Familiarity-based Collaborative Team Recognition in Academic Social Networks

Shuo Yu
School of Software
Dalian University of Technology
Dalian, China
y_shuo@outlook.com

Feng Xia
School of Engineering, IT and Physical Sciences
Federation University Australia
Ballarat, Australia
f.xia@ieee.org

Chen Zhang
School of Software
Dalian University of Technology
Dalian, China
chen.zhang07@outlook.com

Haoran Wei
School of Software
Dalian University of Technology
Dalian, China
willieying@outlook.com

Kathleen Keogh
School of Engineering, IT and Physical Sciences
Federation University Australia
Ballarat, Australia
k.keogh@federation.edu.au

Honglong Chen
College of Control and Science Engineering
China University of Petroleum
Qingdao, China
chenhl@upc.edu.cn

ABSTRACT

Collaborative teamwork is key to major scientific discoveries. However, the prevalence of collaboration among researchers makes team recognition increasingly challenging. Previous studies have demonstrated that people are more likely to collaborate with individuals they are familiar with. In this work, we employ the definition of familiarity and then propose MOTO (faMiliarity-based cOllaborative Team recOgnition algorithm) to recognize collaborative teams. MOTO calculates the shortest distance matrix within the global collaboration network and the local density of each node. Central team members are initially recognized based on local density. Then MOTO recognizes the remaining team members by using the familiarity metric and shortest distance matrix. Extensive experiments have been conducted upon a large-scale data set. The experimental results show that compared with baseline methods, MOTO can recognize the largest number of teams. The teams recognized by MOTO possess more cohesive team structures and lower team communication costs compared with other methods. MOTO utilizes familiarity in team recognition to identify cohesive academic teams. The recognized teams are in line with real-world collaborative teamwork patterns. Based on team recognition using MOTO, the research team structure and performance are further analyzed for given time periods. The number of teams that consist of members from different institutions increases gradually. Such teams are found to perform better in comparison with those whose members are from the same institution.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

KEYWORDS

Academic social networks, familiarity, team recognition, network motif, collaboration.

ACM Reference Format:
Shuo Yu, Feng Xia, Chen Zhang, Haoran Wei, Kathleen Keogh, and Honglong Chen. 2022. Familiarity-based Collaborative Team Recognition in Academic Social Networks. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Scientific discovery in the 21st century relies on global interactions and collaborations between colleagues. Individual research has increasingly been replaced by collaborative teamwork. Scientific collaboration has been regarded as one of the most effective ways to solve complicated scientific research problems [29, 32]. Information technology has greatly facilitated communication between scholars. A variety of tools and methods enable collaboration between scholars so that the team can be distributed across various locations. Understanding and being aware of the inner patterns of collaborative academic teamwork may improve team efficiency and the quality of scientific research. It is also possible to gain insights by studying the inner patterns of collaborative teams at both micro and meso-levels. The scale of collaborative teamwork as well as the quality of teamwork collaboration have increased gradually [10]. A branch of science entitled "Science of Scientific Team Science (SSTS)" [39] has been proposed to study collaborative teamwork, with the aims of enhancing scientific collaboration within teams and improving transdisciplinary research.

Studies of teamwork in management science and psychological science exist, and most of these focus on collaborative teamwork patterns [23, 30], team recognition [44] and team performance enhancement [15]. Other studies include those that investigate attitudes to teamwork. Network science approaches have been proposed to effectively analyze, study, and explore collaborative teamwork patterns [27]. Many large-scale academic networks based on large collaboration and citation relationships can be constructed from easily accessed digital libraries such as DBLP, CiteSeerX and

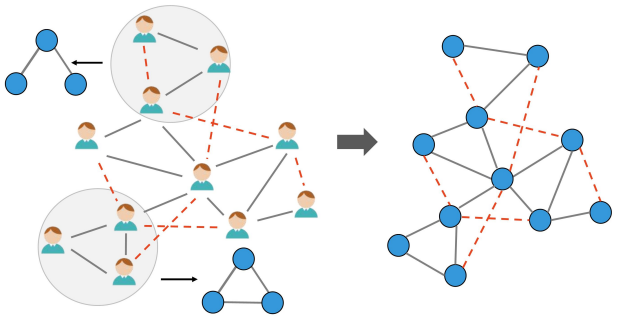

Figure 1: Triangle closures in academic social networks. The left figure shows the academic social network in reality, and the arrow represents the types of motifs that exist in the network. The right figure shows the topology of the network. The blue nodes represent scholars, the gray edges represent the cooperative relationship between them, and the red dashed lines represent possible links.

MAG. Academic networks provide more complex opportunities for scientific team research [47] than traditional research methods that commonly employ case studies or small samples. However, collaborative teamwork data are contained in these networked data in implicit ways, which makes it difficult for scholars to directly employ collaborative team data in their studies [16]. Therefore, one of the most fundamental research problems is to automate the recognition of collaborative teams within academic networks.

Related studies focusing on team recognition have also been conducted. Yu et al. [44] propose a network-based approach to identify collaborative teams from academic networks. They propose an index entitled CII (Collaboration Intensity Index) that qualifies collaboration intensity. CII is employed to identify if a certain relationship belongs to a team. Some related studies also use community detection methods to identify collaborative teams [4, 12]. Some community detection methods will recognize clusters of scholars. However, communities are generally identified on a broader basis than collaborative teams. Though some community detection methods can recognize fine-grained communities, scale is not the only difference between academic communities and collaborative teams. Stability is also a feature of significance in the academic collaborative team. It has been shown that a team will achieve better performance when the collaboration patterns among members tend to be stable [11, 42]. To be specific, members would like to work with those they are more familiar with. A relatively stable communication pattern as well as work pattern will help improve teamwork performance [17]. It is therefore important to consider familiarity as a contribution to stability when recognizing collaborative teams. It is difficult, however to qualify the degree of familiarity between members, especially in large scale academic networks.

An academic collaboration network is generally constructed by analyzing groups of scholars who have produced research publications together. This network is constructed based on authorship and co-authorship relationships. In the constructed network, nodes represent scholars and edges represent the collaboration relations between certain scholars. The weights of edges are generally qualified by the number of publications two scholars published, which reflects the collaboration intensity between two scholars to some extent. It has been discovered that triangle closures are common in academic social networks. These can be represented by network motifs (shown in Fig.1) [37]. Network motifs are induced subgraphs that appear more frequently in real-world networks, which generally represent certain social patterns that have real meaning [48].

