

Article

Interpretable Academic Team Formation on Heterogeneous Information Networks: Constructive Heuristics with Explicit Organizational Affiliation

Nuri Özdemir *  and Hadi Gökçen 

Graduate School of Informatics, Gazi University, Ankara 06500, Türkiye; hgokcen@gazi.edu.tr

* Correspondence: nuri.ozdemir@hotmail.com

Abstract

Assembling expert teams under strict skill-coverage and communication-distance constraints is a fundamental challenge in collaborative knowledge work. Existing learning-based approaches excel at probabilistic link prediction but cannot reliably enforce hard logical constraints or provide interpretable justifications. This study presents a constructive heuristic framework for team formation on Heterogeneous Information Networks (HINs), integrating authors, papers, departments, and organizations into a unified graph database. Seven algorithms exploit distinct structural features—topological proximity, co-authorship history, organizational affiliation, citation impact, and temporal recency—and guarantee constraint satisfaction by construction. Experiments on a subset of the AMiner DBLP dataset ($\approx 625,000$ nodes, $973,000$ edges) covering $12,045$ formation requests across 147 configurations show that algorithm choice is the dominant determinant of runtime, while skill frequency governs feasibility: success rates decline from 81.2% under abundant keywords to 14.9% in the long tail. Algorithms further form statistically distinct clusters in communication cost, and specialized heuristics operate in nearly disjoint solution spaces—supporting a toolbox approach over single-algorithm deployment. These results provide actionable selection guidance: proximity-based algorithms for communication-efficient teams; citation- or recency-aware algorithms when impact matters; cohesion-based algorithms when internal collaboration is the priority.

Keywords: team formation; heterogeneous information networks; constructive heuristics; graph database; organizational affiliation; interpretability; academic collaboration



Academic Editor: Olga Kurasova

Received: 18 April 2026

Revised: 28 June 2026

Accepted: 29 June 2026

Published: 6 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Team formation is a critical challenge across diverse domains, from optimizing student project groups [1] to assembling expert reviewers [2] and sports teams [3]. As collaborative tasks grow in scope and complexity, the need for analytical methods that can systematically identify compatible, high-performing combinations of experts has become increasingly clear [4].

While early approaches focused on simple skill matching, modern research leverages social network analysis to model the communication costs and collaborative potential of candidate teams. Recent work on Heterogeneous Information Networks (HINs) often adopts embedding-based machine learning models (e.g., GNNs, contrastive learning) [5–8]. Although these methods are powerful for probabilistic link prediction, they compress relational semantics into latent vectors, making team recommendations difficult to interpret.

Furthermore, probabilistic models can struggle to enforce the strict and project-dependent hard constraints required for practical team formation, such as guaranteeing specific skill coverage or bounding maximum communication distance.

Motivated by the evidence that co-authorship relations and organizational affiliations are strong predictors of academic collaboration [9,10], this work focuses on a constructive HIN framework with essential entities (author, paper, department, and organization). Unlike “black-box” approaches, our graph-database-driven heuristic design ensures that every recommendation is topologically valid and semantically explainable. This structure also serves as an extensible foundation, allowing future integration of new nodes such as patents or projects without requiring opaque feature extraction.

We position this work as a foundational step towards a unified, parametric team-formation model. Rather than claiming a single globally optimal solution, we introduce a *toolbox* of seven constructive heuristic algorithms. Concretely, the toolbox accepts a set of required skill keywords and a maximum communication distance as inputs, and returns—for each selected algorithm—a team of authors represented as a subgraph of the HIN, where author nodes cover all requested skills and their inter-author distances satisfy the specified bound. The seven algorithms differ in which network features they prioritize (topological proximity, co-authorship history, organizational affiliation, citation impact, or temporal recency), giving practitioners a menu of strategies to match their specific context. Because every recommendation is derived from explicit graph traversal, the rationale for each team member’s inclusion remains fully auditable.

A key practical challenge in this setting is skill scarcity: when required expertise is rare in the network, no algorithm can guarantee team completion regardless of its traversal strategy. Quantifying how skill frequency and network sparsity interact with algorithmic choice—and identifying which heuristics degrade most gracefully under sparse conditions—is therefore as important as algorithm design itself. To this end, we evaluate all seven algorithms across a factorial design of 12,045 team-formation attempts spanning 147 configurations of keyword frequency, team size, and algorithm.

The main contributions of this work are as follows:

- **Scale and scope:** Building on prior work that introduced organizational affiliations into HIN-based team formation at national scale [11], this study extends the framework to an international corpus (AMiner DBLP, $\approx 625,000$ nodes, 973,000 edges), capturing cross-institutional and cross-national collaboration patterns.
- **Algorithm design:** Seven independently operable heuristic algorithms, collectively spanning topological, relational, temporal, and impact-based dimensions of the network, are designed so that any returned team is guaranteed to satisfy all specified constraints by construction.
- **Interpretability:** Every recommendation is produced by explicit graph traversal, making the inclusion of each team member fully traceable—without latent embeddings or black-box inference.
- **Empirical validation:** A large-scale experiment (12,045 formation attempts, 147 configurations) with two-way ANOVA, Tukey HSD post-hoc tests, and pairwise Jaccard analysis provides statistically grounded selection guidance across algorithm, team-size, and skill-frequency dimensions.
- **Extensibility:** The graph database schema accommodates new node types (e.g., patents, funded projects) without altering existing traversal logic, supporting incremental enrichment of the network.

To guide the empirical evaluation, we formulate the following research questions:

- **RQ1 (Feasibility and Parameter Sensitivity):** Can constructive heuristics on HINs generate complete teams under strict skill-coverage and distance constraints? How does skill frequency constrain feasibility and success rates across algorithms?
- **RQ2 (Algorithmic Profiles):** How do the seven heuristics compare in communication cost and runtime as team size grows?
- **RQ3 (Structural Diversity):** Do the algorithms converge on similar team compositions, or do they exploit distinct regions of the solution space—and which algorithm combinations offer the most complementary coverage for parallel deployment?

The remainder of this paper is organized as follows. Section 2 reviews related work on team formation and HIN-based approaches. Section 3 describes the dataset, graph construction, algorithm design, and experimental setup. Section 4 reports and discusses the empirical findings. Section 5 concludes with limitations and directions for future work.

2. Related Work

Team formation literature examines methods for composing teams that increase the chances of project success across various domains, including student groups [12,13], player selections in multiplayer sports [3,14], collaborating academics [15,16], reviewer groups [2], and engineering or construction project teams [17,18]. Factors such as experience level, workload, ease of communication, and geographical distance influence how teams are assembled. Traditional manual team formation is often subjective and prone to human error or bias [4,19,20]. Consequently, many studies [21,22] argue for analytical methods that can systematically explore the space of possible teams to obtain diverse, higher-performing combinations while explicitly incorporating technical and social factors. Within this broader line of work, an important strand focuses on exploiting network structure to model communication and compatibility between potential team members.

Initial models [23,24] focused primarily on matching the skills of potential members with task requirements. Some studies [25–29] incorporated social factors but did not model the team as a connected structure within the social graph. Network parameters were first integrated into an analytical team-formation model in [30], which treated a team as a subset of a weighted co-authorship network and minimized communication-cost functions (network diameter and minimum spanning tree) using the RarestFirst, CoverSteiner, and EnhancedSteiner algorithms. Subsequent work generalized this approach to variable members per skill [31], expertise levels [32], and top- k ranking cost functions [33].

Subsequent research progressively refined team-formation constraints, addressing fairness and its trade-off with communication cost [34,35], as well as workload capacity [36], subgraph-density objectives [37], distance thresholds [38], and interpersonal trust [39].

The Multiple Team Formation Problem (MTFP) was introduced in [40], where individuals split their time across projects and interpersonal relations are modeled. Similarly, Selvarajah et al. [41] integrated communication cost, expertise level, collective trust, profile similarity, and geographical proximity into the TFP. A two-stage approach was proposed in [42], which first generates candidate teams using a heuristic and then selects the most suitable ones, rather than aiming for global optimality from the outset. More recently, Mahajan et al. [43] incorporated privacy attributes into the team formation process.

A reliability parameter was introduced in [44] to model member dropout risk, proposing a structure consisting of main and backup team members. Furthermore, research from the IDEA Lab [5,45–47] utilized graph-based approaches to preserve or improve existing teams—through replacement, refinement, expansion, or shrinkage—rather than forming them from scratch. This body of work ranges from user-driven team restructuring based

on graph similarity to reinforcement learning for dynamic optimization and the use of clustering-based Graph Neural Networks for sub-team replacement.

Several studies have focused on balancing multiple, often conflicting, structural and performance-related objectives in team formation. For example, Yin et al. [48] focused on social diversity by selecting members from different cliques to reduce bias in tasks such as proposal reviews. To address sparsity and scalability, Hamidi Rad et al. [49] introduced Variational Bayesian Neural Networks (VBNNs) to learn mappings between required skills and experts. Balancing the trade-offs between team diversity and communication cost, Gómez-Zará et al. [50] adopted a multi-objective evolutionary approach based on NSGA-II to yield a Pareto front of non-dominated solutions. Similarly, Juárez et al. [51] utilized NSGA-II with specialized 'grey-box' crossover operators to maximize the expertise and collaborative density of recommended teams. In a similar vein, Curiac et al. [16] applied the NSGA-II algorithm to interdisciplinary research networks, formulating a multi-objective optimization model that balances candidates' technical expertise and interpersonal collaboration history extracted from rich bibliometric metadata. A hybrid model combining Graph Convolutional Networks (GCN) and Mixed-Integer Linear Programming (MILP) was proposed in [52] to form balanced teams from a homogeneous network. More recently, Dara et al. [6] introduced a retrieval-augmented neural model for balanced and cohesive team recommendations using a learning-based framework.

Some recent studies [53–56] represent experts as network nodes but derive communication costs from skill overlap rather than co-authorship history. Specifically, an improved particle swarm optimizer using a novel swap operator was proposed in [53], while the Jaya algorithm was adapted with a modified swap operator in [54]. Furthermore, Rehman et al. [56] employed a search space reduction strategy. More recently, El-Ashmawi et al. [55] applied a modified Golden Eagle Optimizer to jointly optimize communication cost, team cardinality, and member similarity.

While many earlier studies relied on homogeneous graphs, the use of Heterogeneous Information Networks (HINs) [57,58], multi-relational networks [59], or multigraphs [60] enables more expressive representations of scientific collaborations. By modeling multiple entity and relation types, these structures support richer formulations of the team formation problem. For instance, Yang et al. [61] were among the first to formulate team formation in a heterogeneous setting by modeling skills as explicit nodes connected to authors. Although introducing skills as nodes does not fundamentally alter the problem definition, the authors argue that this heterogeneous structure improves computational efficiency. Building on this work, Yang et al. [62] further extended heterogeneous modeling by incorporating additional node types (e.g., venues, topics, and papers) and relation types (e.g., authorship and citation), calculating relationship strength based on meta-paths connecting author nodes.

More recent approaches have utilized learning-based representations and embedding methods. For instance, transfer learning was applied in [7] to improve the representation of experts and skills by considering past collaboration patterns within heterogeneous networks. Similarly, Hamidi Rad et al. [8] proposed a subgraph representation learning approach that treats team discovery as a representation learning task, capturing both local and global relationships among various node types such as authors, papers, venues, and skills. To facilitate inter-institutional team formation, an embedding framework was proposed in [63] that explicitly incorporates the research interests and strategic purposes of scholars' parent institutions through a heterogeneous bibliographic network.

