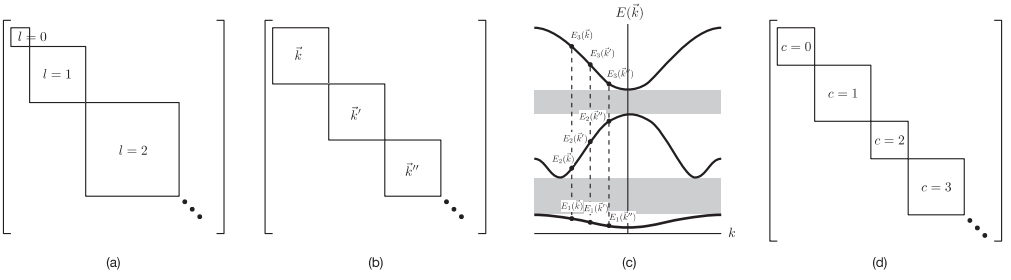


Figure 5. (**a**) The matrix representation of the angular momentum operator L^2 is block-diagonal when it is written in the basis of eigenstates of L_z. Within a diagonal block, all states $|l\ m\rangle$ have eigenvalue $l(l+1)\hbar^2$ for L^2, and eigenvalue $m\hbar$, $m = -l, \ldots, 0, \ldots, l$ for L_z. (**b**) For the Hamiltonian matrix of a crystal in momentum space, we find one diagonal block associated with each wave vector $\vec{k}$. (**c**) When we diagonalize the block-diagonal matrix shown in (**b**), we find the eigenvalues organized into bands $E_n(\vec{k})$ separated by band gaps (gray). (**d**) For a network with community structure, the adjacency matrix A or the Laplacian matrix L is also block diagonal, with each block associated with a different community with community index c.

For a network with adjacency matrix A, we have $A_{ij} = 1$ if node i is linked to node j, or $A_{ij} = 0$ otherwise. In general, nodes in the network need not have the same degree k, i.e., $k_i \neq k_j$ for nodes $i \neq j$. These node degrees can be computed from A, as $k_i = \sum_{j=1}^{N} A_{ij}$, and thereafter organized into a degree matrix $K = \mathrm{diag}(k_1, \ldots, k_N)$. In terms of A and K, the *graph Laplacian* can be defined as $L = K - A$. In Figure 5d, we show the adjacency matrix A (or equivalently the Laplacian matrix L) of a network with well-defined communities (no overlaps between communities). For such a network, A or L would also be block diagonal. The diagonal block associated with community c would be $N(c) \times N(c)$, where $N(c)$ is the number of nodes in community c. Treating the Laplacian L as the Hamiltonian of the network, this block-diagonal structure tells us that there is an observable C that commutes with L, i.e., $[L,\ C] = 0$, and thus the community structure represents some sort of symmetry. More importantly, given the block-diagonal structure of L, its eigenvalues would also be organized into bands separated by band gaps. One of the first to observe these bands of Laplacian eigenvalues separated by a gap was Arenas [29].

3.3. Analysis of Spectral Sequence, Overlapping Communities, Persistent Structures

In the filtration process of a given data set, we vary ϵ to obtain networks with different link densities. When ϵ is small, we expect isolated data points and small clusters of data points. The network is largely unconnected, and therefore we obtain a distribution of eigenvalues for small clusters. As ϵ increases, larger clusters start to form, looking initially like star networks, but eventually becoming complete networks. From spectral graph theory, which is the study of the properties of a network in terms of its characteristic polynomial, eigenvalues $\{\lambda_i\}$, and eigenvectors $\{\vec{u}_i\}$ of L [30,31], we know that any connected component will have one eigenvector with $\lambda = 0$. If a network of N nodes consists of M connected components, then each of the components would contribute one zero eigenvector, i.e., $\lambda = 0$ would be M-fold degenerate. Over and above the zero eigenvalue, special networks such as a star network with N nodes has $N - 2$ unit eigenvalues $\lambda = 1$, and one eigenvalue

$\lambda = N$, whereas a complete network with N nodes has instead $N - 1$ eigenvalues $\lambda = N$. For real networks with intermediate link densities, we then expect the unit eigenvalues $\lambda = 1$ to shift progressively to $\lambda = N$ as the link density increases. This tells us that as link density increases, the distribution of eigenvalues becomes more concentrated at larger eigenvalues. This is illustrated in Figure 6.

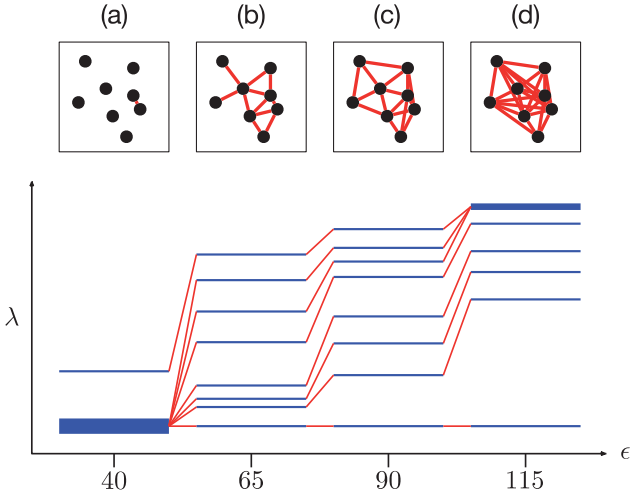


Figure 6. A set of eight data points going through the filtration process, and the resulting Laplacian spectra (blue horizontal lines) for filtration parameters (**a**) $\epsilon = 40$, (**b**) $\epsilon = 65$, (**c**) $\lambda = 90$, and (**d**) $\lambda = 115$. Thin blue lines tell us that the eigenvalues are nondegenerate, whereas thick blue lines indicate that the eigenvalue is degenerate. We use red lines to connect $\lambda_n(\epsilon)$ to $\lambda_n(\epsilon')$, for successive filtration parameters $\epsilon' > \epsilon$.

When $\epsilon = 40$ in Figure 6a, only the two data points closest to each other are linked, whereas the rest of the data points remain isolated. Here, we find the Laplacian eigenvalue $\lambda = 0$ being seven-fold degenerate, and the eigenvalue $\lambda = 2$ for the cluster with two nodes. When the filtration parameter is increased to $\epsilon = 65$ in Figure 6b, the eight data points become a fully connected network. However, the connectivity is not uniform across the network, and part of it looks like a less-densely linked star, while the other more-densely linked part consists of connected 2-simplices. For this oddly shaped network, and also the one shown in Figure 6c when $\epsilon = 90$, the nonzero eigenvalues are distributed between $\lambda_{\min}$ and $\lambda_{\max}$. Finally, when the filtration parameter reaches $\lambda = 115$ in Figure 6d, three nodes attain the maximum degree of $k_{\max} = 7$. This is why the maximum eigenvalue $\lambda_{\max} = 8$ is three-fold degenerate. We call the spectral space visualization $\{\lambda_n(\epsilon)\}$ shown in Figure 6 as ϵ is varied in the filtration process a *spectral sequence*. In the following subsections, we show the spectral sequences of different simple configurations of data points, to identify the relevant features characterizing these configurations. We also show how these features change with separation between clusters, during a fusion process, and in the presence of noise of different strengths.

3.3.1. One Cluster

As a benchmark, let us examine the spectral sequence of a single cluster of data points sampled from a two-dimensional Gaussian distribution. As we can see from Figure 7, there is no prominent band gap in the spectral sequence. We use spectral graph theory to explain this in two limits. First, in the limit of small ϵ, the simplicial complex consists of multiple connected components of different sizes n_i. Furthermore, if these connected components are networks intermediate between star and complete networks, their eigenvalues would

be distributed between $\lambda = 0$ and $\lambda = n_i$. When we superimpose these spectra, we find a "continuous" distribution of eigenvalues between $\lambda = 0$ and $\lambda = \max_i n_i$. Second, in the limit of large ϵ, the simplicial complex consists of a single connected component intermediate between star and complete networks of size N. Therefore, the nonzero Laplacian eigenvalues of such a simplicial complex would also be "continuously" distributed between $\lambda_{\min} > 1$ and $\lambda_{\max} < N$.

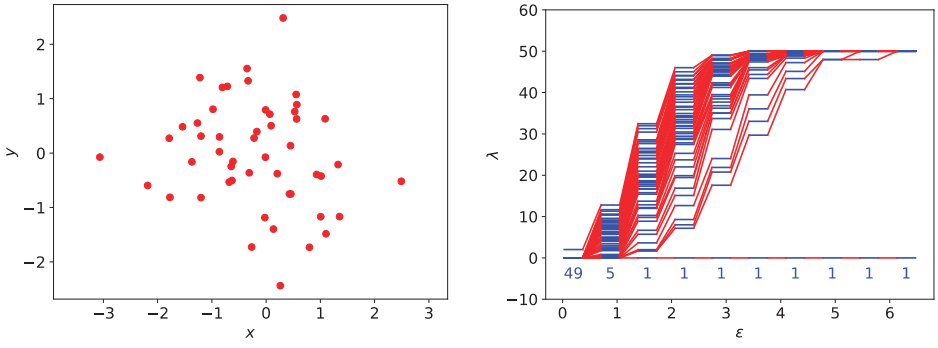


Figure 7. (**left**) A cluster of 50 data points sampled from a two-dimensional normal distribution $p(x_1, x_2) = \frac{1}{2\pi|\Sigma|} \exp\left[-\frac{1}{2}\left(\vec{x} - \vec{\mu}\right)^T \Sigma^{-1}\left(\vec{x} - \vec{\mu}\right)\right]$, where $\vec{\mu} = (\mu_1, \mu_2) = (0,\ 0)$ and Σ is a diagonal covariance matrix with diagonal matrix elements $\sigma_{11}^2 = 1$ and $\sigma_{22}^2 = 1$. (**right**) The spectral sequence (i.e., the distribution of eigenvalues $\{\lambda_n(\epsilon)\}$ of the Laplacian matrix L at different filtration parameter ϵ) of this cluster. In this figure, the numbers of zero eigenvalues for different ϵ are indicated below $\lambda = 0$.

3.3.2. Two Clusters

The simplest example of a data set with community structure would be one with two clusters, as shown in Figure 8. The barcode of this data set was shown in Figure 4, where we saw that this two-cluster structure is persistent with respect to changes in length scale. From Figure 8, we see that there are two zero eigenvalues from $\epsilon \approx 0.7$ to $\epsilon \approx 1.8$. This range of filtration parameter is comparable to the one found from the barcode in Figure 4. However, the spectral signature ($\Delta\lambda = \max_i\{\lambda_{i+1} - \lambda_i\}$, shaded yellow in Figure 8) for this persistent structure is far more prominent, suggesting that the two clusters remain distinguishable even after links start to form between them (overlapping communities). In particular, when $\epsilon = 2.8653$ and $\lambda_1 = 12.9498$, there are 261 links between the two clusters, but $\Delta\lambda$ remains larger than level spacings elsewhere in the spectrum.

Starting at $\epsilon = 1.9254$, the two clusters become linked, and there is only one zero eigenvalue $\lambda_0 = 0$. At this filtration parameter, the first nonzero eigenvalue is $\lambda_1 = 0.6055$. For a smooth manifold M, Jeff Cheeger first proved that $\lambda_1 \geq h^2(M)/4$, where λ_1 is the first nonzero eigenvalue of the Laplace–Beltrami differential operator on M, while the Cheeger constant $h(M)$ is the smallest area of a hypersurface that cuts M into two [32]. This result carries over to discrete networks. Suppose λ_1 is the first nonzero eigenvalue of the Laplacian of network G, which can be split into two networks A (with N_A nodes) and B (with N_B nodes) by cutting the smallest number of links n_{AB}, then $\lambda_1 \geq h^2(G)/4$, where $h(G) = n_{AB}/\max(N_A, N_B)$ [33–35]. This tells us that λ_1 increases with the size of the neck linking networks A and B.

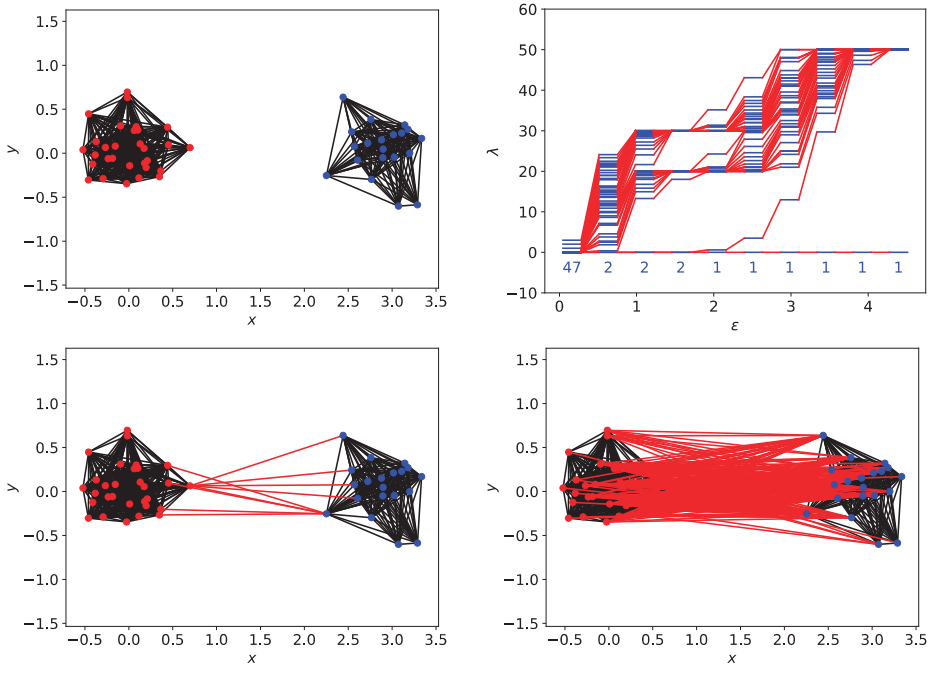


Figure 8. (**top right**) The spectral sequence for two clusters, one with 30 red data points, the other with 20 blue data points. In this figure, the persistent spectral gap that corresponds to this spatially persistent two-cluster structure is shaded yellow. (**top left**) The simplicial complex of the two clusters at $\epsilon = 0.9855$, which consists of the two nearly complete networks that are not connected. (**bottom left**) The simplicial complex of the two clusters at $\epsilon = 1.9254$, showing how the red cluster is connected to the blue cluster by 9 links, between 5 red nodes and 5 blue nodes. (**bottom right**) The simplicial complex of the two clusters at $\epsilon = 2.8653$. At this length scale, the two clusters are connected by 261 links. In these figures, intra-cluster links are black, while inter-cluster links are red.

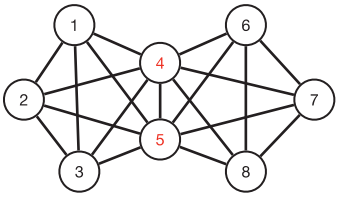
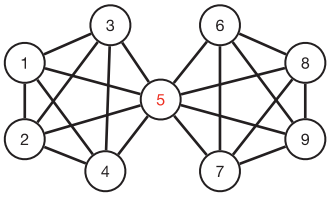
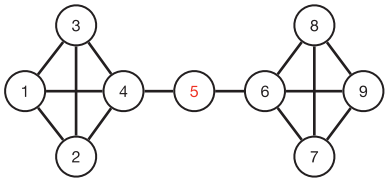
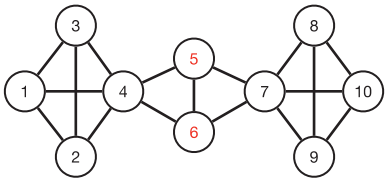
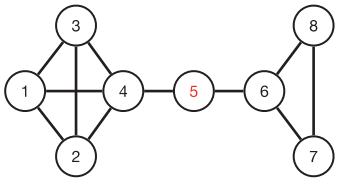
3.4. Analysis of Eigenvectors, and Identification of the Neck from the Fiedler Vector

The *Fiedler vector* $\vec{u}_1$ associated with λ_1 also allows us to identify nodes that are part of the neck [26,36]. In this subsection, we show how this can be done, by first showing the results from toy networks before we analyze the Fiedler vector and other low-lying eigenvectors in the spectral sequence examples shown in Section 3.3 and Supplementary Information Section B.

3.4.1. Toy Networks

In Table 1, we show a sequence of toy networks in which two distinguishable subnetworks are connected by necks of various natures. In the first two networks, the clusters share an edge or a corner, and thus the neck consists of the nodes making up the edge or the corner. In the next two networks, the clusters are bridged by a single node or an edge, and thus the neck consists of the bridging node(s). Nodes in the neck can be identified as zero components in the Fiedler vector. We can also distinguish the two clusters, because the components in one of them is positive, while the other is negative. In the last network, the two clusters are not balanced, with one consisting of four nodes, and the other three nodes. In its Fiedler vector, the weight of the neck (node 5) is not zero, but still significantly smaller than the weights of the other nodes. Through these examples, we realized that the neck consists of nodes with weights close to zero, or significantly smaller than the clustered nodes in the Fiedler vector.

Table 1. The neck (nodes colored in red) between clusters in simple networks, and how they can be identified from the Fiedler vector, which is the eigenvector $\vec{u}_1$ associated with the first non-zero eigenvalue λ_1 of the graph Laplacian L.

	$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$	$\vec{u}_1 = \begin{bmatrix} 0.408 \\ 0.408 \\ 0.408 \\ 0 \\ 0 \\ -0.408 \\ -0.408 \\ -0.408 \end{bmatrix}$
	$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$	$\vec{u}_1 = \begin{bmatrix} 0.354 \\ 0.354 \\ 0.354 \\ 0.354 \\ 0 \\ -0.354 \\ -0.354 \\ -0.354 \\ -0.354 \end{bmatrix}$
	$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$	$\vec{u}_1 = \begin{bmatrix} 0.371 \\ 0.371 \\ 0.371 \\ 0.294 \\ 0 \\ -0.294 \\ -0.371 \\ -0.371 \\ -0.371 \end{bmatrix}$
	$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$	$\vec{u}_1 = \begin{bmatrix} 0.383 \\ 0.383 \\ 0.383 \\ 0.247 \\ 0 \\ 0 \\ -0.247 \\ -0.383 \\ -0.383 \\ -0.383 \end{bmatrix}$
	$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$	$\vec{u}_1 = \begin{bmatrix} 0.353 \\ 0.353 \\ 0.353 \\ 0.268 \\ 0.048 \\ -0.353 \\ -0.463 \\ -0.463 \end{bmatrix}$

3.4.2. Filtration Sequence for Two Clusters

In Section 3.3 and Supplementary Information Section A, we analyzed spectral sequences resulting from the filtration of different data sets, to identify tell-tale signatures for different numbers of clusters. For the spectral sequence shown in Figure 3 for two clusters of data points, let us focus on the Fiedler vectors for $\epsilon = 1.9254$ ($\lambda_1 = 0.6055$) and $\epsilon = 3.3353$ ($\lambda_1 = 29.704$). At $\epsilon = 1.9254$, the two clusters are connected by 9 links, between 5 nodes from cluster 1, and 5 nodes from cluster 2. These 10 nodes, identified from their *smaller absolute weights*, form the neck between clusters 1 and 2. For $\epsilon = 3.3353$, there are 21 nodes with zero weights. All 21 nodes have the maximum degree $k_i = 49$ in a network of 50 nodes and are members of a *bloated neck*.

Just to be careful, we also look at the node in cluster 2 with the minimum degree $k_i = 33$. This node has the largest absolute weight in $\vec{u}_1$, and is linked to all cluster-2 nodes, but only to 14 cluster-1 nodes. Out of these 14 cluster-1 nodes, 13 of them belong to the neck. In addition, we find that set of 10 neck nodes when $\epsilon = 1.9254$ is a subset of the set of 21 neck nodes when $\epsilon = 3.3353$. This tells us that in the filtration process, instead of a simple fusion $A + B \to C$, TDA suggests the process $A + B \to A + n + B \to a + N + b \to \mathcal{N} = C$. In other words, the fusion between clusters A and B begin with the creation of a small neck n. This neck continues to absorb members of A and B to become the bigger neck N (at the expenses of clusters $A \to a$ and $B \to b$ shrinking), until all original members of clusters A and B become absorbed into $\mathcal{N}$, which we can now call cluster C.

In this example, we examine the filtration process involving two clusters. However, we expect the picture to hold even for the filtration processes at different times for two clusters merging into one, since the neck should be present until the two clusters completely fuse together. However, the smaller neck at an earlier time may not be embedded within the larger neck later. This is because even necks can lose or gain nodes, and all processes described in the raindrop analogy apply.

3.4.3. Quasi-Degeneracies and Multiple Necks

Finally, we consider the situation where the data points are connected by more than one neck at some stage in the filtration process. The simplest situation where this occurs is when we have three clusters along a straight line, as shown in Figure 9a. In Figure 9b, we see that when $\epsilon = 0.668$, the three clusters are not linked, and we find three zero eigenvalues. When the filtration parameter is increased to $\epsilon = 1.328$, the three clusters forms a single cluster, with one neck connecting the green cluster to the blue, and another neck connecting the blue cluster to the red. When this occurs, $\lambda_1 = 1.067$ and $\lambda_2 = 2.776$ become non-zero, but remain close to each other. In Figure 9c,d, we show that the eigenvectors associated with λ_1 and λ_2 are the antisymmetric and symmetric combinations of the green and red clusters respectively.

In the symmetric combination $\vec{u}_2$, components of the green and red clusters have the same sign, thus forcing components of the middle blue cluster to have the opposite sign. Components of these three clusters can only have the same sign in $\vec{u}_0$, the eigenvector associated with $\lambda_0 = 0$. In the antisymmetric combination $\vec{u}_1$, components of the green and red clusters have opposite signs, and thus components of the middle blue cluster must be close to zero. In this sense, although a cluster in its own right, in the antisymmetric combination $\vec{u}_1$ the blue cluster plays the role of a neck. Because of this dual role, we call the blue cluster a *bridging cluster*. Because of these differences in signs and magnitudes, if we divide the component $u_{1,i}$ by $u_{2,i}$, this ratio would be close to zero if i is a member of the blue cluster (the bridging cluster, $0 \leq i < 40$), or has an absolute value close to one if i is a member of the green cluster ($40 \leq i < 70$) or the red cluster ($70 \leq i < 100$). Indeed, this is what we see in Figure 9e. In Figure 9e, we also see four absolute ratios that are exceptionally large. This can only happen if $u_{2,i}$ is close to zero, but $u_{1,i}$ is not. From Figure 9d, we see that these four are true members of the two necks connecting the three clusters. Indeed, when we go to $\epsilon = 1.987$ in Figure 9f, where $\lambda_1 = 10.645$ and $\lambda_2 = 23.791$

are very different, $r_i \approx 0$ continues to help us identify the bridging cluster, while $|r_i| \gg 1$ helps us identify the neck, which is thicker at this filtration parameter.

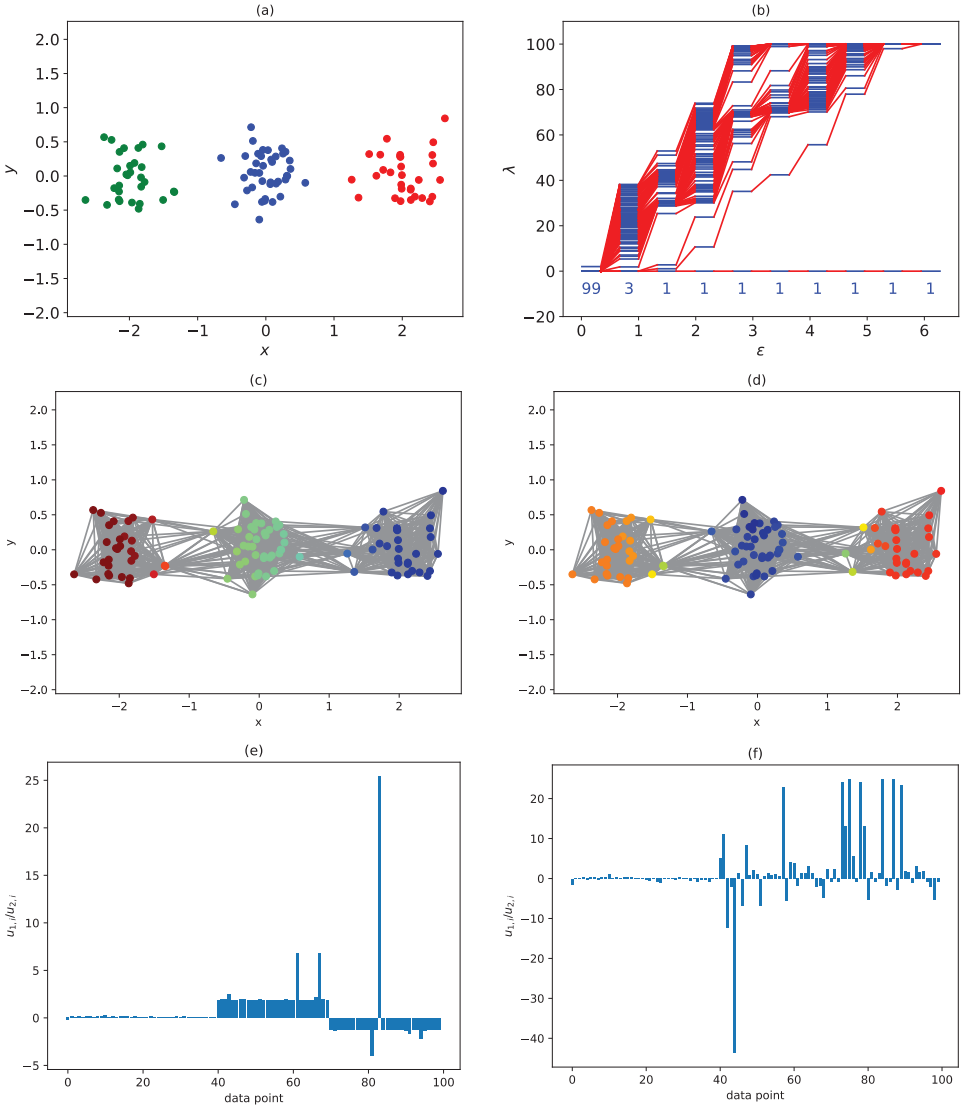


Figure 9. (**a**) Three clusters of data points arranged in a straight line. The blue cluster contains 40 data points, while the red and green clusters each contain 30 data points. (**b**) The spectral sequence of the three clusters of data points. Note the sudden change from a three-cluster description at $\epsilon = 0.668$ to a one-cluster description at $\epsilon = 1.328$. Note also the pair of small, closely spaced eigenvalues $\lambda_1 = 1.067$ and $\lambda_2 = 2.776$ at $\epsilon = 1.328$. (**c**) The data points are colored according to their components in $\vec{u}_1$, the eigenvector associated with λ_1. A similar example was shown by Servedio et al. in Refs. [37,38]. (**d**) The data points are colored according to their components in $\vec{u}_2$, the eigenvector associated with λ_2. (**e**) Ratio of components in $\vec{u}_1$ to components in $\vec{u}_2$ when $\epsilon = 1.328$. (**f**) Ratio of components in $\vec{u}_1$ to components in $\vec{u}_2$ when $\epsilon = 1.987$.

3.5. Spectral Definition of Persistent Structure

Summarizing our findings from Sections 3.3 and 3.4 and Supplementary Information Section B, we realized that persistent structures are accompanied by persistent gaps $\Delta\lambda = \max_i(\lambda_{i+1} - \lambda_i)$ in the spectral sequence. These persistent gaps should not be confused with non-persistent ones that appear when the persistent structures have a discrete spectrum of sizes. From Figure 8, we see that this persistent gap arises because λ_2 increases more rapidly than λ_1 (which can remain zero) when ϵ was first increased, before λ_1 increases rapidly after ϵ exceeded the characteristic gap between the most persistent clusters. In particular, when λ_1 starts rising, the persistent structures are already connected by necks, but they remain distinguishable, i.e., we can talk about $A + n + B$ (a thin neck n connecting two large clusters A and B) or $a + N + b$ (a thick neck N connecting two small clusters a and b). We think of the persistent structures A and B as having vanished only after they are completely absorbed by the neck $\mathcal{N}$, at which time we can identify it as a new persistent structure C where all nodes from A and B have become a complete network. This picture is confirmed by our analysis of the Fiedler eigenvector $\vec{u}_1$ (corresponding to $\lambda_1 > 0$). From the eigenvector perspective, the persistent structures remain distinguishable from the neck since nodes in the neck have zero or smaller absolute weights in the Fiedler vector compared to nodes in the clusters.