By exploring the concept of network motifs, this work uses familiarity [41] based on one's network structure and then proposes a team recognition method. We first calculate the lengths of possible paths between two nodes in the network. Then we calculate the local density and distinguishable distance between each node to draw a decision figure. Nodes with greater local density and a greater distinguishable distance value will be chosen as central nodes for each team. The remaining nodes are linked with a particular central node based on their shortest distances to the central node, local density, and familiarity. The contributions of this paper are summarized below:

- **Collaborative team recognition:** We propose MOTO (faMiliarity-based cOllaborative Team recOgnition algorithm) to recognize collaborative teams in academia. The proposed approach employs network motifs to qualify familiarity among scholars and recognize teams with a local density as well as a distinguishable distance between nodes.
- **Higher-order familiarity qualification:** We employ the qualification metric of higher-order familiarity among scholars. As the structural property of a certain node, the number of motifs can reflect the collaboration familiarity among scholars. Based on this metric, an academic team is recognized more precisely.
- **Real-world data verification:** We use a real-world data set to recognize collaborative teams. Microsoft Academic Graph Computer Science data set is employed in the experiments. The experimental results show that our proposed method can recognize collaborative teams effectively. These teams are then analyzed in more detail.

The remainder of this paper is organized as follows. Section 2 introduces related work, including network structure and team recognition. Section 3 introduces preliminaries including the definition of pairwise familiarity, higher-order familiarity, and network motif. Section 4 illustrates the details of MOTO algorithm. Section 5 analyzes experimental results. Finally, Section 6 concludes the paper.

2 RELATED WORK

2.1 Network Structure

The definition of network motifs and the algorithm for mining them was first provided by Milo et al. [20]. They were attempting to identify patterns in complex networks and found frequent occurrences of subgraph connection patterns that would not be found in equivalent numbers in random networks. These recurring, significant patters of interconnections were given the title of network motifs and Milo and his colleagues found examples in biochemistry, neuorbiology, ecology and engineering networks. Since that time,

researchers have continued to find applications for network motifs in multiple research areas [35].

The discovery and counting of network motifs has become an important research area. Ahmed et al. [1] proposed a fast and efficient parallel counting method for three-point and four-point subgraph patterns. The method can calculate the accurate number of subgraph patterns and significantly reduce the calculation time. Ma et al. [18] explored solutions to the problem of counting motifs on uncertain graphs and proposed two algorithms named Possible Graph Sampling (PGS) and Linking and Counting (LINC). PGS samples some possible worlds from the graph and then runs a deterministic modal counting algorithm on each possible world. LINC is an improvement of PGS; it can effectively calculate the difference in the motif count of two different possible worlds. Different algorithms for the discovery of network motifs have also been studied. Yu et al. [40] classified and summarized the discovery algorithms of network motifs and compared the running time of different algorithms. They also discussed the application of these algorithms in various scenarios.

Many studies use network motifs to analyze the characteristics of different types of networks. Milo et al. [19] analyzed the distribution of triangle motifs and four-order motifs in different networks, and classified networks in different fields according to the statistical importance in the distribution curve of the number of motifs. They identified that the statistical importance of the triangle motif in social networks is significantly higher than in others networks. Zhao et al. [49] studied network motifs in social networks and proposed a method called Motif-based PageRank (MPR), which considers first-order and higher-order relations for user ranking in social networks. They computed the motif-based adjacency matrix and combined it with the edge-based adjacency matrix to re-weigh the links between users. They also studied the performance of other types of motifs. Paranjape et al. [22] defined and studied motifs in time series networks.

Some research has combined an application of clustering and motifs with large-scale networks. Benson et al. [2] developed a general framework based on network clustering of high-order motifs. They showed that there are rich high-order motifs in large-scale networks, such as the information dissemination unit of neural networks and the network hub structure of traffic networks. To solve the dynamic local motif clustering problem, Fu et al. [8] proposed a model called Local Motif Clustering on Time-Evolving Graphs (L-MEGA). L-MEGA mainly used edge filtering, motif push operation, and incremental sweep cut to track the temporal evolution of the local motif cluster.

More recent studies have continued to systematically review network motifs. Yu et al. [45] summarized the definition and related concepts of network motifs. They analyzed network motifs in biological networks, social networks, academic networks, and infrastructure networks. They provided insights into motif discovery, motif technology, motif clustering methods, and network motif applications in different fields. Xia et al. [35] classified network motif measures into structural measures and statistical measures according to the calculation method of the measurement indicators. They analyzed the application of these measures in the discovery, counting, analysis, and clustering of the network motif.

These studies all support the notion that network motifs can reveal the basic structure of most networks and play an important role in various network applications. However, most of the existing research ignores the influence of the network structure and the familiarity between team members and its influence on recognition. To address these issues, we comprehensively consider these factors and use the existing familiarity metric to quantitatively describe the familiarity between scholars. We propose MOTO based on this metric, which makes the identified team more cohesive and lowers the cost of team communication.

2.2 Team Recognition

Academic team recognition algorithms have evolved corresponding to changes in the nature of academic teams over time. Before the large-scale development of the Internet and social networks, academic teams were the same as scientific research institutions, referring to scholars engaged in scientific research in the same institution. Traditional academic team recognition methods have used artificial methods such as questionnaire surveys; these methods have low efficiency, are high time consuming and costly, and are limited by the available samples produced. With the rapid development of social networks, scholars have been able to cooperate remotely and a large number of interagency and interdisciplinary academic teams have emerged. However, the concept of what an academic team actually is, is not settled. For example, some studies regard the co-author of a paper as a member of a research team and have used this definition to explore the macro issues of team science [5, 21]. Some researchers regard a team with two to ten scientists as a scientific team and a team with more than ten as a large team [7]. Some researchers have provided their own definition based on the classic definition. Some studies use visual tools to show networks and combine cliques to identify academic teams [26]. However, in reality, the members of these identified teams may not directly collaborate. Calero et al. [3] proposed a new bibliometric method to identify research teams in specific research fields by combining bibliometric methods and network analysis. Yu et al. [44] proposed a team recognition method called TRAC based on the Collaboration Intensity Index. The method uses a top-down approach to delete edges with a cooperation intensity less than a threshold, and finally, uses the derived small connected networks to define academic teams.

Community detection algorithms can be used to discover the community structure in networks [34]. However, when classic community detection algorithms that deal only with the structure of social networks or detect communities using only node attributes e.g. age, gender and interests are used in isolation, the results may be limited [6]. Team recognition tasks are more fine-grained, the team members have different attributes and this makes relationships complicated. Therefore, the following research improves the community detection algorithm to identify academic teams. Savić et al. [25] proposed a method for community detection in research collaboration networks. They set frequent collaborators as the core of the research team and determined them through w-core. W-core is a graph traversal algorithm, which mainly assigns each node so that the two nodes from the same w-core have the same label, while the two nodes from different w-core have different labels. Villarreal

et al. [31] proposed a clustering algorithm for cooperative scientific networks, where attempts are made to cluster on both sides of a bipartite graph in order to obtain the cluster of authors and articles. The proposed method not only detects research teams, but also describes and visualizes the detected teams.