While our primary focus is team formation, the literature on expert finding and collaborator recommendation is inherently linked to this domain, as these tasks often share identical network representations and learning paradigms. For example, Liu et al. [64] constructed a heterogeneous information network to learn node embeddings using meta-

path-based random walks and applied cosine similarity to recommend the top-N collaborators for a target author. To reduce the computational cost associated with heavy text processing, Gao et al. [65] proposed a transfer learning framework that identifies topic-sensitive experts by leveraging structural similarities to a small set of seed experts within a heterogeneous academic network, generating embeddings from higher-order network relationships and ranking candidates based on their proximity to a prototype embedding derived from seed experts.

More recently, a transformer-based framework for scientific collaborator recommendation on a dynamic, homogeneous network was introduced in [66], integrating temporal features and contextual semantics while distinguishing co-author contributions via directed, text-rich edges. Although Wang et al. [67] did not adopt a heterogeneous information network, their use of a heterogeneous hypergraph with multiple relation types highlights the importance of modeling relational heterogeneity in collaborator recommendation and team-related studies. Finally, a heterogeneous network of papers, patents, authors, and inventors was analyzed in [68]. Although that study does not focus on recommendation, it demonstrates that specific structural and relational configurations are strongly associated with team novelty and impact. This finding reinforces the practical importance of the heterogeneous attributes utilized in our team-formation toolbox.

Positioning Relative to Existing Methods

Most optimization-based team formation studies build on the foundational work in [30], which defined communication-cost objectives, such as diameter, on collaboration networks. Subsequent research [61,62] extended this to Heterogeneous Information Networks (HINs) by incorporating skill, venue, and topic nodes. However, these studies largely neglected *organizational affiliation*, which remains a decisive factor in the assembly and success of real-world academic teams.

In a preliminary study, the present authors addressed this gap [11] by constructing a HIN from national academic data that includes both co-authorship and institutional affiliation ties. That study demonstrated that incorporating organizational links improves team completeness rates and reduces communication costs. However, the analysis was limited to a single national corpus at a relatively modest scale and evaluated only three algorithmic variants. The present work significantly extends this framework by utilizing a large-scale international dataset (AMiner DBLP, $\approx 625,000$ nodes, 973,000 edges), expanding the algorithm suite to seven heuristics covering a broader range of structural priorities, and providing a comprehensive statistical comparison across all configurations.

While institutional nodes were integrated in [63], that approach—similar to most recent state-of-the-art (SOTA) methods [6,8,66]—relies on embedding-based learning. As summarized in the qualitative comparison provided in Section 4.2, such probabilistic models are difficult to interpret and often cannot ensure that all hard constraints, such as exact skill coverage, are met. In contrast, our work establishes a transparent, constructive framework that exploits a multi-relational HIN structure—explicitly modeling departmental hierarchies—and implements heuristic algorithms. These algorithms ensure that whenever a team is returned, it represents a topologically valid solution satisfying all specified constraints. In cases where the network structure makes a complete team impossible under the given constraints, the system explicitly reports failure rather than implicitly relaxing the requirements.

We also focus on keyword scarcity, which arises from the ‘long tail’ distribution of expert skills [20]. Preliminary analysis suggests that rare skills frequently limit the success of strict team formation, particularly when subjecting the network to the hard constraints enforced by constructive heuristics [30], capacity-limited models [36], or Integer Linear

Programming [44,69]. Consequently, we systematically investigate the response of our algorithms to skill scarcity. As noted previously, the impact of skill scarcity is often masked by the probabilistic relaxation inherent in modern embedding-based frameworks.

3. Methods

3.1. Dataset and Network Construction

Despite the advantages they propose, many team formation studies rely on synthetic [39–41,69], semi-synthetic [43] or real-small datasets [42], because the key variables they use (e.g., interpersonal trust) are often unavailable in public corpora and the underlying optimization problems are computationally hard. Within the subfield of academic team formation and expert/collaborator recommendation, most work based on real data [5,8,30] uses the DBLP dataset, due to its clean, curated metadata and simple schema, with fewer studies drawing on alternatives such as AMiner [64,67], Scopus [63] or ACM Digital Library data [65]. A detailed comparison of datasets and network scales used in selected studies is provided in Supplementary Material S1.

In this study, we use the AMiner DBLP–Citation–Network V14 dataset [70,71]. The decisive factor for selecting this source over standard DBLP releases is the inclusion of author affiliation strings. While standard DBLP XML snapshots provide robust bibliographic metadata, they typically lack the institutional affiliation data necessary to construct the organizational layer of our heterogeneous network. AMiner addresses this gap by enabling the creation of affiliation edges and the implementation of institution-aware algorithms. Secondary motivations include AMiner’s broader coverage of papers and the availability of paper-to-paper citation links, which we consider valuable for potential future extensions of this work.

Because the affiliation strings in AMiner are unformatted, department and organization names are extracted and canonicalized via rule-based cleaning and reference lookups to reduce duplication and noise. This process enables the creation of a realistic organizational structure where researchers within the same department are positioned closer to one another, while those in different departments of the same organization are separated accordingly in distance-based calculations.

Heterogeneous information network (HIN) and schema: We model the bibliographic data as a heterogeneous information network $G = (V, E)$, where V is the set of nodes and E is the set of edges. The node set V consists of four types of entities: Authors (A), Papers (P), Organizations (O), and Departments (D). The edge set E contains three types of relationships: WRITTEN, WORKS_AT, and INCLUDES. The schema of the constructed HIN is illustrated in Figure 1.

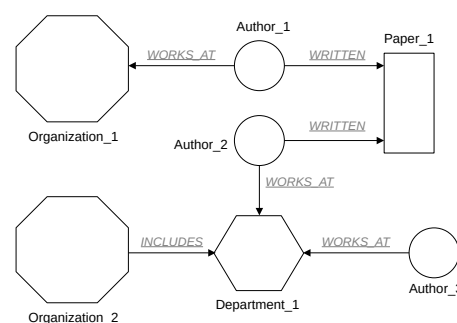


Figure 1. Conceptual model of the property graph network.

To preserve the finest structure available, we apply the following rules when constructing network:

- WRITTEN edges connect Author (A) nodes to Paper (P) nodes

$$A \xrightarrow{\text{WRITTEN}} P.$$

- If an affiliation record contains both department and organization fields, create Department node D and Organization node O ; connect authors to department, and department to organization via

$$A \xrightarrow{\text{WORKS_AT}} D, \quad O \xrightarrow{\text{INCLUDES}} D.$$

- If only an organization is available in affiliation record, create Organization node O and add

$$A \xrightarrow{\text{WORKS_AT}} O.$$

- An author may have multiple WORKS_AT edges (multiple affiliations); we keep all links derived from the records.

Node and edge attributes:

- Author: author_id, author_name, skills.
- Paper: paper_id, paper_title, keywords, n_citation (citation count), years_passed (years since publication; may be NULL).
- Organization/Department: identifiers and names.
- Edges include no properties; temporal and citation filters are enforced by inspecting paper node attributes.

Author distance in the HIN: Distance between two authors $A_i, A_j \in A$ is measured as the shortest path length in G under the chosen edge-type restriction:

In our Neo4j encoding, an author–paper–author co-authorship walk has length 2; analogously, two authors affiliated to the same department are 2 steps away from each other. When an algorithm restricts to co-authorship only, shortest-path searches consider only WRITTEN meta-paths ($A \rightarrow P \leftarrow A$). When restricted to co-working, meta-paths use affiliation links ($A \rightarrow D \leftarrow A$, $A \rightarrow O \leftarrow A$, $A \rightarrow D \rightarrow O \leftarrow A$, $A \rightarrow D \rightarrow O \leftarrow D \leftarrow A$, etc., as appropriate).

Constructing author skills and keyword filtering: To limit the corpus to a coherent domain and to construct author skill sets, we proceeded as follows:

1. We assembled a list of ≈ 600 Industrial-Engineering domain keywords (provided as Supplementary Material S2).
2. For each paper we extracted 10 top terms from the abstract using scikit-learn's TfidfVectorizer [72], and merged these with title words and author-supplied keywords to form combined_keywords.
3. We computed cosine similarity between each paper's combined_keywords and the domain keywords vector; papers with similarity above an empirical threshold (0.08) were retained in the filtered corpus.
4. For each retained paper we propagated its keywords to all co-authors; an author's skills attribute is the union of keywords from their retained papers.

After filtering AMiner dataset and constructing the HIN, the graph comprises nearly 625,000 nodes and 973,000 edges (see Table 1), placing it within the upper range of dataset sizes commonly used in prior network-based team formation and collaborator recommendation studies (see Supplementary Material S1).

Table 1. Summary of node and edge counts in the constructed heterogeneous information network.

Type	Name	Count	Total
Node	Author	350,404	625,776
Node	Paper	166,198	
Node	Organization	56,231	
Node	Department	52,943	
Edge	WRITTEN	533,352	973,239
Edge	WORKS_AT	386,944	
Edge	INCLUDES	52,943	

3.2. Heuristic Algorithms

We present a suite of seven constructive heuristic algorithms designed to progress incrementally from simple topological searches to complex, multi-objective retrieval strategies. The algorithms are classified into two categories: *Connectivity-Based* (focusing on graph topology) and *Attribute-Based* (incorporating node/edge metadata).

To ensure reproducibility, detailed pseudocode for all algorithms is provided in Supplementary Material S3.

3.2.1. Initialization and Search Space Reduction

To address the computational intractability of scanning the entire network, all algorithms begin with a **Heuristic Seed Selection** phase. This step acts as a space-reduction mechanism as in [56], pruning the search space to a manageable subset of high-potential candidates ($f \ll N$) before the constructive phase begins. We implement three distinct filtering policies:

1. **Global Connectivity Filter: Filter Authors By Degree** Used by ACET, OAT, COT, and TAT. This policy selects seed authors who possess the first required skill (k_1) and have the highest node reachability in $d_{\max}$ steps, across *all* edge types. It prioritizes “central hubs” who have diverse connections (co-authors, colleagues), maximizing the probability of finding valid paths to subsequent team members.
2. **Co-Authorship Filter: Filter Authors By Written Degree** Used by CAT and PRT. This filter restricts the reachability calculation to WRITTEN edges only. It ensures that the starting seed authors are central within the collaboration graph and possess a substantial record of published research.
3. **Citation Impact Filter: Filter Authors By High Citation** Used by CIT. Instead of topological degree, this policy ranks candidates by the cumulative citation count of their papers containing skill k_1 . It prioritizes “Star” experts, assuming that high-impact authors act as effective attractors for other qualified members.

By decoupling the seed selection logic from the traversal logic, we ensure that each algorithm starts with more appropriate candidates for its specific heuristic strategy (e.g., CIT starts with high-impact nodes, while CAT starts with high-collaboration nodes). On the other hand, for specific needs, the seed selection can be swapped or modified without altering the core algorithm.

Temporal Missing-Data Strategy

For algorithms incorporating temporal constraints (TAT and CIT), the framework provides a configurable mechanism to handle missing publication-date metadata. Users can select from three distinct strategies:

1. **Optimistic:** Missing values are assigned a recency score within the validity threshold ($t_{\text{imputed}} \leq t_{\text{max}}$). It prioritizes recall, treating temporally unverified connections as valid to maximize the candidate pool.