Through Sections 3.3 and 3.4 and Supplementary Information Section B, we now have a deeper appreciation of the raindrop analogy described in Section 2. Clearly, when two persistent structures A and B are not connected, their individual descriptions are continuous in time. Such descriptions would involve an equation for the rate of change of the mass of A, another for the rate of change of the center of mass (CM) of A, one more for the rate of change of the CM velocity of A, and a last one governing how the shape of A changes. We also find a similar set of equations for B. Once they become connected, we need a single description that is continuous in time, but we do not completely discard the earlier descriptions of A and B. Instead, we think of the single description of $A + n + B$ as being obtained by introducing one more set of equations for the neck n (which will eventually become the persistent structure C) and impose constraints on these equations. For example, $m_A + m_B + m_n$ must now be approximately conserved, and similarly for the momentum. In this merging stage, it is actually inconvenient to use only one set of equations for $C = A + B$, because too many things are changing simultaneously. It is convenient to use one set of equations for C only after A and B are completely absorbed by the neck.

4. Results and Discussion

4.1. Data

The daily prices of 671 Taiwan Stock Exchange (TWSE) stocks from 1 April 2018 to 30 September 2020 (Figure 10b), 530 Singapore Exchange (SGX) stocks from 31 August 2019 to 30 April 2021 (Figure 10a), and 504 component stocks of the S&P 500 from 1 June 2019 to 31 December 2020 (Figure 10c) were downloaded from Yahoo! Finance using Python's pandas_datareader module. We then post-processed the financial time series as follows. First, "NaNs" were replaced with "0s". Moreover, if the time series contains more than 50% "0s", we remove this ticker symbol from the list. For the remaining stocks, we applied standardization, and also computed their returns. For SGX, some delisted stocks were downloaded manually from the investing.com website. Similarly, a few S&P 500 component stocks changed during the period of study, and so we downloaded both new and old component stocks from the investing.com website.

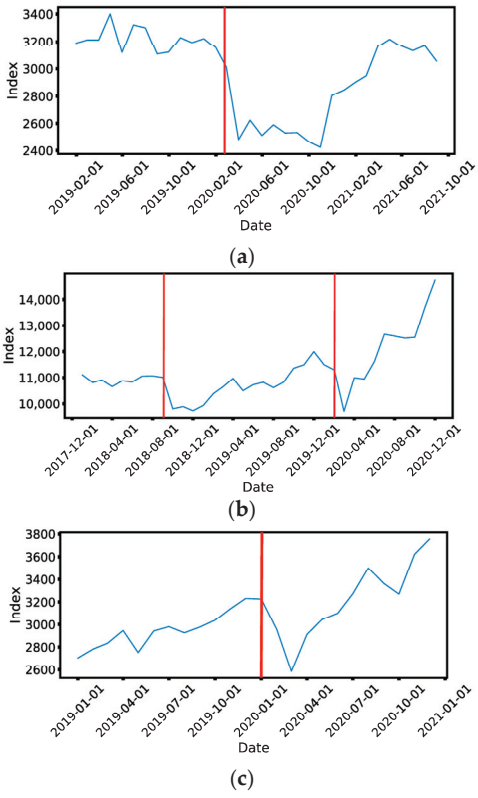


Figure 10. Monthly values of (**a**) the Straits Times Index (STI) of the SGX, (**b**) the Taiwan Capitalization Weighted Stock Index (TAIEX) of the TWSE, and (**c**) the Standard & Poor's 500 (S&P 500) between 1 January 2019 to 31 August 2021, 1 January 2018 to 31 December 2020, and 1 June 2019 to 31 December 2020 respectively. We are specifically interested in two market crashes (Sep 2018 and March 2020) on the TWSE, one market crash (Mar 2020) on the SGX, and one market crash (Mar 2020) for the S&P 500. In these figures, these are shown as red vertical lines.

4.2. Methods

First, we identified four periods, each with a market crash (on TWSE, SGX, or S&P 500) in the middle, as shown in Table 2. We then computed the Pearson cross correlations

$$C_{ij} = \frac{\sum_{t=1}^{N}(r_{i,t} - \overline{r}_i)(r_{j,t} - \overline{r}_j)}{\sqrt{\sum_{t=1}^{N}(r_{i,t} - \overline{r}_i)^2}\sqrt{\sum_{t=1}^{N}(r_{j,t} - \overline{r}_j)^2}} \tag{3}$$

of the daily returns $r_{i,t}$ and $r_{j,t}$ within a six-month time window with $N+1$ trading days, which we advanced one week at a time. Here, $\overline{r}_i$ and $\overline{r}_j$ are the average returns of stocks i and j within each six-month time window. For each time window, we further convert the pairwise cross correlations C_{ij} into pairwise ultrametric distances $0 \leq d_{ij} = \sqrt{2(1 - C_{ij})} \leq 2$.

Table 2. The start and end dates of the four periods used to study the September 2018 mini-crash and March 2020 crash on the TWSE, the March 2020 crash on the SGX, and the February 2020 crash of the S&P 500.

Crash	Start Date	End Date
Sep 2018 TWSE mini-crash	1 April 2018	30 April 2019
Mar 2020 TWSE crash	1 August 2019	30 September 2020
Mar 2020 SGX crash	1 August 2019	30 April 2021
Mar 2020 S&P 500 crash	1 June 2019	31 December 2020

Next, for a given market crash and each of its distance matrices, we perform the TDA filtration process by varying the filtration parameter ϵ. Two stocks, i and j, are linked if $d_{ij} \leq \epsilon$. Therefore, for a given time window at filtration parameter ϵ, we constructed an adjacency matrix A_{ij} whose matrix elements are $A_{ij} = 1$ if $d_{ij} \leq \epsilon$, and $A_{ij} = 0$ otherwise. Using the adjacency matrix, we then computed the degree matrix whose diagonal elements are

$$K_{ii} = k_i = \sum_{j \neq i} A_{ij}, \tag{4}$$

and whose off-diagonal elements are $K_{ij} = 0$. Finally, we constructed the graph Laplacian $L(\epsilon) = K - A$ to obtain its eigenvalues and eigenvectors. Over a judicious choice of filtration parameters, $\epsilon = 0.5$, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, we then visualize the spectral sequence $\{\lambda_i(\epsilon)\}$ for each time window, but analyzed the spectral sequences and Fiedler vectors for the selected time windows.

4.3. March 2020 TWSE Crash

We start by analyzing the spectral sequences for the March 2020 TWSE crash, which was said to be caused by the start of the COVID-19 pandemic [39,40]. The complete series of spectral sequences can be found in Supplementary Figure C1 in the Supplementary Information. Here, we show in Figure 11 the spectral sequences for only four time windows: (1) 1 August 2019–31 January 2020, (2) 22 September 2019–22 March 2020, (3) 15 October 2019–15 April 2020, and (4) 1 April 2020–30 September 2020. The first time window is before the March 2020 TWSE crash, while the fourth time window is after the March 2020 TWSE crash. The March 2020 TWSE crash occurred at the end of the second time window, and in the middle of the third time window.

From Supplementary Figure C1 in the Supplementary Information, we see that the spectral sequences changed very rapidly from the 15 September 2019–15 March 2020 time window (that just missed the March 2020 TWSE crash) to the 22 September 2019–22 March 2020 time window (that first that included the March 2020 TWSE crash). For the first seven time windows that do not include the March 2020 TWSE crash, their spectral sequences resemble that of the 1 August 2019–31 January 2020 time window shown in Figure 11a, which in turn resembles that of a single cluster of points shown in Figure 7 of Section 3.3.1. For the 21 time windows overlapping the March 2020 TWSE crash, their spectral sequences are similar to those of the 22 September 2019–22 March 2020 (Figure 11b) and 15 October 2019–15 April 2020 (Figure 11c) time windows. These spectral sequences bear similarities to those shown in Figure 8 of Section 3.3.2, Supplementary Figures A1 and A2 and Supplementary Information Sections A1 and A2, where we find prominent persistent gaps near the ends of the spectral sequences. Finally, the last four time windows shown in Supplementary Figure C1 have spectral sequences similar to the first seven time windows, as well as Figure 11d, suggesting that the TWSE had recovered from the March 2020 crash. These observations are consistent with the suspicion by econophysicists that a market crash is a critical transition. They also suggest that the March 2020 TWSE crash was short, lasting only for the first two weeks of March 2020 (indeed this is the time to go from the TAIEX high of 11,321 on 1 March 2020 to the low of 9234 on 15 March 2020), and seen as the

"V"-shape feature in Figure 10b. They also agree with the picture of a market crash being the result of the fragmentation of a giant cluster.

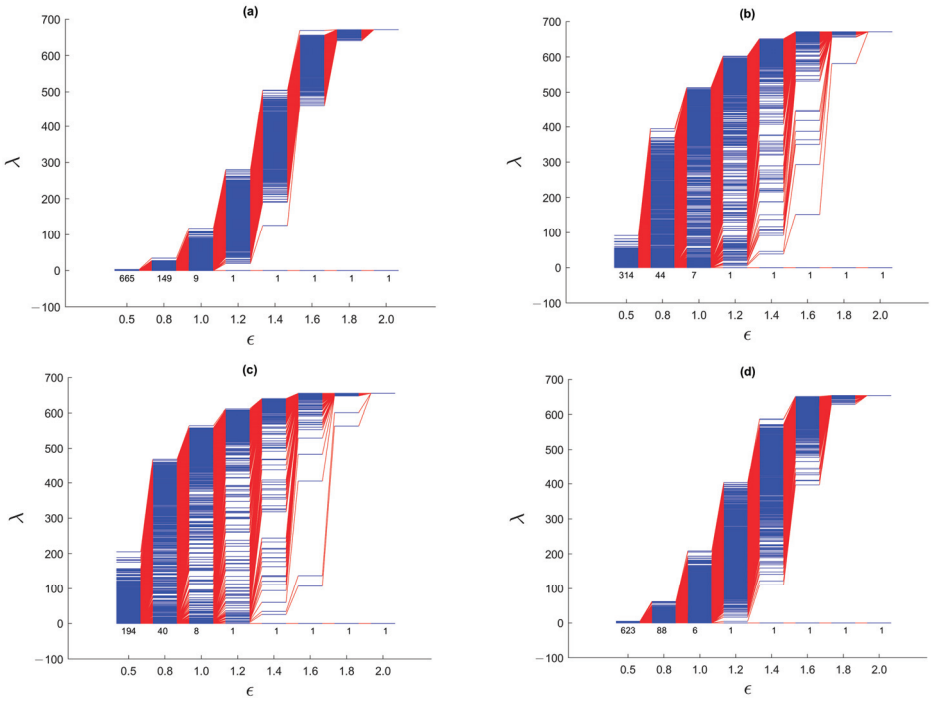


Figure 11. The spectral sequences of the TWSE for $\epsilon = 0.5, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0$ over the six-month time windows: (**a**) 1 August 2019–31 January 2020 (671 stocks), (**b**) 22 September 2019–22 March 2020 (671 stocks), (**c**) 15 October 2019–15 April 2020 (655 stocks), and (**d**) 1 April 2020–30 September 2020 (654 stocks). During the March 2020 TWSE crash, the TAIEX fell from a high of 11,321 on 1 March 2020 to a low of 9234 on 15 March 2020.

For the first seven time windows and the last four time windows, we find a narrow band of eigenvalues ($0 \leq \lambda < 10$) for the smallest filtration parameter $\epsilon = 0.5$. This tells us that at this scale, most of the clusters are small, and therefore the total number of clusters is comparable to the total number of stocks on the TWSE. The narrow bandwidth at $\epsilon = 0.5$ is consistent with only localized random walks on small, disconnected components. On the other hand, for the 21 time windows whose spectral sequences show prominent gaps, there is a broad band of eigenvalues ($0 \leq \lambda < 300$) for $\epsilon = 0.5$. This suggests that at this scale, there is a broad distribution of cluster sizes, including a few strongly correlated ones with up to 300 stocks during the market crash. For these time windows, the broad bandwidth at $\epsilon = 0.5$ is consistent with the delocalization of random walkers on larger connected components.

Moreover, before and after the March 2020 TWSE crash, λ_1 rose rapidly at $\epsilon \approx 1.3$, whereas during the market crash, λ_1's rapid rise only began at $\epsilon \approx 1.7$. This delay in the rapid rise of λ_1 suggests that during the market crash, the gap in correlations between clusters is 30–40% larger than the standard deviation in the continuous distribution of correlations within the giant cluster prior to its fragmentation. Indeed, the persistent gap is most pronounced at $\epsilon = 1.6$, although in some time windows, this persistent gap can also be observed at $\epsilon = 1.4$ or $\epsilon = 1.8$. Finally, as we elucidate the picture of March 2020 TWSE crash as a strongly correlated giant cluster fragmenting into a few strongly correlated clusters, let us clarify that this need not involve all stocks. Unaffected stocks then form a

noisy background, whose effect is to obfuscate the persistent gap. Based on our analysis in Supplementary Information Section A.5, the persistent gap can nevertheless be identified from the late rise in λ_1. This is indeed what we observed.

From Section 3.4, we understand that when two clusters become first connected by a thin neck, components of the two clusters have opposite signs in $\vec{u}_1$, while components of the neck have significantly smaller or zero weights. However, as the neck becomes thicker with increasing ϵ, components become distributed about zero, and few members of the two clusters remain distinguishable. With these in mind, let us start our eigenvector analyses with the time window 1 August 2019–31 January 2020, which was before the March 2020 COVID-19 crash. From Supplementary Figure C2(a), we see that $\lambda_0 = 0$ is non-degenerate for $1.2 \leq \epsilon \leq 2.0$, and λ_1 changes most sharply between $\epsilon = 1.4$ and $\epsilon = 1.6$. Let us therefore examine the Fiedler vector $\vec{u}_1$ for $1.2 \leq \epsilon \leq 1.8$. For $\epsilon = 1.2$, $\lambda_1 = 19.538$, most of the Fiedler components have an absolute value of around 10^{-3}, except for one component whose value is 0.991. To check for non-overlapping distributions that represent Fiedler components from the two clusters, we therefore limit ourselves to bins between -0.005 and $+0.005$ to plot high-resolution histograms in Figure 12. The distributions of Fiedler components at different filtration parameters for this time window are indeed consistent with there being just one giant cluster in the market before the crash.

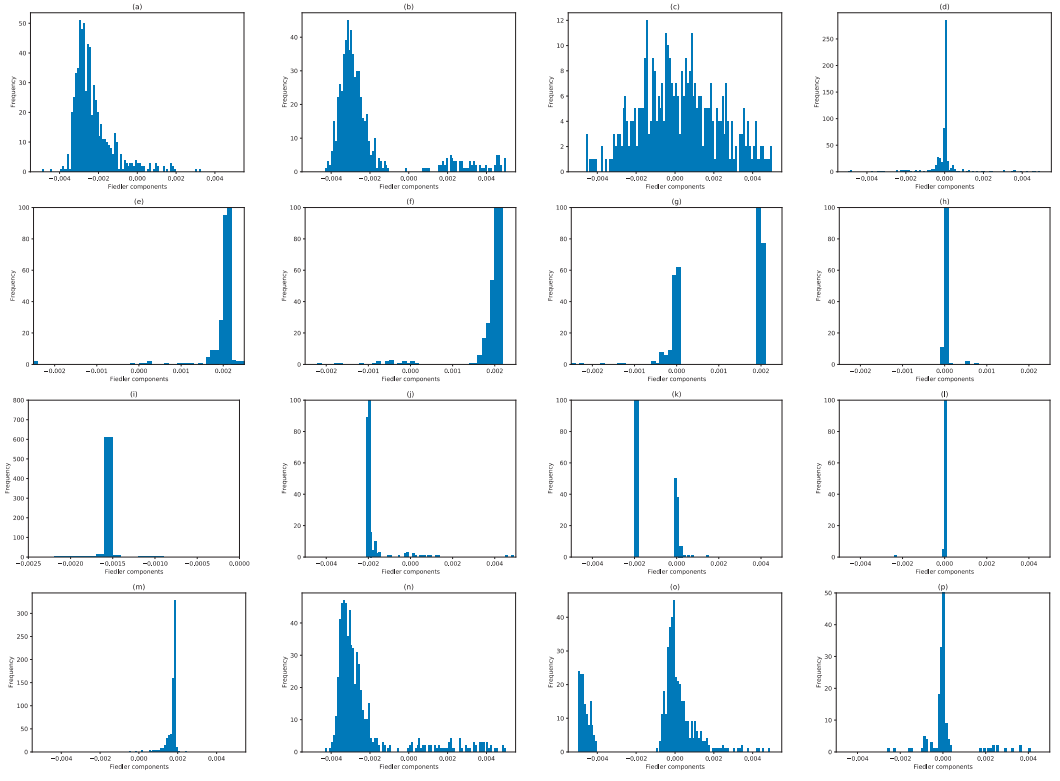


Figure 12. Distribution of Fiedler components at different filtration parameters: (first row) (**a**) $\epsilon = 1.2$, (**b**) $\epsilon = 1.4$, (**c**) $\epsilon = 1.6$, and (**d**) $\epsilon = 1.8$ for the 1 August 2019–31 January 2020 time window; (second row) (**e**) $\epsilon = 1.2$, (**f**) $\epsilon = 1.4$, (**g**) $\epsilon = 1.6$, and (**h**) $\epsilon = 1.8$ for the 22 September 2019–22 March 2020 time window; (third row) (**i**) $\epsilon = 1.2$, (**j**) $\epsilon = 1.4$, (**k**) $\epsilon = 1.6$, and (**l**) $\epsilon = 1.8$ for the 15 October 2019–15 April 2020 time window, and (fourth row) (**m**) $\epsilon = 1.2$, (**n**) $\epsilon = 1.4$, (**o**) $\epsilon = 1.6$, and (**p**) $\epsilon = 1.8$ for the 1 April 2020–30 September 2020 time window.

While the picture for $\epsilon = 1.2$ is not clear in Figure 12a, for $\epsilon = 1.4$ it is clear from Figure 12b that most of the stocks (with negative components) were organized into a giant cluster, while most of the rest (with positive components) were organized into a minor cluster (shown in Supplementary Table G1). As expected, when $\epsilon \geq 1.6$ (Figure 12c,d), the distribution of Fiedler components become unimodal, and centered around zero. Nevertheless, in Figure 12d we see that remnants of the two clusters are still visible at $\epsilon = 1.8$, with 76 components larger than 0.001, and 79 components less than -0.001. The cluster with 79 components is shown in Supplementary Table G1 in the Supplementary Information.

Moving on to the 22 September 2019–22 March 2020 time window (Supplementary Figure B2(b), which includes the first week of the crash), $\lambda_0 = 0$ is again non-degenerate for $1.2 \leq \epsilon \leq 2.0$, and λ_1 now changes most sharply between $\epsilon = 1.6$ and $\epsilon = 1.8$. At $\epsilon = 1.2$ and $\epsilon = 1.4$, there is a noticeable gap in the distribution of Fiedler components, between those that are negative, and those that are positive. The smaller of these groups are shown in Supplementary Table G1. However, we must be careful interpreting all of these components as part of the minor cluster, as we can see from Figure 12e,f that some of these components are close to zero, and might be part of the neck instead. Since λ_1 changes most rapidly between $\epsilon = 1.6$ and $\epsilon = 1.8$, we therefore expect the distribution of Fiedler components at $\epsilon = 1.6$ to be similar to those at $\epsilon = 1.4$. Indeed, two clusters can be identified, but there is now a larger neck with close-to-zero components. Based on the small gap at around -0.001 (Figure 12g), we identified members of the minor cluster, as shown in Supplementary Table G1. Finally, at $\epsilon = 1.8$, most of the components have become zero, suggesting that the neck (566 stocks) has grown to dominate the two clusters. Roughly 100 stocks of the major cluster and 5 stocks of the minor cluster remain distinguishable, as we can see from Figure 12h. As we can see from Supplementary Table G1, a smaller minor cluster identified at a given ϵ is almost perfectly embedded in the larger minor cluster identified at the preceding or succeeding ϵ. This self-consistency at different length scales helps us to reliably identify the two clusters within a given time window.

Moreover, from the spectral sequence of this time window, we see that there is a pair of nearly degenerate eigenvalues $\lambda_1 = 38.693$ and $\lambda_2 = 46.046$ at $\epsilon = 1.4$. This pair of eigenvalues came from a larger group of nearly degenerate eigenvalues at $\epsilon = 1.2$. Unlike the situation shown in Figure 9, where two smaller clusters merge first before merging later with a third cluster, having two small eigenvalues suggests the formation of two thin necks, as shown in Figure 9. When we examine the ratio $r_i = u_{1,i}/u_{2,i}$ at $\epsilon = 1.4$ (Figure 13(left)), we find a group of 622 components that can be distinguished from the remaining 49 components. These 622 components contain the major cluster, whose components have ratios around $r_i = 1.4$. Of the 49 components that do not belong to the major cluster, 14 has $-0.5 < r_i < 0.5$, and are the most likely candidate for the bridging cluster shown in Section 3.4.3. From Section 3.4.3, we also understood that components with very large absolute ratios $r_i = u_{1,i}/u_{2,i}$ are members of the necks. Specifically, those with $r_i < -2$ have been identified alongside the minor cluster at $\epsilon = 1.4$ and $\epsilon = 1.6$ in Supplementary Table G1. The rest are likely members of the neck that link the major cluster to the bridging cluster, as shown schematically in Figure 14.

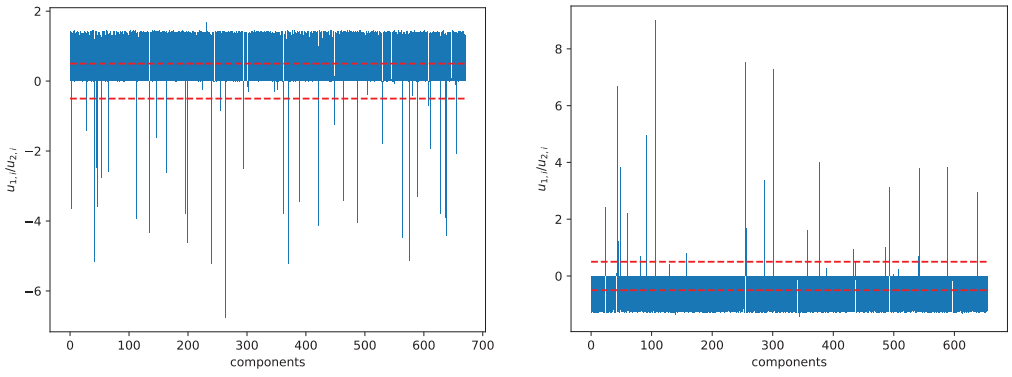


Figure 13. Bar plot of the ratio $u_{1,i}/u_{2,i}$ of components of the eigenvectors $\vec{u}_1$ and $\vec{u}_2$ associated with the smallest non-trivial eigenvalues λ_1 and λ_2 of the graph Laplacian obtained at filtration parameter $\epsilon = 1.4$, in (**left**) the 22 September 2019–22 March 2020 time window, and (**right**) the 15 October 2019–15 April 2020 time window. In this figure, components between the two red dashed lines are likely to be members of a bridging cluster.

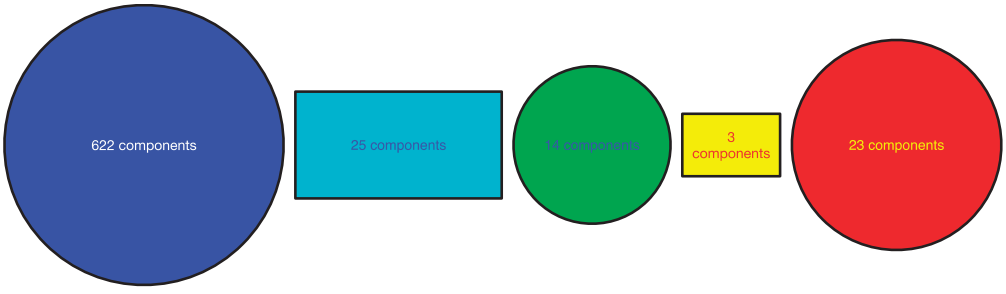


Figure 14. Schematic figure showing the major cluster (blue, 622 components) being linked to the minor cluster (red, 23 components) through a bridging cluster (green, 14 components) in the 22 September 2019–22 March 2020 time window. In the cyan neck between the blue and green clusters, there are 25 components. In the yellow neck between the green and red clusters, there are three components.

Next, let us move on to the 15 October 2019–15 April 2020 time window (Supplementary Figure B2(c)), which covers both weeks of the crash. In this time window, $\lambda_0 = 0$ is non-degenerate over $1.2 \leq \epsilon \leq 2.0$, while λ_1 changes most rapidly between $\epsilon = 1.6$ and $\epsilon = 1.8$. At $\epsilon = 1.2$, $\lambda_1 = 2.979$, 651 of the stocks are in the major cluster, while the minor cluster contains the 4 stocks shown in Supplementary Table G1. At this filtration parameter, there are no components close to zero. When we go to $\epsilon = 1.4$, $\lambda_1 = 25.758$, we find three sub-distributions of components. The first sub-distribution, containing 625 components, represents the major cluster. The second sub-distribution, containing 18 components close to zero, represents either the neck, or a bridging cluster. The third sub-distribution, containing the 12 components shown in Supplementary Table G1, represents the minor cluster. At $\epsilon = 1.6$, $\lambda_1 = 106.97$, the sub-distribution centered about zero becomes well defined. Nevertheless, there are 5 components remaining in the minor cluster, as shown in Supplementary Table G1. Here, we see that the minor cluster for $\epsilon = 1.6$ is completely embedded in the minor cluster for $\epsilon = 1.4$. Finally, when $\epsilon = 1.8$, $\lambda_1 = 563$, nearly all components are close to zero, but remnants of the major cluster (136 components) and minor cluster (6 components) can still be seen.

Just like the 22 September 2019–22 March 2020 time window, in this time window there is also a pair of nearly degenerate eigenvalues λ_1 and λ_2. Unlike for the 22 September 2019–22 March 2020 time window, where the near degeneracy occurs only at $\epsilon = 1.4$, in the 15 October 2019–15 April 2020 time window this near degeneracy can be seen for $1.4 \leq \epsilon \leq 1.8$. At $\epsilon = 1.4$, $\lambda_1 = 25.758$, $\lambda_2 = 34.013$, we see from Figure 13(right) 621 narrowly distributed components associated with the major cluster in $r_i = u_{1,i}/u_{2,i}$. Of the remaining ratios, 11 have absolute values close to zero, and may be associated with the bridging cluster, while if we set the threshold to $r_i > 1.0$, we find 18 neck components. λ_1 and λ_2 are also quasi-degenerate at $\epsilon = 1.6$ and $\epsilon = 1.8$, but the bridge and neck components identified from these two filtration parameters are different from those identified at $\epsilon = 1.4$.

Finally, let us analyze Fiedler vectors in the 1 April 2020–30 September 2020 time window, which has no overlap with the March 2020 TWSE crash. As we can see from Figure 11d, $\lambda_0 = 0$ is non-degenerate for $1.2 \leq \epsilon \leq 2.0$, while λ_1 changes most rapidly between $\epsilon = 1.4$ and $\epsilon = 1.6$. At $\epsilon = 1.2$, $\lambda_1 = 4.882$, we see in Figure 12m that there is a single distribution of Fiedler components. When $\epsilon = 1.4$, $\lambda_1 = 110.37$, we see from Figure 12n that there are now two sub-distributions of Fiedler components. The first represents a major cluster, while the second (shown in Supplementary Table G1), extending from zero to 0.005, probably includes both the neck and the minor cluster. When $\epsilon = 1.6$, the larger sub-distribution of Fiedler components is the one about zero (Figure 12o), even though the remnant sub-distribution associated with the major cluster is still sizeable. The sub-distribution of the minor cluster (shown in Supplementary Table G1) overlaps with that of the neck, making it difficult to isolate. Finally, at $\epsilon = 1.8$, we find a narrow sub-distribution of Fiedler components about zero (Figure 12p), and two weak sub-distributions away from zero. The latter represent remnants of the major and minor clusters.