Yu et al. [46] use a slightly different approach by defining an academic team as an academic cooperative team composed of leaders, core team members and non-core members. Their research teams identification method identifies team leaders based on the centrality measure and uses 2-clique to identify core members. However, their approach does not take into consideration the degree of relationship between team members, such as the closeness and familiarity of the connection, which makes it difficult to identify academic teams efficiently.

Most of existing community detection methods are complicated and have high computational costs. Therefore, in this work, we propose a team recognition method by exploring cluster centers. Compared to the community detection approaches, the design of our method is straightforward, which can recognize clusters regardless of their shape and the dimensionality of the space in which they are embedded. Moreover, academic teams are generally with "core+extended" structure[36]. Our proposed method can better recognize teams with such structure.

3 PRELIMINARIES

In keeping with the objectives of this paper, we now in this section provide a definition of familiarity, which is used to measure the overall familiarity between scholars and other team members. We also provide a more formal definition of the concept of network motifs.

3.1 Familiarity

Yu et al. [41] first proposed the definition of familiarity. They divided it into Pairwise familiarity and Higher-order familiarity. The following is the specific calculation formula.

3.1.1 Pairwise Familiarity. Pairing familiarity refers to the number of team members who have established a cooperative relationship with the scholar, i.e., there are edges in the cooperative network. The formula is shown in Eq.(1).

$$\|F\|_1(i,T) = \sum_{j \in T, j \neq i} PairwiseCol_{ij} \tag{1}$$

where T refers to the team. When there is an edge between i and j, $PairwiseCol_{ij} = 1$, otherwise $PairwiseCol_{ij} = 0$. For scholars outside the team T, the more people they have worked with in the team, the more familiar the scholar is with the team. The communication cost of cooperation is even lower.

3.1.2 Higher-order Familiarity. Higher-order familiarity refers to the number of team members who have established a relatively stable cooperative relationship with the scholar. Relatively stable means they have established a triangle motif. The formula is shown in Eq.(2).

$$\|F\|_n(i,T) = \sum_{j \in T, j \neq i} MultiCol_{ij} \tag{2}$$

where $MultiCol_{ij} = 1$ means that i and j appear in at least one triangle motif, $MultiCol_{ij} = 0$ means that i and j never formed a triangle motif. $\|F\|_n(i,T)$ indicates that the number of members in team T who form a triangle motif with i. The higher of $\|F\|_n$, the more familiar i is with team T.

3.2 Network Motif

Milo et al. [20] first proposed the definition of network motif. They proposed that network motifs are interconnections patterns of the subgraph that repeatedly appears in the original network, which appears more frequently than in the similar random network. The distribution of node degree in random networks and real networks should be consistent.

Let $G = \{V, E\}$ be a network, where V is the node set, E is the edge set. $G_k \subset G$ means the subgraph of G whose size is k. Given a network G, a set of parameters $\{P, U, D, N\}$ and a set of N similar random networks. The network motif is defined as an induced subgraph appearing in the real network that meets the following three conditions:

$$p(\bar{f}_{rand}(G_k) > f_{real}(G_k)) \leq P$$

$$f_{real}(G_k) \leq U$$

$$f_{real}(G_k) - \bar{f}_{rand}(G_k) > D \times \bar{f}_{rand}(G_k)$$

where $f_{real}(G_k)$ is the occurrence of the subgraph in the real network, $\bar{f}_{rand}(G_k)$ is the average occurrence of the subgraph in all random networks. P is the probability threshold determined by N similar random networks. The first condition is to ensure the motif did appear with much higher frequency in real-world network comparing to random network. U is the unique cutoff value of the frequency of network motif in the real network. The second condition is to limit the appearing frequency of motif that appears in real-world network. In the third condition, D is the minimum difference cutoff ratio to ensure the minimum difference between $f_{real}(G_k)$ and $\bar{f}_{rand}(G_k)$. According to the experience, the parameters $\{P, U, D, N\}$ are generally set as $\{0.01, 4, 0.1, 1000\}$.

Fig. 2 shows all possible directed 3-motifs. Here we only list motifs with two or more edges to make sure that all the listed motifs are connected. At present, many studies have identified motifs with distinctive characteristics in different types of networks. For example, the triangle fully connected motif appears more frequently than other motifs in social networks [19]. These triangle fully connected motifs are demonstrated in subfigures 9 to 13 of Fig. 2. Such findings have motivated researchers to take advantage of the characteristics of motifs in relevant research. In particular, collaboration relationships are recognized as two-way edges. Consequently, in this work, motifs are regarded as being undirected in line with undirected collaboration networks.

4 THE DESIGN OF MOTO

In an academic team, the familiarity between members is an important feature of the team. Some studies have shown that people are more inclined to cooperate with familiar people, and team members are more familiar with their team than others [9]. In this section, we

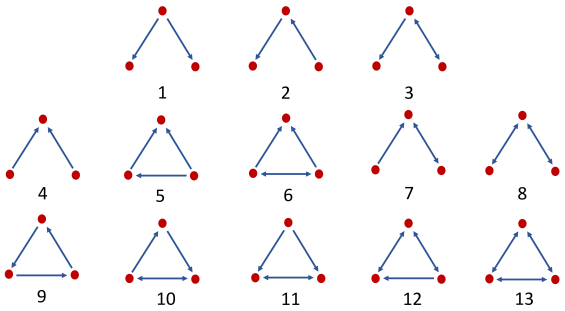

Figure 2: All possible directed 3-motifs.

propose MOTO, which is based on the CFSFDP algorithm proposed by Alex et al. [24].

The main steps of MOTO are shown in Fig.3. Firstly, we add weights to the edges in the academic collaboration network, and then calculate the shortest distance matrix between any two nodes. Next, calculate the local density of each node and the minimum distance between each node and all nodes with higher local density. Determine the cluster center node and the number of clusters based on the local density of each node and the maximum distance from the high-density node. After that, assign other nodes except the center nodes to the nearest cluster center node so as to complete the preliminary team recognition. Then divide the edge area of each team based on the familiarity and determine the threshold of the local density and team familiarity of the team members. Filter the team members according to the two thresholds to identify the academic team. Finally, we divide these academic teams with academic institutions and get the academic teams within the academic institutions. The following sections will describe major steps of MOTO in detail.

4.1 Calculation of Node Pair Distance

In step one, we calculate the distance between all scholars in the network. G is an undirected weighted graph. The weight of an edge is the cooperation distance between two nodes, which is shown in Eq.(3).

$$d_{ij} = 1 - \frac{|P_i \cap P_j|}{|P_i \cup P_j|} = 1 - \frac{\cot_{ij}}{pn_i + pn_i - \cot_{ij}} \tag{3}$$

where P_i and P_j represent the paper set of scholar i and j, respectively. $|P_i \cap P_j|$ is the number of papers co-authored by scholars i and j. $|P_i \cup P_j|$ is the number of non-repeated papers written by the two scholars. The minimum value of d_{ij} is 0, and the maximum value is 1. In order to improve the efficiency of calculation, we simplify it. We represent $|P_i \cap P_j|$ as cot_{ij}, which refers to the times of cooperation between scholar i and j. pn_i and pn_j represent the number of papers by scholar i and j, respectively.