2. **Pessimistic:** Missing values are assigned a score exceeding the threshold ($t_{imputed} > t_{max}$). This conservative approach prioritizes precision, strictly filtering out any connection that lacks verifiable temporal data.
3. **Stochastic:** Missing values are assigned a random year difference distributed within a plausible range (e.g., $U(0, 30)$). This option allows users to evaluate algorithmic robustness under conditions of random data noise.

Note: In the experimental design, we exclusively utilize the deterministic strategies (Optimistic for TAT, Pessimistic for CIT) to provide clear baseline comparisons without stochastic variance.

3.2.2. Connectivity-Based Algorithms

These algorithms minimize communication cost by exploiting the graph structure. They differ primarily in which edge types they are permitted to traverse.

ACET (All-Connections-Equal Team Formation)

As the baseline, ACET operates on the unweighted projection of the HIN. It employs a greedy Breadth-First Search (BFS) to find the nearest candidate possessing the required skill, treating co-authorship and organizational links identically. It minimizes pure hop distance but ignores the semantic nature of the relationship.

CAT (Co-Authorship Team Formation)

CAT restricts the search space exclusively to WRITTEN edges. By enforcing direct co-authorship paths, it prioritizes candidates with a documented history of collaboration. While CAT emphasizes established social ties, it often fails in sparse networks where no direct co-authorship path exists between the required experts.

OAT (Organizational Affiliation Team Formation)

OAT prioritizes institutional proximity by traversing only WORKS_AT and INCLUDES edges. It exploits the hierarchical structure (Author–Department–Organization) to identify candidates. OAT distinguishes between colleagues in the same department (distance 2) and those in the same university but different departments (distance 4), preserving the granular structure of large institutions.

3.2.3. Attribute and Objective-Based Algorithms

These algorithms introduce higher-order logic to address specific team quality requirements, such as cohesion, recency, or impact.

PRT (Prioritized Relationship Team Formation)

PRT reflects the real-world preference hierarchy of researchers: direct co-authors first, then organizational neighbors, then global BFS as a fallback. This three-phase cascade—(1) co-authors ($d = 2$), (2) organizational peers ($d = 2$), (3) global BFS up to d_{max} —avoids exhaustive network search whenever a closer match exists, yielding the fastest runtimes among the heterogeneous-path algorithms while still guaranteeing completeness when a solution exists.

COT (Cohesion-Optimized Team Formation)

COT selects each new member by maximizing a weighted *Cohesion Score* rather than minimizing distance. The score aggregates shared papers and common organizational affiliations between the candidate and every existing team member, normalized by topological distance. This favors candidates with dense pre-existing ties to the emerging team, trading some topological tightness for stronger internal cohesion. Full pseudocode and weight parameters are given in Supplementary Material S3.

TAT (Time-Aware Team Formation)

TAT restricts recruitment to direct co-authors ($d = 2$) and selects the candidate whose most recent skill-matching paper minimizes *Years_Passed*, subject to a recency threshold $t_{\max}$. Organizational links are excluded because affiliation records lack start/end dates and cannot be verified for activeness. This design maximizes temporal reliability at the cost of coverage: in sparse or metadata-poor regions of the graph, the pool of qualifying candidates shrinks rapidly, which explains TAT's lower success rates relative to heterogeneous-path algorithms.

CIT (Citation-Impact Team Formation)

CIT follows the same three-phase cascade as PRT but selects, within each phase, the candidate with the highest cumulative citation count on skill-matching papers that satisfy the recency cutoff. This prioritizes local academic authority over global distance: the most cited co-author is preferred even if a higher-impact expert exists further away. The joint citation-and-recency filter makes CIT effective at identifying high-impact recent specialists, but also the most selective algorithm—success rates fall sharply when citation metadata or recent publications are sparse.

3.2.4. Complexity Analysis

An important aspect of this work is evaluating how these heuristic choices impact computational cost. Table 2 summarizes the worst-case time complexity for each algorithm. The algorithmic procedures (provided in Supplementary Material S3) indicate that while attribute-based methods (COT, CIT) incur additional costs for pairwise checks or metadata aggregation, prioritized methods (PRT, TAT) theoretically mitigate the need for deep BFS expansions, explaining the substantially lower effective runtime observed in practice.

Table 2. Comparison of algorithm computational complexity. $C_{\text{loc}} \approx b_w^2 + b_o^2$ represents prioritized local checks. All bounds omit the common seed filtering cost $O(n \cdot C_{\text{filter}})$.

Algorithm	Constraint	Worst-Case Complexity	Dominant Factor
ACET	Topology (All)	$O(S \cdot K \cdot b^{d_{\max}})$	Global BFS ($b^{d_{\max}}$)
CAT	Co-authorship	$O(S \cdot K \cdot b_w^{d_{\max}})$	Co-author b_w
OAT	Organization	$O(S \cdot K \cdot b_o^{d_{\max}})$	Org branching b_o
PRT	Priority Rule	$O(S \cdot K \cdot (C_{\text{loc}} + b^{d_{\max}}))$	Local check/BFS
COT	Cohesion	$O(S \cdot K \cdot b^{d_{\max}} \cdot (K + C_{\text{dist}}))$	Pairwise checks (K^2)
TAT	Temporal	$O(S \cdot K \cdot b_w \cdot P)$	Recency Filter (P)
CIT	Citation	$O(S \cdot K \cdot (C_{\text{loc}} + b^{d_{\max}}) \cdot (P + C_{\text{dist}}))$	Metadata agg. (P)

Each algorithm's cost consists of a one-time seed filtering step followed by K member-selection rounds, each expanding up to S seeds. The dominant factors differ by algorithm: ACET, OAT, and COT are governed by the global BFS branching factor $b^{d_{\max}}$; CAT and TAT benefit from restricting traversal to WRITTEN edges ($b_w \ll b$); PRT and CIT amortize the BFS cost through early-exit priority phases, paying only $O(C_{\text{loc}})$ when a close match exists. COT adds a per-candidate cohesion scoring step ($O(K)$ per candidate), and CIT adds citation aggregation over P papers. These theoretical differences align closely with the empirical runtime patterns reported in Section 4.

3.3. Experimental Design and Metrics

3.3.1. Overall Structure of the Graph Data

Before running the team-formation algorithms, we examined the global structure of the graph. The network, containing 625,776 nodes and 973,239 edges, is fragmented into

17,039 connected components. The largest connected component (giant component; GC) includes 552,076 nodes ($\approx 88\%$ of the network); the second-largest component contains only 83 nodes. Thus, almost all viable team formations are expected to occur inside the GC.

The graph comprises 350,404 authors and 166,198 papers connected by 533,352 WRITTEN relationships, yielding on average 1.52 papers per author and 3.21 authors per paper. From a co-authorship perspective, the network is extremely sparse: the co-authorship density—the fraction of unordered author pairs that have co-authored at least one paper (i.e., are connected by an author–paper–author two-step path)—is below 0.0001.

Conversely, the graph exhibits robust institutional coverage with 386,944 WORKS_AT relationships, yielding an average of 1.10 affiliations per author. However, this connectivity is non-uniform: approximately **15.3%** (53,458) of authors lack any affiliation links, due to limitations in the original metadata and its extraction rather than a true absence of affiliations. In addition, **13.1%** (45,981) of authors are affiliated with more than one department or organization and thus act as critical inter-institutional bridges.

3.3.2. Implementation and Sampling Motivation

Early experimental runs showed that many random keyword sets failed to produce complete teams, mainly due to extremely rare keywords. To quantify this effect, for each keyword, we counted the number of associated papers; The full frequency table (`keyword_frequencies.csv`) is provided as Supplementary Material S4. The rank–frequency distribution is highly skewed (see Figure 2): roughly one million unique keywords occur only a few times, about 331,000 distinct terms appear in a single paper, and a small core of keywords appear in over 20,000 papers. Coverage of keyword mentions by the top- k keywords is 35.00% (top 1000), 60.74% (top 10,000), and 80.73% (top 100,000). Based on these cutoffs, we defined three experimental sampling intervals: `most_1` (ranks 1–1000), `most_2` (ranks 1001–10,000), and `all` (ranks 1–100,000).

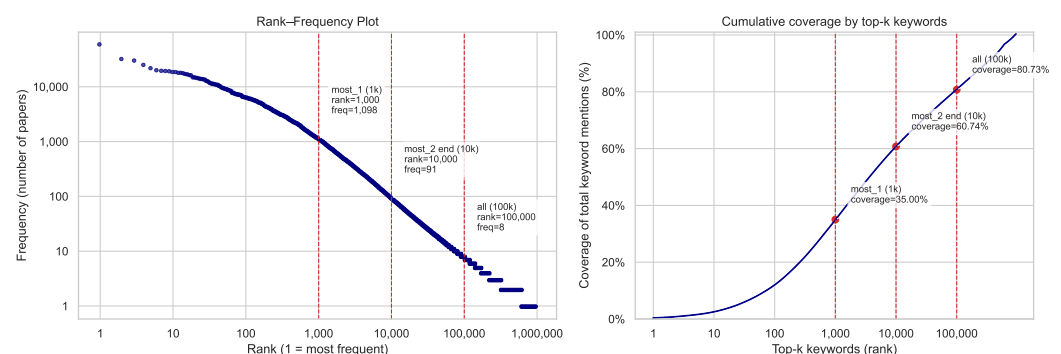


Figure 2. Keyword frequency diagnostics. (Left): Rank–frequency distribution (log–log) with vertical lines marking the sampling cutoffs. (Right): Cumulative coverage of mentions by top- k unique keywords, where red points correspond to the cutoffs.

3.3.3. Evaluation Metrics

Given the complexity of assembling teams from a heterogeneous information network, we prioritize operational feasibility over task-level evaluation. To directly address the research questions posed in Section 1, we evaluate **algorithm performance** using the following four quantitative metrics:

- **Success Rate (RQ1).** Because of limited network connectivity and the maximum allowed search depth ($d_{\max}$), algorithms cannot always construct a full team. We define the *success rate* as the ratio of *complete* teams to total attempts. A team is regarded as complete only if all target skills are covered; partial solutions are treated as failures. This metric serves as the primary indicator of **Feasibility and Parameter**

Sensitivity (RQ1), capturing both whether the algorithm can form a complete team and how performance degrades as skill rarity increases.

- **Computation Time (RQ1, RQ2)**. Given the NP-hard nature of the team-formation problem, traversing a large heterogeneous collaboration network is computationally expensive. This metric measures the wall-clock time required to form a single team. It establishes the **Algorithmic Profiles (RQ2)** of different heuristics and, in conjunction with RQ1, quantifies sensitivity to team size, helping to determine which methods remain practical for larger queries.
- **Normalized Team Distance (RQ2)**. We quantify communication cost using the sum of shortest-path distances via which new members attach to the existing team (analogous to a greedy Steiner-tree approximation). To allow fair comparison across varying team sizes (N), we report *normalized* team distance by dividing the total cost by the number of edges ($N - 1$). This metric quantifies the “communication cost” dimension of the **Algorithmic Profiles (RQ2)**, enabling us to compare how tightly connected the resulting teams are across different traversal strategies.
- **Team Similarity (RQ3)**. To interpret algorithmic differences beyond mere performance numbers, we measure the Jaccard similarity between teams produced by different algorithms for identical inputs. This metric serves as the quantitative basis for **Structural Diversity (RQ3)**, determining whether different heuristics converge on the same local optima or successfully exploit the network’s heterogeneity to yield structurally diverse solutions.