4.4. September 2018 TWSE Mini-Crash

After the detailed analyses shown in Section 4.4, the natural question that comes to mind is how much of what we have found there is universal, i.e., applies to all market crashes, and how much of these are peculiar to the March 2020 TWSE crash. To answer this question, we repeated our spectral sequence and Fiedler vector analyses for two other market crashes. The first such crash is the September 2018 TWSE mini-crash in this section, so that we can ascertain universal features of market crashes over at least two crashes on the TWSE. The second such crash is the March 2020 SGX COVID-19 crash in Section 4.6, and also Section 4.6 so that we can confirm universal features of the same market crash (COVID-19 crash) at least across three different markets.

To do this, let us start with the gross features seen in the spectral sequences in Figure 15. From Supplementary Figure D1 in the Supplementary Information, we see that the spectral sequences of the first three time windows and the last four time windows resemble that from a single cluster of data points, whose spectral sequence is characterized by the absence of persistent gaps. These suggest that the TWSE was in a *gapless* normal phase prior to the September 2018 mini-crash, and returned to the normal phase after the mini-crash. For time windows overlapping the mini-crash, the spectral sequences are characterized by persistent gaps at $\epsilon = 1.4$ and/or $\epsilon = 1.6$. The persistent gaps (appearing over a broad range of ϵ) for this mini-crash appear to be weaker than the ones seen for the COVID-19 crash, but they are qualitatively similar. Therefore, a gapped spectral sequence appears to be a universal feature associated with a *gapped* market crash phase, even though the strength of the gap may vary from crash to crash. A closely related (dilation) universal feature that we can identify from the spectral sequences of the two TWSE crashes is the filtration parameter value ϵ at which λ_1 changes most rapidly. For both crashes, λ_1 rises sharply between $\epsilon = 1.2$ and $\epsilon = 1.4$ in the normal phase, but is delayed till to between $\epsilon = 1.6$ and $\epsilon = 1.8$ in the crash phase. Moreover, during the mini-crash, which lasted about four months according to the number of spectral sequences with persistent gaps (agreeing with the "U"-shaped feature seen in Figure 10b), the most pronounced change occurred when we go from the 1 October 2018–28 February 2019 time window to the

8 October 2018–8 March 2019 time window, where the clear gap at $\epsilon = 1.6$ seen in the former completely disappeared in the latter. Therefore, unlike the COVID-19 crash, where the transition into the crash phase is sharp but not the transition out of the crash, the September 2018 TWSE mini-crash shows the opposite behavior, whereby the transition into the crash phase is not sharp, but the transition out of the crash is.

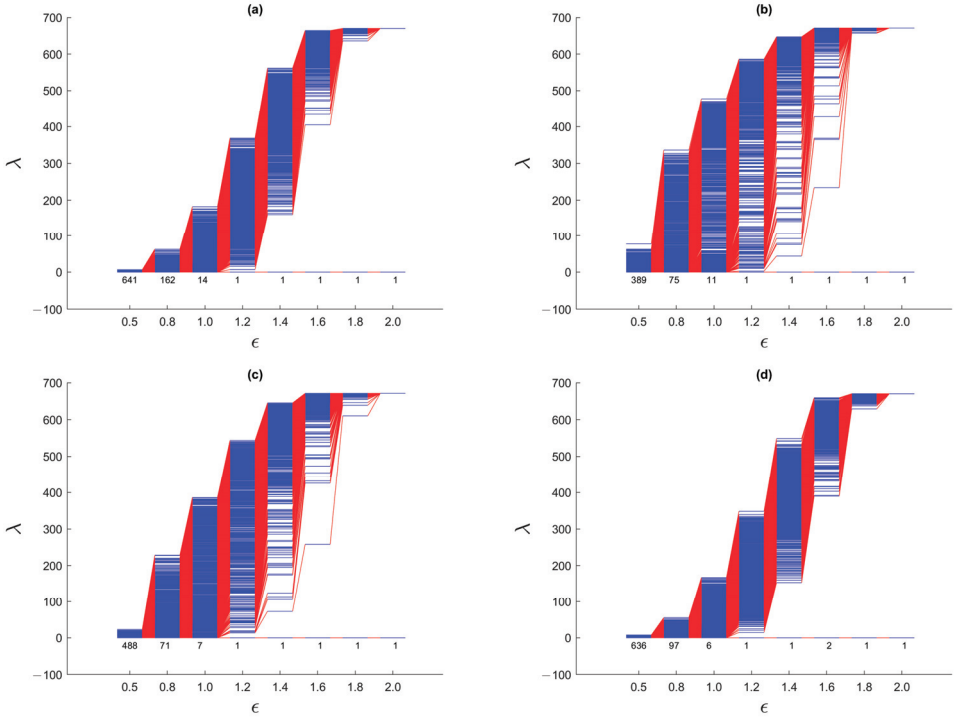


Figure 15. The spectral sequences of the TWSE for $\epsilon = 0.5, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0$ over the six-month time windows: (**a**) 8 April 2018–8 October 2018, (**b**) 22 August 2018–22 February 2019, (**c**) 1 October 2018–15 April 2019, and (**d**) 1 November 2018–30 April 2019. During the September 2018 TWSE mini-crash, the TAIEX fell from a high of 11,006 on 16 September 2018 to a low of 9489 on 21 October 2018. The TAIEX remained low, reaching 9382 on 30 December 2018, before it started rising again.

We also analyzed the Fiedler components at different ϵ over the four selected time windows associated with the September 2018 TWSE mini-crash. Their histograms are shown in Figure 16. As in Figure 13, we find the same evolution from bi-modal to unimodal distributions of Fiedler components as we increase ϵ. Just as for the March 2020 TWSE COVID-19 crash, the Fiedler vector points to the existence of a major cluster, comprising nearly all the stocks in the TWSE, and a minor cluster. The sub-distribution of Fiedler components associated with this minor cluster is weak in the time windows before and after the crash, and strong in the time windows overlapping the crash. However, unlike in the March 2020 COVID-19 crash, where λ_1 and λ_2 are quasi-degenerate at $\epsilon = 1.4$, suggesting the presence of two necks and necessitating the use of the ratio $r_i = u_{1,i}/u_{2,i}$ to identify a bridging cluster between the major and minor clusters, for the September 2018 mini-crash λ_1 and λ_2 are clearly different in the two time windows containing the crash, at $\epsilon = 1.4$. There is also a tri-modal feature in the distribution of Fiedler components at $\epsilon = 1.6$ (Figure 16o) in the time window right after the September 2018 mini-crash. We do not understand the meaning behind this feature, which is also absent from other

time windows of the September 2018 mini-crash, and all time windows of the March 2020 COVID-19 crash.

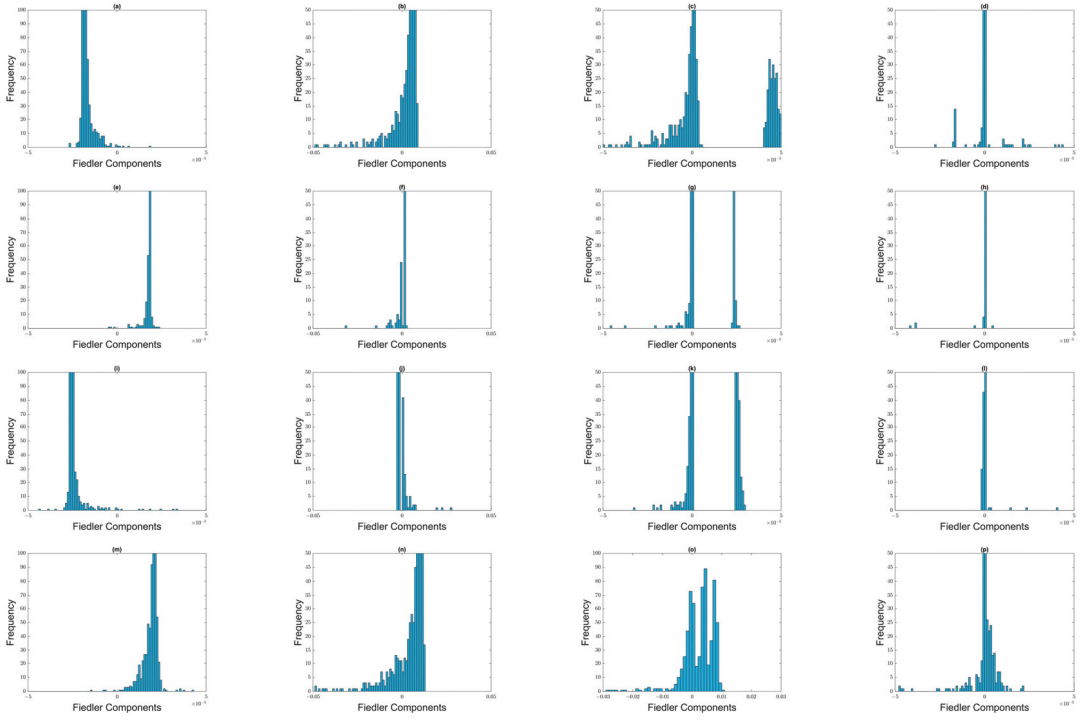


Figure 16. Distribution of Fiedler components at different filtration parameters: (first row) (**a**) $\epsilon = 1.2$, (**b**) $\epsilon = 1.4$, (**c**) $\epsilon = 1.6$, and (**d**) $\epsilon = 1.8$ for the 8 April 2018–8 October 2018 time window; (second row) (**e**) $\epsilon = 1.2$, (**f**) $\epsilon = 1.4$, (**g**) $\epsilon = 1.6$, and (**h**) $\epsilon = 1.8$ for the 22 August 2018–22 February 2019 time window; (third row) (**i**) $\epsilon = 1.2$, (**j**) $\epsilon = 1.4$, (**k**) $\epsilon = 1.6$, and (**l**) $\epsilon = 1.8$ for the 1 October 2018–1 March 2019 time window; and (fourth row) (**m**) $\epsilon = 1.2$, (**n**) $\epsilon = 1.4$, (**o**) $\epsilon = 1.6$, and (**p**) $\epsilon = 1.8$ for the 1 November 2018–1 April 2019 time window.

4.5. March 2020 SGX Crash

For the SGX, we computed spectral sequences for 69 time windows in total, and show these as Supplementary Figure E1 in the Supplementary Information. We included this many time windows for the SGX, because the COVID-19 crash on this market has a long "U"-shape (see Figure 10a), unlike the short "V"-shaped COVID-19 crash on the TWSE (see Figure 10b). This difference is due to the different COVID-19 pandemic trajectories in the two regions: where Taiwan managed to keep COVID-19 at bay over nearly the whole of 2020 (hence the "V"-shaped crash), Singapore succumbed to the pandemic and had to enact strict population health measures starting April 2020 (hence the "U"-shaped crash). To avoid missing the actual recovery from the crash, we made sure that our spectral sequences cover the whole "U"-shaped period. Out of these time windows, the first six and the last 39 spectral sequences are reminiscent of the spectral sequence of a single cluster. This finding on the SGX further supports our universality hypothesis that the gapless normal phase of a stock market consists of a single undifferentiated cluster, based on our findings on the TWSE in Section 4.4 and Section 4.5. The remaining 24 spectral sequences were found to be gapped, suggesting that these 24 time windows overlapped with the March 2020 SGX COVID-19 crash. Based on the Straits Times Index (STI), the SGX reached a high on 9 February 2020, but according to the spectral sequences the crash only started on or

after 8 March 2020, and surprisingly returned to normal on or after 15 March 2020. This tells us that the spectral sequence method is sensitive to the difference between the normal and crash phases of a stock market, and can therefore be used to time the start and end of crashes, instead of using the index value.

The dilation feature seen during the September 2018 and March 2020 TWSE crashes is even more pronounced during the March 2020 SGX crash. This supports our hypothesis that this dilation feature, which is closely associated with the opening of the spectral gap, is universal across markets. Just as for the TWSE, λ_1 changes most sharply between $\epsilon = 1.4$ and $\epsilon = 1.6$ in the normal phase, but between $\epsilon = 1.6$ and $\epsilon = 1.8$ in the crash phase of the SGX. In fact, in Figure 17c, λ_1 changes most sharply between $\epsilon = 1.8$ and $\epsilon = 2.0$. During the SGX COVID-19 crash, we found non-persistent gaps appearing at ϵ = 1.4, 1.6, and 1.8, which suggest that the size distribution of large persistent clusters is discrete, just like it was for TWSE. The main difference between the TWSE and the SGX, is the SGX having between 3 and 6 zero eigenvalues at $\epsilon = 1.2$.

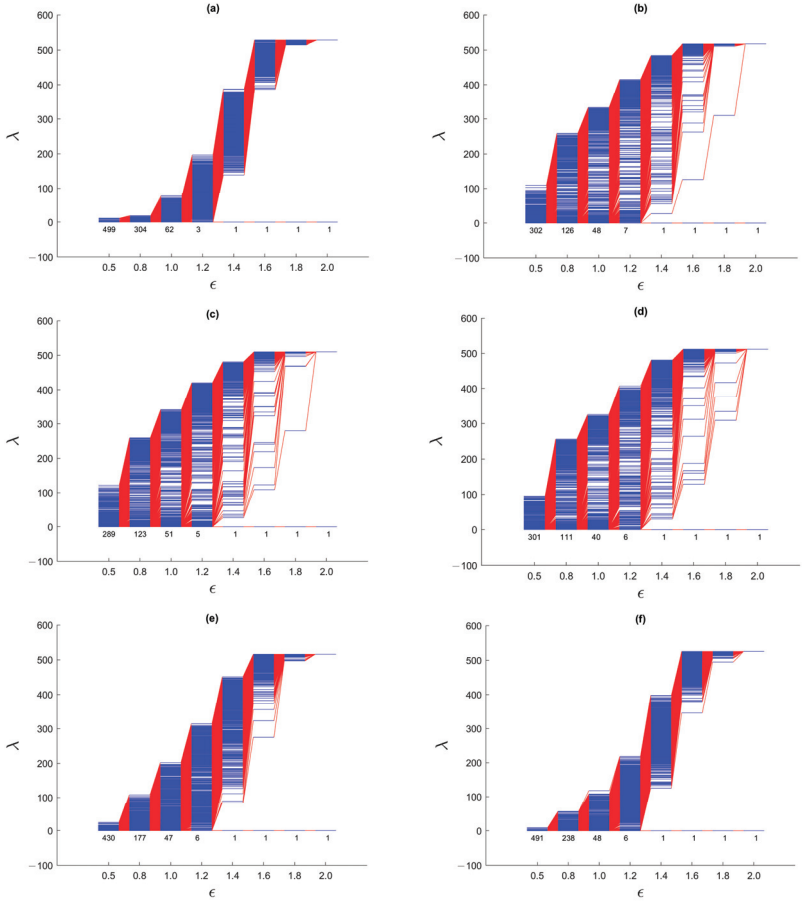


Figure 17. The spectral sequences of the TWSE for $\epsilon = 0.5, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0$ over the six-month time windows: (**a**) 1 August 2019–31 December 2019, (**b**) 8 October 2019–8 March 2020, (**c**) 8 November 2019–8 April 2020, and (**d**) 22 February 2020–22 August 2020, and (**e**) 8 March 2020–8 September 2020, and (**f**) 8 April 2020–8 October 2020. During the March 2020 STI crash, the STI fell from a high of 3220 on 9 February 2020 to a low of 2410 on 15 March 2020.

Moving on, we showed the histograms of Fiedler components over the March 2020 SGX COVID-19 crash in Figure 18. For the (first row) 1 August 2019–31 December 2019 and (last row) 8 April 2020–8 October 2020 time windows, the distributions of Fiedler components are mostly unimodal, except for $\epsilon = 1.6$. This agrees with our observation for both crashes on the TWSE. For the (second row) 8 October 2019–8 March 2020, (third row) 8 November 2019–8 April 2020, (fourth) 22 February 2020–22 August 2020, and (fifth row) 8 March 2020–8 September 2020 time windows, the distributions of Fiedler components are bimodal, even up to $\epsilon = 1.8$. Again, this agrees with our observation for both crashes on the TWSE.

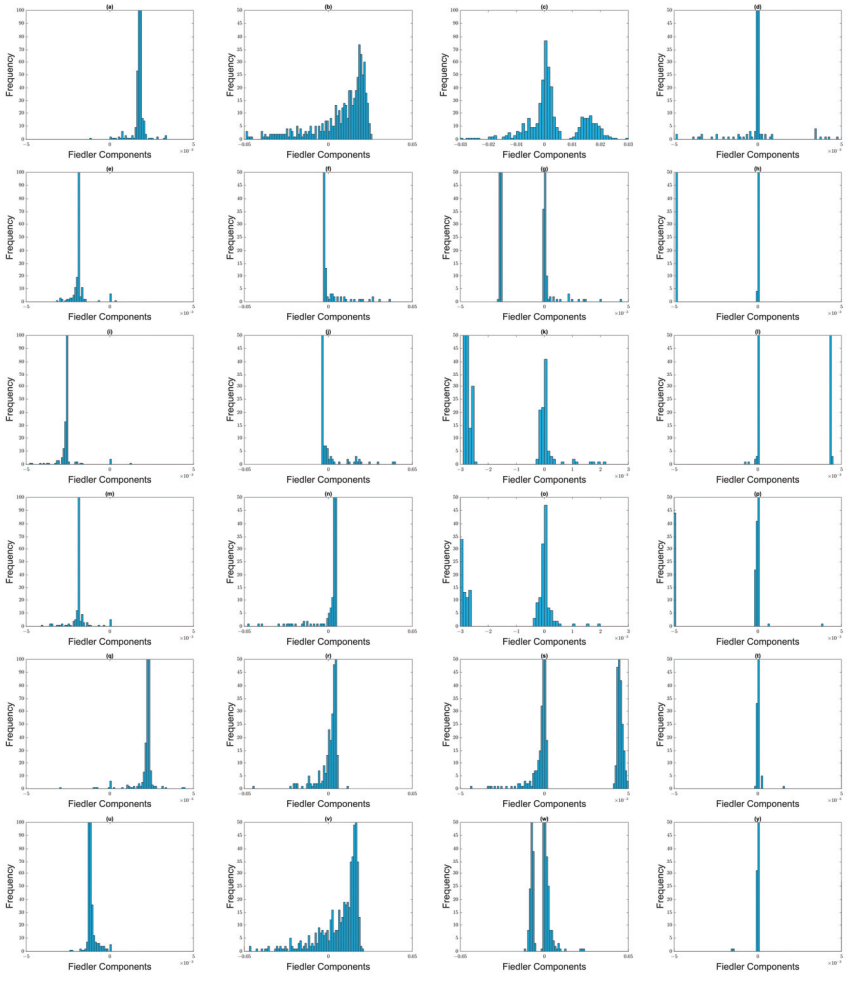


Figure 18. Distribution of Fiedler components at different filtration parameters: (first row) (**a**) $\epsilon = 1.2$, (**b**) $\epsilon = 1.4$, (**c**) $\epsilon = 1.6$, and (**d**) $\epsilon = 1.8$ for the 1 August 2019–1 January 2020 time window; (second row) (**e**) $\epsilon = 1.2$, (**f**) $\epsilon = 1.4$, (**g**) $\epsilon = 1.6$, and (**h**) $\epsilon = 1.8$ for the 8 October 2019–8 April 2020 time window; (third row) (**i**) $\epsilon = 1.2$, (**j**) $\epsilon = 1.4$, (**k**) $\epsilon = 1.6$, and (**l**) $\epsilon = 1.8$ for the 8 November 2019–8 May 2020 time window; (fourth row) (**m**) $\epsilon = 1.2$, (**n**) $\epsilon = 1.4$, (**o**) $\epsilon = 1.6$, and (**p**) $\epsilon = 1.8$ for the 22 February 2020–22 August 2020 time window; (fifth row) (**q**) $\epsilon = 1.2$, (**r**) $\epsilon = 1.4$, (**s**) $\epsilon = 1.6$, and (**t**) $\epsilon = 1.8$ for the 8 March 2020–8 September 2020 time window; and (sixth row) (**u**) $\epsilon = 1.2$, (**v**) $\epsilon = 1.4$, (**w**) $\epsilon = 1.6$, and (**y**) $\epsilon = 1.8$ for the 8 April 2020–8 October 2020 time window.

4.6. March 2020 S&P500 Crash

In addition to emerging markets such as TWSE and SGX, we also investigated the component stocks of S&P 500 from 1 June 2019 to 31 December 2020. Our target is again the COVID-19 crash, which occurred between 1 and 8 March 2020 in the S&P 500 according to the spectral sequences shown in Supplementary Figure F1, compared to 1–15 March 2020 in TWSE and 8–15 March 2020 in SGX. Compared to the March 2020 TWSE crash (whose beginning was sharp, but whose ending was not) and the March 2020 SGX crash (whose beginning and ending were both sharp), the beginning and end of the March 2020 S&P 500 crash were both not sharp. In fact, according to conventional indicators, the S&P 500 attained a high of 3380 on 14 February 2020, and a low of 2304 on 20 March 2020.

As expected, the spectral sequence of the S&P 500 stocks is gapless in the normal phase, and gapped in the crash phase. This suggests strongly that the existence of a persistent spectral gap distinguishes the crash phase from the normal phase, whether or not the stock market is emerging or mature. The difference between the S&P 500 (measuring the mature US markets) and the TWSE/SGX is the extent of the persistent spectral gap (going into smaller length scales) in the spectral sequences of the S&P 500, compared to those in TWSE and SGX. For the S&P 500, there is also a persistent description in terms of two clusters. We suspect this is because of the significant fraction of S&P 500 component stocks are traded on National Association of Securities Dealers Automated Quotations (NASDAQ), while the remainder are traded on the New York Stock Exchange (NYSE).

Next, we show the histograms of Fiedler components over the February 2020 S&P 500 COVID-19 crash in Figure 19. In agreement with what we found in the SGX and TWSE, the distributions of Fiedler components go from bimodal to unimodal as we increase ϵ. This transition coincides with λ_1 changing most rapidly with ϵ.

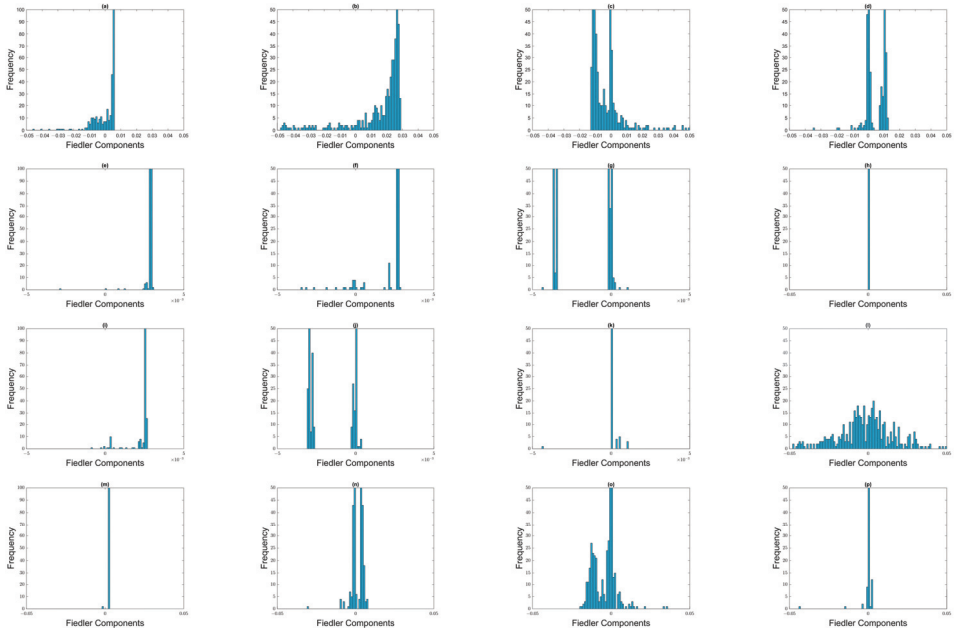


Figure 19. Distribution of Fiedler components at different filtration parameters: (first row) (**a**) $\epsilon = 1.2$, (**b**) $\epsilon = 1.4$, (**c**) $\epsilon = 1.6$, and (**d**) $\epsilon = 1.8$ for the 1 June 2019–30 November 2019 time window; (second row) (**e**) $\epsilon = 1.2$, (**f**) $\epsilon = 1.4$, (**g**) $\epsilon = 1.6$, and (**h**) $\epsilon = 1.8$ for the 08 September 2019–08 March 2020 time window; (third row) (**i**) $\epsilon = 1.2$, (**j**) $\epsilon = 1.4$, (**k**) $\epsilon = 1.6$, and (**l**) $\epsilon = 1.8$ for the 15 February 2020–15 August 2020 time window; and (fourth row) (**m**) $\epsilon = 1.2$, (**n**) $\epsilon = 1.4$, (**o**) $\epsilon = 1.6$, and (**p**) $\epsilon = 1.8$ for the 22 June 2020–22 December 2020 time window.

5. Conclusions

In summary, we explained using a simple raindrop analogy the concept of persistent structures, and why they are useful as mesoscopic variables for describing the dynamics of complex systems (for example, market crashes) in terms of continuous and discontinuous changes. We then drew inspiration from the connection between (1) the symmetries $[A, H] = 0$ of a quantum system, and (2) the block-diagonal structure of the Hamiltonian, leading ultimately to (3) the organization of energy eigenvalues into bands separated by band gaps, to approach the problem of overlapping communities obtained during the filtration process in TDA. Instead of trying to identify such communities in real space, we should therefore look for signatures of community structure in spectral space. For this work, the graph Laplacian $L = K - A$ (A being the adjacency matrix, and K being the (diagonal) degree matrix) plays the role of the Hamiltonian.

To check its feasibility, we tested the spectral approach on a series of toy models, from a single cluster of data points to two or more well-defined clusters of data points, characterized by gaps of different length scales, in the absence or presence of a noisy background. We then introduced the spectral sequence as a novel tool to visualize how the Laplacian spectra of different ϵ change over increasing filtration parameter ϵ. For a single cluster of data points, the spectral sequence is gapless, whereas for multiple well-defined clusters the spectral sequence contain gaps that persist over a wide range of ϵ. Connecting the real-space TDA and the analysis of Laplacian spectra, we proposed for these persistent gaps to be used as signatures of persistent structures in the data. Spectral gaps that are persistent with respect to changes in length scale tend to be persistent with respect to changes in time, and are robust with respect to background noise. We also analyzed the Fiedler vector $\vec{u}_1$ associated with the first non-zero Laplacian eigenvalue $\lambda_1 > 0$, which is well-known to contain information on community structure in the data. In the case of two merging clusters, we confirmed earlier studies that their components have different signs in $\vec{u}_1$, but within each cluster, components have roughly the same value. We also developed a new understanding that components with significantly smaller (or even zero) absolute magnitudes are members of the neck, a structure that must be considered as distinct from the clusters it connects. We also understood for the first time how there can be near degeneracy between $\lambda_1 > 0$ and $\lambda_2 \approx \lambda_1$, when three clusters are arranged in a linear configuration, with two necks forming roughly around the same length scales. Members of the bridging cluster can be distinguished from members of the two necks by examining the ratios of their components in $\vec{u}_1$ and $\vec{u}_2$.

Finally, we tested this spectral approach to unravel persistent structures on the daily prices of 671 stocks of the TWSE, 530 stocks of the SGX, and 504 component stocks of the S&P 500. Based on the toy model studies, we realized that this approach is ideal for analyzing topological and geometrical changes in stock markets when they crash (fragmentation), and also when they recover (agglomeration). Therefore, we identified two time windows (1 April 2018 to 30 April 2019, and 1 August 2019 to 30 September 2020) associated with two crashes on the TWSE, one time window (1 August 2019 to 30 April 2021) associated with the crash on the SGX, and one time window (1 June 2019 to 31 December 2020) associated with the crash on the S&P 500. We then computed the Pearson cross correlations C_{ij} between stocks in six-month windows, converted C_{ij} into pairwise distances D_{ij} for the TDA filtration process, before sliding the time window one week at a time. We found universally across market crashes and stock markets that (1) the spectral sequence is gapless (absence of persistent or non-persistent gaps) when the time window is entirely within the normal phase, (2) the spectral sequence is gapped (presence of persistent gaps at large length scales) when the time window overlaps with the market crash phase, (3) the most rapid change in λ_1 is delayed in the crash phase relative to the normal phase, (4) in the normal phase, the distribution of Fiedler components is predominantly uni-modal, (5) in the crash phase, the distribution of Fiedler components change from bi-modal to uni-modal at the filtration parameter where λ_1 changes most rapidly. These are all results not previously known.