In the subsequent clustering step, we need to calculate the distance between any two nodes, which is the sum of the distance between two nodes of the shortest path. It is represented as

$$dis(v_i, v_j) = shortestPathLength(v_i, v_j) \tag{4}$$

We choose the Dijkstra algorithm to calculate the distance. The Dijkstra algorithm is the shortest path algorithm from one node to the other nodes. It is applicable to both directed and undirected graphs, and it requires the weight to be non-negative. Due to the large scale of the academic collaboration network, the exploration range value of the shortest path can be set in the calculation process. This can reduce the complexity whilst calculating the distance required for clustering.

4.2 Calculation of Local Density and Distinguishable Distance

In step two, we calculate the local density ρ of each node within the cutoff distance d_c. The ρ of a scholar in the network measures the density of scholars who have a certain degree of close cooperation with the scholar. d_c is the only hyper-parameter in the algorithm, which represents the range of ρ. For each node $i \in V$ in G, ρ_i refers to the number of other nodes in the network whose distance from node v_i does not exceed the range of d_c except for node v_i. It can be calculated using the following equation:

$$\rho_{v_i} = \sum_{v_j \in V, v_j \neq v_i} \chi\left(dis\left(v_i, v_j\right) - d_c\right) \tag{5}$$

$$\chi(x) = \begin{cases} 1, & x < 0 \\ 0, & x \geq 0 \end{cases} \tag{6}$$

where the value of function $\chi(x)$ is 1 when the distance between v_i and v_j is less than d_c, otherwise the value is 0. It should be noted that we believe that the ρ of the cluster center node is very high. This is specifically reflected in the fact that the scholar keeps collaboration with more people in the team, rather than that the scholar is the leader of the team. Then we sort all nodes in descending order according to their ρ. The high-density node set of node v_i is $VP_{v_i} = \left\{v_j \mid \rho_{v_j} > \rho_{v_i}\right\}$.

Next, we calculate the shortest distance between each node and the high-density node, i.e., the distinguishable distance δ. We use the distance between two nodes to distinguish between two teams so if we assume that a node is a cluster center node, the node with greater ρ than this node is either the center of another team or the node closer to the center in the same team. In other words, in the cluster where a central node is located, there should be no nodes with higher ρ than it. Therefore, when determining the cluster center, in order to ensure that the distance between the clusters is significant, the cluster center should be further away from all the higher density nodes than those within its cluster so that the two clusters will not merge into one cluster. The δ of node v_i is the minimum distance between v_i and VP_{v_i}, defined as

$$\delta_{v_i} = \begin{cases} \min\limits_{v_j \in VP} \operatorname{dis}\left(v_i, v_j\right), \rho_{v_i} \neq \max_{-} \rho \\ \max\limits_{v_j \in V, v_j \neq v_i} \operatorname{dis}\left(v_i, v_j\right), \rho_{v_i} = \max_{-} \rho \end{cases} \tag{7}$$

where $\max_{-} \rho$ is the maximum ρ of all nodes. For the node with the highest ρ, its distinguishable distance is the maximum distance from any other node.

4.3 Determine Cluster Center

In step three, we use the local density and distinguishable distance of each node to draw the cluster center decision graph, as shown in Fig.4. The horizontal axis represents the local density, and the

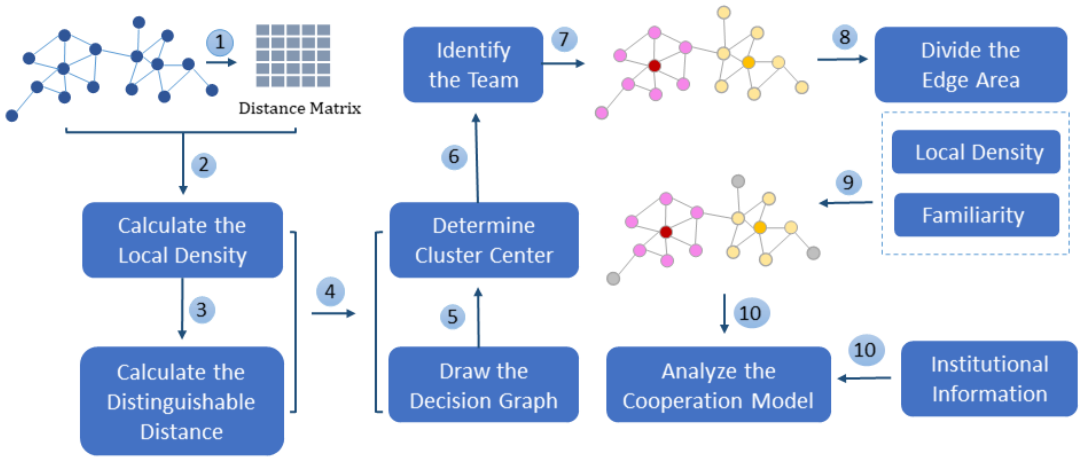

Figure 3: The flowchart of MOTO.

vertical axis implies the distinguishable distance. The decision graph is divided into four areas, and the nodes in each area have different characteristics:

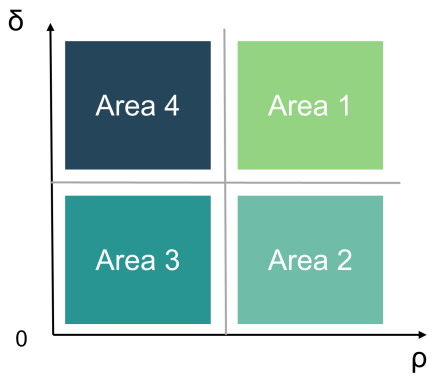

Figure 4: The regional division of decision graph.

- **Area one:** The nodes have high ρ and high δ, which accords with the characteristics of high ρ of cluster center nodes and are some distance from other possible cluster center nodes. So the nodes in this area are cluster center nodes. This area can be clearly distinguished from other areas.
- **Area two:** The nodes have high ρ but low δ. These nodes' cutoff distances contain nodes with higher ρ, which are nodes close to the center but not the center in the team. The specific roles of these nodes in the team need to be analyzed in combination with experimental results.
- **Area three:** The nodes have low ρ and low δ. These points are located in relatively sparse locations in the network, and comparatively further away from all nodes with high ρ. They may be located in the middle of different circles, or the collaborators are relatively scattered. Each specific situation should be analyzed based on the experimental results.
- **Area four:** The nodes have low ρ but high δ, which means that they are relatively isolated nodes. Such scholars have few collaborators.