3.3.4. Experimental Design and Algorithm-Specific Parameters

To study how parameters affect team-assembly success, we created a Team Size Scalability Test Suite that covers all seven algorithms across the three keyword-frequency intervals and sweeps team size (number of keywords) from 2 to 8. Crucially, to ensure results are not biased by the peculiarities of any single keyword combination, we generated 10 distinct keyword sets for each size&keyword-frequency configuration, and for each keyword set we selected up to 10 distinct seed authors as alternative starting points. This design yields $7 \times 3 \times 7 = 147$ high-level configurations and 14,700 planned team-formation attempts (147 configurations $\times$ 10 keyword sets $\times$ 10 seed authors).

Algorithm-specific parameters were held at default values to avoid introducing additional degrees of freedom. In particular:

- In the heterogeneous graph, the theoretical minimum distance is $d = 2$ (a direct connection via a Paper or Department/Organization node). Thus, we constrained the searches to $d_{\max} = 5$, for all algorithms except TAT, as paths substantially longer than 2–3 intermediate hops likely represent too indirect a relationship for effective coordination. For TAT, we set $d_{\max} = 2$ because temporal information is mostly reliable for immediate co-authorship.
- Project keywords are drawn from the pool of author skills. Thus, each selected keyword appears for at least one author in the corpus. A run can still fail if no candidate satisfying the connectivity and distance constraints exists.
- For COT: `paper_weight` = 2.0, `org_weight` = 1.0, `cohesion_weight` = 0.7, `distance_weight` = 0.3. This configuration prioritizes historical ties over topological distance, with a specific emphasis on direct co-authorship over organizational affiliation. These weights were set by the authors as a default based on design judgment, not tuned; the assumption is that prior co-authorship signals collaboration more strongly than shared affiliation.
- Handling of missing `Years_Passed` values (see Section 3.2 for the full description of imputation strategies):

- TAT uses the **Optimistic** strategy: missing values are imputed as $t_{\text{imputed}} = 14$, which falls within the recency threshold ($14 \leq t_{\text{max}} = 15$), so temporally unverified connections are admitted to maximise the candidate pool.
- CIT uses the **Pessimistic** strategy: missing values are imputed as $t_{\text{imputed}} = 11$, which exceeds the recency threshold ($11 > t_{\text{max}} = 10$), so connections lacking verifiable date metadata are excluded to prioritise precision.

3.3.5. Computational Infrastructure and Parallelization

To ensure statistical robustness across 14,700 runs, the framework was parallelized and deployed on the a high-performance computing cluster (24 cores, 256 GB RAM per node). While preliminary benchmarking on a consumer workstation (Intel Core i9, 32 GB RAM) confirmed that individual algorithms are viable for desktop use, the HPC environment was necessary to handle the aggregate computational load and to prevent latencies during memory-intensive graph traversals. This parallel setup reduced the wall clock time for the full experiment to approximately 8 h and 13 min, which would have required weeks of serial execution.

4. Results and Discussion

Our goal is not to propose a single optimal solver or to explain social mechanisms in collaboration networks, but to provide a configurable *toolbox* of team-formation heuristics and to characterize their operating regimes: when they can form complete teams, how fast they run, and the communication cost of the resulting teams (subgraphs). The algorithms are intended as starting points that can later be tuned or specialized for particular user and project needs. Throughout the analysis, ACET—which applies unweighted BFS over all edge types without metadata filters—serves as the topological baseline against which the attribute-based heuristics are compared.

4.1. General Overview of Experimental Results

Although the experimental design targeted 14,700 runs, Table 3 summarizes outcomes from the realized dataset of 12,045 runs, as some keyword configurations—especially in the *most_2* and *a11* intervals—yielded no or fewer distinct seed authors, resulting in fewer runs than planned. The table reports the number of completed teams, empirical success rate, execution time, and normalized intra-team distance for each algorithm–strategy pair. For each algorithm–strategy pair, the mean and standard deviation are computed in two stages: we first average each metric within every team size (2–8), then report the mean and standard deviation across these per-size averages. The standard deviation is therefore undefined when successful teams span only a single team size—as in the CIT/*a11* configuration, whose 8 successful teams all occur at one keyword count. The full per-run metrics (file *runs.csv*) and team compositions (file *members.xlsx*) are provided as Supplementary Material S5.

Under the high-frequency *most_1* regime, ACET, COT, OAT, and PRT achieve near-perfect success (1.00), while CAT and especially TAT already suffer from their stricter constraints (0.63 and 0.15, respectively). As keywords become sparser, all algorithms deteriorate, but at very different rates. In *most_2*, ACET, COT, OAT, and PRT remain above 0.8, whereas CAT drops to 0.06, CIT to 0.24, and TAT practically to zero (0.02). In the rarest *a11* regime, heterogeneous traversal methods (ACET, COT, OAT, PRT) retain non-trivial success (≈ 0.27 – 0.28), while CAT and TAT fail entirely and CIT succeeds in only 2% of runs.

Algorithms that emphasize recency or citation impact (TAT, CIT) are substantially more selective. TAT succeeds only when there exists a dense recent co-authorship neighborhood within two hops; such regions are common for frequent keywords but almost absent for rare

ones. CIT, by contrast, can exploit longer paths, yet its additional citation and recency filters mean that in sparse settings the intersection of “reachable” and “high-impact” candidates is often empty.

Table 3. Summary of team formation results across keyword frequencies and algorithms. The table aggregates performance across team sizes 2–8. ‘-’ denotes an undefined value (no teams formed, or a standard deviation that cannot be computed; see text); ‘0.00’ denotes a strictly zero standard deviation.

Algorithm	Strategy	Num Teams	Success Rate		Avg. Time Per Team		Norm. Team Dist.	
		Sum	Mean	Std	Mean	Std	Mean	Std
ACET	most_1	700	1.00	0.00	3.47	3.09	2.03	0.03
	most_2	563	0.84	0.13	3.84	3.34	2.21	0.12
	all	94	0.28	0.14	3.56	2.69	2.42	0.23
CAT	most_1	444	0.63	0.08	0.06	0.05	2.63	0.25
	most_2	34	0.06	0.10	0.12	0.12	3.03	0.17
	all	0	0.00	-	-	-	-	-
OAT	most_1	700	1.00	0.00	3.92	3.27	2.04	0.03
	most_2	558	0.83	0.13	30.56	39.03	2.21	0.13
	all	92	0.27	0.13	49.57	94.88	2.39	0.23
PRT	most_1	700	1.00	0.00	1.78	1.34	2.00	0.00
	most_2	558	0.86	0.13	33.69	46.69	2.01	0.01
	all	85	0.27	0.15	76.75	154.89	2.04	0.07
COT	most_1	700	1.00	0.00	3.94	3.31	2.03	0.03
	most_2	563	0.84	0.13	4.46	3.73	2.22	0.13
	all	94	0.28	0.14	4.07	3.05	2.42	0.25
TAT	most_1	103	0.15	0.12	0.04	0.03	2.00	0.00
	most_2	9	0.02	0.04	0.13	0.17	2.00	0.00
	all	0	0.00	-	-	-	-	-
CIT	most_1	631	0.90	0.07	13.68	5.17	2.43	0.37
	most_2	159	0.24	0.19	5.78	3.19	2.51	0.34
	all	8	0.02	0.05	2.39	-	2.81	-

TAT’s low feasibility comes mainly from its distance constraint ($d_{\max} = 2$) rather than from its temporal imputation, since recruitment is limited to direct co-authors. CIT is highly selective: although it shares PRT’s three-phase cascade and differs mainly in its citation-and-recency filter, its overall success rate (0.46) is well below PRT’s (0.80), and the gap is largest under most_2 (0.24 vs. 0.86).

One structural caveat concerns organization-dependent traversal. About 15.3% of authors lack any affiliation link, so they are invisible to organizational meta-paths. This disproportionately affects methods that rely on WORKS_AT edges. OAT is the most exposed, since it recruits through organizational edges alone and cannot reach such authors at all; its success rates are therefore conservative lower bounds. Algorithms that combine organizational and co-authorship traversal (ACET, COT, PRT) can still reach these authors via WRITTEN paths and are more robust.

Runtime behavior also varies strongly by algorithm but only weakly by sampling strategy once we condition on success. For example, PRT is consistently among the fastest methods where teams exist (mean ≈ 1.78 s under most_1), whereas OAT and PRT under all exhibit heavy tails (standard deviations larger than their means), reflecting occasional very expensive searches when the BFS is forced to explore much of the reachable component. Attribute-based methods such as CIT show the opposite pattern: when keyword sets are rare, the initial filters admit very few candidates, so per-run times can actually decrease despite lower success.

Communication-cost results (Figure 3) broadly match the summary statistics: across all strategies ACET, COT, OAT, and especially PRT form relatively tight teams with normalized distances around 2.0–2.4, i.e., close to the theoretical lower bound $d = 2$. PRT and TAT cluster almost exactly at this bound; TAT attains similarly low costs but only on the small subset of instances where it successfully forms teams. In contrast, CIT and CAT consistently produce looser teams (normalized distances 2.4–3.0), especially under sparser strategies, because they prioritize citation impact or co-authorship exclusivity over pure topological proximity.

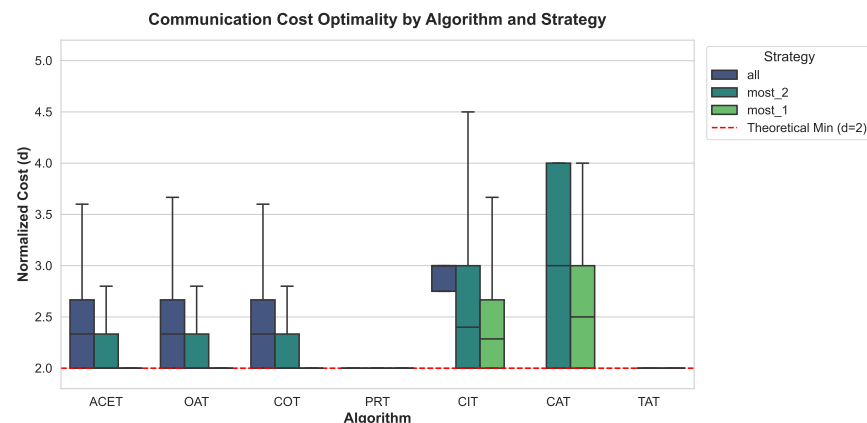


Figure 3. Normalized communication cost (team distance) by algorithm and sampling strategy.

These differences illustrate the trade-offs available in the toolbox: some heuristics favor low distance, others accept greater distance in exchange for feature-based constraints. Aggregating over algorithms, we observe a sharp decline in feasibility: success rates drop from 81.2% (3978/4900) under the abundant *most_1* regime to 52.6% (2444/4648) in *most_2*, and finally to just 14.9% (373/2497) under *all*, where topological sparsity renders heuristics largely ineffective (**RQ1: Feasibility and Parameter Sensitivity**). A Tukey HSD test on per-configuration success rates confirms that all three strategy pairs differ significantly from one another ($p < 0.001$ in all pairwise comparisons), with mean differences of 0.66 (*most_1* vs. *all*), 0.38 (*most_2* vs. *all*), and 0.29 (*most_1* vs. *most_2*). This establishes that the frequency regime is a statistically verified determinant of feasibility, not merely a descriptive trend.