Together, our spectral analyses of toy models and real-world stock market data suggests that two clusters A and B do not become a single cluster AB the moment they are linked by a neck, but continue to retain their distinct identities until their members are completely absorbed by the growing neck. This can be summarized by the fusion process $A + B \to A + n + B \to a + N + b \to \mathcal{N}(= C)$. Within this new perspective, $n \to N \to \mathcal{N}$ represents the thickening of the neck, while $A \to a$ and $B \to b$ represent the absorption of A and B by the neck. This ternary fusion picture is useful regardless of whether the fusion is a result of increasing the length scale during the filtration process, or a result of interactions that bring the two clusters closer to each other over time. In terms of this ternary fusion process, we can explain many mysteries observed in real-world data that cannot be explained using only binary fusion processes.

Finally, we explored only the graph Laplacian, its spectrum, and its eigenvectors in this paper. However, we know that the graph Laplacian is the simplest member of a hierarchy of Hodge Laplacians, which we plan to explore in our future works.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/e25060846/s1, Section A: Spectral Sequences of Additional Toy Models; Section B: Fiedler Eigenvector Analysis of Additional Toy Models; Section C: Spectral Sequences for Mar 2020 TWSE Crash; Section D: Spectral Sequences for Sep 2018 TWSE Mini-Crash; Section E: Spectral Sequences for Mar 2020 SGX Crash; Section F: Spectral Sequences for Jan 2020 S&P 500 Crash; Section G: Neck and Bridging Components for the Mar 2020 TWSE COVID-19 Crash.

Author Contributions: Conceptualization, K.X. and S.A.C.; Methodology, P.T.-W.Y., K.X. and S.A.C.; Formal analysis, P.T.-W.Y., K.X. and S.A.C.; Writing—original draft, P.T.-W.Y. and S.A.C.; Writing—review & editing, P.T.-W.Y., K.X. and S.A.C.; Visualization, P.T.-W.Y. and S.A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All Python and Matlab scripts are available at https://doi.org/10.21979/N9/UCEELS, along with instructions on how to use them. This include reading the processed market data and perform the necessary computations to give the final results.

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

Abbreviations

TWSE	Taiwan Stock Exchange
SGX	Singapore Stock Exchange
TDA	topological data analysis
RCA	Ricci curvature analysis
MST	minimal spanning tree
PMFG	planar maximally filtered graph
CM	center of mass
STI	Straits Times Index
TAIEX	Taiwan Capitalization Weighted Stock Index
S&P 500	Standard & Poor's 500
NYSE	New York Stock Exchange
NASDAQ	National Association of Securities Dealers Automated Quotations

References

1. Akutsu, T.; Hayashida, M.; Ching, W.-K.; Ng, M.K. Control of Boolean networks: Hardness results and algorithms for tree structured networks. *J. Theor. Biol.* **2007**, *244*, 670–679. [CrossRef] [PubMed]
2. Kauffman, S.A. *The Origins of Order: Self-Organization and Selection in Evolution*; Oxford University Press: New York, NY, USA, 1993.

3. Galas, D.J.; Sakhanenko, N.A.; Skupin, A.; Ignac, T. Describing the complexity of systems: Multivariable "set complexity" and the information basis of systems biology. *J. Comput. Biol.* **2014**, *21*, 118–140. [CrossRef] [PubMed]
4. Lee, J.A.; Verleysen, M. *Nonlinear Dimensionality Reduction*; Springer: Berlin/Heidelberg, Germany, 2007; Volume 1.
5. Ma, Y.; Fu, Y. *Manifold Learning Theory and Applications*; CRC Press: Boca Raton, FL, USA, 2012; Volume 434.
6. D'Addese, G.; Casari, M.; Serra, R.; Villani, M. A fast and effective method to identify relevant sets of variables in complex systems. *Mathematics* **2021**, *9*, 1022. [CrossRef]
7. Villani, M.; Roli, A.; Filisetti, A.; Fiorucci, M.; Poli, I.; Serra, R. The search for candidate relevant subsets of variables in complex systems. *Artif. Life* **2015**, *21*, 412–431. [CrossRef]
8. Yen, P.T.-W.; Cheong, S.A. Using topological data analysis (TDA) and persistent homology to analyze the stock markets in Singapore and Taiwan. *Front. Phys.* **2021**, *9*, 20. [CrossRef]
9. Yen, P.T.-W.; Xia, K.; Cheong, S.A. Understanding Changes in the Topology and Geometry of Financial Market Correlations during a Market Crash. *Entropy* **2021**, *23*, 1211. [CrossRef]
10. Medina-Mardones, A.M.; Rosas, F.E.; Rodríguez, S.E.; Cofré, R. Hyperharmonic analysis for the study of high-order information-theoretic signals. *J. Phys. Complex* **2021**, *2*, 035009. [CrossRef]
11. Haken, H. *The Science of Structure: Synergetics*; Van Nostrand Reinhold: Washington, DC, USA, 1984.
12. Haken, H. *Synergetics: An Introduction Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology*; Springer: Berlin, Germany, 2012.
13. Mantegna, R.N. Hierarchical structure in financial markets. *Eur. Phys. J. B* **1999**, *11*, 193–197. [CrossRef]
14. Onnela, J.-P.; Chakraborti, A.; Kaski, K.; Kertiész, J. Dynamic asset trees and portfolio analysis. *Eur. Phys. J. B* **2002**, *30*, 285–288. [CrossRef]
15. Bonanno, G.; Caldarelli, G.; Lillo, F.; Mantegna, R.N. Topology of correlation-based minimal spanning trees in real and model markets. *Phys. Rev. E* **2003**, *68*, 046130. [CrossRef]
16. Miccichè, S.; Bonanno, G.; Lillo, F.; Mantegna, R.N. Degree stability of a minimum spanning tree of price return and volatility. *Phys. A Stat. Mech.* **2003**, *324*, 66–73. [CrossRef]
17. Onnela, J.-P.; Chakraborti, A.; Kaski, K.; Kertesz, J.; Kanto, A. Dynamics of market correlations: Taxonomy and portfolio analysis. *Phys. Rev. E* **2003**, *68*, 056110. [CrossRef] [PubMed]
18. Zhang, Y.; Lee, G.H.T.; Wong, J.C.; Kok, J.L.; Prusty, M.; Cheong, S.A. Will the US economy recover in 2010? A minimal spanning tree study. *Phys. A Stat. Mech.* **2011**, *390*, 2020–2050. [CrossRef]
19. Cheong, S.A.; Fornia, R.P.; Lee, G.H.T.; Kok, J.L.; Yim, W.S.; Xu, D.Y.; Zhang, Y. The Japanese economy in crises: A time series segmentation study. *Economics* **2012**, *6*. [CrossRef]
20. Tumminello, M.; Aste, T.; Di Matteo, T.; Mantegna, R.N. A tool for filtering information in complex systems. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 10421–10426. [CrossRef]
21. Aste, T.; Shaw, W.; Di Matteo, T. Correlation structure and dynamics in volatile markets. *New J. Phys.* **2010**, *12*, 085009. [CrossRef]
22. Song, D.-M.; Tumminello, M.; Zhou, W.-X.; Mantegna, R.N. Evolution of worldwide stock markets, correlation structure, and correlation-based graphs. *Phys. Rev. E* **2011**, *84*, 026108. [CrossRef]
23. Pozzi, F.; Di Matteo, T.; Aste, T. Spread of risk across financial markets: Better to invest in the peripheries. *Sci. Rep.* **2013**, *3*, 1665. [CrossRef]
24. Massara, G.P.; Di Matteo, T.; Aste, T. Network filtering for big data: Triangulated maximally filtered graph. *J. Complex Netw.* **2017**, *5*, 161–178. [CrossRef]
25. Donath, W.E.; Hoffman, A.J. Lower bounds for the partitioning of graphs. *IBM J. Res. Dev.* **1973**, *17*, 420–425. [CrossRef]
26. Fiedler, M. Algebraic connectivity of graphs. *Czechoslov. Math. J.* **1973**, *23*, 298–305. [CrossRef]
27. Spielman, D.A.; Teng, S.-H. Spectral partitioning works: Planar graphs and finite element meshes. In Proceedings of the 37th Conference on Foundations of Computer Science, Burlington, VT, USA, 14–16 October 1996; pp. 96–105.
28. Von Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **2007**, *17*, 395–416. [CrossRef]
29. Arenas, A.; Diaz-Guilera, A.; Pérez-Vicente, C.J. Synchronization reveals topological scales in complex networks. *Phys. Rev. Lett.* **2006**, *96*, 114102. [CrossRef] [PubMed]
30. Chung, F.R. *Spectral Graph Theory*; American Mathematical Soc.: Providence, RI, USA, 1997; Volume 92.
31. Cvetković, D.M.; Doob, M.; Sachs, H. *Spectra of Graphs: Theory and Application*; Academic Press: Cambridge, MA, USA, 1980.
32. Cheeger, J. *A Lower Bound for the Smallest Eigenvalue of the Laplacian, Problems in Analysis (Papers Dedicated to Salomon Bochner, 1969)*; Princeton University Press: Princeton, NJ, USA, 1970.
33. Alon, N.; Spencer, J.H. *The Probabilistic Method*; John Wiley & Sons: Hoboken, NJ, USA, 2016.
34. Dodziuk, J. Difference equations, isoperimetric inequality and transience of certain random walks. *Trans. Am. Math. Soc.* **1984**, *284*, 787–794. [CrossRef]
35. Hoory, S.; Linial, N.; Wigderson, A. Expander graphs and their applications. *Bull. Am. Math. Soc.* **2006**, *43*, 439–561. [CrossRef]
36. Fiedler, M. Laplacian of graphs and algebraic connectivity. *Banach Cent. Publ.* **1989**, *1*, 57–70. [CrossRef]
37. Capocci, A.; Servedio, V.D.; Caldarelli, G.; Colaiori, F. Detecting communities in large networks. *Phys. A Stat. Mech.* **2005**, *352*, 669–676. [CrossRef]

38. Servedio, V.; Colaiori, F.; Capocci, A.; Caldarelli, G. Community structure from spectral properties in complex networks. In Proceedings of the AIP Conference Proceedings, Aveiro, Portugal, 29 August–2 September 2004; American Institute of Physics: College Park, MD, USA, 2005; pp. 277–286.
39. Kao, S.-C.; Hsu, C. Virus Jitters Send TAIEX Plummeting. Available online: https://www.taipeitimes.com/News/biz/archives/2021/05/13/2003757292 (accessed on 9 January 2023).
40. Lee, K.-J.; Lu, S.-L. The impact of COVID-19 on the stock price of socially responsible enterprises: An empirical study in Taiwan stock market. *Int. J. Environ. Res. Public Health* **2021**, *18*, 1398. [CrossRef]

entropy

MDPI

Article

A Goodwin Model Modification and Its Interactions in Complex Networks

Francisco Yáñez Rodríguez [1] and Alberto P. Muñuzuri [1,2,*]

[1] Group of NonLinear Physics, University of Santiago de Compostela, 15706 Santiago de Compostela, Spain; francisco.yanez@rai.usc.es
[2] Galician Center for Mathematical Research and Technology (CITMAga), 15782 Santiago de Compostela, Spain
* Correspondence: alberto.perez.munuzuri@usc.es

Abstract: The global economy cannot be understood without the interaction of smaller-scale economies. We addressed this issue by considering a simplified economic model that still preserves the basic features, and analyzed the interaction of a set of such economies and the collective emerging dynamic. The topological structure of the economies' network appears to correlate with the collective properties observed. In particular, the strength of the coupling between the different networks as well as the specific connectivity of each node happen to play a crucial role in the determination of the final state.

Keywords: econophysics; nonlinear interactions; dynamical systems; complex networks

1. Introduction

Much attention has been devoted to the synergetic behaviors of sets of coupled dynamical systems and their nonlinear interactions. Examples are widely observed, ranging from biology [1–4], chemistry [5–10], social systems [11,12], and, of course, economy [13–15]. In all these cases, the collective phenomena observed are more than just the addition of the individuals, and new collective dynamics emerge.

The structure of the network of connections has shown to be determinant in the collective phenomena observed. With this manuscript, we plan to show the role played by the network topology on the synergetic properties observed in the economic models. For that, we considered a set of simple economic models coupled.

Although there are many models that describe different aspects of an economy, such as the Harrod–Domar model [16,17], the Solow–Swan model [18], and the Philips curve model [19], one of the most known models in economics is the Goodwin model [20], which focuses on predicting salary and employment. This model is based on the very-well-known Lotka–Volterra model, used to predict the relation between prey and predators [21] in population dynamic studies. In the Goodwin model, salary plays the role of the predator whereas employment is the prey, exhibiting periodic behaviors. This model combines many of the properties of some of the previous models [22] and its equations exhibit a relatively simple dynamic so that the outcome of coupling several such economies becomes more apparent.

This simple model has been extended in many different contributions (a summary of these can be found in [23–25]) aiming to expand its applicability by considering different scenarios. In these extensions, different dynamics were included and even empirical data seem to be described. With this manuscript, we stuck to the original version of the model and only included a minor modification that prevents the appearance of non-physical divergent solutions. Thus, the long-term divergent behaviors (exponential growths) were erased in order to analyze the global dynamics in a medium-to-long-term temporal scale. This guarantees that all the variables do not diverge and also adds more possible solutions and behaviors to the economies predicted. In this way, the focus of this contribution is on the collective phenomena arising as several economies interact together, exchanging

Citation: Rodríguez, F.Y.; Muñuzuri, A.P. A Goodwin Model Modification and Its Interactions in Complex Networks. *Entropy* **2023**, *25*, 894. https://doi.org/10.3390/e25060894

Academic Editor: José F. F. Mendes

Received: 26 March 2023
Revised: 17 May 2023
Accepted: 30 May 2023
Published: 2 June 2023

resources in a network. Note that this is particularly relevant nowadays as globalization has forced economies to interact with each other, but we are not yet in a global world with one single macroeconomy and thus the network structure and strength play a significant role.

The paper is organized as follows. The methods section presents the Goodwin model with the modifications introduced and a description of the dynamical behaviors that can be observed. The second part of the methods section presents the different network topologies that we considered. The results and discussions of our study are presented in the next section and the manuscript is ended with some conclusions.

2. Materials and Methods

2.1. Goodwin Model and Corrections

The original Goodwin model is an attempt to describe the behavior of salary and employment [20,22]. With that purpose, it is necessary to define the following variables:

- $c(t)$ is the capital invested in the production of a good.
- $q(t)$ is the quality of the final product. A linear relation with the capital invested is assumed, so we can write $c(t) = \kappa \cdot q(t)$.
- $n(t)$ is the population involved in the production of a good.
- $p(t)$ is the productivity of employees.
- $w(t)$ is the salary that a single subject receives when working in the production of a good, called unit wage in the following.
- $l(t) \equiv q(t)/p(t)$ is the ratio of quality and productivity. It relates both variables and provides information about the effectiveness of the jobs from the point of view of the producers, which we will name employment level.
- $u(t) \equiv w(t)/p(t)$ is the ratio between salary and productivity. We will refer to it as the salary directly.
- $v(t) \equiv l(t)/n(t)$ is the ratio between the level of employment and the number of employees, or the employment rate.

The set of Equation (1) describes the original model first introduced by Richard Goodwin [20]:

$$\begin{array}{cc} \frac{dp}{dt} = \alpha\, p & \frac{dq}{dt} = q\left(\frac{1-u}{\kappa}\right) \\ \frac{dw}{dt} = w(\rho\, v - \gamma) & \frac{dl}{dt} = l\left(\frac{1-u}{\kappa} - \alpha\right) \\ \frac{dn}{dt} = \delta\, n & \frac{dv}{dt} = v\left(\frac{1-u}{\kappa} - \alpha - \delta\right) \\ \frac{dc}{dt} = q(1-u) & \frac{du}{dt} = u(\rho\, v - \gamma - \alpha) \end{array} \tag{1}$$

As we can see in Equation (1), Richard Goodwin's original model considers that population and productivity exponentially grow with time and thus most of the model variables also tend to infinite for sufficiently long times.

In this set of equations, the variables are those described above, and the parameters introduced are $\rho, \gamma, \kappa, \alpha$, and δ. Note that ρ and γ already introduce some limitations to the salary values as they tend to prevent the exponential growth (this is not fully achieved as the productivity might exponentially grow and thus force the whole system to diverge). The parameters δ and α will determine how fast productivity and population grow, as they control the exponential growth in the equations. On the other hand, ρ and γ are parameters that characterize the time evolution of the unitary salary w and thus the salary u. Finally, ρ and γ quantify the dependence of the employment on the evolution of the salary, playing a similar role to that in the previously mentioned Philips curve [19]. Finally, κ is the relation between capital and quality mentioned above. All these parameters are defined as positive. Some examples of the dynamical behaviors observed are presented in the Supplementary Materials.

To avoid the exponential growths observed in the original Goodwin model, we introduced a second-order correction term in the sense of the logistic equation, i.e., when the values of the variables start to significantly grow, the added terms prevent divergencies by

assuming a limited number of resources. The new set of equations, which will be used in this manuscript, is given by

$$\begin{array}{c} \frac{dn}{dt} = 0 \\ \frac{dp}{dt} = \alpha\, p\left(1 - \frac{p}{\sigma}\right) \\ \frac{dw}{dt} = (w - \beta_u)(\rho\, v - \gamma) \\ \frac{dc}{dt} = (q - \beta_v)(1 - \lambda\, u) \end{array} \tag{2}$$

where the parameters α and σ control the logistic growth, σ determines the value of the productivity compatible with the considered economy, and α is the productivity growth rate. We also introduced the parameters β_u and β_v, which are related to some physical constraints in the salary and employment variables that we will discuss in more details below. λ is another parameter introduced to control the evolution of the capital and avoid divergencies. Finally, ρ and γ take the same role as in the original model. In addition, as in the previous model, all the parameters introduced by us were constrained to positive values. Note that this approach to limiting the exponential growth of the system is the simplest that can be considered, and it just limits the resources at our disposal.

Supposing that the relation between capital and quality is linear ($c(t) = \kappa \cdot q(t)$), the other variables in the model are then given by the following Equation (3):

$$\begin{array}{c} \frac{dq}{dt} = (q - \beta_v)\frac{1-\lambda u}{\kappa} \\ \frac{dl}{dt} = \left(l - \frac{\beta_v}{p}\right)\frac{1-\lambda u}{\kappa} - \alpha l\left(1 - \frac{p}{\sigma}\right) \\ \frac{dv}{dt} = \left(v - \frac{\beta_v}{np}\right)\frac{1-\lambda u}{\kappa} - \alpha v\left(1 - \frac{p}{\sigma}\right) \\ \frac{du}{dt} = \left(u - \frac{\beta_u}{p}\right)(\rho v - \gamma) - \alpha u\left(1 - \frac{p}{\sigma}\right) \end{array} \tag{3}$$

Figure 1 presents numerical simulations of the original Goodwin model (Figure 1a) and the modified model (Figure 1b), both obtained using an explicit 4th-order Runge–Kutta integration scheme [26]. It is possible to observe that the oscillatory behavior remains stable even for long periods of time whereas the original model steadily diverges. Note that the original Goodwin model in Figure 1a presents an exponential growth of the variables as time evolves (this is clearly seen in the orange curve, employment). On the other hand, once the modification is included, the exponential growth is avoided and the oscillatory behavior can be observed for very long periods of time (notice the orange curve in Figure 1b that saturates as expected from Equation (3)).

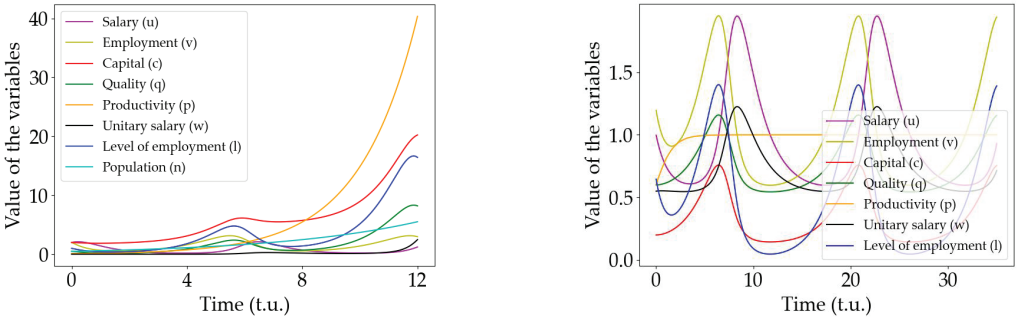


Figure 1. Comparison between the predictions of the original model and our modification (given by Equations (2) and (3)). (**a**) Solution to the equations of the original Goodwin model. Note that the exponential growth of the productivity (the orange curve), makes the oscillations amplitude of the other variables diverge. (**b**) Solution of the modified model. Note that the oscillations remain stable in time due to the logistic behavior of the productivity variable (in color orange).

Note that the oscillatory behavior observed is the same as that predicted in the original Goodwin model; however, in our case, the productivity variable does not diverge and, consequently, the oscillations remain stable, with time allowing for long-time observations and thus interactions between different economies.

From now on, we will focus on the behaviors of the variables describing wages (u) and employment (v) in a case where the productivity is constant with value $p = \sigma$ for simplicity, i.e., the productivity for the time considered remains almost constant. Note that the modified set of equations presents a stable point at $p = \sigma$; thus, we assume that the productivity has reached the expected stationary value and we consider modifications from this moment. Thus, our set of equations becomes

$$\begin{aligned} \frac{du}{dt} &= f(u,v) = \left(u - \frac{\beta_u}{\sigma}\right)(\rho v - \gamma) \\ \frac{dv}{dt} &= h(u,v) = \left(v - \frac{\beta_v}{n\sigma}\right)\frac{1-\lambda u}{\kappa} \end{aligned} \tag{4}$$

2.2. Linear Stability Analysis and Phase Diagrams

In this section and using linear stability analysis [27], we studied the different behaviors that the set of Equation (4) exhibit. The system presents two fixed points:

$$\left(u_0^1 = \rho/\gamma, v_0^1 = 1/\lambda\right); \left(u_0^2 = \beta_u/\sigma, v_0^2 = \beta_v/n\sigma\right) \tag{5}$$

In Equation (5), we recover the fixed point $\left(u_0^1, v_0^1\right)$ that was already present in the original Goodwin model, while $\left(u_0^2, v_0^2\right)$ depends on the new parameters introduced (β_u and β_v) and that were not present in the original model. Analyzing the sign of the eigenvalues e_i of the Jacobian matrix for our system, we can classify the different behaviors of the system. The Jacobian matrix is given by,

$$\left|\mathcal{J}\left(\vec{x}\right) - e\mathbb{I}\right| = \left| \begin{matrix} \frac{\partial f(u,v)}{\partial u} - e & \frac{\partial f(u,v)}{\partial u} \\ \frac{\partial h(u,v)}{\partial u} & \frac{\partial h(u,v)}{\partial v} - e \end{matrix} \right|_{(u_0,v_0)} \tag{6}$$

Thus, the eigenvalues for each fixed point are given by the expressions

$$\begin{aligned} \left(u_0^1, v_0^1\right) &\to e_{1,2}^1 = \pm i\sqrt{\frac{\rho\lambda}{\kappa}\left(\frac{1}{\lambda} - \frac{\beta_u}{\sigma}\right)\left(\frac{\gamma}{\rho} - \frac{\beta_v}{n\sigma}\right)} \\ \left(u_0^2, v_0^2\right) &\to e_1^2 = \frac{\rho\beta_v}{n\sigma} - \gamma; e_2^2 = \frac{\sigma - \lambda\beta_u}{\sigma\kappa} \end{aligned} \tag{7}$$

Analyzing the results of Equation (7), we notice that β_u and β_v determine the sign of both eigenvalues for each fixed point, resulting in four different behaviors. Note that the sign of the eigenvalues determines the stability of each fixed point and thus the dynamics of the solutions around it. To visualize the different dynamics observed depending on the value of the parameters β_u and β_v, we present Figure 2. Here, for the different values of β_u and β_v, we present the flow diagrams for the salary (u) and the employment (v). Note that there are some critical values for the parameters $\beta_u = \sigma/\lambda$ and $\beta_v = \sigma\gamma n/\rho$ that signal a change in the dynamical behavior and thus a bifurcation point. These values mark the frontier between the different states of our economies.

Figure 3 shows a summary of all the behaviors observed with the modified Goodwin model, integrating the model equations with some initial conditions, where the parameters of the model are equal to 1 except for the beta parameters β_u and β_v. In this case, it is trivial to see that the bifurcation is located on $\beta_u = 1$ and $\beta_v = 1$.

Figure 3a is the temporal evolution of the u and v variables for the model parameters $\beta_u = 0.5$ and $\beta_v = 1.5$. This is an unstable situation and, after some transient, the values of the variables diverge; thus, it represents a non-realistic situation without equivalent within the economic context.

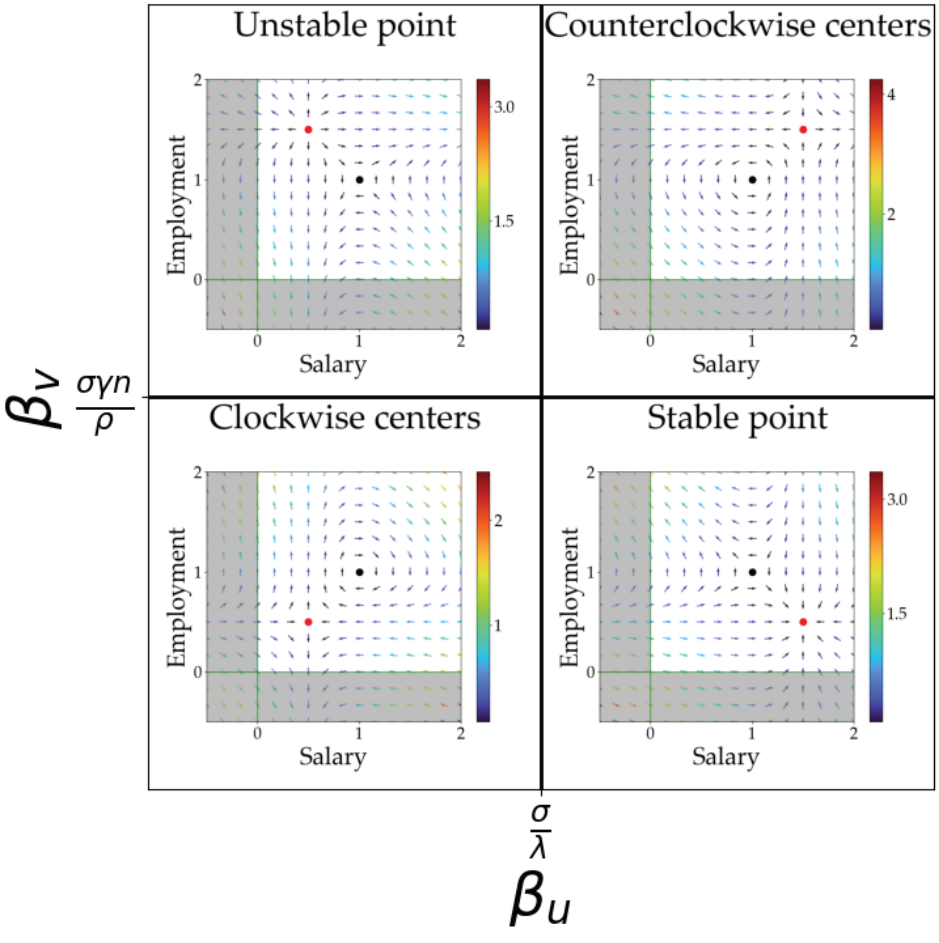


Figure 2. Phase diagram for the modified Goodwin model depending on the parameters (β_u and β_v). The black spot in each phase diagram marks the position of the first fixed point (u_0^1, v_0^1) (always in $(1,1)$) whereas the red spot marks the second fixed point (u_0^2, v_0^2). Shaded regions correspond to negative values for u and/or v and thus do not provide solutions with a physical meaning. All the parameters of the model (including population) are set as equal to 1. The color of the arrows in the diagram reflects the magnitude of the flow vector in the phase diagram (the colder the color, the smaller its magnitude, as indicated in the color bar).