Due to the larger scale of the academic collaboration network, it is still necessary to apply a more intuitive judgment method when observing the decision graph to get the local density of the cluster center area and the boundary of the distinguishable distance. We calculate the product of ρ and δ of each node, i.e., $\gamma_{v_i} = \rho_{v_i} \times \delta_{v_i}$. Therefore, according to γ and the decision graph, we can get a set $Center = \{c\}$ consisting of nodes with relatively high ρ and δ.

4.4 Team Recognition

In this step, after having identified the cluster center nodes in step three, we divide the entire network into $|\ Center\ |$ clusters, i.e., academic teams. For each non-cluster center node v_i, we calculate the distance $dis(v_i, c)$ between the node and each central node, and place the node into the cluster where the nearest cluster center node is located.

Next, we identify the set of eligible nodes in each cluster. Academic team members are closely connected with the team and maintain a certain degree of familiarity with the team members but are relatively sparsely connected with other teams . That is, the cooperative behavior of a team member should have the familiarity and closeness of connecting with other members of the team. Team familiarity means the sense of participation in the team, i.e., the member has direct collaboration or high-level collaboration relationship with the team members. Closeness measures the local density degree of members in the collaboration network. A node with higher closeness is more probably to be recognized as central node. Therefore, Closeness is employed to recognized cluster centers and familiarity is employed to recognize other team members corresponding to centers. In summary, academic team members should meet the following two conditions:

1) **Closeness:** ρ is higher than the threshold ρ'.

2) **Familiarity:** the team familiarity is higher than the threshold $\|F\|'_n$.

Determining the values of ρ' and $\|F\|'_n$ is an important step in filtering. Because the team size and the closeness of cooperation are different, it is necessary to determine each team's threshold based on the characteristics of each team. First, divide the edge area of the team based on the condition that there are member nodes of different teams within the neighborhood of the node's cutoff distance, then the edge area $border(T)$ of the team T is represented as:

$$\text{border}(T) = \left\{v_i \mid \exists \,\text{dis}\left(v_i, v_j\right) < d_c, v_i \in T, v_j \notin T\right\} \tag{8}$$

Algorithm 1 MOTO

Input: $G = (V, E, W)$, which is the academic collaboration graph with edge weight (collaborative distance between scholars), cutoff distance d_c;
Output: academic team list T_c;

ρ_{v_i},δ_{v_i} = Calculation of Local density ρ_{v_i} and distinguishable distance δ_{v_i} (G, d_c)
$Cluster$ = Division of Academic Teams Algorithm (G, ρ_{v_i}, δ_{v_i})
for each T of $Cluster$ **do**
  for v_i in V **do**
    calculate $\|F\|_n(v_i, T)$;
  end for
  $border(T) = \left\{v_i \mid \exists \,\text{dis}\left(v_i, v_j\right) < d_c, v_i \in T, v_j \notin T\right\}$
  $\rho' = \max \rho, \rho \in border(T)$
  $\|F\|'_n = \max \|F\|_n, \|F\|_n \in border(T)$
  $T_c = \left\{v_i \left| \rho_{v_i} \geq \rho', \right| |F| \right|_n (v_i, T) \geq ||F||'_{n'} v_i \in T\right\}$
end for
return T_c

Algorithm 2 Calculation of Local density ρ_{v_i} and distinguishable distance δ_{v_i}

Input: $G = (V, E, W)$, which is the academic collaboration graph with edge weight (collaborative distance between scholars), cutoff distance d_c;
Output: local density ρ_{v_i}, distinguishable distance δ_{v_i} ;

for $pair(v_i, v_j)$ in V **do**
  calculate $dis(v_i, v_j)$;
end for
for v_i in V **do**
  $\rho_{v_i} = \sum_{v_j \in V, v_j \neq v_i} \chi\left(dis\left(v_i, v_j\right) - d_c\right)$;
end for
for v_i in V **do**
  $VP_{v_i} = \left\{v_j \mid \rho_{v_j} > \rho_{v_i}\right\}$;
end for
for v_i in V **do**
  if $\rho_{v_i} \neq \max_{-} \rho$ **then**
    $\delta_{v_i} = \min_{v_j \in VP} \text{dis}\left(v_i, v_j\right)$;
  else
    $\delta_{v_i} = \max_{v_j \in V, v_j \neq v_i} \text{dis}\left(v_i, v_j\right)$
  end if
end for
return ρ_{v_i},δ_{v_i}

Algorithm 3 Division of Academic Teams Algorithm

Input: $G = (V, E, W)$, which is the academic collaboration graph with edge weight (collaborative distance between scholars);
Output: cluster set $Cluster$;

for v_i in V **do**
  $\gamma_{v_i} = normalized(\rho_{v_i} \times \delta_{v_i})$
end for
Select the cluster center node set $Center$
for v_i in V **do**
  $Cluster_{v_i} = \arg\min_{c_j \in Center} dis(v_i, c_j)$
end for
return $Cluster$

The nodes in the team edge area are the nodes where the cooperation between the team and other teams is not significant. Additionally the local density of these nodes is not strong enough to be a cluster center, nor are they isolated scholars. Team familiarity is similar to the situation of these nodes. Therefore we choose the maximum local density and team familiarity of the node in the edge region as the thresholds for ρ' and $\|F\|'_n$.

The next step is to filter all members of the team according to ρ' and $\|F\|'_n$. Nodes with local density and familiarity above the threshold are identified as team members. The team edge nodes are the boundary part of multiple teams or a single team, which can be successfully identified by the above filtering methods. The set of teams obtained is expressed as T_c:

$$T_c = \left\{v_i \left| \rho_{v_i} \geq \rho', \right| |F| \right|_n (v_i, T) \geq ||F||'_{n'} v_i \in T\right\} \tag{9}$$

Finally, the division of nodes according to institutional attributes and the division of academic teams obtained by clustering are shown in the network. An academic team TI_c^i in $Institution_k$ is represented as:

$$TI_c^i = \left\{v_j \left| v_j \in T_c, \right| v_j \in \text{Institution}_k\right\} \tag{10}$$

These are the main processes of MOTO, and the specific pseudocodes are shown in Algorithm 1. The parameter setting varies according to the different parameter values of the academic collaboration network.

5 EXPERIMENTS AND RESULTS

The focus of this section is to introduce the dataset used in the experimentation, the data preprocessing process, the network statistics overview and the experimental settings. In order to evaluate our proposed algorithm, we also introduce a number of baseline methods and analyze the experimental results.