Figure 4 shows that this scarcity effect is accompanied by large dispersion within strategies. Under *most_1*, per-configuration success rates are tightly clustered around 0.8 with almost no outliers; under *all*, the median drops to zero, yet the spread increases significantly, with outcomes ranging from complete failure to moderate success. These results suggest that, in this dataset, skill scarcity plays a more dominant role than algorithmic choice in constraining success. Algorithm choice determines *how* quickly performance degrades as we move from *most_1* to *all*, but not the inevitability of the decline itself. This divergence is clear in Table 3: while robust heuristics like ACET fade gradually ($1.00 \rightarrow 0.84$), selective methods like CIT experience a sharper drop ($0.90 \rightarrow 0.24$) under moderate scarcity.

4.1.1. Analysis of Computational Cost (Efficiency)

To address **RQ2 (Algorithmic Profiles)** and **RQ1 (Feasibility and Parameter Sensitivity)**, we restricted attention to successful runs ($N = 6795$) for computation-time analysis. A two-way ANOVA on $\log(\text{Time})$ confirms that Algorithm is the dominant determinant of cost ($F_{6,6776} = 42.45$, $p < 10^{-50}$, $\eta^2 = 0.04$), whereas the global main effect of Sampling Strategy is not statistically significant ($F_{2,6776} = 2.05$, $p \approx 0.13$, $\eta^2 = 0.001$). This lack of a global strategy effect reflects *divergent algorithmic responses* to sparsity: opposing trends

cancel out when averaged across algorithms, even though individual means differ substantially (e.g., OAT increases from ≈ 4 s in *most_1* to ≈ 50 s in *all*). PRT and OAT exploration methods slow down markedly as valid candidates become rare, whereas Attribute-based methods (such as CIT) may speed up in sparse regimes because few candidates survive the initial filters, reducing the number of expensive scoring operations. As a result, the global average runtime remains relatively stable even though individual algorithms react strongly to data density.

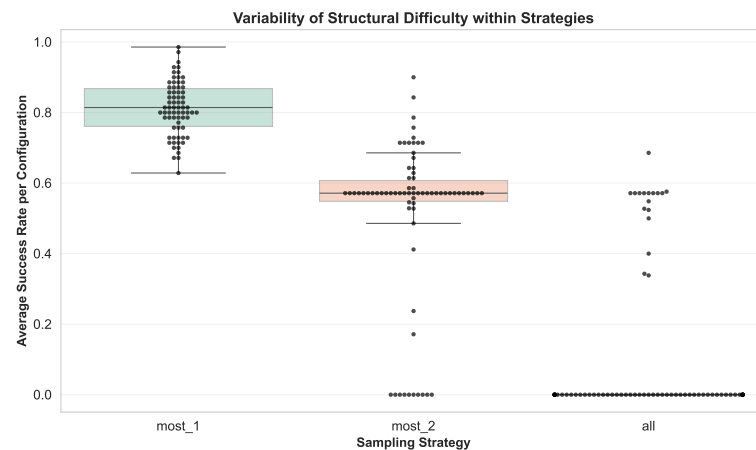


Figure 4. Variability of structural difficulty within strategies. Each point represents one of the 147 unique keyword configurations.

Figure 5 illustrates sensitivity to team size (**RQ1 (Feasibility and Parameter Sensitivity)**) under the *most_1* condition. TAT and PRT show almost flat execution-time curves as team size grows, reflecting bounded neighborhoods and prioritized early stopping. In contrast, ACET, COT, OAT, and especially CIT exhibit higher baselines and steeper growth, consistent with their deeper neighborhood exploration and more costly per-candidate scoring. These patterns highlight complementary options in the toolbox: some algorithms trade time for robustness, while others are faster but more fragile.

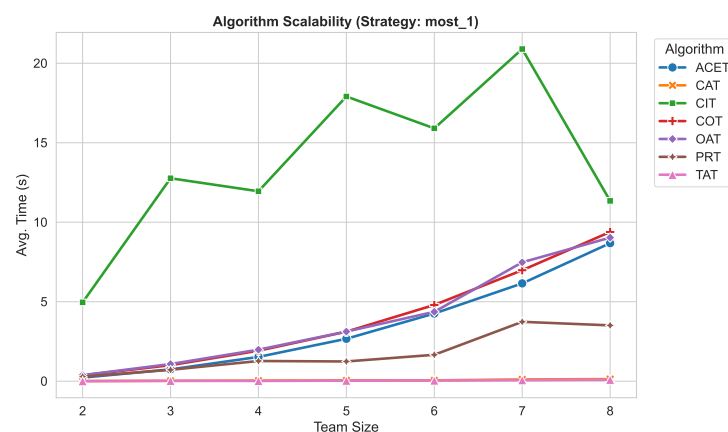


Figure 5. Scalability of algorithms under the *most_1* strategy (execution time vs. team size).

The HPC infrastructure provides some runtime advantages. In our preliminary desktop benchmarks, ‘stable’ heuristics like ACET and COT averaged ≈ 13.5 s per team (approx. $4 \times$ slower than HPC), which remains fully viable for interactive use. However, memory-intensive outliers such as PRT or TAT occasionally exceeded 300 s on the desktop due to relatively limited available RAM (32 GB). The HPC environment (256 GB) eliminated these bottlenecks, providing a cleaner measure of algorithmic complexity.

4.1.2. Analysis of Communication Cost (Objective Function)

Normalized team distance is our proxy for communication cost, quantifying the quality dimension of **RQ2 (Communication Cost)**. A two-way ANOVA on normalized distance shows that Algorithm is the dominant determinant of communication cost ($F_{6,6776} = 169.2$, $p < 10^{-100}$, $\eta^2 = 0.13$), with a smaller but significant main effect of Sampling Strategy ($F_{2,6776} = 15.3$, $p < 10^{-6}$, $\eta^2 = 0.004$), whereas the Algorithm $\times$ Strategy interaction is not statistically significant ($F_{12,6776} = 1.12$, $p \approx 0.29$, $\eta^2 = 0.002$). This indicates that the ability to approach the theoretical minimum is governed primarily by algorithm choice rather than by any specific algorithm–strategy combination.

To identify which algorithms differ significantly from one another, we conducted a Tukey HSD post-hoc test on normalized distance across the $N = 6795$ successful runs (all algorithms combined). The results identify three statistically distinct clusters ordered by increasing communication cost:

- **Cluster A—Minimum distance:** PRT $\approx$ TAT ($p = 1.0$, mean difference < 0.007), both approaching the theoretical lower bound $d = 2$. Their proximity-first design explicitly minimises hop distance at each selection step.
- **Cluster B—Moderate distance:** ACET $\approx$ COT $\approx$ OAT ($p = 1.0$, mean difference < 0.004), with means around 2.2–2.4. All three perform full neighbourhood exploration without hard distance prioritisation, leading to slightly looser but still compact teams. Cluster B is significantly higher-cost than Cluster A ($p < 0.001$, mean difference ≈ 0.17).
- **Cluster C—High distance:** CAT and CIT both differ significantly from all other algorithms ($p < 0.001$ in all pairwise comparisons), with means reaching 2.6–3.0. By prioritising co-authorship exclusivity or citation impact over shortest-path distance, they deliberately trade topological tightness for feature-based quality.

These findings confirm that algorithm choice has a structured, statistically verifiable effect on communication cost that goes beyond the omnibus ANOVA result.

Figure 3 confirms these three clusters across all strategies, with PRT and TAT sitting closest to $d = 2$, ACET/COT/OAT in the middle range, and CIT/CAT consistently producing the largest team distances.

Overall, the feasible solutions produced by all algorithms remain within the allowed slack ($d \leq 5$), but there is a clear strategic trade-off: PRT and TAT emphasize distance minimization, whereas CIT and CAT exploit the available slack to optimize non-topological objectives. In practice, this means: **choose PRT or ACET** when tight coordination and low communication overhead are the primary concern; **choose CIT or TAT** when the goal is to surface high-impact or recently active specialists, accepting a larger team footprint; and **choose COT** when prior collaborative density among members is the priority.

4.1.3. Impact of Structural Topology (Keyword Sensitivity)

To investigate **RQ1 (Feasibility and Parameter Sensitivity)** beyond simple frequency counts, we analyzed how the specific topology induced by a keyword set affects performance. We quantified this “structural difficulty” by analyzing variance across the 147 unique keyword configurations while controlling for Algorithm and Sampling Strategy.

A linear mixed model on execution time for successful runs, with Unique_Config_ID as a random intercept, yielded an intraclass correlation coefficient (ICC) close to 0%. The fixed effects (especially team size, $\beta = 0.29$, $p < 0.001$) explained almost all variance. This suggests that, conditional on success, the particular topology induced by a keyword set adds no measurable runtime overhead: if a team can be formed, execution time is driven primarily by team size and the chosen algorithm, not by which specific keywords define the query.

4.1.4. Analysis of Team Composition Diversity

A key practical question for a toolbox is whether different algorithms return similar or different teams—addressing **RQ3 (Structural Diversity)**. We measured pairwise team similarity using the Jaccard index (J) and cosine similarity, and we counted for how many problem instances (N) both algorithms in a pair successfully formed a team. Table 4 provides per-algorithm aggregates. On average, Jaccard scores range from 0.23 (COT) to 0.47 (PRT), indicating moderate overlap in the teams produced by different methods. Zero-overlap cases are absent across the board (0.00%); this is structural rather than algorithmic, as all methods initialize the teams with the same/similar seed authors for a given problem instance. Thus, the variation in Jaccard scores arises entirely from how differently each heuristic recruits the *remaining* members. The detailed pairwise matrix in Figure 6 reveals how the toolbox diversifies solutions. Some patterns emerge:

Table 4. Pairwise team similarity by algorithm (full dataset).

Algorithm	Avg. Jaccard	Avg. Cosine	Zero Overlap (%)	Avg. Common Members	Comparisons
COT	0.23	0.34	0.00	1.30	4592
CIT	0.25	0.37	0.00	1.24	409
OAT	0.34	0.46	0.00	2.01	4578
CAT	0.36	0.48	0.00	1.99	1955
TAT	0.41	0.52	0.00	1.55	554
ACET	0.41	0.52	0.00	2.35	4592
PRT	0.47	0.57	0.00	2.54	4388

1. **ACET and PRT** have the highest pairwise similarity, with average Jaccard $J = 0.81$ and 1237 shared successful runs. This confirms PRT’s design intent: its fallback phase effectively defaults to ACET’s topology-driven logic when prioritized connections are unavailable. In most feasible instances, they converge on nearly identical teams.
2. **CIT and CAT** illustrate a sharp divergence in strategy. While they share a number of successful runs (35), their solution similarity is remarkably low ($J = 0.21$). This indicates that even when a problem allows both algorithms to succeed, they construct fundamentally different teams: CAT recruits direct collaborators (local cohesion), while CIT prioritizes high-impact experts (including organizational connections), resulting in distinct team compositions for the same tasks.
3. **CIT and TAT** share only 4 problem instances. This negligible overlap confirms that they solve disjoint subsets of the problem space: TAT requires dense, recent co-authorship (cliques), while CIT searches for metric-driven stars who may be topologically distant. They essentially operate in mutually exclusive structural regimes.
4. **COT** exhibits low similarity to other global search methods ($J \approx 0.18$ – 0.22 vs. ACET, CIT, and PRT), confirming that maximizing subgraph density selects experts fundamentally different from those chosen by distance-minimizing (ACET), impact-maximizing (CIT), or priority-based (PRT) heuristics. Interestingly, COT finds its closest structural analogue in **TAT** ($J = 0.38$); this suggests that TAT’s temporal filter (restricting search to recent co-authors) acts as a proxy for cohesion, implicitly selecting the same dense, clique-like structures that COT explicitly maximizes.