Increasing β_u above the bifurcation point (at $\beta_u = 1$) produces a completely different dynamic that is plotted in the right upper corner in Figure 2 ($\beta_u = 1.5$ and $\beta_v = 1.5$). A counterclockwise center in the flow diagram that translates into a periodic oscillatory behavior is shown in Figure 3b. This situation corresponds to some cyclic dynamic in the economy considered. In economic terms, these are solutions with the same economic meaning as the equivalent predicted by Goodwin in the original model, but with the variables' roles swapped, i.e., in Goodwin model, the salary is the predator and employment the prey, thinking in the analogue of a Lotka–Volterra model, whereas, in the situation described by Figure 3b, it is the opposite: salary is the prey and employment is the predator, which is something totally different from the previous ideas held in this type of economic theory [22], but with economic meaning. Another important detail is that these oscillations may pass through negative values if the centers are far from the fixed point. However, in

these situations, the variations in salary and employment would be enormous in small fractions of time, so these situations are non-realistic. Note that the model proposed is just an approximation of reality within a certain range of validity. In this case, large values of the salaries or employment produce non-realistic solutions and thus we should include additional higher-order terms in the equations if we were to control it. In summary, near the fixed point, the results describe physically and economically sound situations (within the range of validity of the model) that will be analyzed in the coming sections.

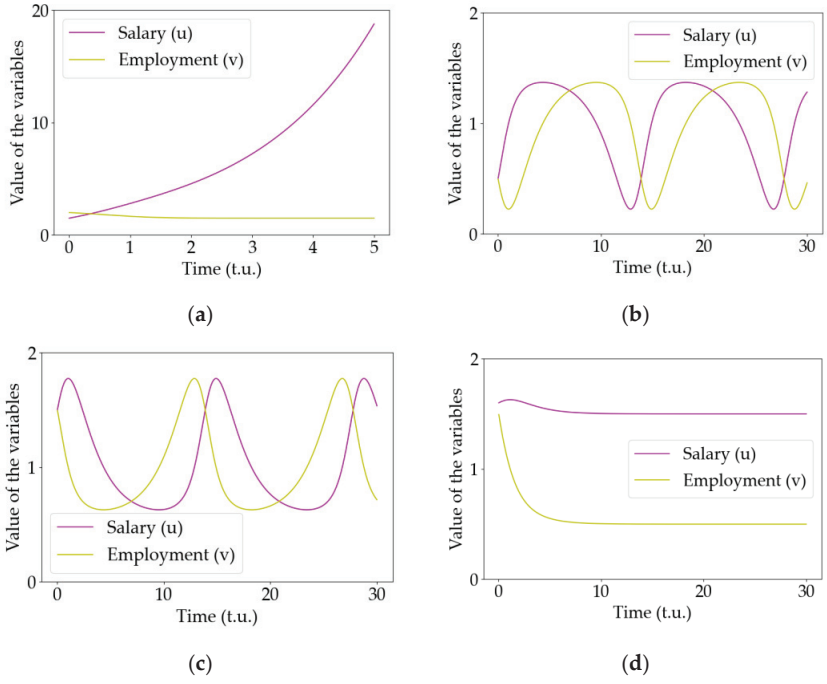


Figure 3. Temporal evolution of the model variables (the salary u and the employment v) for the four different configurations plotted in Figure 1. (**a**) Unstable point with $\beta_u = 0.5$ and $\beta_v = 1.5$ (note that the scale in the vertical axis is much larger than those in the others to stress the exponential behavior of this solution). (**b**) Counterclockwise (CCW) periodic center with $\beta_u = 1.5$ and $\beta_v = 1.5$. (**c**) Clockwise (CW) center with $\beta_u = 0.5$ and $\beta_v = 0.5$, corresponding to the original solution to the Goodwin model. (**d**) Stable point with $\beta_u = 1.5$ and $\beta_v = 0.5$. The remaining parameters are set as equal to 1.

Crossing the bifurcation line below $\beta_v = 1$, different dynamics are observed. In the bottom left corner of Figure 2 ($\beta_u = 0.5$ and $\beta_v = 0.5$.), a clockwise center is observed that represents a periodic behavior of the model variables but with completely different period and geometry, analogous to those oscillations predicted by Goodwin in the original model, as shown in Figure 3c. In addition, notice that, with this configuration, the other unstable fixed point prevents these oscillations from crossing into negative values for the variables.

Finally, the bottom right corner of Figure 2 shows the dynamic for $\beta_u = 1.5$ and $\beta_v = 0.5$. Here, the system exhibits a stable fixed point, represented in Figure 3d. The economic interpretation is straightforward: the economy will tend to stability and, once it is reached, it remains there without any further dynamic.

From the observation of the phase diagrams in Figure 2, we notice that some other unstable dynamics can be expected, but we will not focus on them as they correspond to unrealistic non-physical situations.

Although the parameters β_u and β_v were introduced for mathematical reasons, their meaning in economic terms can be understood based on the dynamics observed. In the Goodwin model solutions (Figures 2 and 3c), they are directly related to minimum values for the two variables: wages (u) and employment (v). This is related to economies where neither the employment nor the salaries can drop to zero. It does not seem a very strong restriction but, nevertheless, it significantly enriches the range of behaviors observed.

2.3. Network Topologies

In the following, we considered N economies that are interconnected following a given network structure. We can describe this system using a complex network [28–30] where the nodes are each of the N economies considered and the links connecting two nodes describe the existence of an influence relationship between both.

The structure of the interaction network between the different economies considered is fully encoded in the adjacency matrix [28–30], where $a_{ij} = 1$ if there exists an interaction between economy i and economy j, and 0 otherwise. Since the interaction between two economies is mutual, the adjacency matrix must be symmetric; thus, $a_{ij} = a_{ji}$. We recovered the activity of each node by summing up its interactions in what is called the connectivity degree, or simply degree, k_i, for each node.

In this manuscript, we considered different types of networks [28–30] corresponding to different patterns of influence among the economies:

- A scale-free network (commonly known as Barabasi–Albert). The BA graph mixes network growth and preferential attachment to generate a power-law connectivity distribution. In this case, a few of the nodes are connected with many economies whereas the majority have a significantly reduced number of connections, and a few of the economies are widely connected whereas most of them only have a few connections, mostly with those dominating economies.
- A Watts–Strogatz network. The WS graph is built from an initial chain with k nearest neighbors from which connections are then rewired randomly with a given probability, creating shortcuts along the network. In this case, economies are just connected with neighboring ones with an additional probability to connect to far-away economies. In this configuration, all the economies are connected with the same degree. Most of the connections are local except for a few long-range connections that still describe a proximity economy with some attempts to become global.
- Mean field network. All nodes interact with an imaginary node that is just the average of all the nodes in the network. The specific relationship between all the economies involved is not clear; nevertheless, they all interact through this imaginary node with the mean value. This is a global situation where economies are so interconnected that, finally, they just feel the average among all of them.

The adjacency matrices for all the cases considered were generated using the graph generators of the Python library networkx (specifically the functions nx.watts_strogatz_graph() and nx.barabasi_albert_graph()). It is important to mention that we simulated random networks using a Watts–Strogatz network with $p = 1$ due to the fact that, in a random network, the connectivity for each node is also random, which is something we do not desire. In addition, in this work, we considered that the Watts–Strogatz network has a value of p equal to 0.05, unless stated otherwise. Examples of the networks used and some details are shown in the Supplementary Materials (Section S2).

2.4. Economies Connected via a Network

In this section, we present the modifications that were introduced in the model to account for the influence of the other economies. We considered a set of N economies, each of them described by the modified Goodwin model presented in Equation (4). Now, the evolution of the variables for each economy will depend on the internal dynamic but also

on the values of the variables of the economies directly connected. This is represented by the following set of equations:

$$\begin{aligned} \frac{du_i}{dt} &= \left(u_i - \frac{\beta_{u_i}}{\sigma_i}\right)(\rho_i v_i - \gamma_i) + g\sum_{j=1}^{N} a_{ij}(u_j - u_i) \\ \frac{dv_i}{dt} &= \left(v_i - \frac{\beta_{v_i}}{n_i \sigma_i}\right)\frac{1-\lambda_i u_i}{\kappa_i} + g\sum_{j=1}^{N} a_{ij}(v_j - v_i) \end{aligned} \tag{8}$$

where $i = 1 \ldots N$ denotes each of the N economies considered.

The first term of Equation (8) is given by the internal dynamic of each economy that is described in Equation (4). However, the second term carries the information of the network. a_{ij} is an element of the adjacency matrix as described in the previous section that equals 1 when the economies i and j are connected and equals 0 when there is no such connection. The specific shape of the network term is directly derived from Fick's law, which describes the tendency to displace the excess from one node to a connected deficitary one. Within the economic context of the present model, this means that economies with higher wages or employment rates will influence their connected nodes by raising their equivalent variables. This, as a first approximation, seems reasonable as the well-being of economies tends to positively influence those economies connected (and vice versa). If the employment rates are high in one economy, this will most likely result in an increase in the employability in the connected economies, as well as the salaries. The simplest way to include this in the equations is through the term $u_j - u_i$, meaning that if the variable u_j in the j-node has a higher value, the variable u_i will increase (as the derivative $\frac{du_i}{dt} > 0$). On the other hand, considering the j-node, the situation is the opposite and $\frac{du_j}{dt} < 0$; thus, the variable u_j will be reduced. This is also reasonable as large salaries in an economy surrounded by economies with lower salaries will tend to lower to compensate for the possibility of moving the economy to those neighboring economies. At the same time, the salaries in those neighboring economies will tend to rise until some equilibrium is reached.

The parameter g controls the weight of the neighboring economies (the smaller this term is, the less relevant the network interaction will be). The limiting case with $g = 0$ corresponds to a set of independent economies that do not interfere with each other. Note that those nodes connected with a larger number of nodes will experience a stronger network influence as the number of terms from the summation that are non-null will be larger. This is also reasonable as those economies that interact with a large number of other economies (nodes) will also experience a larger influence from them.

The mathematical description of the network presented is the simplest that can be considered. More complicated interdependencies can be introduced even considering asymmetric relationships between the economies involved [31,32] but this configuration is the simplest that still produces some meaningful results. Note that each node receives a different contribution from the network term as its connectivity might be different.

When a mean field interaction between the economies is considered, the set of equations presented in Equation (8) becomes

$$\begin{aligned} u_i' &= \left(u_i - \frac{\beta_{u_i}}{\sigma_i}\right)(\rho_i v_i - \gamma_i) + g(\langle u \rangle - u_i) \\ v_i' &= \left(v_i - \frac{\beta_{v_i}}{n_i \sigma_i}\right)\frac{1-\lambda_i u_i}{\kappa_i} + g(\langle v \rangle - v_i) \end{aligned} \tag{9}$$

where $\langle u \rangle$ and $\langle v \rangle$ are the average value over all the considered economies of the u and v variables, respectively. Note that these equations describe the interaction of each node with an imaginary node endowed with the average properties of all the nodes described above.

The set of Equation (8) or (9) with $N = 50$ was solved using an explicit 4th-order Runge–Kutta scheme [13], properly adding the network term. We used this number of nodes because the Goodwin model is a macroeconomic model and, in the real world, there

are not thousands of macroeconomies interacting between them, so a small network is closer to the types of economies that we are trying to characterize.

3. Results and Discussion

The mathematical model and the main parameters were introduced in the previous section. In particular, the intrinsic dynamics for each economy are controlled by the values of β_u and β_v. On the other hand, we have several parameters controlling the type of interaction between the different economies. g controls the weight of the network on each economy and the type of network considered (BA, WS, random, or mean field). These are the parameters that we analyze below.

3.1. Phase Diagram for a Complex Network Model

Figure 4 summarizes all the behaviors observed for this model. Each panel in the figure corresponds to a phase diagram where the weight of the network term g and the connectivity k of the network are varied (with the exception of the BA, where the parameter is m, which is the number of connections that a node must have when the network is being created out of a random network; see the Supplementary Materials for details). Sixteen of such diagrams are presented, each corresponding to a different type of network (WS, random, BA, and mean field) as we move along the horizontal direction, whereas, when moving along the vertical direction, we explore different values of the parameters β_u and β_v that control the internal dynamic of each economy.

For these parameters, we considered the three dynamical behaviors described in Figures 2 and 3 that present non-divergent behaviors, namely CW-oscillations or clockwise oscillations ($\beta_u = 0.5$ and $\beta_v = 0.5$), where all the economies exhibit this behavior in the absence of the network influence; stable point ($\beta_u = 1.5$ and $\beta_v = 0.5$), where all economies are in a fixed steady point in the absence of a network; CCW-oscillations or counterclockwise oscillations ($\beta_u = 1.5$ and $\beta_v = 1.5$), where all economies oscillate in the absence of a network. The last configuration considered and named as mixed state in Figure 4 corresponds to a situation where each individual economy has a different value of β_u and β_v randomly chosen from the three configurations described above.

The collective behaviors observed are color-coded in Figure 4. Regions in blue denote those parameter values that result in non-physical configurations (divergent trajectories mostly). The rest of the collective behaviors are classified depending on the collective behavior of the system variables. For some parameter values (marked in orange), we observe that all variables converge to a common fixed steady state. Under some other circumstances, each economy or node in our network converges with time to a different steady state (marked in green). Two other states are observed involving some transient oscillations in the process of reaching a stationary state. In red, we mark those parameter values that produce a solution where all the economies oscillatory synchronize and collectively tend to a single fixed point (spiral stable state). In addition, those parameter values that induce several spiral states, different for each node, are marked in purple. Finally, marked in brown are those parameter values that result in a complicated behavior where each node follows the dynamics of a fixed point or a stable spiral. For those cases where we obtained a non-physical solution, we repeated the simulations to discard numerical instabilities.

Note that the effect of the network connectivity is almost negligible for all the cases considered. The network weight, given by parameter g, on the other hand, determines the dynamic of the solution.

Four different network topologies were considered, and the results are independent of them. Only the mean field case (described by (Equation (9)) exhibits a different behavior. For the three other networks (described by Equation (8)), the salary always goes to the fixed point predicted by the linear approximation; on the other hand, and for the main field case (Equation (9)), we observe multiple fixed points different for each node. As we will discuss later, the fixed point reached by the economies can be related to the specific connectivity degree of each node.

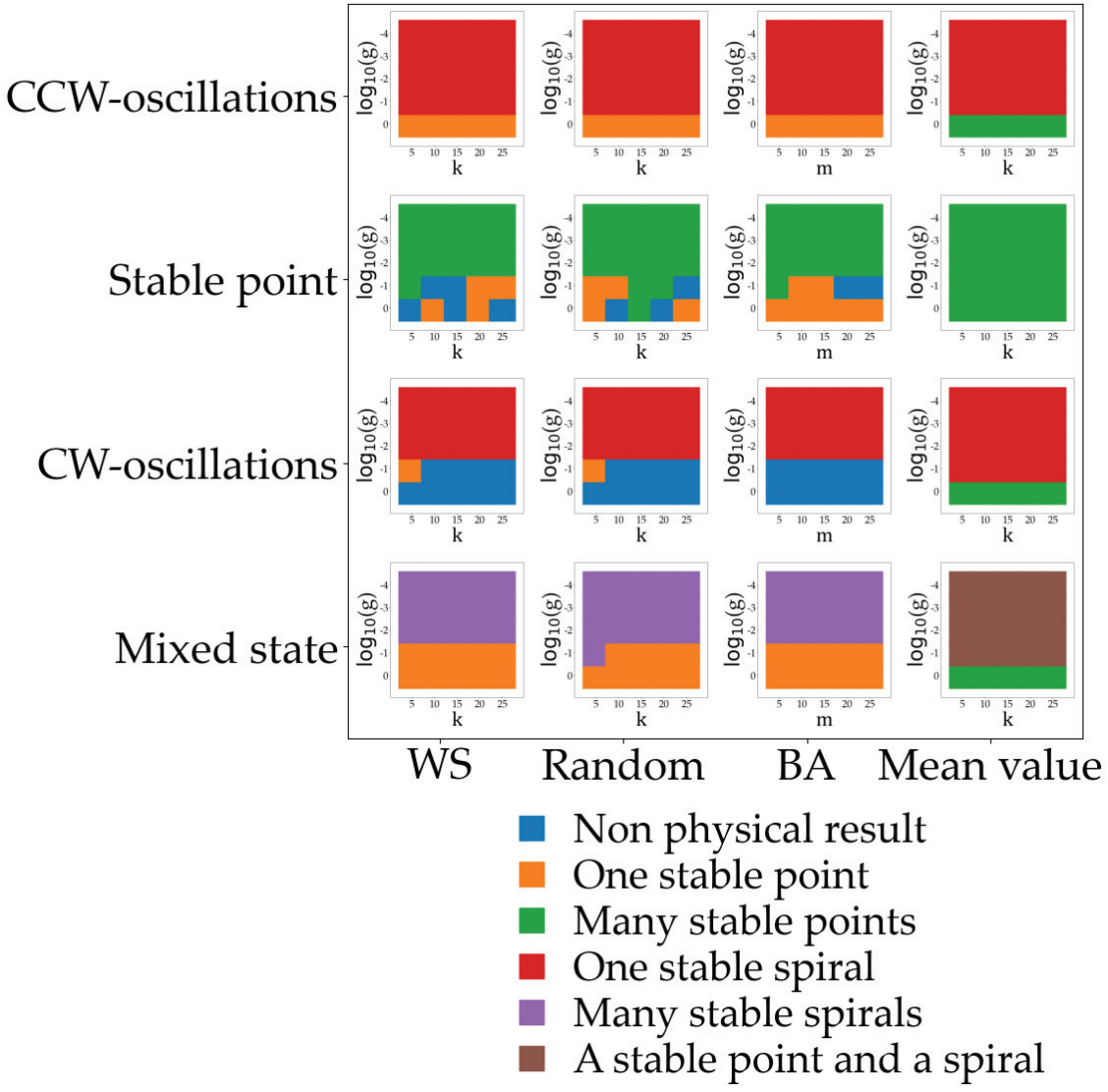


Figure 4. Multiple phase diagrams for the relevant variables in our model. Color-coded are marked the final configurations after the transient for the different values of the relevant parameters considered. (All other parameters are set equal to 1 as in previous results). Note that for each point in the phase diagrams multiple simulations were performed in order to assure the result as the specific network considered depends on the simulation (between 5 and 10 simulations were typically run).

The small number of economies considered ($N = 50$) might be the reason behind this lack of sensitivity toward the selected network. A larger number of involved economies are expected to be more sensitive to the network configuration although, for the present study, we stuck to smaller numbers that better describe the interaction of a limited number of macroeconomies.

The internal dynamic of each economy turned out to be determinant in the selection of the final configuration. It is interesting to note that economies endowed with a counterclockwise initial condition evolve into a steady state due to the network influence.

In Figure 5, we present an example for each of the dynamical behaviors represented in Figure 4. Going from left to right and up to down in Figure 5, the first row presents the dynamics observed for the two model variables when a mixed state is considered as the initial condition, many stable spirals (Figure 5a), and a stable point and a spiral (Figure 5b). In this last case, the fixed point is always $(u = 1.5, v = 0.5)$, and the stable spiral node is $(u = 1, v = 1)$. In both cases, the final state is a fixed point that is reached via some temporal non-coherent damped oscillations. Note that even those economies that were initially already in a steady state experience some oscillations before reaching the stability.

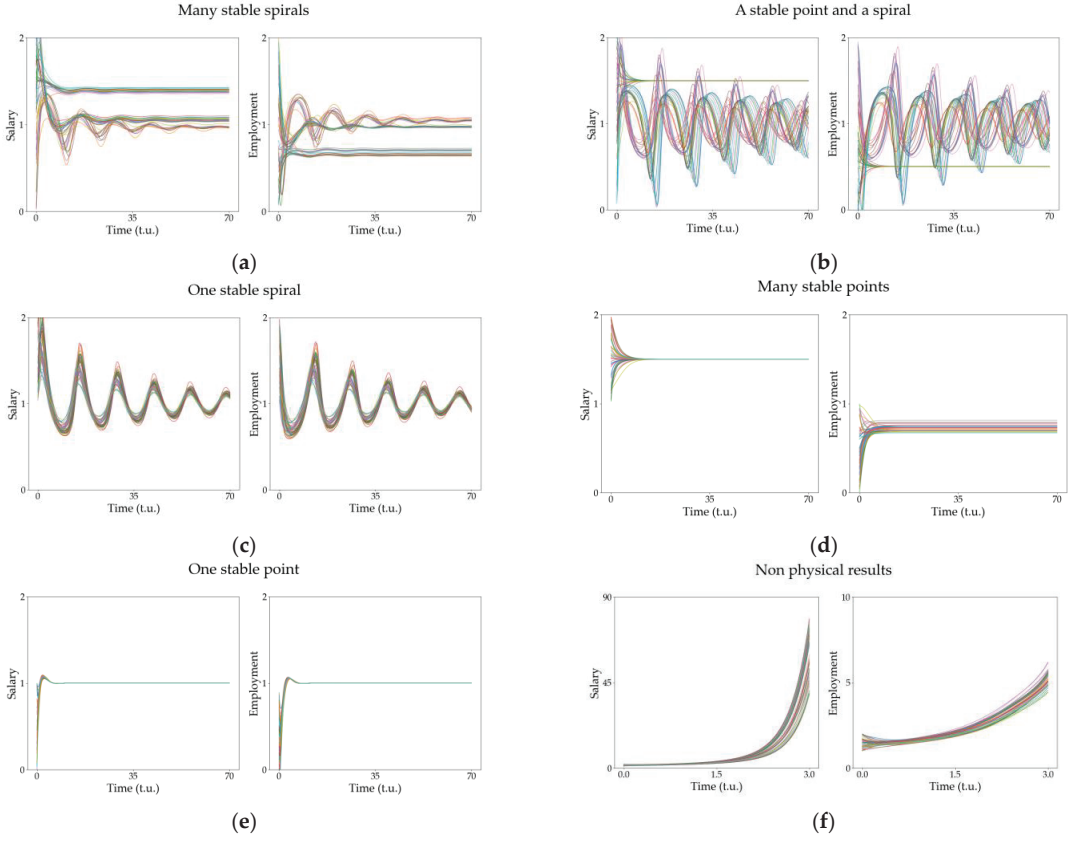


Figure 5. Evolution of the model variables for the different configurations described in Figure 4. The temporal evolution of the two model variables is plotted when the initial configuration of the nodes is (**a**) many stable spirals, (**b**) a stable point and a spiral, (**c**) one stable spiral, (**d**) many stable points, (**e**) one stable point and (**f**) corresponds with a non-physical result. Note that, in (**f**), the scale in the vertical axis is significantly larger than in the rest of the figures.

Figure 5c shows the dynamic of the system when all economies converge to a fixed point after several damped oscillations when the initial state is CW-oscillations. Note that the cycles (or centers) of the Goodwin model disappear due to the network influence stabilizing a fixed point. In Figure 5d, all economies steadily move to a fixed point that changes depending on the particular node (this case will be analyzed in more detail below).

In Figure 5e, all economies exhibit a single fixed point after some transient that does not involve damped oscillations. Note that this configuration is observed for large values of g, i.e., when the weight of the connected economies is important. From an economic point of view, all the economies are so interconnected that the final solution is common to

all of them; they describe a common global economy. Finally, Figure 5f is an example of a non-physical solution where all economies diverge to infinity due to the high coupling with the network. This non-physical solution is observed when the parameter quantifying the weight of the network (g) is significantly large.

Note that the system becomes more sensitive to the initial conditions for each economy and the network topology when the network coupling is very strong (cases where $\log_{10} g \leq -1$). When the economies are less coupled, the structure of the network becomes less relevant. In addition, the oscillatory behavior observed in our solutions is the natural approach for reaching the stationary solution in our model and it seems a plausible solution in a more realistic context. However, it is interesting to note that the direct relaxation to the stationary solution is also a solution in the model (Figure 5d).

In the following sections, we analyze the main collective phenomena observed in the network of connected economies, specifically the synchronized oscillations of the solutions in Figure 5c and the relation of the fixed points in employment with the connectivity in the solutions of Figure 5d.

3.2. Synchronization of the Spiral Solutions

Figures 4 and 5 show that, for some parameter values, all economies may lead to a fixed value but, during the transient, they may experience synchronized damped oscillations. We analyze this behavior in this section. Figure 6 shows a summary of the results observed. Figure 6a shows the evolution of the variable u (salary) for all the economies considered and the evolution of the variable v (employment) is shown in Figure 6b. Note that, although the initial values for the simulations are randomly chosen, all the economies synchronize almost immediately and coherently oscillate. The lower row in Figure 6 (Figure 6c–f) shows the histograms of the delays between the different economies and model variables shortly after the beginning of the simulation and at some later stage. Note that, in both instances, the histogram has a narrow bell-shaped distribution, although the distribution becomes narrower as time passes and the different economies interact for longer periods (Figure 6e,f). Figure 6g,h show the evolution of variables u and v when the network weight is one order of magnitude smaller. In this case, the synchronization is not evident, and the histograms presented in the following row (Figure 6i–l) show a much broader distribution (that becomes even broader as time evolves), reflecting the non-coherent state of synchronization between the economies considered.

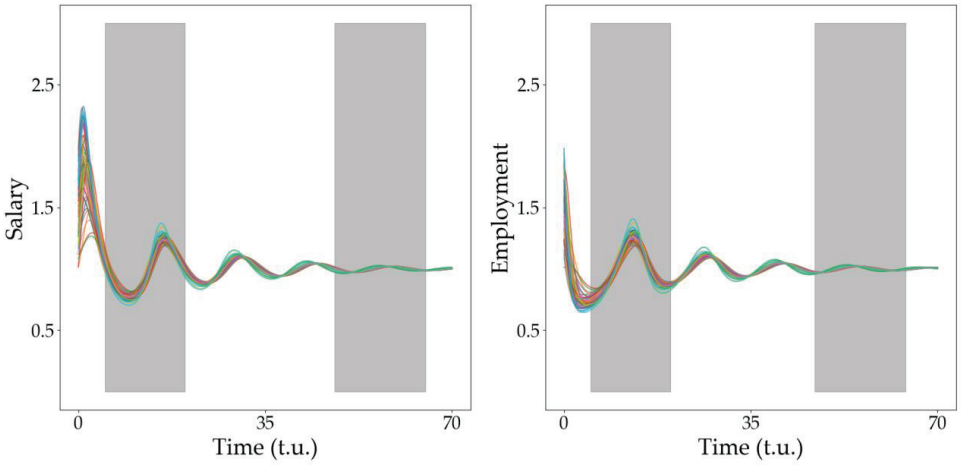


Figure 6. *Cont.*

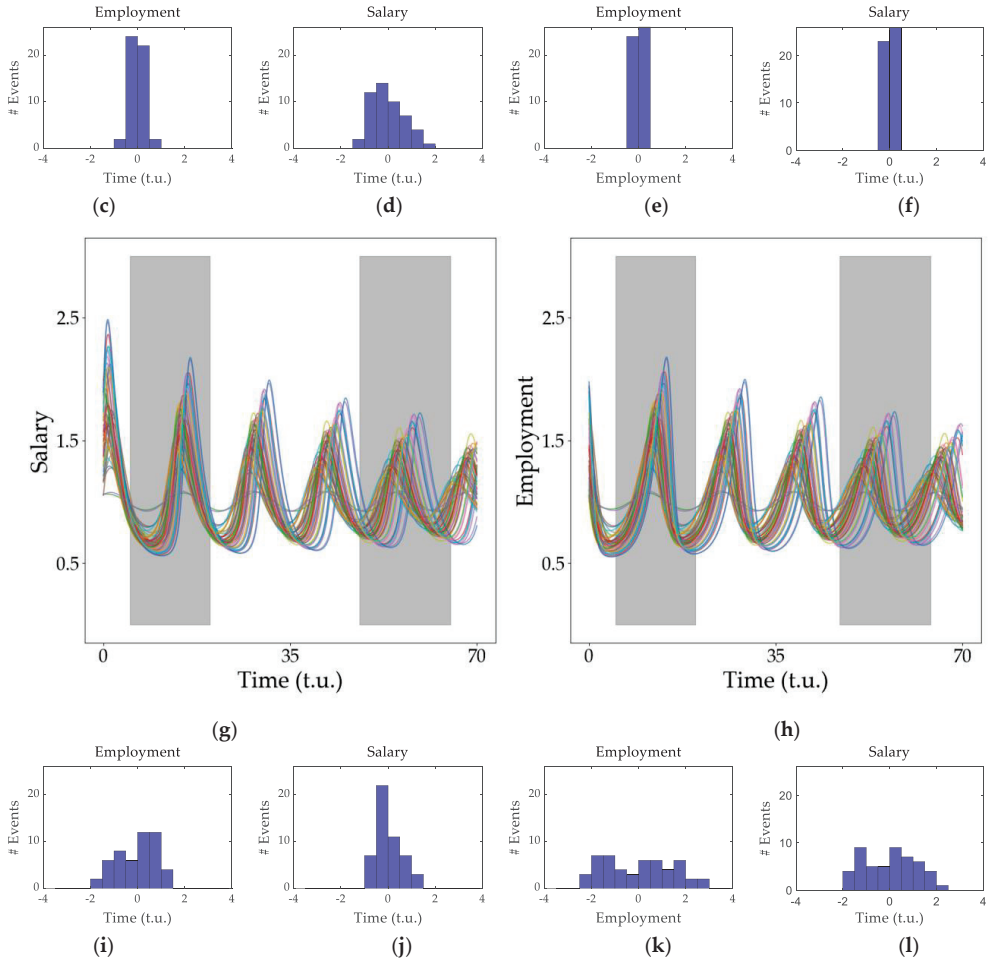


Figure 6. Synchronization in spiral dynamics. (**a**,**b**) Evolution of our economic network variables for a large network weight ($\log_{10} g = -2$). The shaded areas are the regions where we will analyze the local maxima and study the difference in phases between economies. The histograms with the time delay between all economies are shown in (**c**) for the u variable in the first shadowed region and (**d**) for the v variable in the first shadowed region. The histograms at the end of the simulation are shown in (**e**) for the u variable and (**f**) for the v variable. Evolutions of (**g**) the employment variable (u) and (**h**) the salary variable (v) are shown for a low value of the network weight ($\log_{10} g = -3$). The behavior of all the economies is less synchronized as shown in the histograms in the lowest row. The histograms with the time delays between the economies whose dynamics are plotted in figures (**g**,**h**) are shown in (**i**) for the u variable in the first shadowed region and (**j**) for the v variable in the first shadowed region. The histograms at the end of the simulation are shown in (**k**) for the u variable and (**l**) for the v variable. The network used for this simulation is a WS with $k = 15$ and $p = 0.05$.