5.1 Dataset Collection and Data Preprocessing

We conducted extensive experiments using data from Microsoft Academic Graph (MAG)[1]. It is an open academic dataset that contains more than 200 million scientific research literature publication records and citation relationships between the literature since 1800 [28]. MAG includes six entities: publications, authors, institutions, journals, conferences, and fields of study. The relationships between entities are shown in Fig.5.

[1]https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

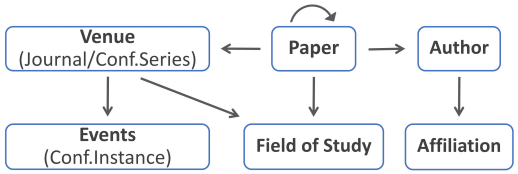

Figure 5: The relationship of entities in MAG data set.

For the MAG dataset, we perform the following processing operations to obtain the experimental data.

(1) Firstly, we extract all papers in computer-related fields. We select papers in these fields from 2006 to 2017 as experimental data. MAG contains 34 sub-areas in Computer Science. According to the attribute fieldOfStudy of the paper, we extract 12,923,247 papers related to the above 35 fields. We then extract papers in the desired year. Yu et al. [44] showed that if two scholars collaborated and did not cooperate again in the next four years, they would not cooperate again. So the experiment focused on selecting periods of four years. Due to the evolutionary behavior of the team, each period is separated by two years. After filtering papers with missing information, we extracted 1,066,628 papers from 2006 to 2009, 1,258,318 papers from 2008 to 2011, 1,477,560 papers from 2010 to 2013, 1,602,827 papers from 2012 to 2015, and 2,827,671 papers from 2014 to 2017.
(2) Secondly, we filtered the authors. Although students are also involved in scientific research, most of them will end their academic lives within 5 years and do not constitute the backbone of the academic team. Therefore, we select scholars with an academic life of 5 years or more as the research objects. Finally, we obtained 291,188 scholars who meet the above requirements in 12 years. Then, we constructed the academic collaboration network by the cooperative relationship between these scholars.

After constructing the academic collaboration network, we found that the network contained many connected pieces and nodes. To more easily control the experiment, we selected the largest connected piece in the collaboration network in each period to use in subsequent experiments. The profile of the academic collaboration network is shown in Tab.1. As the network grew in size over time, the average number of collaborations between scholars and the average degree of nodes increased. This indicates that the number of each scholar's collaborators also increased.

An important qualification that we needed to make was that the institution a scholar belonged to, was not necessarily fixed over time i.e. they may not belong to the same institution across different periods. So we extracted the author's institution and removed duplicates to get the author's full institutional attributes. If a scholar had more than one institution in a certain period, the scholar was considered to belong to both those institutions at the same time. The institutional attributes of scholars are mainly used during cooperation mode analysis.

Table 1: The profile of the academic collaboration network.

Properties	Nodes	Edges	Avg Co-times
2006-2009	105,721	298,768	2.163
2008-2011	140,241	429,253	2.092
2010-2013	167,535	544,986	2.033
2012-2015	179,773	602,301	2
2014-2017	197,001	832,248	2.449
Properties	**Avg degree**	**Triangles**	**CCF**
2006-2009	5.652	416,454	0.38
2008-2011	6.121	648,055	0.391
2010-2013	6.505	1,368,684	0.394
2012-2015	6.701	1,146,961	0.397
2014-2017	8.45	1,843,896	0.385

5.2 Experimental Settings

There is only one important parameter, the cutoff distance d_c, which can be set by experience so that the number of nodes in each node's d_c neighborhood is 1%-2% of the total number of network nodes [24]. We calculate the distribution from 0.0 to 3.5 in the experiment and select the center node based on the observed value. The statistical results show that different d_c values have no obvious effect on the distinguishable distance of the cluster center but influence the δ of the cluster center. Therefore, when determining the cluster center, ρ and δ are both standardized. Tab.2 shows the impact impact of team recognition results in different d_c values. We can see that d_c within 1.5-2.5 has no obvious influence on the experimental results, and the robustness is good. In our experiment, d_c is set as 1.6 for 2006-2009, 1.5 for 2008-2011, 1.5 for 2010-2013, 1.5 for 2012-1015, and 1.4 for 2014-2017.

Table 2: The impact of team recognition result with d_c.

d_c	**0.5**	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**
2006-2009	14,566	14,867	14,988	14,988	14,988	14,593	14,354
2008-2011	21,946	22,003	22,279	22,280	22,280	22,014	22,003
2010-2013	27,017	27,127	27,628	27,628	29,628	27,319	27,278
2012-2015	27,143	27,274	28,207	28,207	28,205	27,920	27,674
2014-2017	28,179	28,380	29,530	29,530	29,530	28,739	28,240

5.3 Evaluation Metrics

To evaluate and analyze the effectiveness of MOTO for recognizing team results, we used five metrics: the number of recognized teams, team size, team communication cost, the number of triangle motifs, and separation degree. These metrics are introduced in detail below.

Number of teams recognized and team size: The number of recognized teams is one of the most basic metrics used. Team size is an important structural variable of a team, which can not be ignored. Appropriate team size is not only conducive to team communication, but also can improve team efficiency, which is the basic guarantee for completing research tasks of a scientific research team.

Team communication cost: It is an important indicator of whether or not the team cooperates effectively. The lower the communication cost, the more effective the team collaboration is. To measure this, we use Communication Cost Radius (CCR), i.e., the diameter of the induced subgraph of team members, which is the maximum length of the shortest path between any two nodes[14, 33]. The calculation formula is shown in Eq. (11).

$$CCR = \max_{i,j \in T} shortestPathLength(i, j) \tag{11}$$

Number of triangle motifs: Triangle motifs are a connection mode that exist widely in social networks, which is also defined as triadic closure in social networks. We use it as an indicator to evaluate the team structure. The more triangle motifs in the team, the closer the cooperation between team members.

Separation degree: It is used to measure the closeness of team members to the external and internal connections of the team. The greater the separation degree of the team, the closer the connection between team members and people outside the team; the smaller the separation degree, the closer the internal connection of the team. The measure can be calculated using the following equation

$$Separability(T) = \frac{Out_T}{All_T} \tag{12}$$

where Out_T is the number of connections between members of team T and people outside the team. All_T is the number of connections between members of team T and the inside and outside of T.

5.4 Baseline Approaches

We use four methods as baseline approaches for comparison with our proposed algorithm: Team Recognition Algorithm based on CII (TRAC), Team Identification Based on iterative Centrality Ranking (TIBCR), Cluster Affiliation Model for Big Networks (BIGCLAM), and Discovering Community Cores (DCC).

(1) **TRAC** [44] is a team identification algorithm based on the Collaboration Intensity Index (CII). It is a network edge weight filtering method. The first step is to set the edge weights in the collaboration network as CII. The second step is to screen network nodes according to Partnership Ability Index (PHI). The third step is to set the cooperation constraint coefficient W, delete the edge whose CII is lower than W, and delete the node without edges.