Overall, these findings support a “toolbox” interpretation: algorithms such as ACET and PRT mainly trade off efficiency versus exhaustive search while returning structurally similar teams, whereas methods like CIT, TAT, and COT explore different regions of the solution space and therefore provide genuinely diverse alternatives when feasible. None of the heuristics is intended to be universally optimal; rather, each provides a different balance of feasibility (success rate), runtime, and structural properties that practitioners can match to their specific constraints and objectives. Concretely, the low pairwise overlap

between COT and CIT ($J \approx 0.18$) means that running both algorithms on the same query and presenting the two resulting teams to a decision-maker nearly doubles the effective solution space at negligible additional cost—an argument for deploying the full toolbox rather than committing to a single algorithm.

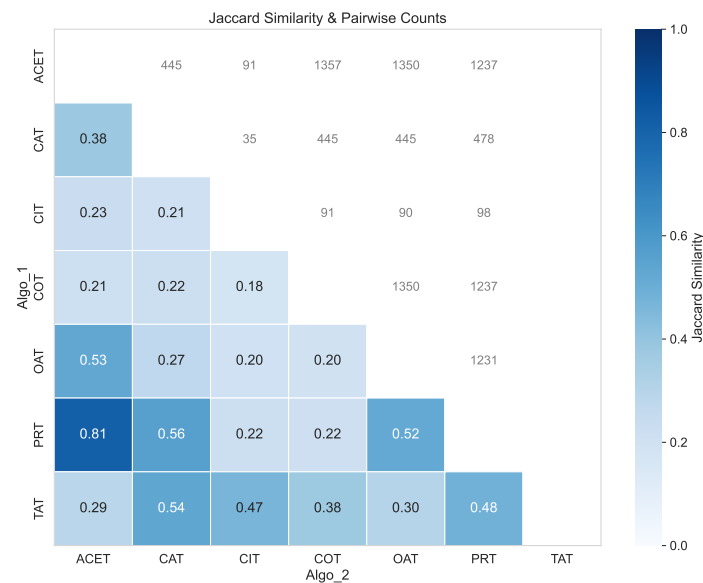


Figure 6. Pairwise comparison of algorithm outputs: Jaccard similarity of team members (lower triangle) and count of shared successful runs (upper triangle).

4.2. Comparison with Prior Work

Wang et al. [73] proposed a unified benchmark to empirically evaluate and compare 10 TF algorithms on a common platform using real-world datasets. They concluded that no single algorithm is perfect for all scenarios, considering perspectives such as distance and computational efficiency. Their benchmark results on the DBLP network (7159 expert nodes, 15,110 co-authorship edges) showed an average added cost of approximately 1.0–2.5 co-authorship steps per skill. As shown in Table 3 and Figure 3, most of our algorithms achieve average normalized team distances between 2.0 and 2.5. In our HIN model, a normalized distance of 2.0 corresponds to an Author–Paper–Author connection, which is topologically equivalent to a single “co-authorship step” in homogeneous networks. Thus, our algorithms effectively map to the theoretical lower bound of ≈ 1.0 co-authorship step, thereby aligning well with standard benchmarks while extending the capability to incorporate complex node attributes. Notably, we observed the highest communication costs for CAT. Since CAT restricts connections strictly to direct co-authorship (simulating a homogeneous topology), this finding implies that exploiting heterogeneous paths enables the formation of tighter teams than is possible under strict co-authorship constraints.

Regarding scalability, ref. [73] showed that as the number of skills increases, runtime scales either gently or exponentially depending on the algorithm. In their worst case, runtime reached 12 s for 20 skills, while the best cases remained at millisecond levels. Although differences in datasets and setups prevent a direct comparison, our results in Table 3 and Figure 5 similarly show that algorithms dependent solely on co-authorship relations (TAT and CAT) maintain near-zero average runtimes, while those traversing heterogeneous relations incur higher computational costs.

Yang et al. [61,62] present numerical results for runtimes and communication costs, claiming improvements over prior baselines. Although both studies utilize heterogeneous information networks, direct comparisons are not feasible due to differences in datasets, node types (e.g., topic, venue), and edge definitions. Furthermore, recent state-of-the-art

(SOTA) methods [5,6,8,66,67] predominantly employ information retrieval and similarity metrics (e.g., NDCG, recall) rather than optimizing the graph-based communication cost functions typical of our work. Consequently, direct distance-based benchmarking is not applicable.

While techniques like Graph Neural Networks (GNNs) and Transformer-based embeddings excel at latent link prediction, they differ fundamentally from the constructive heuristics proposed here. Learning-based approaches offer high predictive accuracy on soft constraints but function as “black boxes” and often struggle to enforce the strict logical constraints required in administrative settings (e.g., “Must be from Dept A”). In contrast, our HIN-based heuristics provide deterministic guarantees, full interpretability (the “why”), and zero training overhead, making them more suitable for interactive systems where user-defined constraints are non-negotiable. Table 5 highlights the paradigmatic differences between our approach and SOTA learning models. In Supplementary Material S1, we provide additional comparison tables summarizing the key features (dataset, node/edge counts, network types, methodologies) of selected studies.

Table 5. Qualitative comparison between the proposed HIN heuristics and current SOTA embedding-based methods.

Feature	Proposed Heuristics	SOTA Learning Models
Methodology	Constructive traversal (BFS/DFS)	Latent embeddings & GNNs
Output	Explicit teams (deterministic)	Probabilistic scores/ranked lists
Constraints	Hard (logical guarantees)	Soft (loss functions)
Interpretability	Transparent (explicit paths)	Opaque (latent features)
Cost	No training (query-time only)	High (pre-training required)
Goal	Attainment (strict compliance)	Prediction (prioritize likely links)

Our experimental design also offers a refinement to existing benchmarks. While prior studies often report average success rates for the entire network, our analysis of skill frequency (Section 4.1.3) demonstrates that team formation becomes significantly more challenging when skills are rare. This analysis exposes structural bottlenecks where topological sparsity, rather than algorithmic choice, acts as the primary constraint on success. To our knowledge, this is the first study to treat skill frequency as a primary variable, suggesting that future benchmarks must explicitly account for distributional sparsity to avoid inflating success metrics.

Finally, we distinguish our work by integrating multi-level organizational affiliation relations into the formation process. Although Lee et al. [63] incorporate institutional context, their approach relies on latent embeddings within a limited scope of two institutions. In contrast, our approach utilizes explicit links across a broad organizational network (over 50,000 distinct departments) to satisfy hard logical constraints, such as requiring membership in a target department. This makes our work the first to combine multi-level affiliation relations with constructive heuristics for explicit, interpretable team assembly.

4.3. Limitations

This study is subject to certain limitations regarding scope and methodology:

- **Single-domain evaluation:** All experiments were conducted on a subset of the AMiner-DBLP v14 academic collaboration graph, filtered using keywords relevant to Industrial Engineering. Collaboration patterns may differ in different domains.
- **Heuristic nature:** Our algorithms are constructive heuristics designed for runtime efficiency (query-time execution). Unlike exhaustive search or Integer Linear Programming (ILP), they do not guarantee global optimality. Consequently, potentially “tighter” teams might be missed by our greedy strategies.

- **Topological dependence:** As evidenced by our sparsity analysis, a method's success is heavily constrained by the underlying network density. We observed a significant structural bottleneck where rare keyword combinations often correspond to disconnected or sparse subgraphs, rendering many team formation requests unsolvable by topology-driven methods regardless of the algorithm used. These results suggest that performance is highly contingent on network density, and outcomes may differ in more densely connected or alternative structures.
- **Data quality and affiliation coverage:** Since the source data includes affiliations in an unformatted string format, we performed extensive cleaning to extract over 50,000 organization and 50,000 department entities. We were unable to manually verify all entities, so some inconsistencies may remain in the affiliation graph. In addition, about 15.3% of authors lack any affiliation link in the source metadata, which disproportionately constrains organization-dependent methods such as OAT; we discuss the implications for success-rate interpretation in Section 4.
- **Imputation strategy not isolated:** For the time-aware algorithms (TAT and CIT), the missing-date imputation strategies were fixed as parameters (Optimistic for TAT, Pessimistic for CIT) to give deterministic baselines. We did not measure their individual effect on success. The CIT-PRT comparison reported above is only an indirect proxy, since it also reflects CIT's citation filter. Measuring the temporal effect on its own would need a separate study.
- **Fixed scoring weights:** COT's composite weights were chosen as defaults, not derived from a systematic search. Their effect was not studied here; future work could tune them through a parameter sweep.
- **Structural metrics only:** We evaluate feasibility (success rate), runtime, and graph-based communication cost, but we *do not* assess whether the suggested teams are substantively "good" in terms of productivity, personality compatibility, etc.

The results describe the behavior of a flexible toolbox of heuristics, not universal prescriptions for optimal team composition.

5. Conclusions and Future Work

This work investigates expert team formation on a large, heterogeneous information network through a systematic evaluation of constructive graph search strategies. Rather than proposing a single optimal solver or a normative notion of an ideal team, we focus on how distinct algorithmic choices behave under varying conditions of skill frequency and network sparsity.

Our main contributions are:

- **Constructive HIN framework.** We build a heterogeneous information network that integrates authors, papers, departments, and organizations into a unified graph database. Unlike black-box embedding models, this framework relies on explicit topological paths, ensuring that every formed team is logically valid, structurally connected, and semantically interpretable.
- **A toolbox of seven heuristics.** We implement a diverse suite of constructive algorithms, each designed to exploit a different structural feature of the network. These include topological proximity (ACET, OAT), strict co-authorship (CAT), prioritized multi-phase search (PRT), structural cohesion (COT), temporal recency (TAT), and citation impact (CIT).
- **Empirical characterization at scale.** On a large academic collaboration graph, we systematically evaluate feasibility (success rate), efficiency (runtime), communication cost (normalized distance), and solution diversity (Jaccard similarity). We show that ACET, COT, OAT, and PRT can reliably form complete teams in realistic settings,

whereas CIT, CAT, and TAT are more selective and succeed mainly in dense local regions. This divergence stems from the broader heterogeneous connectivity captured by the former group. Notably, the superior performance of OAT (which leverages organizational paths) compared to CAT (which relies on co-authorship) aligns with prior findings that shared organizational affiliations create a significantly denser connectivity than co-authorship links alone.

- **Trade-offs and diversity.** The results reveal clear performance profiles:
 - **PRT and TAT** achieve the lowest communication costs (tightest teams) when they succeed.
 - **ACET, COT, and OAT** form slightly looser but still compact teams with high feasibility.
 - **CIT and CAT** incur higher topological costs (longer paths) to satisfy metadata constraints like citation impact or strict co-authorship.
 - **COT** explores cohesive subgraphs, often recruiting experts distinct from those selected by purely distance-based strategies.
 - Pairwise similarity analysis confirms that while robust solvers (e.g., ACET, PRT) often converge on similar solutions, specialized heuristics (e.g., CIT, TAT) explore largely distinct regions of the expertise space, identifying fundamentally different expert configurations.

Collectively, these distinct performance profiles demonstrate the practical utility of the toolbox approach. Rather than relying on a ‘one-size-fits-all’ black box, users can select the heuristic that best matches their constraints: **PRT** serves as the recommended default for forming tight, cohesive teams; **ACET** provides a fallback for maximum reachability in sparse networks; and specialized solvers (**CIT, TAT**) offer targeted solutions when specific traits like temporal activity or citation impact are critical.