A narrow distribution reflects that the economies oscillate in synchrony whereas a wider-spread histogram corresponds to a set of unsynchronized economies. An interesting parameter for characterizing this is the standard deviation from the mean value, σ. We analyzed the variation in σ (that, in our context, gives an inverse measurement of the synchronization degree) as we varied the weight of the network on the dynamics, g. These results are shown in Figure 7. Note that, as expected, as the weight of the network

becomes more important, the degree of synchronization among the economies becomes higher (and thus σ becomes smaller). Figure 7a shows that the variation in g produces unsynchronous oscillations.

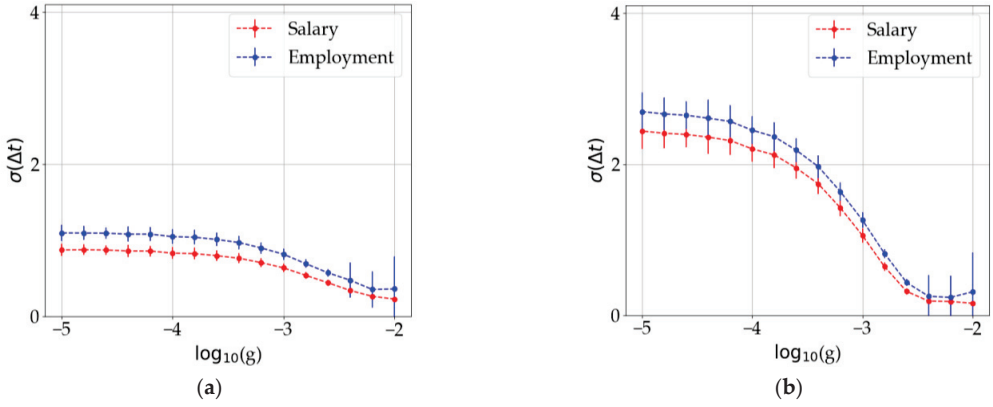


Figure 7. Dispersion of the economies' variables as a function of the network weight g. (**a**) Standard deviation at the beginning of each simulation. (**b**) Same magnitude calculated at the end of the simulation. Error bars show the dispersion of this parameter over 100 runs of the same simulation.

Our results also show some dependence on the average degree of the connectivity for each type of network considered, but the results are not conclusive. Nevertheless, as they might help to illustrate the global behavior, we include them in the Supplementary Information.

Different network topologies produce similar results to those in Figure 7 (not shown in the text).

3.3. Dispersion of the Steady States

Another interesting phenomenon observed in Figures 4 and 5 (specifically, in Figure 5d) is observed when the economies tend toward a steady value but each economy (node in the network) reaches a significantly different steady state. This is observed in Figure 8. In Figure 8a, we can notice that, as the variable salary reaches the same value in all nodes, the other variable, employment, differs depending on the node (Figure 8b). In order to enlighten the origin of this deviation, we measured the connectivity of each node (economy) and plotted the corresponding value of the employment variable at the steady state. The results are plotted in Figure 8b,c. There is a clear linear dependence between both variables and thus we conclude that the final value of the variables linearly depends on the node connectivity. This means that those economies more connected to others will benefit from larger levels of employment while the salaries will remain unchanged but equal to all economies in the network.

As in the previous section, when we reduce the value of g, the interaction with the network becomes weaker, which implies that the steady states merge together and the distance between them becomes negligible. Note that the results presented in this section correspond to a random network, but equivalent results are observed with the other types of networks considered (see the results in the Supplementary Materials). This lack of sensitivity to the network structure was already expected from Figure 4, where the effect of the type of network on the type of behavior observed is demonstrated to be almost null. Note that, from an economic point of view, our results indicate that increasing the connectivity of each node (economy) results in a rise in the employment variable as well, meaning that very global economies are more prone to rising salaries in all economies involved.

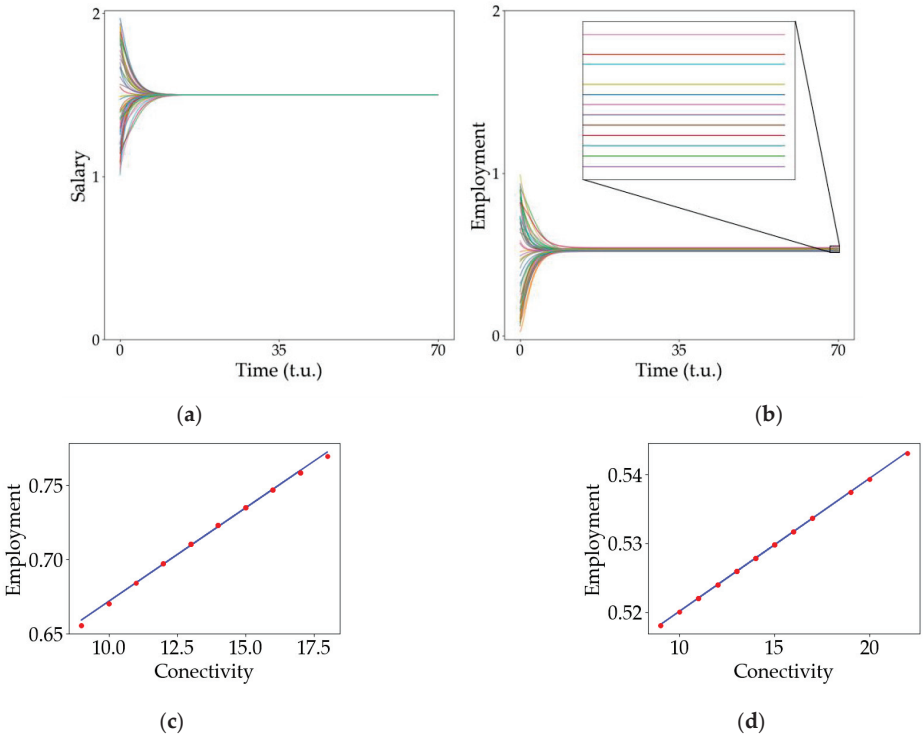


Figure 8. Relationship between connectivity and the stable point. Evolution of (**a**) variable u, salary, and (**b**) variable v, employment, for all the economies in the network. In (**b**), there is a zoom on the squared section to better appreciate the distance between the stable points. Dependence of the employment final state with the node connectivity for (**c**) a large value of the network weight ($\log_{10} g = -2$) and for (**d**) a smaller value of the network weight ($\log_{10} g = -3$). All simulations used a random network with $k = 15$. The size of the vertical axis is different for each figure to improve visibility of the results.

4. Conclusions

The modified Goodwin model, which introduces some limiting values for the salary and employment variables, results in an adequate mathematical description for a simplified economy and long-time observations. In particular, it prevents divergent solutions. The analysis of a reduced number of different economies coupled via some network of connections showed interesting collective properties. The different parameters relevant for this study were analyzed and the importance of the neighboring economies was stressed. The specific topology of the network considered appears to be less relevant than expected, probably due to the small number of economies considered. Nevertheless, the connectivity of the nodes plays an important role in determining the levels of employment and salary achieved. The strength of the network coupling also proves to be determinant in controlling the dynamics. Two phenomena are described in more detail: the synchronous oscillatory behavior that strongly depends on the strength of the interactions between the nodes of the network and the dispersion of the final steady states that is determined by the connectivity degree of each node. The economic descriptors of each economy are strongly influenced by the other economies connected. It is interesting to note the case of strongly coupled economies that result in a global improvement of the model variables.

It is important to note that, although the model describing each economy is a simple one that lacks the complexity of the real economies, the aim of this contribution focuses on

the synergetic behavior of a collection of economies interconnected. In fact, the existence of a network topology reflects the fact that we are not in a global fully interconnected economy and allows us to examine the role of the different parts of the structure on the variables describing each economy. Within the structure described here, it is possible to imagine modifications of the present model for a more realistic description of the reality. In any case, the basic consequences derived from our contribution remain valid, i.e., the network structure formed by the different economies strongly influence the dynamic of the whole economic system.

In summary, we can conclude that the interactions between economies, rather than being negligible, may become the source of the dynamic for the entire system, and that the structure of the network also plays a significant role in the collective behavior.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/e25060894/s1.

Author Contributions: Conceptualization, A.P.M. and F.Y.R.; Methodology, A.P.M.; Software, F.Y.R.; Validation, F.Y.R.; Formal analysis, F.Y.R.; Investigation, A.P.M. and F.Y.R.; Resources, F.Y.R.; Data curation, F.Y.R.; Writing—original draft, F.Y.R.; Writing—review & editing, A.P.M.; Visualization, F.Y.R.; Project administration, A.P.M.; Funding acquisition, A.P.M. All authors have read and agreed to the published version of the manuscript.

Funding: We gratefully acknowledge financial support by the Spanish Ministerio de Economía y Competitividad and European Regional Development Fund under contract RTI2018-097063-B-I00 AEI/FEDER, UE, and by Xunta de Galicia under Research Grant No. 2021-PG036. Authors are part of the CITMAga Strategic Partnership. All these programs are co-funded by FEDER (UE).

Acknowledgments: We acknowledge useful conversations with Xesús Pereira López and their colleagues at the Facultade de Economía from Universidade de Santiago de Compostela. FYR acknowledges conversations with Fernando Vadillo that sparked his interest in early stages of the model.

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

References

1. White, J.A.; Chow, C.C.; Rit, J.; Soto-Treviño, C.; Koppel, N. Synchronization and Oscillatory Dynamics in Heterogeneous, Mutually Inhibited Neurons. *J. Comput. Neurosci.* **1998**, *5*, 5–16. [CrossRef] [PubMed]
2. Waters, C.M.; Bassler, B.L. QUORUM SENSING: Cell-to-Cell Communication in Bacteria. *Annu. Rev. Cell Dev. Biol.* **2005**, *21*, 319–346. [CrossRef] [PubMed]
3. Buck, J. Synchronous rhythmic flashing of fireflies. II. *Q. Rev. Biol.* **1988**, *63*, 265–289. [CrossRef] [PubMed]
4. Lutherm, S.; Fenton, F.H.; Kornreich, B.G.; Squires, A.; Bittihn, P.; Hornung, D.; Zabel, M.; Flanders, J.; Gladuli, A.; Campoy, L.; et al. Low-energy control of electrical turbulence in the heart. *Nature* **2011**, *475*, 235–239. [CrossRef]
5. Tinsley, M.; Taylor, A.; Huang, Z.; Wang, F.; Showalter, K. Dynamical quorum sensing and synchronization in collections of excitable and oscillatory catalytic particles. *Phys. D Nonlinear Phenom.* **2010**, *239*, 785–790. [CrossRef]
6. Tinsley, M.; Taylor, A.; Huang, Z.; Wang, F.; Showalter, K. Dynamical quorum sensing and synchronization in large populations of chemical oscillators. *Science* **2009**, *323*, 614–617.
7. Tinsley, M.; Taylor, A.; Huang, Z.; Wang, F.; Showalter, K. Phase clusters in large populations of chemical oscillators. *Angew. Chem. Int. Ed.* **2011**, *123*, 10343–10346.
8. Kiss, I.Z.; Zhai, Y.; Hudson, J.L. Emerging coherence in a population of chemical oscillators. *Science* **2002**, *296*, 1676–1678. [CrossRef]
9. Toiya, M.; González Ochoa, H.O.; Vanag, V.K.; Fraden, S.; Epstein, I.R. Synchronization of chemical micro-oscillators. *J. Phys. Chem. Lett.* **2010**, *1*, 1241–1246. [CrossRef]
10. García-Selfa, D.; Ghoshal, G.; Bick, C.; Pérez-Mercader, J.; Muñuzuri, A.P. Chemical oscillators synchronized via an active oscillating medium: Dynamics and phase approximation model. *Chaos Solitons Fractals* **2021**, *145*, 110809. [CrossRef]
11. Néda, Z.; Ravasz, E.; Brechet, Y.; Vicsek, T.; Barabási, A.L. The sound of many hands clapping. *Nature* **2000**, *403*, 849–850. [CrossRef]
12. Morales, A.J.; Vavilala, V.; Benito, R.M.; Bar-Yam, Y. Global patterns of synchronization in human communications. *J. R. Soc. Interface* **2017**. [CrossRef] [PubMed]
13. Acemoglu, D.; Ozdaglar, A.; Tahbaz-Salehi, A. Network, Shocks, and Systemic Risk. In *Oxford Handbook of the Economics of Networks*; Section 21.1.1; Oxford University Press: Oxford, UK, 2016.
14. Mantegna, R.N.; Stanley, H.E. *An Introduction to Econophysics*; Cambridge University Press: Cambridge, UK, 2000.
15. Jackson, M.O.; Watts, A. The Evolution of Social and Economic Networks. *J. Soc. Econ. Netw.* **2002**, *106*, 265–295. [CrossRef]

16. Harrod, R.F. An Essay in Dynamic Theory. *Econ. J.* **1939**, *49*, 14–33. [CrossRef]
17. Domar, E.D. Capital Expansion, Rate of Growth, and Employment. *Econometrica* **1946**, *14*, 137–147. [CrossRef]
18. Solow, R.M. A Contribution to the Theory of Economic Growth. *Q. J. Econ.* **1956**, *70*, 65–94. [CrossRef]
19. Phillips, A.W. The Relation between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom 1861–1957. *Economica* **1958**, *25*, 283–299. [CrossRef]
20. Goodwin, R.M. *A Growth Cycle: Socialism, Capitalism and Economic Growth*; Cambridge University Press: Cambridge, UK, 1967.
21. Murray, J.D. *Mathematical Biology: I. An Introduction (Chapters 2–3)*, 3rd ed.; Springer: Berlin/Heidelberg, Germany, 1993.
22. Nell, E.J.; Semmler, W. *Nicholas Kaldor and Mainstream Economics. Confrontation or Convergence? (Chapters 5, 19, 22 & 32)*, 1st ed.; Palgrave Macmillan: London, UK, 1991.
23. Santos, J.F.C.; Araújo, R.A. Using Non-Linear Estimation Strategies to Test an Extended Version of the Goodwin Model on the US Economy. *Rev. Keynes. Econ.* **2020**, *8*, 268–286. [CrossRef]
24. Araujo, R.A.; Moreira, H.N. Testing a Goodwin's model with capacity utilization to the US economy. In *Nonlinearities in Economics: An Interdisciplinary Approach to Economic Dynamics, Growth and Cycles*; Orlando, G., Pisarchik, A.N., Stoop, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2021.
25. Araújo, R.A.; Dávila-Fernández, M.J.; Moreira, H.N. Some new insights on the empirics of Goodwin's growth-cycle model. *Struct. Change Econ. Dyn.* **2019**, *51*, 42–54. [CrossRef]
26. Press, W.H.; Flannery, B.P.; Teukolsky, S.A.; Vetterling, W.T. *Numerical Recipes*; Cambridge University Press: Cambridge, UK, 1988.
27. Strogatz, S.H. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*; Studies in Nonlinearity; Addison-Wesley Publishing: Boston, MA, USA, 1994.
28. Albert, R.; Barabási, A.L.; Pósfai, M. *Network Science*; Cambridge University Press: Cambridge, UK, 2016.
29. Albert, R.; Barabási, A.L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **2002**, *74*, 47–97. [CrossRef]
30. Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D.U. Complex networks: Structure and dynamics. *Phys. Rep.* **2006**, *424*, 175–308. [CrossRef]
31. Gonçalves, J.; Matsushita, R.; Da Silva, S. The Asymmetric Brazilian Input–Output Network. *J. Econ. Stud.* **2021**, *48*, 604–615. [CrossRef]
32. Vanag, V.K.; Epstein, I.R. Cross-diffusion and pattern formation in reaction–diffusion systems. *Phys. Chem. Chem. Phys.* **2009**, *11*, 897–912. [PubMed]

entropy

MDPI

Article

Transaction Entropy: An Alternative Metric of Market Performance

Hua Zhong, Xiaohao Liang and Yougui Wang *

School of Systems Science, Beijing Normal University, Beijing 100875, China;
201931250002@mail.bnu.edu.cn (H.Z.); 202221250004@mail.bnu.edu.cn (X.L.)
* Correspondence: ygwang@bnu.edu.cn

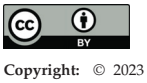
Abstract: Market uncertainty has a significant impact on market performance. Previous studies have dedicated much effort towards investigations into market uncertainty related to information asymmetry and risk. However, they have neglected the uncertainty inherent in market transactions, which is also an important aspect of market performance, besides the quantity of transactions and market efficiency. In this paper, we put forward a concept of transaction entropy to measure market uncertainty and see how it changes with price. Transaction entropy is defined as the ratio of the total information entropy of all traders to the quantity of transactions, reflecting the level of uncertainty in making successful transactions. Based on the computational and simulated results, our main finding is that transaction entropy is the lowest at equilibrium, it will decrease in a shortage market, and increase in a surplus market. Additionally, we make a comparison of the total entropy of the centralized market with that of the decentralized market, revealing that the price-filtering mechanism could effectively reduce market uncertainty. Overall, the introduction of transaction entropy enriches our understanding of market uncertainty and facilitates a more comprehensive assessment of market performance.

Keywords: market uncertainty; transaction entropy; market performance; price filtering mechanism; willingness price

Citation: Zhong, H.; Liang, X.; Wang, Y. Transaction Entropy: An Alternative Metric of Market Performance. *Entropy* **2023**, *25*, 1140. https://doi.org/10.3390/e25081140

Academic Editor: Stanisław Drożdż

Received: 30 June 2023
Revised: 21 July 2023
Accepted: 28 July 2023
Published: 30 July 2023



1. Introduction

The concept of the market holds great significance in economics and serves as a fundamental basis for research of economics [1]. Understanding the mechanisms by which markets function has profound implications for decision making, policy formulation, and economic development. The research on market primary functioning has long been focused on two key aspects: price formation and market efficiency.

In the context of market price, this is determined by the interaction between sellers and buyers. In a perfect competitive market, the market price is deemed to be at the cross point of the supply and demand curves. Therefore, the factors that influence these curves, such as the willingness of the market participants and information dissemination, have impacts on the level of the market price. Regarding market efficiency, market surplus is usually used to measure it. Market surplus represents the total welfare generated by transactions between sellers and buyers. An increase in market surplus signifies an improved efficiency in market transactions and a more optimal allocation of resources.

However, most analyses of price formation and market efficiency are typically conducted under the assumption of ideal conditions, without accounting for the uncertainties faced by the participants in a market. Several studies have verified the existence of uncertainties in market transactions [2,3]. During an actual transaction process, each participant has limited access to information and cannot obtain complete knowledge about the counterparty's information or the overall market situation. This inherent imperfection in information significantly influences the decision making and behavior of both parties

involved, thereby increasing transactional uncertainty, which subsequently impacts market prices and market surplus. Therefore, understanding the mechanism of price formation and improving market efficiency have become the cornerstones of economic analyses, and we contend that incorporating uncertainty into a market can not only provide a more accurate description of market performance, but also enrich our understanding of market functioning.

Financial markets serve as the primary focus of market uncertainty analyses. In financial markets, the market participants make investment decisions based on expectations, which inherently carry a certain level of uncertainty. Therefore, uncertainty is a common and essential aspect of financial markets. These uncertainties, in turn, exert negative effects on market information efficiency [4]. To enhance market efficiency, it is crucial to understand and measure financial market uncertainty [5]. Information entropy is commonly used to measure such a kind of uncertainty, which is developed from information theory [6]. As the amount of available information increases, the uncertainty decreases, resulting in a decrease in entropy. Conversely, when there is less information and a higher uncertainty, this entropy increases [7].

Firstly, as Eugene Fama argued, when uncertainty arises in financial markets, it is challenging for participants to assess and respond to information accurately, resulting in price volatility [8,9]. The uncertainty arising from market volatility is closely related to fluctuations in unpredictable asset prices, highlighting the dynamic and uncertain nature of price movements, which can significantly impact investment decisions and the overall market sentiment. To measure the market uncertainty related to price volatility, various variants of entropy have been proposed based on information entropy. Claudiu Vinte introduced the approach of cross-sectional intrinsic entropy to estimate the uncertainty in stock markets [10]. This approach takes into account the trading volume and price movements of various assets, allowing for a more holistic understanding of market dynamics. As the understanding of market volatility gets deeper, some researchers have recognized its transmission effect, which can give rise to various forms of market uncertainty within the same classification. Thomas Dimpfl employed transfer entropy to quantitatively assess the transmission of volatility between different financial markets [11]. Understanding this transmission of volatility can be crucial for investors and policymakers in making informed decisions and conducting effective risk management.

Furthermore, uncertainty in portfolio selection is related to investors' asset allocation. Investors aim to achieve objectives through the rational allocation of different types of assets. However, there are randomness and fuzziness factors in markets that prevent investors from fully predicting the returns and values of assets, which leads to uncertainty in portfolio selection [12,13]. Philippatos and Wilson were among the first to apply the concept of entropy to portfolio selection [14]. They proposed a mean entropy approach to measure the uncertainty in the asset allocation process. Their pioneering research shed light on the importance of considering uncertainty in portfolio management, leading to a paradigm shift in how investors make decisions about asset allocation. Building upon their work, more generalized forms of entropy, such as incremental entropy, were formulated. Compared to the traditional portfolio selection theory, the theory based on incremental entropy emphasizes that there is an optimal portfolio for a given probability of return [15]. Xu et al. introduced the concept of hybrid entropy and utilized it to measure the asset risk caused by both randomness and fuzziness [16]. Using information entropy to measure the level of uncertainty in portfolio selection can effectively assist investors in evaluating and optimizing their asset allocation strategies.

Finally, uncertainty in the option-pricing process is related to the impact of uncertain factors such as underlying asset price volatility and interest rates. Options are financial instruments whose value and returns depend on the price movements of the underlying assets and other factors. Les Gulko introduced the entropy pricing theory (EPT), which can provide valuation results similar to the Sharpe–Lintner capital asset-pricing model and the Black–Scholes formula [17]. His research was also extended to stock option pricing [18]

and bond option pricing [19], using the EPT to measure collective market uncertainty. The EPT model demonstrates simplicity and user-friendliness, aligning with the principles of the Efficient Market Hypothesis [5]. Based on the previous analysis, it is evident that information entropy is a comprehensive tool for measuring the uncertainty in financial markets. In comparison to traditional tools, it provides a better reflection of these uncertainties that exist in financial markets.

Although much attention has been paid to the uncertainty in financial markets, it is worth noting that there are other forms of uncertainties that exist across various markets. Information asymmetry is an important factor that leads to quality uncertainty [20,21], which may lead to issues such as adverse selection [22,23]. Additionally, there is economic policy uncertainty (EPU) present in the market, referring to the impact of exogenous shocks related to economic policies that introduce unpredictability and uncertainty into the market [24]. Lots of empirical studies have shown that EPU shocks can lead to stock market turbulence [25,26]. These uncertainties are not limited to financial markets, and they can actually occur in all kinds of markets.

All the uncertainties mentioned above are important and have indeed been extensively studied in the literature. However, it should be noted that there is another significant type of uncertainty that has not received sufficient attention. This particular form of uncertainty stems from the mismatch between the quantities of desired exchanges in market transactions. In a market where buyers and sellers engage in trading activities, an equilibrium is achieved when the quantity supplied equals the quantity demanded. However, in reality, markets often operate in a disequilibrium state, where the quantities supplied and demanded are not equal with each other. This condition implies that traders may face uncertainty in their transactions. In this paper, we focus on this specific type of uncertainty and aim to put forward a metric to measure and analyze it.

Based on the foregoing analyses, the current applications of entropy in measuring the uncertainty caused by incomplete information, as well as its utilization in characterizing asset portfolios and risk assessment in financial markets, does not provide a comprehensive understanding of the mechanisms of market operation. In particular, there is no equivalent concept of entropy to express the uncertainty of participants' trading in the market. Thus, we come up with the concept of transaction entropy to represent the uncertainty of market trade. We investigate how this transaction entropy changes with price. The results show that the equilibrium market has the lowest entropy. Additionally, we also compare the total entropy between centralized and decentralized markets, where the key distinction lies in the presence of a price-filtering mechanism. The result shows that the total entropy is lower in a centralized market than that in a decentralized market. This finding highlights the effectiveness of price filtering in reducing market uncertainty and emphasizes the importance of integrating a price-filtering mechanism in the trading process to ensure market transaction stability.

The contributions of this paper can be summarized as follows: (1) the proposal of a concept of "transaction entropy" to measure the level of uncertainty in the process of transactions. By introducing this concept, we are able to better understand and quantify market uncertainty, providing a new perspective for in-depth analyses of the mechanism of market function; (2) the addition of an alternative metric for market performance based on the existing framework of market function analyses. Through investigating the variation in transaction entropy with respect to price changes, we find that the state of market equilibrium not only corresponds to the highest volume of transaction and the maximum market surplus, but also the lowest entropy; (3) a comparison of the levels of total entropy between centralized and decentralized markets, revealing that the presence of a price-filtering mechanism enhances successful transactions and reduces market uncertainty; (4) a comparison of computational and simulation results in terms of various aspects, including the quantity of transactions, market surplus, transaction entropy, and the total entropy in centralized and decentralized markets, to verify the theoretical analysis; and (5) a clarification of the

limitations of traditional market equilibrium analyses, while emphasizing the importance of transaction uncertainty.

The remaining sections of this paper are organized as follows. Section 2 formulates the functions of supply and demand based on the concept of willingness price. In Section 3, we analyze market performance, including transaction quantity, market surplus, and transaction entropy, using the rationing rate. Additionally, we compare the total entropy in centralized and decentralized markets and discuss policy implications based on the comparison results. Section 4 presents the simulation settings and results, demonstrating the generating process of each variable that characterizes market performance. In Section 5, we discuss the importance of market transaction uncertainty by highlighting the shortcomings of the Walrasian general equilibrium and Marshall partial equilibrium approaches. We also discuss the plausible applications of transaction uncertainty analyses in real-world scenarios. Section 6 draws the conclusions.