(2) **TIBCR** [46] is a team leader and team identification algorithm based on iterative centrality ranking. The first step is to calculate and rank the intermediate centrality of each node in the academic collaboration network. Then, the 2-clique method is used to identify the core team members. Based on the team leader and core team members, the snowball method is used to identify the general team members.

(3) **BIGCLAM** [38] is a model-based overlapping community detection method suitable for large networks. It can detect densely overlapping, hierarchically nested, and non-overlapping communities in massive networks. It first calculates the attribution vector of each node, and then uses a method based on matrix factorization to divide the community. The algorithm constructs a bipartite affiliation graph to simulate the structure of the community. Based on the new bipartite graph, it uses the graph adjacency matrix to maximize the affiliation matrix of the node.

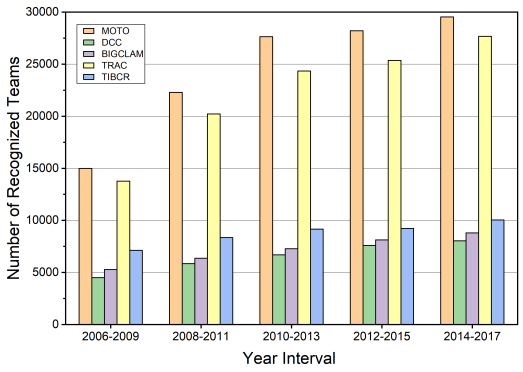

Figure 6: The number of recognized teams in time intervals.

(4) **DCC** [13] improves Speaker-listener Label Propagation Algorithm (SLPA). In SLPA, speaker-listener labels are allocated to different nodes according to the information transmission process. The labels are then spread among nodes according to the previous and current iteration information of the nodes. Finally, the labels are used to aggregate the nodes and form a community. DCC expands on this by setting the weight of the network to Intimacy.

TRAC is designed based on network edge weight filtering. This method is straightforward but neglects the team structure. Moreover, TRAC cannot recognize overlaps. TIBCR is an iterative method, which can respectively recognize team leader, core members, and other members. However, the iteration process is time-consuming. Therefore, it is not suitable for team recognition in large-scale networks. DCC improves the process of iteration to simplify the label update. Therefore, DCC is a more effective iterative method and meanwhile can recognize teams with overlaps. However, due to the randomness of label propagation, the recognition results of DCC are not stable. BIGCLAM considers community structure and membership strength to detect communities with overlap. This method utilizes coordinate ascending method to optimize non-negative matrix factorization. Therefore, it has obvious shortage in complexity, scalability, and linear model expression ability.

5.5 Experimental Results

In this section, we compare the result of our proposed algorithm with the baseline methods. Experiments were conducted using high-order familiarity (MOTO-H) and pairwise familiarity (MOTO-P), respectively.

Number of teams recognized and team size: Fig.6 shows that the number of teams recognized by MOTO and comparison algorithm for different periods. The x-axis represents the period, and the y-axis represents the number of teams. This figure shows, that over time, the number of recognized teams increased. With respect to the overall network, the number of network nodes also increases over time, as shown in Tab. 1. We determine that the size of the collaboration network has become larger in recent years, so the number of academic teams will also increase. This increase is consistent with our real world expectations. In all time periods, MOTO identifies the greatest number of academic teams when compared with baseline methods chosen for comparison.

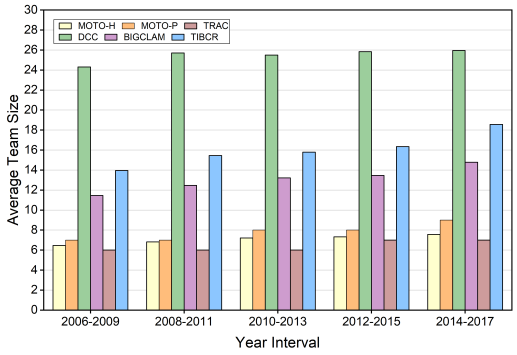

Figure 7: The average team size in time intervals.

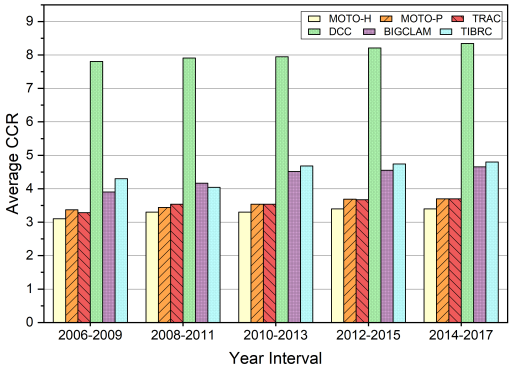

Figure 8: The average CCR of teams in time intervals.

Fig.7 shows that the size of team recognized by MOTO-H and MOTO-P is the second and third smallest among all algorithms, respectively. MOTO-H and MOTO-P team sizes range between 7 to 10 people approximately. Therefore, MOTO recognizes more teams, but meanwhile the teams recognized by MOTO are with regular number of team members. That is to say, in the recognition process, MOTO does not split the teams. The team size recognized is slightly different with different familiarity. The high-order familiarity requires the team members to establish more triangle motifs representing cooperative relationships with other members, which is smaller than the team size obtained by pairwise familiarity. The number of teams recognized by the TRAC is also small but the difference to MOTO is less than about 14%. The teams identified by TRAC are smaller, with approximately 6 or 7 members. The recognition results are closer to the average degree of nodes. However, the total number of teams is not the highest of all methods, because some nodes will become isolated nodes while deleting edges. DCC recognizes the least number of teams and the average size of the team is large, with 24-26 people, which does not match expectations based on real world observations. The team size recognized by TIBCR and DCC is between 11-19 members, which is a medium-sized team. The reason for the larger team size under TIBCR and DCC is that these two methods do not remove some members who do not cooperate closely.

In summary, the recognition result of MOTO is closer to our expectations of the team cooperation situation in reality. MOTO also loses minimal information when compared to other baseline methods. Choosing a higher level of familiarity will identify smaller teams. Over time, the size of the teams gradually increases.

Team communication cost: Fig.8 shows the average CCR in each time interval. In each time interval, the results of MOTO-H and MOTO-P are the two lowest apart from the average CCR of MOTO-P in 2006-2009 which was less than 0.1 higher than TRAC. Overall therefore the performance of MOTO is similar to or better than TRAC. Similarly, the average team size recognized by MOTO-H and MOTO-P was larger than TRAC. The communication costs of other algorithms are significantly higher than MOTO and TRAC. which is mainly because the size of the teams recognized by these three algorithms is significantly larger than MOTO and TRAC. It can be concluded that the teamwork recognized by MOTO-H and MOTO-P is more efficient than other methods. The difference between the CCR of the two MOTO algorithms is minimal (less than 0.3 across each time interval). We suggest that this is due to the team size identified by MOTO-H being slightly smaller than MOTO-P. These results also indicate that the cost of high-order familiarity relationship cooperation is lower.