Building on this foundation, we see several concrete directions for future research:

1. **User-centered evaluation and feedback:** We plan to conduct user studies (e.g., with researchers or managers) to collect feedback on the perceived quality and plausibility of suggested teams, using this data to refine or re-weight the heuristics.
2. **Richer network schema:** The current HIN focuses on authors, papers, departments, and organizations. A natural extension is to integrate additional entities such as project grants and patents. Additionally, we aim to incorporate venue–paper links to infer qualitative information about publication outlets.
3. **Alternative formation objectives:** This study focuses on forming experienced and well-connected (expert) teams. Future work could explore alternative objectives, such as forming referee committees (conflict avoidance) or teams of newcomers (mentorship and growth).
4. **Data quality and scale:** We plan to enhance the network’s resolution by refining the entity extraction for departments and organizations and integrating temporal information into affiliation edges. Concurrently, we aim to deploy the framework on the full AMiner corpus and diverse disciplinary datasets. This will allow us to test scalability and verify whether the observed structural patterns (such as the sharp decline in success rates within the long tail) generalize across different scientific fields.
5. **Hybrid architectures:** While this study contrasts constructive heuristics with learning-based models, a promising frontier lies in their integration. Future work could explore hybrid frameworks that combine the semantic richness of latent embeddings with the structural guarantees of explicit graph traversal. This would aim to combine the predictive power of deep learning with the logical interpretability and strict constraint satisfaction required for administrative decision support.

In summary, the present work offers a transparent, extensible toolbox for graph-based team formation and a detailed empirical characterization of its behavior. Future extensions that incorporate user feedback, richer network structure, improved temporal modeling, and larger datasets will move this toolbox closer to a practical decision-support system for real-world team assembly.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/informatics13070107/s1>.

Author Contributions: Conceptualization, N.Ö. and H.G.; methodology, N.Ö.; software, N.Ö.; validation, N.Ö. and H.G.; formal analysis, N.Ö.; investigation, N.Ö.; resources, N.Ö.; data curation, N.Ö.; writing—original draft preparation, N.Ö.; writing—review and editing, N.Ö. and H.G.; visualization, N.Ö.; supervision, H.G. 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: The AMiner DBLP–Citation–Network V14 dataset used in this study is publicly available at <https://open.aminer.cn/open/article?id=655db2202ab17a072284bc0c> (accessed on 12 November 2025). The derived network, experimental results (`runs.csv`, `members.xlsx`), and analysis scripts are available from the corresponding author upon reasonable request.

Acknowledgments: The numerical experiments reported in this study were performed on TRUBA (Turkish National e-Science e-Infrastructure), provided by TÜBİTAK ULAKBİM, High Performance and Grid Computing Center.

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

Abbreviations

The following abbreviations are used in this manuscript:

HIN	Heterogeneous Information Network
GNN	Graph Neural Network
BFS	Breadth-First Search
ACET	All-Connections-Equal Team Formation
CAT	Co-Authorship Team Formation
OAT	Organizational Affiliation Team Formation
PRT	Prioritized Relationship Team Formation
COT	Cohesion-Optimized Team Formation
TAT	Time-Aware Team Formation
CIT	Citation-Impact Team Formation
DBLP	Digital Bibliography & Library Project
ANOVA	Analysis of Variance
HSD	Honestly Significant Difference (Tukey)
ICC	Intraclass Correlation Coefficient
ILP	Integer Linear Programming
MST	Minimum Spanning Tree
NDCG	Normalized Discounted Cumulative Gain
SOTA	State of the Art

References

- Farland, M.Z.; Feng, X.; Behar-Horenstein, L.S.; Beck, D.E. Impact of Team Formation Method on Student Team Performance Across Multiple Courses Incorporating Team-based Learning. *Am. J. Pharm. Educ.* **2019**, *83*, 7030. [\[CrossRef\]](#) [\[PubMed\]](#)
- Mirzaei, M.; Sander, J.; Stroulia, E. Multi-Aspect Review-Team Assignment Using Latent Research Areas. *Inf. Process. Manag.* **2019**, *56*, 858–878. [\[CrossRef\]](#)
- Tavana, M.; Azizi, F.; Azizi, F.; Behzadian, M. A Fuzzy Inference System with Application to Player Selection and Team Formation in Multi-Player Sports. *Sport Manag. Rev.* **2013**, *16*, 97–110. [\[CrossRef\]](#)
- Stavrou, G.; Adamidis, P.A.; Papathanasiou, J.B.; Tarabanis, K.A. Team Formation: A Systematic Literature Review. *Int. J. Bus. Sci. Appl. Manag.* **2023**, *18*, 17–34. [\[CrossRef\]](#)
- Hu, C.; Zhou, Q.; Tong, H. Genius: Subteam Replacement with Clustering-based Graph Neural Networks. In Proceedings of the 2024 SIAM International Conference on Data Mining (SDM), Houston, TX, USA, 18–20 April 2024; pp. 10–18. [\[CrossRef\]](#)
- Dara, M.; Rad, R.H.; Zarrinkalam, F.; Bagheri, E. Retrieval-Augmented Neural Team Formation. In *Advances in Information Retrieval*; Hauff, C., Macdonald, C., Jannach, D., Kazai, G., Nardini, F.M., Pinelli, F., Silvestri, F., Tonellotto, N., Eds.; Springer Nature: Cham, Switzerland, 2025; Volume 15574, pp. 362–371. [\[CrossRef\]](#)
- Hamidi Rad, R.; Bagheri, E.; Kargar, M.; Srivastava, D.; Szlichta, J. Retrieving Skill-Based Teams from Collaboration Networks. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 11–15 July 2021; Sigir '21; pp. 2015–2019. [\[CrossRef\]](#)
- Hamidi Rad, R.; Nguyen, H.; Al-Obeidat, F.; Bagheri, E.; Kargar, M.; Srivastava, D.; Szlichta, J.; Zarrinkalam, F. Learning Heterogeneous Subgraph Representations for Team Discovery. *Inf. Retr. J.* **2023**, *26*, 8. [\[CrossRef\]](#)
- Hall, K.L.; Vogel, A.L.; Huang, G.C.; Serrano, K.J.; Rice, E.L.; Tsakraklides, S.P.; Fiore, S.M. The Science of Team Science: A Review of the Empirical Evidence and Research Gaps on Collaboration in Science. *Am. Psychol.* **2018**, *73*, 532–548. [\[CrossRef\]](#) [\[PubMed\]](#)
- Smith, T.B.; Vacca, R.; Krenz, T.; McCarty, C. Great Minds Think Alike, or Do They Often Differ? Research Topic Overlap and the Formation of Scientific Teams. *J. Inf.* **2021**, *15*, 101104. [\[CrossRef\]](#) [\[PubMed\]](#)
- Özdemir, N.; Gökçen, H. Time Sensitive Scientific Team Recommendation on Heterogeneous Information Network. *J. Inf. Syst. Manag. Res.* **2026**, *8*, 38–52. [\[CrossRef\]](#)
- Jiang, D.; Dahl, B.; Du, X. A Systematic Review of Engineering Students in Intercultural Teamwork: Characteristics, Challenges, and Coping Strategies. *Educ. Sci.* **2023**, *13*, 540. [\[CrossRef\]](#)
- Vasiljević, J.; Lavbič, D. A Data-Driven Approach to Team Formation in Software Engineering Based on Personality Traits. *Electronics* **2023**, *13*, 178. [\[CrossRef\]](#)
- Caldeira, N.; Lopes, R.J.; Fernandes, D.; Araujo, D. From Optical Tracking to Tactical Performance via Voronoi Diagrams: Team Formation and Players' Roles Constrain Interpersonal Linkages in High-Level Football. *Sensors* **2022**, *23*, 273. [\[CrossRef\]](#) [\[PubMed\]](#)
- Taramasco, C.; Cointet, J.P.; Roth, C. Academic Team Formation as Evolving Hypergraphs. *Scientometrics* **2010**, *85*, 721–740. [\[CrossRef\]](#)
- Curiac, C.D.; Micea, M.V.; Ploscă, T.R.; Curiac, D.I.; Doboli, A. Optimized Interdisciplinary Research Team Formation Using a Genetic Algorithm and Publication Metadata Records. *AI* **2025**, *6*, 171. [\[CrossRef\]](#)
- Maric, A.; Cajner, H. Effects of Personality Type Tools and Problem-Solving Methods on Engineering Company Project Success. *Sustainability* **2025**, *17*, 11185. [\[CrossRef\]](#)
- Alahmadi, M. Multi-Objective Combinatorial Optimization for Dynamic Inspection Scheduling and Skill-Based Team Formation in Distributed Solar Energy Infrastructure. *Systems* **2025**, *13*, 822. [\[CrossRef\]](#)
- Juárez, J.; Santos, C.P.; Brizuela, C.A. A Comprehensive Review and a Taxonomy Proposal of Team Formation Problems. *ACM Comput. Surv.* **2022**, *54*, 1–33. [\[CrossRef\]](#)
- Saeedi, M.; Hosseini, H.; Wong, C.; Fani, H. A Survey of Subgraph Optimization for Expert Team Formation. *ACM Comput. Surv.* **2025**, *57*, 1–40. [\[CrossRef\]](#)
- Sanchez-Anguix, V.; Alberola, J.M.; Del Val, E.; Palomares, A.; Teruel, M.D. Comparing Computational Algorithms for Team Formation in the Classroom: A Classroom Experience. *Appl. Intell.* **2023**, *53*, 23883–23904. [\[CrossRef\]](#)
- Wang, Q.; Wang, Z.; Cui, Z.; Lu, Y. A Method for Forming New-Type Construction Project Management Teams Using CSCD-NSGA-II. *Buildings* **2026**, *16*, 816. [\[CrossRef\]](#)
- Zakarian, A.; Kusiak, A. Forming Teams: An Analytical Approach. *IIE Trans.* **1999**, *31*, 85–97. [\[CrossRef\]](#)
- Baykasoglu, A.; Dereli, T.; Das, S. Project team selection using fuzzy optimization approach. *Cybern. Syst.* **2007**, *38*, 155–185. [\[CrossRef\]](#)
- Gaston, M.E.; Simmons, R.; desJardins, M. Adapting Network Structure for Efficient Team Formation. In *AAAI Technical Report (WS-04-02)*; Technical Report; AAAI: Menlo Park, CA, USA, 2004; pp. 1–8.
- Fitzpatrick, E.L.; Askin, R.G. Forming Effective Worker Teams with Multi-Functional Skill Requirements. *Comput. Ind. Eng.* **2005**, *48*, 593–608. [\[CrossRef\]](#)