2. The Expression of Demand and Supply with Willingness Price

A partial equilibrium analysis (PEA) is a widely used tool for understanding market performance. It argues that supply and demand collectively represent two sides of traders in a market, making it simple to analyze the consequence of their interaction by tracing the equilibrium point and social welfare implications [27]. However, the PEA also needs to be improved, since it fails to clearly identify how sellers and buyers constitute supply and demand curves correspondingly. To solve this problem, Wang and Stanley introduced the concept of willingness price and formulated supply and demand functions to restate the PEA in a goods market [28]. The major advantage of this approach is that the laws of supply and demand can be derived directly, and the efficiency of market equilibrium can be strictly proved.

In this paper, we follow their approach to describe the supply and demand in a goods market. We assume that each trader is willing to make a trade of one unit of goods and has a willingness price before participating in the trade. For one seller, their willingness price is defined as the minimum price that they are willing to sell one unit of goods. On the other side, the willingness price of a buyer is defined as the maximum price that they are willing to spend for one unit of goods. Supposing that a seller with a willingness price v^s meets a buyer with a willingness price v^b, their deal can be made only if $v^b \geq v^s$ is valid. Although we cannot identify all traders' willingness prices in real markets, we know that they exist there and govern whether a deal can be made or not.

As all participants' willingness prices are exogenously given, the willingness prices of sellers and buyers must have a distribution correspondingly. It is reasonable to assume that willingness prices spread over the domain of $(0, +\infty)$. This spread can be characterized by probability density functions, $f_s(v)$ and $f_B(v)$ for sellers and buyers, respectively. Supposing that the numbers of the sellers and buyers are given exogenously, denoted as N_S and N_B, respectively, then we can use $F_s(v) = N_s \times f_s(v)$ and $F_B(v) = N_B \times f_B(v)$ to characterize such distributions. From the normalization condition, we have the integrals of $F_s(v)$ and $F_B(v)$ over the whole region of willingness prices, which are N_s and N_B, respectively,

$$\int_0^\infty F_s(v)dv = N_s, \tag{1}$$

$$\int_0^\infty F_B(v)dv = N_B. \tag{2}$$

For any one seller, given a market price of p, they will make their choice by comparing the willingness price and market price, that is to say, the necessary condition for the seller to sell one unit of goods can be expressed as

$$p \geq v_S. \tag{3}$$

Otherwise, the seller will withdraw their offer.

Equation (3) implies that only the sellers whose willingness price is not greater than the actual market price are willing to sell their goods. Combining (1) and (3), we can obtain the supply function with a given market price $Q_S(p)$, which can be written as

$$Q_S(p) = \int_0^p F_s(v)dv. \tag{4}$$

The above rationale can also be applied to derive the demand function. For a buyer, only if his willingness price v_B is higher than or equal to the market price p, will he buy one unit of goods in a market, i.e.,

$$p \leq v_B. \tag{5}$$

Otherwise, he will give up on his purchase. Combining (2) and (5), we can obtain the demand function with a given market price $Q_D(p)$ of the market, which is given by,

$$Q_D(p) = \int_p^\infty F_B(v)dv. \tag{6}$$

As is well known, there are many factors that can affect the supply and demand in a market. From the expressions of supply and demand given by Equations (4) and (6), the implicit governing factor of supply and/or demand is the willingness prices of the market participants. Thus, we can infer that most relevant factors take their effects through the willingness prices of sellers and buyers. As a result, any change in any variable that impacts these willingness prices will have an impact on the supply and demand of the goods. In addition, the extent of a market determines the total quantities of the goods demanded and supplied, which also has an impact on the supply and demand functions.

Another important inference of supply and demand functions is that we can prove the laws of supply and demand by taking a derivative of these two formulas. The first derivatives of the supply and demand functions can be expressed, respectively, as the following,

$$\frac{dQ_S}{dp} = F_s(p) > 0, \tag{7}$$

$$\frac{dQ_D}{dp} = -F_B(p) < 0. \tag{8}$$

The results show that the relationship between the quantity supplied and the market price is positive. In other words, the higher market price, the more goods supplied in the market. On the contrary, the relationship between the quantity demanded and the market price is negative. Fewer goods are demanded as the price rises.

The interaction between supply and demand determines the equilibrium price level and quantity of transactions. Combining Equations (4) and (6), we can obtain the equilibrium price $p = p^*$.The equilibrium transaction quantity T^* can be derived directly, which can be expressed as,

$$T^* = \int_0^{p^*} F_s(v)\, dv = \int_{p^*}^\infty F_B(v)dv. \tag{9}$$

Figure 1 illustrates the supply and demand curves in a commodity market. The supply curve is upward sloping, and the demand curve is downward sloping. The cross-point of these two curves specifies the market equilibrium, which corresponds to the equilibrium quantity and market-clearing price of the market.

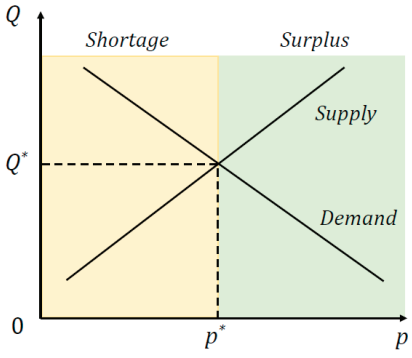


Figure 1. A simplified diagram of supply and demand curves in a market. The shortage region is marked in yellow color, while the surplus region is in green color.

3. The Market Performance with Formulated Supply and Demand Functions

In this section, our primary focus is on evaluating various aspects of market performance using the newly formulated supply and demand functions. Specifically, we analyze three key dimensions: transaction quantity, market surplus, and market uncertainty caused by a quantity mismatch of the supply and demand in a disequilibrium market. To quantify this uncertainty, we propose the concept of transaction entropy, which is derived from information entropy.

3.1. The Quantity of Transactions

Supply and demand represent two parties of a goods market, and their interaction determines not only the market price, but also the quantity of transactions. In this section, we set the market price as being given exogenously, and investigate how transaction quantity is determined by supply and demand as the price varies.

The state of a market depends on the level of given price. The market is in equilibrium when the price makes the market clear. Otherwise, the market is in disequilibrium. This disequilibrium can be divided into two cases, one is shortage and the other is surplus. When the price is lower than the equilibrium level, it corresponds to a state of shortage, where there is more quantity demanded than the quantity supplied in the market. When the price is higher than the equilibrium level, it corresponds to a state of surplus, where there is more quantity supplied than the quantity demanded in the market. As shown in Figure 1, the regions of shortage and surplus are marked in yellow color and green color, respectively.

According to the short-side principle, the realized quantity of transactions is determined by the short side. The short side refers to the trading party with fewer willing exchanges, and those with more are at the long side. At equilibrium, the quantity supplied is equal to the quantity demanded. In this case, the quantity of realized transactions T^* given by Equation (9) is equal to the quantity supplied and demanded.

In a shortage market, the quantity demanded exceeds the quantity supplied. Therefore, the quantity of realized transactions is determined by the quantity supplied. The expression of the realized quantity of transactions in a shortage market $T_{ST}(p)$ can be expressed as,

$$T_{ST}(p) = \int_0^p F_s(v)dv \quad p < p^*. \tag{10}$$

For a surplus market, the quantity demanded is less than the quantity supplied. In contrast, the quantity of realized transactions in a surplus market $T_{SP}(p)$ can be written as follows,

$$T_{SP}(p) = \int_p^\infty F_B(v)dv \quad p > p^*. \tag{11}$$

Based on the preceding analyses, the transaction quantity in the various states of a market can be given by,

$$T(p) = \begin{cases} \int_0^p F_s(v)dv & p < p^*, \\ \int_0^{p^*} F_s(v)dv = \int_{p^*}^{\infty} F_B(v)dv & p = p^*, \\ \int_p^{\infty} F_B(v)dv & p > p^*. \end{cases} \tag{12}$$

Figure 2 shows the computational results of the relationship between transaction quantity and market price based on Expression (12), represented by the blue line. Obviously, the quantity of transactions increases with an increase in market price when $p < p^*$, and decreases when $p > p^*$. The quantity of transactions reaches its maximum when the market price attains its equilibrium level.

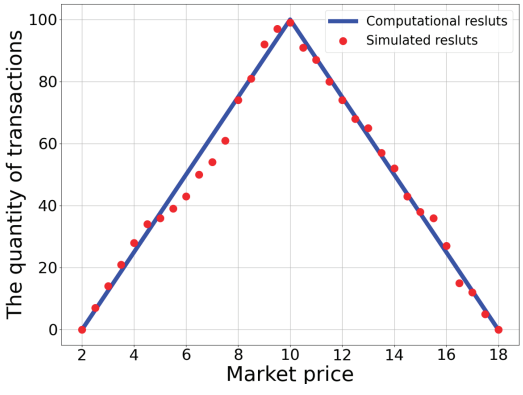


Figure 2. The relationship between transaction quantity and market price. The blue line represents the computational results, while red dots denote the simulation results. The participants' willingness prices and market prices are in the range of [2,18], and the market reaches equilibrium at a price of $p^* = 10$. For further details about the simulation settings, see the section of Simulation Results.

3.2. Market Surplus

3.2.1. The Rationing Rates

According to the short-side principle, we know that all the participants at the long side are willing to make transactions, nevertheless, some of them cannot achieve their desired outcome. Thus, we define the rationing rate as the ratio of the quantity of actual transactions to the quantity of desired exchanges. The sellers' and buyers' rationing rates can be used in the following analysis of market surplus and transaction entropy. Their expressions (G_s and G_B) are given as follows, respectively,

$$G_s = \frac{T}{Q_S}, \tag{13}$$

$$G_B = \frac{T}{Q_D}. \tag{14}$$

It is obvious that G_s and G_B are in the range of [0, 1]. The quantities supplied and demanded will change with a variation in the market price. Therefore, the level of rationing rate will be altered as the market price varies. When the market price equals the equilibrium one, the rationing rates of either sellers or buyers equal one. Thus, we obtain,

$$G_s(p^*) = G_B(p^*) = 1. \tag{15}$$

In the shortage region, i.e., $p < p^*$, all sellers can fulfill their willing exchanges, where only a portion of buyers can successfully match with the sellers and achieve their desired

transactions. As a result, the sellers' rationing rate is one, while the buyers' rationing rate would be less than 1. Thus, we obtain,

$$G_s(p) = 1, \tag{16}$$

$$G_B(p) < 1. \tag{17}$$

Meanwhile, with an increasing market price, there are more commodities supplied and less demanded. The rationing rate of sellers remains constant with the increase in price, while the rationing rate of buyers increases. We then obtain,

$$\frac{dG_s(p)}{dp} = 0, \tag{18}$$

$$\frac{dG_B(p)}{dp} > 0. \tag{19}$$

In contrast, the above rationale can also be applied to the surplus region, where $p > p^*$. The rationing rate of sellers is lower than 1, and the buyers' rationing rate is one. Then, we obtain,

$$G_S(p) < 1, \tag{20}$$

$$G_B(p) = 1. \tag{21}$$

The relationship between the rationing rate and market price in a surplus market can also be derived. In this case, as the market price increases, sellers are less likely to obtain their rations, because the quantity supplied increases while the quantity demanded decreases. Meanwhile, the rationing rate of the buyers will not change. The derivatives of the rationing rates of sellers and buyers have the following properties,

$$\frac{dG_s(p)}{dp} < 0, \tag{22}$$

$$\frac{dG_B(p)}{dp} = 0. \tag{23}$$

Figure 3 depicts the dependence of these rationing rates on market price. As shown in this figure, when a market is in a shortage, the rationing rate of buyers is less than one, whereas the rationing rate of sellers is equal to one. In contrast, the rationing rate of sellers is smaller than 1, while the buyers' rationing rate equals one when a market is in surplus. When a market is in equilibrium, the rationing rates of either the sellers or buyers are 1.

3.2.2. The Formulation of Market Surplus

Market surplus, used to measure market efficiency, is another essential component of traditional market performance analyses. The surplus of one seller (buyer) can be defined as the difference between the actual (willingness) price and the willingness (actual) price. In the transactions of a goods market, only a portion of participants will be able to realize their willing exchanges, and a surplus will be generated. Therefore, it is reasonable to take rationing rates into account when formulizing the surplus of a market.

For sellers, given a market price p, the total realized surplus of these sellers (Z_{sr}) in the market could be calculated as follows,

$$Z_{sr}(p) = \int_0^p F_s(v)(p - v)G_s(p)dv. \tag{24}$$

On the other side, given a market price p, the total realized surplus of the buyers (Z_{Br}) in the market could be given by,

$$Z_{Br}(p) = \int_p^{\infty} F_B(v)(v-p)G_B(p)dv. \tag{25}$$

The total realized market surplus for a price $Z_r(p)$ is the sum of them, i.e.,

$$Z_r(p) = \int_0^{p} F_s(v)(p-v)G_s(p)dv + \int_p^{\infty} F_B(v)(v-p)G_B(p)dv. \tag{26}$$

Taking the first derivatives of Equation (26), the expression of the relationship between the derivation of surplus and market price can be expressed as,

$$\begin{array}{c} \frac{\partial Z_r(p)}{\partial p} = \int_0^{p} F_s(v)G_s(p)dv - \int_p^{\infty} F_B(v)G_B(p)dv \\ + \int_0^{p} F_s(v)(p-v)\frac{\partial G_s(p)}{\partial p}dv + \int_p^{\infty} F_B(v)(v-p)\frac{\partial G_B(p)}{\partial p}dv. \end{array} \tag{27}$$

Combining Equations (4)–(6), (13) and (14), Equation (27) can be rewritten as,

$$\frac{\partial Z_r(p)}{\partial p} = \int_0^{p} F_s(v)(p-v)\frac{\partial G_s(p)}{\partial p}dv + \int_p^{\infty} F_B(v)(v-p)\frac{\partial G_B(p)}{\partial p}dv. \tag{28}$$

When the market is in a shortage, we can obtain the following expression by combining Equations (18), (19) and (28),

$$\frac{\partial Z_r(p)}{\partial p} > 0. \tag{29}$$

When the market is in surplus, we can obtain the following expression by combining Equations (22), (23) and (28),

$$\frac{\partial Z_r(p)}{\partial p} < 0. \tag{30}$$

Figure 4 depicts the relationship between market surplus and market price. From this figure, we can find that the market surplus increases when $p < p^*$ and decreases when $p > p^*$. When the market is at equilibrium, the market surplus attains its maximum.

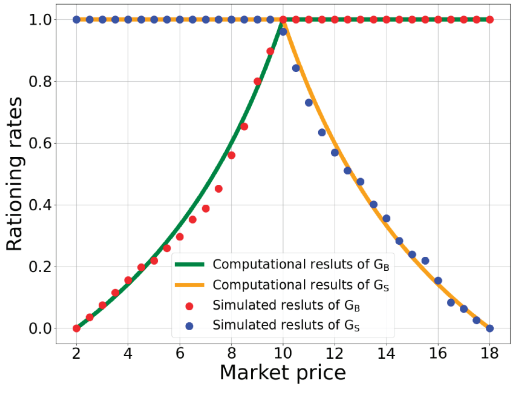


Figure 3. The relationship between rationing rates and market price. The green line and red dots represent the computational and simulation results of $G_s(p)$, respectively, while the orange line and blue dots are computational and simulation results of $G_B(p)$, respectively. For details about the simulation settings, see the section of Simulation Results.

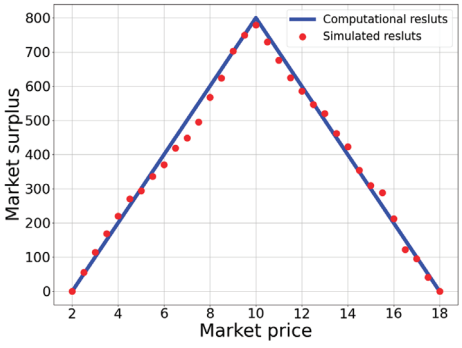


Figure 4. The relationship between market surplus and market price. The blue line represents the computational results, while red dots are the simulation results.

3.3. Market Uncertainty

Except for traditional market performance, which focuses on transaction quantity and market surplus, we also consider market uncertainty as an additional dimension of market performance.

3.3.1. Transaction Entropy

Information entropy is a commonly used tool for measuring the level of disorder and uncertainty, and its extension has been widely applied in the fields of economics and finance [5,29,30]. In this section, we introduce a new kind of information entropy, named transaction entropy, to characterize market uncertainty and investigate how transaction entropy changes as market price varies.

To figure out the information entropy of one participant, we need to identify the transaction procedure, which is shown in Figure 5. At first, the participant has to make sure whether they satisfy the price-filtering mechanism given by Equations (3) and (5). At this stage, there are only two filtering results for the participants: remain in or exit the market. The exiting participants refer to ones whose willingness prices does not satisfy the condition of trade in the market, while the remaining participants refer to those who satisfy the trading conditions. It is worth noting that it is possible to fail in the trade for the remaining participants. Only the traders in the short side can make a deal.

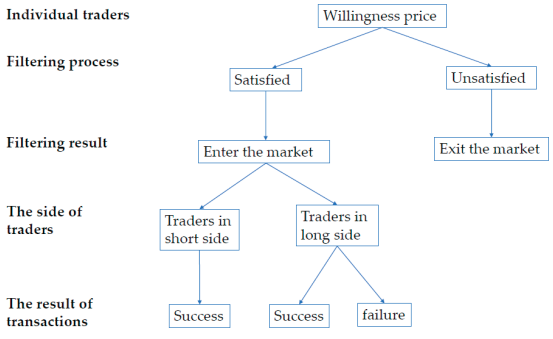


Figure 5. The flow chart of the transaction process for one participant's trade in a market.

In one word, the possibility of the participant making a deal is uncertain. Therefore, we can use the information entropy proposed by Shannon to present the uncertainty of the

trader's transaction in the market [4]. The definition of information entropy for individual traders $H(E)$ can be written as,

$$H(E) = -[E * lnE + (1 - E) * \ln(1 - E)], \tag{31}$$

where E is the possibility of a successful trade. It should be noted that the possibility of a successful trade in this case is the rationing rate, referred to in the former subsection. Therefore, the respective information entropy of one seller H_s and one buyer H_B can be given by, respectively,

$$H_s(p) = -[G_s(p) * lnG_s(p) + (1 - G_s(p)) * ln(1 - G_s(p))], \tag{32}$$

$$H_B(p) = -[G_B(p) * lnG_B(p) + (1 - G_B(p)) * ln(1 - G_B(p))] \tag{33}$$

We assume that a trader is willing to make an exchange with one unit of goods, so Equations (32) and (33) can also present the information entropy of their willingness exchange quantity. The willingness exchange quantities of the remaining sellers and buyers are denoted as Q_S and Q_D. Combining the supply and demand functions given by Equations (4) and (6), we obtain the total information entropy TS as follows,

$$TS = \int_0^p F_s(v)H_s(p)dv + \int_p^\infty F_B(p)H_B(p)dv. \tag{34}$$

From Equations (16) and (17), we find that the rationing rate of sellers $G_s(p) = 1$ and rationing rate of buyers $G_B(p) < 1$ when $p < p^*$. As a result, we can derive that $H_s(p) = 0$ and $H_B(p) \neq 0$ directly from Equations (32) and (33). The total information entropy of the market equals the information entropy of buyers. When $p > p^*$, the rationing rate of sellers $G_s(p) < 1$ and the rationing rate of buyers $G_B(p) = 1$ is based on Equations (20) and (21), so $H_B(p) = 0, H_s(p) \neq 0$. In this case, the total information entropy of the whole market equals the information entropy of the sellers, which can be obtained from Equation (34). The rationing rates of the sellers and buyers are equal to one when $p = p^*$, given by Equation (15), and the information entropy of the sellers and buyers is equal to zero. Thus, Equation (34) can be rewritten as,

$$TS = \begin{cases} \int_p^\infty F_B(v)H_B(p)dv & p < p^* \\ 0 & p = p^* \\ \int_0^p F_s(v)H_s(p)dv & p > p^*. \end{cases} \tag{35}$$

The expression indicates that the resulting information entropy contains the contributions of all the actual transactions. To eliminate the effect of the market scale on the information entropy, we define the transaction entropy generated by one transaction to measure the market performance. Then, the transaction entropy takes the following form,

$$S = \frac{TS}{T} = \begin{cases} \frac{H_B(p)}{G_s(p)} & p < p^* \\ 0 & p = p^* \\ \frac{H_s(p)}{G_B(p)} & p > p^*. \end{cases} \tag{36}$$

For the sake of simplicity, we denote that $G(p) = min\{G_s, G_B\}$. When the market price is lower than the equilibrium price, the minimum rationing rate between the sellers and buyers is that of the sellers. When the market price is higher than the equilibrium level, the minimum rationing rate is that of the buyers. Then, Equation (36) can be transformed into the following form,

$$S = -\frac{[G(p) * lnG(p) + (1 - G(p)) * \ln(1 - G(p))]}{G(p)}. \tag{37}$$

It is obvious that the level of transaction entropy S is non-negative. According to L'Hospital's rule, the entropy tends to be positive infinity when the market price tends to positive infinity or zero. That is to say,

$$\lim_{p \to 0} S = \lim_{p \to 0} ln(\frac{1}{G(p)} - 1) = +\infty, \tag{38}$$

$$\lim_{p \to +\infty} S = \lim_{p \to +\infty} ln(\frac{1}{G(p)} - 1) = +\infty. \tag{39}$$

Taking the first derivation of Equation (37), we can obtain,

$$\frac{\partial S}{\partial p} = \frac{G'(p)}{G^2(p)} * \ln(1 - G(p)). \tag{40}$$

When the market price is lower than the equilibrium one, the relationship between the transaction entropy and market price is negative, which can be expressed as,

$$\frac{\partial S}{\partial p} < 0. \tag{41}$$

When the market price is greater than the equilibrium one, the transaction entropy and market price have a positive relation, which can be presented as,

$$\frac{\partial S}{\partial p} > 0. \tag{42}$$

Figure 6 presents the results of the relationship between the transaction entropy and market price. From the figure, we can see that the slope is downward when $p < p^*$, while it is upward in the case of $p > p^*$. Moreover, the single equilibrium transaction entropy corresponds to zero when $p = p^*$.

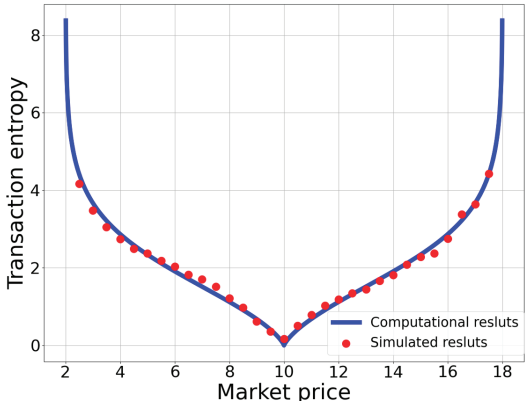


Figure 6. The dependence of transaction entropy on market price. The blue line represents the computational results, while the red dots represent the simulation results. When the market is in equilibrium, the transaction entropy is zero, indicating the absence of transaction uncertainty in the market.

3.3.2. Total Entropy in Centralized and Decentralized Markets

In this section, we redirect our focus from analyzing the entropy generated by one transaction (S) to examining the total entropy (TS) within two distinct market structures: a centralized market and a decentralized market. The difference between these two markets is the presence of price filtering or not. A centralized market can be regarded as having transactions with price filtering, while a decentralized market has transactions without

price filtering. By comparing the entropy in these two market types, we can reveal the role of price filtering in mitigating market uncertainty.

Firstly, we examine the total entropy in a centralized market. The centralized market is characterized by the presence of a central authority or intermediary that sets one order book to collect the bid–ask prices of traders, thereby facilitating all trading activities within the market [31,32]. It is worth noting that, in our previous analysis of transaction entropy, we assumed that a given market price serves as the reference condition for transactions, which is consistent with the key assumption of the centralized market. Therefore, we can conduct an analysis of the total entropy in a centralized market based on the existing results from the previous sections.

For the sake of simplicity, we make the following assumptions: (1) the number of sellers is equal to that of buyers, denoted as N; (2) the willingness prices of the sellers and buyers are in the range of $[a, b]$, and both a and b are positive; and (3) the supply and demand functions are linear. With these assumptions, we can easily obtain $F_S(v) = F_B(v) = k$, where k is a constant variable. As for the total entropy, considering the foregoing assumptions, we can rewrite Equation (35) as follows,

$$TS = \begin{cases} \int_p^b F_B(v)H_B(p)dv & a < p < p^*; \\ 0 & p = p^*; \\ \int_a^p F_s(v)H_s(p)dv & p^* < p < b. \end{cases} \tag{43}$$

Additionally, we can express the supply and demand functions in the centralized market, denoted as $Q_{SC}(p)$ and $Q_{DC}(p)$, respectively, as follows:

$$Q_{SC}(p) = \int_a^p F_S(v)dv = k(p - a), \tag{44}$$

$$Q_{DC}(p) = \int_p^b F_B(v)dv = k(b - p). \tag{45}$$

Taking the first derivations of (44) and (45), we obtain the following results,

$$Q'_{SC}(p) = F_s(p) = k, \tag{46}$$

$$Q'_{DC}(p) = -F_B(p) = -k. \tag{47}$$

By substituting Equations (13), (14), (32) and (33) into Equation (43), the expression of the total entropy can be rewritten as,

$$TS = \begin{cases} -\begin{bmatrix} Q_{SC}(p)lnQ_{SC}(p) + (Q_{DC}(p) - Q_{SC}(p))ln(Q_{DC}(p) - Q_{SC}(p)) \\ -Q_{DC}(p)lnQ_{DC}(p) \end{bmatrix} & a < p < p^*; \\ 0 & p = p^*; \\ -\begin{bmatrix} Q_{DC}(p)lnQ_{DC}(p) + (Q_{SC}(p) - Q_{DC}(p))ln(Q_{SC}(p) - Q_{DC}(p)) \\ -Q_{SC}(p)lnQ_{SC}(p) \end{bmatrix} & p^* < p < b. \end{cases} \tag{48}$$

To clarify the concavity of the total entropy, we can differentiate Equation (48) based on Equations (46) and (47). The results show that $\lim_{p \to a} TS' = +\infty$, $\lim_{p \to p*-} TS' = -\infty$, and $\lim_{p \to p*+} TS' = +\infty$, $\lim_{p \to b} TS' = -\infty$, where TS' is the derivative of TS. Moreover, it can be observed that the second derivative of TS is negative, which is presented in Appendix A, indicating a concave shape. There are three price levels corresponding to the total entropy being down to zero, that is, $p = a$, $p = b$,and $p = p^*$.

Then, we turn our attention to an exploration of the total entropy in a decentralized market. The decentralized market operates without a centralized authority or intermediary, enabling participants to engage in direct transactions with one another [33]. The key characteristic of a decentralized market is the random matching of sellers and buyers for

one period at a time, along with anonymous pairwise meetings involving bargaining [34,35]. The transaction process in the decentralized market is illustrated in Figure 7.

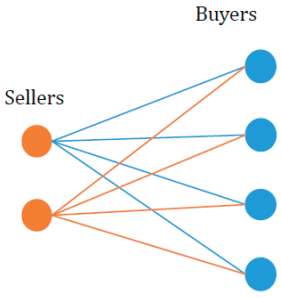


Figure 7. The random matching between sellers and buyers in a decentralized market. The orange circles represent sellers, and the blue circles represent buyers, each having his willingness price.

In this decentralized and random trading environment, every trader has an opportunity to engage in trading with one of the counterparty. A transaction will only be made if the buyer's willingness price surpasses the seller's willingness price; otherwise, the trade will not take place.