27. Backstrom, L.; Huttenlocher, D.; Kleinberg, J.; Lan, X. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006; pp. 44–54. [\[CrossRef\]](#)
28. Cheatham, M.; Cleereman, K. Application of Social Network Analysis to Collaborative Team Formation. In *Proceedings of the International Symposium on Collaborative Technologies and Systems (CTS'06)*, Las Vegas, NV, USA; IEEE: New York, NY, USA, 2006; pp. 306–311. [\[CrossRef\]](#)
29. Wi, H.; Oh, S.; Mun, J.; Jung, M. A Team Formation Model Based on Knowledge and Collaboration. *Expert Syst. Appl.* **2009**, *36*, 9121–9134. [\[CrossRef\]](#)
30. Lappas, T.; Liu, K.; Terzi, E. Finding a Team of Experts in Social Networks. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 28 June–1 July 2009; pp. 467–476. [\[CrossRef\]](#)
31. Li, C.T.; Shan, M.K. Team Formation for Generalized Tasks in Expertise Social Networks. In Proceedings of the 2010 IEEE Second International Conference on Social Computing, Minneapolis, MN, USA, 20–22 August 2010; IEEE: New York, NY, USA, 2010; pp. 9–16. [\[CrossRef\]](#)
32. Dorn, C.; Dustdar, S. Composing Near-Optimal Expert Teams: A Trade-Off between Skills and Connectivity. In *On the Move to Meaningful Internet Systems: OTM 2010*; Meersman, R., Dillon, T., Herrero, P., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6426, pp. 472–489. [\[CrossRef\]](#)
33. Kargar, M.; An, A. Discovering Top-k Teams of Experts with/without a Leader in Social Networks. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management, Glasgow, UK, 24–28 October 2011; pp. 985–994. [\[CrossRef\]](#)
34. Anagnostopoulos, A.; Becchetti, L.; Castillo, C.; Gionis, A.; Leonardi, S. Power in Unity: Forming Teams in Large-Scale Community Systems. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, ON, Canada, 26–30 October 2010; pp. 599–608. [\[CrossRef\]](#)
35. Anagnostopoulos, A.; Becchetti, L.; Castillo, C.; Gionis, A.; Leonardi, S. Online Team Formation in Social Networks. In Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 16–20 April 2012; pp. 839–848. [\[CrossRef\]](#)
36. Majumder, A.; Datta, S.; Naidu, K. Capacitated Team Formation Problem on Social Networks. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 1005–1013. [\[CrossRef\]](#)
37. Gajewar, A.; Das Sarma, A. Multi-Skill Collaborative Teams Based on Densest Subgraphs. In *Proceedings of the 2012 SIAM International Conference on Data Mining*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2012; pp. 165–176. [\[CrossRef\]](#)
38. Rangapuram, S.S.; Bühler, T.; Hein, M. Towards Realistic Team Formation in Social Networks Based on Densest Subgraphs. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13–17 May 2013; pp. 1077–1088. [\[CrossRef\]](#)
39. Awal, G.K.; Bharadwaj, K.K. Team Formation in Social Networks Based on Collective Intelligence—An Evolutionary Approach. *Appl. Intell.* **2014**, *41*, 627–648. [\[CrossRef\]](#)
40. Gutiérrez, J.H.; Astudillo, C.A.; Ballesteros-Pérez, P.; Mora-Melià, D.; Candia-Véjar, A. The Multiple Team Formation Problem Using Sociometry. *Comput. Oper. Res.* **2016**, *75*, 150–162. [\[CrossRef\]](#)
41. Selvarajah, K.; Zadeh, P.M.; Kobti, Z.; Palanichamy, Y.; Kargar, M. A Unified Framework for Effective Team Formation in Social Networks. *Expert Syst. Appl.* **2021**, *177*, 114886. [\[CrossRef\]](#)
42. Daş, G.S.; Altınkaynak, B.; Göçken, T.; Türker, A.K. A Set Partitioning Based Goal Programming Model for the Team Formation Problem. *Int. Trans. Oper. Res.* **2022**, *29*, 301–322. [\[CrossRef\]](#)
43. Mahajan, Y.; Cho, J.h.; Chen, I. Privacy-Preserving and Diversity-Aware Trust-Based Team Formation in Online Social Networks. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–32. [\[CrossRef\]](#)
44. Fathian, M.; Saei-Shahi, M.; Makui, A. A New Optimization Model for Reliable Team Formation Problem Considering Experts' Collaboration Network. *IEEE Trans. Eng. Manag.* **2017**, *64*, 586–593. [\[CrossRef\]](#)
45. Cao, N.; Lin, Y.R.; Li, L.; Tong, H. G-Miner: Interactive Visual Group Mining on Multivariate Graphs. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Republic of Korea, 18–23 April 2015; pp. 279–288. [\[CrossRef\]](#)
46. Li, L.; Tong, H.; Cao, N.; Ehrlich, K.; Lin, Y.R.; Buchler, N. TEAMOPT: Interactive Team Optimization in Big Networks. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, Indianapolis, IN, USA, 24–28 October 2016; pp. 2485–2487. [\[CrossRef\]](#)
47. Zhou, Q.; Li, L.; Tong, H. Towards Real Time Team Optimization. In *Proceedings of the 2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA; IEEE: New York, NY, USA, 2019; pp. 1008–1017. [\[CrossRef\]](#)
48. Yin, X.; Qu, C.; Wang, Q.; Wu, F.; Liu, B.; Chen, F.; Chen, X.; Fang, D. Social Connection Aware Team Formation for Participatory Tasks. *IEEE Access Pract. Innov. Open Solut.* **2018**, *6*, 20309–20319. [\[CrossRef\]](#)

49. Hamidi Rad, R.; Fani, H.; Kargar, M.; Szlichta, J.; Bagheri, E. Learning to Form Skill-based Teams of Experts. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*; ACM: New York, NY, USA, 2021; pp. 2049–2052. [\[CrossRef\]](#)
50. Gómez-Zarà, D.; Das, A.; Pawlow, B.; Contractor, N.S. In Search of Diverse and Connected Teams: A Computational Approach to Assemble Diverse Teams Based on Members' Social Networks. *PLoS ONE* **2022**, *17*, e0276061. [\[CrossRef\]](#) [\[PubMed\]](#)
51. Juárez, J.; Brizuela, C.A.; Terashima-Marín, H.; Coello Coello, C.A. Recombination Operators for the Multi-Objective Team Formation Problem in Social Networks. In *Proceedings of the 2024 IEEE Congress on Evolutionary Computation (CEC), Yokohama, Japan*; IEEE: New York, NY, USA, 2024; pp. 1–8. [\[CrossRef\]](#)
52. Sharaf, M.A.; Alghamdi, T.G. Balanced Team Formation Using Hybrid Graph Convolution Networks and MILP. *Appl. Sci.* **2025**, *15*, 2049. [\[CrossRef\]](#)
53. El-Ashmawi, W.H.; Ali, A.F.; Tawhid, M.A. An Improved Particle Swarm Optimization with a New Swap Operator for Team Formation Problem. *J. Ind. Eng. Int.* **2019**, *15*, 53–71. [\[CrossRef\]](#)
54. El-Ashmawi, W.H.; Ali, A.F.; Słowik, A. An Improved Jaya Algorithm with a Modified Swap Operator for Solving Team Formation Problem. *Soft Comput.* **2020**, *24*, 16627–16641. [\[CrossRef\]](#)
55. El-Ashmawi, W.H.; Sheta, A.F.; Alqadi, B.S.; Abdelminaam, D.S.; Mohammed Alsekait, D.M. A New Version of the Golden Eagle Optimizer Algorithm and Its Application for Solving a Trio-Objective Skillful Team Formation Problem in a Social Network. *J. Artif. Intell. Soft Comput. Res.* **2025**, *15*, 357–384. [\[CrossRef\]](#)
56. Rehman, M.Z.; Zamli, K.Z.; Almutairi, M.S.; Chiroma, H.; Aamir, M.; Kader, M.A.; Mohd Nawi, N. A Novel State Space Reduction Algorithm for Team Formation in Social Networks. *PLoS ONE* **2021**, *16*, e0259786. [\[CrossRef\]](#) [\[PubMed\]](#)
57. Sun, Y.; Han, J. Mining Heterogeneous Information Networks: A Structural Analysis Approach. *ACM SIGKDD Explor. Newsl.* **2013**, *14*, 20–28. [\[CrossRef\]](#)
58. Zhou, L.; Wang, J.; Song, Y.; Wang, L.; Chen, H. Community Search over Heterogeneous Information Networks: A Survey. *ACM Comput. Surv.* **2025**, *58*, 1–37. [\[CrossRef\]](#)
59. Cai, D.; Shao, Z.; He, X.; Yan, X.; Han, J. Community Mining from Multi-Relational Networks. In *Proceedings of the Knowledge Discovery in Databases: PKDD*; Jorge, A.M., Torgo, L., Brazdil, P., Camacho, R., Gama, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2005; pp. 445–452. [\[CrossRef\]](#)
60. Shafie, T. A Multigraph Approach to Social Network Analysis. *J. Soc. Struct.* **2015**, *16*, 1–21. [\[CrossRef\]](#)
61. Yang, J.; Li, M.; Wu, B.; Xu, C. Forming a Research Team of Experts in Expert-Skill Co-Occurrence Network of Research News. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*; Kumar, R., Caverlee, J., Tong, H., Eds.; IEEE: New York, NY, USA, 2016; Asonam '16; pp. 1143–1150. [\[CrossRef\]](#)
62. Yang, X.; Wang, P.; Fang, W. Team Formation with Relationship Strength Based on Meta Path in Heterogeneous Network. In *Proceedings of the 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*; Wenzheng, L., Babu, M., Eds.; IEEE: New York, NY, USA, 2018; IEEE Computer Society; Volume 2018, pp. 797–801. [\[CrossRef\]](#)
63. Lee, O.J.; Hong, S.; Kim, J.t. Interinstitutional Research Team Formation Based on Bibliographic Network Embedding. *Mob. Inf. Syst.* **2021**, *2021*, 6629520. [\[CrossRef\]](#)
64. Liu, X.; Wu, K.; Liu, B.; Qian, R. HNERec: Scientific Collaborator Recommendation Model Based on Heterogeneous Network Embedding. *Inf. Process. Manag.* **2023**, *60*, 103253. [\[CrossRef\]](#)
65. Gao, X.; Wu, S.; Xia, D.; Xiong, H. Topic-Sensitive Expert Finding Based Solely on Heterogeneous Academic Networks. *Expert Syst. Appl.* **2023**, *213*, 119241. [\[CrossRef\]](#)
66. Guan, K.; Huang, W.; Chen, T.; Chan, W.K.V. TEDRec: Transformer-based Scientific Collaborator Recommendation via Textual-Edge Dynamic Network Modeling. *Inf. Process. Manag.* **2025**, *62*, 104283. [\[CrossRef\]](#)
67. Wang, X.; Huang, W.; Zhao, B.; Li, S. Scientific Collaborator Recommendation via Hypergraph Embedding. *Inf. Process. Manag.* **2026**, *63*, 104423. [\[CrossRef\]](#)
68. Liu, J.; Zhang, Q.; Xu, S.; Huang, C.; Gong, X. Innovation Performance Improvement Path in Scientific Research Teams: A Perspective of Network Embeddedness. *Inf. Process. Manag.* **2026**, *63*, 104484. [\[CrossRef\]](#)
69. Campêlo, M.B.; Figueiredo, T.F. Integer Programming Approaches to the Multiple Team Formation Problem. *Comput. Oper. Res.* **2021**, *133*, 105354. [\[CrossRef\]](#)
70. Tang, J.; Zhang, J.; Yao, L.; Li, J.; Zhang, L.; Su, Z. ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, 24–27 August 2008*; pp. 990–998. [\[CrossRef\]](#)
71. AMiner Open Dataset. AMiner DBLP-Citation-Network V14. Version V14; 5,259,858 Papers, 36,630,661 Citations. Available online: <https://open.aminer.cn/open/article?id=655db2202ab17a072284bc0c> (accessed on 12 November 2025).

72. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
73. Wang, X.; Zhao, Z.; Ng, W. A Comparative Study of Team Formation in Social Networks. In *Proceedings of the Database Systems for Advanced Applications*; Renz, M., Shahabi, C., Zhou, X., Cheema, M.A., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 389–404. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.