In order to make a comparison between the levels of total entropy in different markets, we keep the core assumptions presented in the centralized market. We suppose that the traders in the market only trade once at a time with one unit of goods in a random way. The probability of a successful transaction for a buyer with a willingness price of v' is the ratio of the number of sellers with a willingness price lower than v' to the total number of sellers. Similarly, the probability of a successful transaction for a seller with a willingness price of v' is the ratio of the number of buyers with a price higher than v' to the total number of buyers. Therefore, the respective expressions for the probability of a successful transaction for a seller (E_s) and a buyer (E_B) with a willingness price of v' are as follows,

$$E_s = \frac{\int_{v'}^{b} F_B(v)\nu}{\int_{a}^{b} F_B(v)\nu}, \tag{49}$$

$$E_B = \frac{\int_{a}^{v'} F_s(v)\nu}{\int_{a}^{b} F_s(v)\nu}. \tag{50}$$

At this time, the total entropy in the decentralized market with random matching TS_{de} is the sum of the buyers' entropy and sellers' entropy, which can be expressed as,

$$\begin{aligned} TS^{de} &= \int_a^b F_S(v)dv \ * \ H_S(E_S) + \int_a^b F_B(v)dv \ * \ H_B(E_B) \\ &= \int_a^b F_S(v) \ * \ H_S(v)dv + \int_a^b F_B(v) \ * \ H_B(v)dv \end{aligned} \tag{51}$$

The result shows that the total entropy in the decentralized market with random matching is a constant variable, and the detailed calculations can be found in Appendix B. This constant entropy can be expressed as,

$$TS^{de} = k(b-a). \tag{52}$$

This result indicates that the total entropy is closely related to the market scale in this market, with the willingness prices of sellers and buyers not changing due to the assumption of traders only trading once at a time.

3.3.3. Comparison of Total Entropy in Centralized and Decentralized Markets

By investigating the characteristics of centralized and decentralized markets, it is obvious that the most prominent difference between these two market structures lies in the role of price in market transactions. In a centralized market, the difference between the willingness price and market price acts as a criterion for sellers and buyers to enter the market. Conversely, in a decentralized market, there is no market price to guide the market participants, and they trade by random matching. Therefore, the decentralized market can be seen as operating without price filtering.

By comparing the different levels of total entropy in centralized (TS^{ce}) and decentralized markets (TS^{de}), we can shed light on the role of price filtering in the transaction uncertainty of a market. As discussed earlier in the analysis of the centralized market, the total entropy exhibits a symmetrical, double-humped, downward profile. Therefore, there exists two price levels at which the total entropy reaches its maximum. These price levels can be derived by solving the equation for the derivative of the total entropy with respect to price, i.e., $TS'(p) = 0$. The expressions for the resulting prices corresponding to the maximum entropy are as follows, and the detailed derivation can be found in Appendix C,

$$p_1 = \frac{\left(5-\sqrt{5}\right)b + \left(5+\sqrt{5}\right)a}{10}, p_2 = \frac{\left(5+\sqrt{5}\right)b + \left(5-\sqrt{5}\right)a}{10}, \tag{53}$$

By substituting Equation (53) into Equation (48), we can obtain the maximum total entropy in the centralized market $\left(TS^{ce}_{max}\right)$ as follows,

$$TS^{ce}_{max} = \frac{k * (b-a)}{2} * ln\frac{\left(5+\sqrt{5}\right)}{\left(5-\sqrt{5}\right)}. \tag{54}$$

Comparing Equations (52) and (54), we can find that the total entropy of the decentralized market surpasses that of the centralized market for all prices. This result indicates that there is a higher uncertainty in transactions within a random-matching market compared to transactions with price filtering. Thus, it is evident that the filtering mechanism plays an effective role in reducing the transaction uncertainty and ensuring successful trading in the centralized market. Figure 8 illustrates the computational results of the total entropy in the centralized and decentralized markets. The double-humped curve is the total entropy in the centralized market, and the horizontal line on the top is the total entropy in the decentralized market.

Based on the computational and simulation results of the total entropy in centralized and decentralized markets, we can conclude that the price-filtering mechanism plays an effective role in reducing market uncertainty. This yields a direct suggestion for policymakers to mitigate market uncertainty, that is, to make the market price public information during the process of transactions between buyers and sellers.

However, how to form a proper market price is a key challenge for policymakers. If the willingness prices of the market participants are available, as commonly occurs in stock markets, a bid–ask mechanism can generate market prices continuously. When the willingness prices are private information, governments could set a market price to regulate markets. However, the possibility that the exogenously set market price is exactly equal to the equilibrium one is so low that market disequilibria are inevitable. As a result, transaction uncertainty during the trading process will present, i.e., the transaction entropy comes out. In this case, the traders on the "long side," have to face transaction uncertainty in the market. As a response, they will adjust their bargaining prices during the transaction process to fulfill their willingness to trade until the market price converges to the equilibrium, where the transaction uncertainty is minimized.

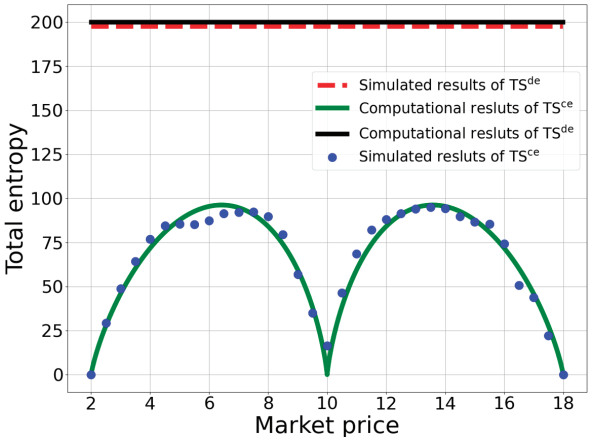


Figure 8. The comparison of the total entropy between centralized and decentralized markets. The green and black lines represent the computational results of centralized and decentralized markets, respectively, while the blue dots and red dash line indicate the simulation results of total entropy in centralized and decentralized markets, respectively. It is observed that the total entropy of the decentralized market is much higher than that of the centralized market at any market price.

In summary, in order to form a public market price when the willingness prices of participants are available, a bid–ask mechanism can work out. Oppositely, when these willingness prices are private, the market participants should be allowed to collectively form a market price by a competitive bargaining process. This self-organized process enables the market price to converge towards the equilibrium one. Although the resulting market price fluctuates over time, transaction uncertainty could be mitigated effectively by this way.

4. Simulation Results

Based on the computational results and theoretical analyses presented above, we develop an agent-based model in this section to simulate the interactions between buyers and sellers in a market and their exchange outcomes. This market system comprises N buyers and N sellers. By enabling them to make transactions, we can observe how some key variables in this market, including the transaction quantity, market surplus, and transaction entropy, change with market price.

At the beginning, we set $N = 200$, and each trader is endowed with a willingness price before trading in the market. The willingness prices of these buyers and sellers are randomly generated within the range of [2,18], following a uniform distribution. To make the simulations meaningful, we set the market price in the model to vary within the range of [2,18]; otherwise, no transactions will occur. By following the change in market price, we can observe the trading behavior of all the traders and the overall market dynamics.

We first perform simulations of a centralized market. The price was set to increase gradually with an increment of 0.5 every step for the simulations, resulting in a total of 33 simulation results corresponding to market prices in the range of [2,18]. With a given market price, buyers and sellers can compare this with their own willingness prices and decide whether they participate in the potential trade or not. Following the rules given by Equations (3) and (5), only the screened participants have a chance to make transactions. According to the short-side principle, some participants may not be able to make a successful deal. The actual quantity of transactions is determined by the short side. Given the initial setup, the simulation results for how the quantity of transactions depends on the market price are plotted as red dots in Figure 2.

To estimate the possibility of successful trades for participants, we conducted 100 random transactions between screened buyers and sellers during the simulation process. Hence, the probability of successful trades for each participant could be computed as a ratio of the number of successful trades to 100 times. Subsequently, by calculating the average ratio of the successful transactions of all the screened buyers, we could obtain the buyers' rationing rate with the given market price. Similarly, by calculating the average ratio of the successful transactions of all the screened sellers, we obtained the sellers' rationing rate corresponding to the given market price. In Figure 3, the red and blue dots represent the simulation results of the rationing rates of the sellers and buyers, respectively. By comparing the rationing rate of the buyers with that of the sellers for each market price, we could further obtain the minimum rationing rates for all given market prices.

Moreover, each successful transaction in the trading process contributes to the market surplus from all the screened participants in once matching. To enhance the reliability of the estimation of the total market surplus, we repeated the random matching of the screened buyers and sellers 20,000 times and took the average as the value of the market surplus for each market price. All the simulation results are represented by the red dots in Figure 4.

Furthermore, for each participant who entered the market through price filtering, we obtained the probability of successful transactions for 100 times of random matching. Based on the calculations of probability for all participants, we could obtain the total entropy for the market. We then performed 200 repetitions of such a calculation of the total entropy and obtained its average value. The simulation results of the total entropy for all market prices are plotted as the blue dots in Figure 8. Then, we could obtain the transaction entropy by dividing the total entropy by the quantity of market transactions. The simulation results of the transaction entropy for all market prices are represented by the red dots in Figure 6.

For the simulation of the entropy in a decentralized market, we followed a similar process as that for obtaining the simulation results of the total entropy in a centralized market. In this kind of market, there is no price-filtering mechanism, so sellers and buyers are directly matched randomly. We first computed the possibility of successful transactions in the market for each participant and then obtained each agent's information entropy accordingly. By summing up all the agents' information entropy, we could obtain the total entropy in the market. We took an average of the total entropy by performing 200 simulations, which is plotted as a dash line in Figure 8. From all the figures mentioned above, we can see that the simulation results are in a high accordance with the computational ones, showing that the theoretical analyses are verified by such an alternative way.

5. Discussion

Market equilibrium is a fundamental concept in economic analyses, and its research involves two primary theories: the Walrasian general equilibrium theory and Marshallian partial equilibrium theory. The Walrasian general equilibrium theory assumes that there is an auctioneer who acts as an information center during the trading process. Prices are gradually adjusted in response to changes in supply and demand until equilibrium is achieved across all markets. However, the existence of the fictional Walrasian auctioneer has been criticized for its inconsistency with reality [36,37]. In contrast, the Marshallian partial equilibrium theory has been widely accepted by economists in market analyses with supply and demand curves. It focuses on individual markets and takes producers and consumers as the market participants, who are matched in a reverse rank during transactions [38,39]. This reverse rank matching refers to willingness bids to buy being typically arranged from high to low in the order book, and willingness asks to sell being arranged from low to high. This way of matching implies that the information of traders' willingness prices is public, leading to transparent transactions and the absence of uncertainty in these transactions. As a result, the concept of transaction entropy is not applicable in this case. However, except for certain call auction markets, the willingness prices of traders are private information in most markets. Therefore, the partial equilibrium theory has limited applications.

In this paper, we argued that traders' willingness prices are private information, and the transaction process can be depicted as a random matching of the market participants in the market. Specifically, in a centralized market, the price maker can just know who has entered the market after setting the market price, but is not aware of the willingness prices of the existing traders. Likewise, in a decentralized market, the willingness prices of traders, which guide them to make decisions, are not known by each other, whether they have successfully made a deal or not. Therefore, we can see that inherent uncertainty exists in actual transactions due to the unavailability of traders' willingness prices. It is necessary to introduce the concept of transaction entropy to characterize this market uncertainty when willingness prices are private information.

Although our work is primarily a theoretical analysis, our findings can be extended to practical applications in various scenarios. Such applications involve many real markets, with stock markets serving as a prime example. In a stock market, the bid–ask mechanism dominates the trading and the market equilibrium can be obtained from the bid–ask prices posed by the market participants, without any transaction uncertainty.

However, stock markets often encounter situations of market disequilibrium, especially when they attain their upper limit or lower limit, leading to transaction uncertainty. The traders would have their responses to this uncertainty, which, in turn, exert significant effects on the market. In normal conditions, when market participants become aware of the presence of uncertainty, they actively adjust their bargaining prices during the bidding process to achieve market-clearing prices. Therefore, transaction uncertainty can enhance traders' sensitivity to market conditions, facilitating more astute investment strategies and accelerating the convergence to an efficient market.

In contrast, in an extreme situation, transaction uncertainty can trigger intense responses and impose negative effects on the market. On one hand, the transaction uncertainty caused by a shortage may engender false prosperity and asset bubbles in the market. Investors, driven by dramatic uncertainty, may engage in excessive speculation, artificially inflating stock prices. However, such a prosperity bubble is unsustainable and could eventually burst, resulting in severe market downturns and financial losses for investors. On the other hand, the transaction uncertainty resulting from a market surplus can lead to market downturns and even cause market panic and crashes. The stock market circuit breakers witnessed during the COVID-19 pandemic are a spirited instance of this. When the market experiences substantial declines and its trading activities exceed the predefined thresholds, a trading halt is automatically executed, with the intention of preventing further market collapse. However, this circuit breaker can exacerbate short-term market panic, intensifying investors' concerns about market instability and risks.

In conclusion, in order to maintain market stability and ensure the positive development of the financial system, we should consider the impacts of transaction uncertainty on markets when formulating risk mitigation measures.

6. Conclusions

Following the statistical approach from Wang and Stanley [28], in which the concept of willingness price was introduced to formulate supply and demand functions, as well as market surplus in a goods market, we expanded the metrics of market performance by introducing a new kind of information entropy to measure the transaction uncertainty in a disequilibrium market.

The first metric of market performance is the realized quantity of transactions. Given a market price in the centralized market, the realized quantity of transactions can be derived from the supply and demand functions. According to the short-side principle, the quantity of transactions is governed by the quantity supplied when the market is in a shortage, while when the market is in a surplus, the realized quantity is governed by the quantity demanded. When the market is at equilibrium, the quantity of transactions is determined by the cross-point of the supply and demand curves. We find that the quantity of transactions

reaches its maximum at equilibrium, and it will decrease when the market price departs from the market-clearing point to a shortage or surplus.

The second metric is market surplus, which is a traditional index of market efficiency. In the calculation of realized market surplus, the rationing rate is indispensable, which is defined as the ratio of the actual transaction quantity to the desired one. Sellers and buyers have their rationing rates, which are dependent on the market status. It can be proved that the realized market surplus is at its highest when the market is at equilibrium, since it increases in a shortage and decreases in a surplus.

We argue that transaction uncertainty is a new dimension of market performance. To measure this kind of uncertainty, we first introduced transaction entropy to reflect the level of uncertainty in individual transactions. When a market is at equilibrium, the transaction entropy is zero. Otherwise, we will have positive transaction entropy when a market is in disequilibrium. It has a decreasing trend in a shortage, but an increasing trend in a surplus. The results indicated that there is no transaction uncertainty at equilibrium, and disequilibrium leads to a higher transaction uncertainty. We then made a comparison of the total entropy in centralized and decentralized markets and found that it is lower in a centralized market than a decentralized market. This means that the price-filtering mechanism plays a key role in reducing market uncertainty.

Finally, we argue that market uncertainty is necessary in analyzing market performance, since willingness prices are private information. Traditional approaches to market equilibrium assume that information of the willingness prices of traders is available, and traders engage in reverse rank matching when they make transactions. However, these assumptions are unrealistic, and the willingness prices of traders can only guide them to choose whether to enter market or not. Once they have entered a market, they are randomly matched to trade with each other, which must incur uncertainty in transactions.

Author Contributions: Conceptualization, H.Z. and X.L.; methodology, X.L.; software, X.L.; formal analysis, H.Z. and Y.W.; investigation, H.Z. and X.L.; writing—original draft preparation, H.Z.; writing—review and editing, X.L. and Y.W.; visualization, H.Z.; supervision, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Appendix A. The Derivation of Concavity of Total Entropy in the Centralized Market

As shown in Equation (49), the total entropy when $a < p < p^*$ can be expressed as:

$$TS = -\begin{bmatrix} Q_{SC}(p) ln Q_{SC}(p) + (Q_{DC}(p) - Q_{SC}(p)) ln(Q_{DC}(p) - Q_{SC}(p)) \\ -Q_{DC}(p) ln Q_{DC}(p) \end{bmatrix}.$$

Taking the first derivation of Equation (49), which can be expressed as:

$$\begin{aligned} TS' &= -\begin{bmatrix} Q'_{SC}(p)(lnQ_{SC}(p) - ln(Q_{DC}(p) - Q_{SC}(p))) \\ +Q'_{DC}(p)(ln(Q_{DC}(p) - Q_{SC}(p)) \\ -lnQ_{DC}(p))] \end{bmatrix} \\ &= Q'_{SC}(p)(ln(Q_{DC}(p) - Q_{SC}(p)) - lnQ_{SC}(p)) \\ &\quad -Q'_{DC}(p)(ln(Q_{DC}(p) - Q_{SC}(p)) \\ &\quad -lnQ_{DC}(p)) \\ &= F_{SC}(p)(ln(Q_{DC}(p) - Q_{SC}(p)) - lnQ_{SC}(p)) \\ &\quad +F_{BC}(p)(ln(Q_{DC}(p) - Q_{SC}(p)) \\ &\quad -lnQ_{DC}(p)) \\ &= 2k * ln(Q_{DC}(p) - Q_{SC}(p)) - k * lnQ_{SC}(p) - k * lnQ_{DC}(p). \end{aligned} \tag{A1}$$

Then, we can identify the concavity of the TS function by taking the second derivative as follows,

$$\begin{aligned} TS'' &= -\frac{(2k)^2}{Q_{DC}(p)-Q_{SC}(p)} - \frac{k^2}{Q_{SC}(p)} + \frac{k^2}{Q_{DC}(p)} \\ &= -\frac{1}{(Q_{DC}(p)-Q_{SC}(p)) * Q_{SC}(p)*Q_{DC}(p)} * [(2k)^2 \\ &\quad *Q_{SC}(p) * Q_{DC}(p) + k^2 * Q_{DC}(p) \\ &\quad *(Q_{DC}(p) - Q_{SC}(p)) - k^2 * Q_{SC}(p) \\ &\quad *(Q_{DC}(p) - Q_{SC}(p))] \\ &= -\frac{k^2}{(Q_{DC}(p)-Q_{SC}(p))*Q_{SC}(p)*Q_{DC}(p)} \\ &\quad *(Q_{SC}(p) + Q_{DC}(p))^2 < 0. \end{aligned} \tag{A2}$$

Therefore, TS is a concave function when $a < p < p^*$. With the similar derivation process, we can deduce that there are two concave curves symmetric about $p = p^*$ within the interval $[a, b]$.

Appendix B. The Total Entrpy in the Decentralized Market

To facilitate obtaining an integral expression of Equation (51), we split it into two components, that is, the sellers' total entropy $TS_S^{de} = \int_a^b F_S(v)H_S(v)dv$ and the buyers' total entropy $TS_B^{de} = \int_a^b F_B(v)H_B(v)dv$. Combining this with Equation (31), we derive the buyers' market entropy first, which can be expressed as,

$$\begin{aligned} TS_B^{de} &= \int_a^b F_B(v)(-1)[E_B lnE_B + (1-E_B)ln(1-E_B)]\nu \\ &= -k\int_a^b (\tfrac{v-a}{b-a} ln \tfrac{v-a}{b-a} + \tfrac{b-v}{b-a} ln \tfrac{b-v}{b-a})\nu \\ &= -\tfrac{k}{b-a}\int_a^b (v-a)ln(v-a) + (b-v)ln(b-v) \\ &\quad -(b-a)ln(b-a)\nu \\ &= -\tfrac{k}{b-a}\int_a^b (v-a)ln(v-a) + (b-v)ln(b-v)dv + k(b-a)ln(b-a). \end{aligned} \tag{A3}$$

Supposing $x = v - a, t = b - v$, Equation (A3) can be rewritten as:

$$\begin{aligned} TS_B^{de} &= \left(-\tfrac{k}{b-a}\right)[\int_0^{b-a} xlnxdx + \int_0^{b-a} tlntdt] \\ &\quad +k(b-a)ln(b-a) \\ &= 2\left(-\tfrac{k}{b-a}\right)\int_0^{b-a} xlnxdx + k(b-a)ln(b-a). \end{aligned} \tag{A4}$$

where $2\int_0^{b-a} xlnxdx = (b-a)^2 ln(b-a) - \frac{1}{2}(b-a)^2$. Therefore, the final total entropy of the buyers in the market with random matching can be derived as:

$$\begin{aligned} TS_B^{de} &= \left(-\tfrac{k}{b-a}\right)\left[(b-a)^2 ln(b-a) - \tfrac{1}{2}(b-a)^2\right] \\ &\quad +k(b-a)ln(b-a) \\ &= \tfrac{1}{2}k(b-a). \end{aligned} \tag{A5}$$

Then, we can derive the final total entropy of the sellers' total entropy through a similar derivation process, given by:

$$TS_S^{de} = \frac{1}{2}k(b-a) \tag{A6}$$

Summing up Equations (A5) and (A6), we obtain the final expression of the total entropy in the decreolization market, as shown in Equation (52).

Appendix C. The Prices Which Correspond the Maximum Total Entropy in the Centralized Market

According to Appendix A, we can know that the condition of the maximum total entropy in the centralized market is $TS' = 0$. According to Equation (A2), we obtain:

$$(Q_{DC}(p) - Q_{SC}(p))^2 = Q_{DC}(p) * Q_{SC}(p). \tag{A7}$$

Combining (A7) with Equations (45) and (46), we can obtain the price when the total entropy is maximized, which is:

$$p_1 = \frac{(5-\sqrt{5})b + (5+\sqrt{5})a}{10},\ a < p < p^*. \tag{A8}$$

With the similar derivation process, we obtain:

$$p_2 = \frac{(5+\sqrt{5})b + (5-\sqrt{5})a}{10},\ p^* < p < b. \tag{A9}$$

References

1. Buzzell, R.D. Market Functions and Market Evolution. *J. Mark.* **1999**, *63*, 61–63. [CrossRef]
2. Jurado, K.; Ludvigson, S.C.; Ng, S. Measuring Uncertainty. *Am. Econ. Rev.* **2015**, *105*, 1177–1216. [CrossRef]
3. Sniazhko, S. Uncertainty in Decision-Making: A Review of the International Business Literature. *Cogent Bus. Manag.* **2019**, *6*, 1650692. [CrossRef]
4. Bouattour, M.; Martinez, I. Efficient Market Hypothesis: An Experimental Study with Uncertainty and Asymmetric Information. *Financ. Contrôle Strat.* **2019**, *22*, 4. [CrossRef]
5. Zhou, R.; Cai, R.; Tong, G. Applications of Entropy in Finance: A Review. *Entropy* **2013**, *15*, 4909–4931. [CrossRef]
6. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
7. Gulko, L. The Entropic Market Hypothesis. *Int. J. Theor. Appl. Financ.* **1999**, *2*, 293–329. [CrossRef]
8. Fama, E.F. Efficient Capital Markets: A Review of Theory and Empirical Work. *J. Financ.* **1970**, *25*, 383–417. [CrossRef]
9. Fama, E.F. Efficient Capital Markets: II. *J. Financ.* **1991**, *46*, 1575–1617. [CrossRef]
10. Vinţe, C.; Ausloos, M. The Cross-Sectional Intrinsic Entropy—A Comprehensive Stock Market Volatility Estimator. *Entropy* **2022**, *24*, 623. [CrossRef]
11. Dimpfl, T.; Peter, F.J. Analyzing Volatility Transmission Using Group Transfer Entropy. *Energy Econ.* **2018**, *75*, 368–376. [CrossRef]
12. Huang, X. A Review of Uncertain Portfolio Selection. *J. Intell. Fuzzy Syst.* **2017**, *32*, 4453–4465. [CrossRef]
13. Huang, X. Portfolio Analysis: From Probabilistic to Credibilistic and Uncertain Approaches. In *Portfolio Analysis: From Probabilistic to Credibilistic and Uncertain Approaches*; Springer: Berlin, Germany, 2010; pp. 117–156.
14. Philippatos, G.C.; Wilson, C.J. Entropy, market risk, and the selection of efficient portfolios. *Appl. Econ.* **1972**, *4*, 209–220. [CrossRef]
15. Ou, J. Theory of portfolio and risk based on incremental entropy. *J. Risk Financ.* **2005**, *6*, 31–39. [CrossRef]
16. Xu, J.P.; Zhou, X.Y.; Wu, D.D. Portfolio Selection Using λ Mean and Hybrid Entropy. *Ann. Oper. Res.* **2011**, *185*, 213–229. [CrossRef]
17. Gulko, L. Dart Boards and Asset Prices: Introducing the Entropy Pricing Theory. *Adv. Econom.* **1997**, *12*, 237–276.
18. Gulko, L. The entropy theory of stock option pricing. *Int. J. Theor. Appl. Financ.* **1999**, *2*, 331–355. [CrossRef]
19. Gulko, L. The entropy theory of bond option pricing. *Int. J. Theor. Appl. Financ.* **2002**, *5*, 355–383. [CrossRef]
20. Akerlof, G.A. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *Q. J. Econ.* **1970**, *84*, 488–500. [CrossRef]
21. Stiglitz, J.E. Information and the Change in the Paradigm in Economics. *Am. Econ. Rev.* **2002**, *92*, 460–501. [CrossRef]
22. Wankhade, L.; Dabade, B.M. Analysis of Quality Uncertainty Due to Information Asymmetry. *Int. J. Qual. Reliab. Manag.* **2006**, 23. [CrossRef]
23. Wankhade, L.; Dabade, B. Quality Uncertainty Due to Information Asymmetry. In *Quality Uncertainty and Perception. Contri-butions to Management Science*; Physica: Heidelberg, Germany, 2010; pp. 13–25.
24. Baker, S.R.; Bloom, N.; Davis, S.J. Measuring Economic Policy Uncertainty. *Q. J. Econ.* **2016**, *131*, 1593–1636. [CrossRef]
25. Kang, W.; Ratti, R.A. Oil Shocks, Policy Uncertainty and Stock-Market Returns. *J. Int. Financ. Mark. Inst. Money* **2014**, *26*, 305–318. [CrossRef]
26. Arouri, M.; Estay, C.; Rault, C.; Roubaud, D. Economic Policy Uncertainty and Stock Markets: Long-Run Evidence from US. *Financ. Res. Lett.* **2016**, *18*, 136–141. [CrossRef]
27. Wigle, R. *Partial Equilibrium Analysis: A Primer*; Wilfrid Laurier University: Waterloo, ON, Canada, 2004.
28. Yougui, Wang; H. E. Stanley. Statistical approach to partial equilibrium analysis. *Phys. A Stat. Mech. Appl.* **2009**, *388*, 1173–1180. [CrossRef]
29. Carruthers, B.G. From uncertainty toward risk: The case of credit ratings. *Socio-Econ. Rev.* **2013**, *11*, 525–551. [CrossRef]
30. Olbryś, J.; Ostrowski, K. An Entropy-Based Approach to Measurement of Stock Market Depth. *Entropy* **2021**, *23*, 568. [CrossRef]
31. Miao, J. A Search Model of Centralized and Decentralized Trade. *Rev. Econ. Dyn.* **2006**, *9*, 68–92. [CrossRef]
32. Peivandi, A.; Vohra, R.V. Instability of Centralized Markets. *Econometrica* **2021**, *89*, 163–179. [CrossRef]
33. Malamud, S.; Rostek, M. Decentralized Exchange. *Am. Econ. Rev.* **2017**, *107*, 3320–3362. [CrossRef]
34. Rubinstein, A.; Wolinsky, A. Decentralized Trading, Strategic Behaviour and the Walrasian Outcome. *Rev. Econ. Stud.* **1990**, *57*, 63–78. [CrossRef]
35. Blouin, M.R.; Serrano, R. A Decentralized Market with Common Values Uncertainty: Non-Steady States. *Rev. Econ. Stud.* **2001**, *68*, 323–346. [CrossRef]

36. De Fraja, G.; Sakovics, J. Walras Retrouvé: Decentralized Trading Mechanisms and the Competitive Price. *J. Political Econ.* **2001**, *109*, 842–863. [CrossRef]
37. Tesfatsion, L. *Walrasian Equilibrium: A Critique, Class Lecture Notes*; Department of Economics, Iowa State University: Ames, IA, USA, 2004.
38. Plott, C.; Roy, N.; Tong, B. Marshall and Walras, Disequilibrium Trades and the Dynamics of Equilibration in the Continuous Double Auction Market. *J. Econ. Behav. Organ.* **2013**, *94*, 190–205. [CrossRef]
39. Hudik, M. The Marshallian Demand Curve Revisited. *Eur. J. Hist. Econ. Thought* **2020**, *27*, 108–130. [CrossRef]

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
www.mdpi.com

Entropy Editorial Office
E-mail: entropy@mdpi.com
www.mdpi.com/journal/entropy

MDPI

Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-0416-0