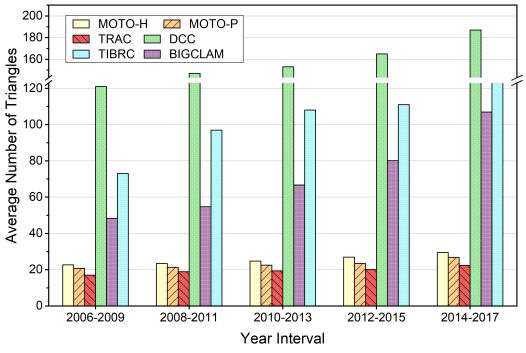

Figure 9: The average number of triangles in time intervals.

Number of triangle motifs: Fig.9 shows the average number of triangle motifs recognized by different algorithms. The team sizes identified by DCC, TIBRC, and BIGCLAM are at least twice that of MOTO-H and MOTO-P, therefore more large triangle motifs will be present. Although the number of tree-order motifs in the team recognized by MOTO is far less than that of these three comparison algorithms, it does not mean that the team collaboration is not sufficiently represented. We comprehensively evaluated it by combining CCR and other indicators. When the team size differs by 1-3 people, the results of MOTO-H and MOTO-P are higher than TRAC by about 17% and 26%, respectively, which is not enough to exclude the influence of team size. Therefore, the recognition result of MOTO-H, MOTO-P, and TRAC is relatively reasonable in the angle of the triangle motifs number.

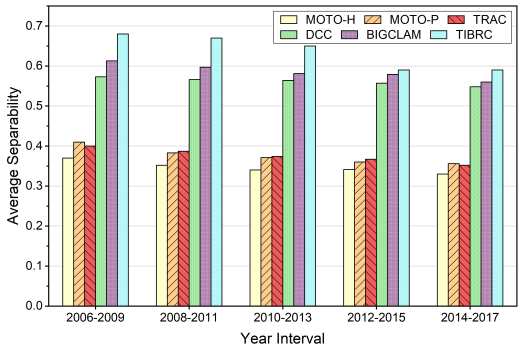

Figure 10: The average separability in time intervals.

Separation degree: Fig.10 shows the average separation degree of each algorithm. The separation degree of DCC, BIGCLAM, and TIBRC are significantly higher than the others, and they are all greater than 0.5. The separation degree greater than 0.5 means that more than half of the team members' connections are connected to nodes outside the team, i.e., the connections inside the team are not tighter than those outside the team, so the recognized team structure does not match the real structure. Comparing MOTO-H, MOTO-P and TRAC, it can be found that the separation degree of MOTO-H is the lowest. The results of MOTO-P and TRAC are similar, but MOTO-P is generally lower.

Based on the above evaluations, MOTO recognizes the largest number of teams. Previous studies have illustrated that real team size varies from 3 to 8 [43] in the computer science discipline. The team sizes are closer to real team sizes, and there are fewer lost network structures. Comparing the results of high-order familiarity and pairwise familiarity, we find that the team that uses high-order familiarity has lower communication costs and a tighter structure. However, considering the higher computational complexity of high-order familiarity, we should select appropriate familiarity according to the actual situation for efficient academic team recognition.

5.6 Analysis of Cooperation Model

This section analyzes the cooperation mode and team performance of academic teams in combination with academic institutions. Academic teams can be divided into interagency teams and intra-institutional teams. According to the indicator evaluation in the previous section, we choose MOTO-H for analysis, because it has the best recognition result.

Firstly, we analyze the trend of scholars' cooperative behavior. Fig.11 shows the number of authors who collaborate on one paper. The number of the legend refers to the number of authors collaborated in one paper. In the legend, the darkest colored segment numbered with "1" refers to the proportion of scholars who authored a paper alone. Likewise, the lighter the color is, the more collaborators in one particular paper are. The specific portion is correspondingly shown on the pie charts. For example, there are totally 39% papers published with individual author during 2006-2009. According to these statistics, more than 60% of the papers are completed cooperatively, and the proportion of co-authored papers is generally higher as time passes. The percentage of cooperation on papers is 3% higher in 2014-2017 compared to 2008-2011. Simultaneously, the number of co-authors also increased over time, which suggests that the team size increases with time in the above recognition results. Secondly, we analyze the proportion of interagency teams. Fig.12 shows the proportion of interagency teams in academic teams with different team sizes in each period. When the team size is more than 20, the proportion of interagency teams is more than 83%. However statistically, the number of teams with size greater than 20 is few and it is therefore difficult to use teams of this size for comparison purposes. Therefore, we select the team size of 2-20 for comparison. When the team size is 2, the proportion of interagency teams is about 30% in all periods, which indicates that when the team has only two people working together, most of them are scholars from the same institution. When the team size is 3, the proportion is about 55%. Teams of 4-8 people accounted for about 79% of the recognition results, and the proportion of interagency teams exceeds 57%. The larger the team size, the higher the proportion of interagency teams. This shows that interagency cooperation has become the main cooperation mode of teamwork. We also can see that when the team size is the same, the proportion of interagency teams has increased significantly over time. In particular, from 2014 to 2017, the proportion of interagency teams was significantly higher than in other periods by more than 10%. When the team size reaches 16 or more, more than 80% of the teams in all time periods involve interagency cooperation.

Finally, we analyze the impact of the increase of interagency academic teams on team performance. The team's performance can be measured by the team's average citations. The calculation formula is:

$$Citation_T = \frac{\sum_{i \in T} citation_i}{n_T} \tag{13}$$

where $citation_i$ is the sum of paper citations of i in a certain period. n_T is the team size.

We sort the teams with the same team size in each period in descending order according to the average number of team references. In order to facilitate comparison, we select the top 20% of teams with a team size of 3-8, as shown in Fig.13 since this selection provided a significant dataset for comparison purposes. When the team size is 3, the performance is relatively low in different time periods. When the team size is 4-6, the proportion of interagency teams and intra-agency teams is basically equal, i.e. their performance is comparable. When the team size is 7-8, the proportion of interagency teams is higher than that of intra-agency teams, i.e., interagency teams perform better than intra-agency teams.

6 CONCLUSION

Collaborative teams are assembled to fill the knowledge gap in academia better. There are a large amount of scientific research problems that demand solutions based on collaborative team work. Multi-variate factors including but not limited to familiarity, ability, team scale, and team composition together have an impact on the output of a team. How to optimize team structure, arrange resources, as well as enhance collaboration, are all fundamental issues that are needed to be solved. Therefore, in the beginning, collaborative

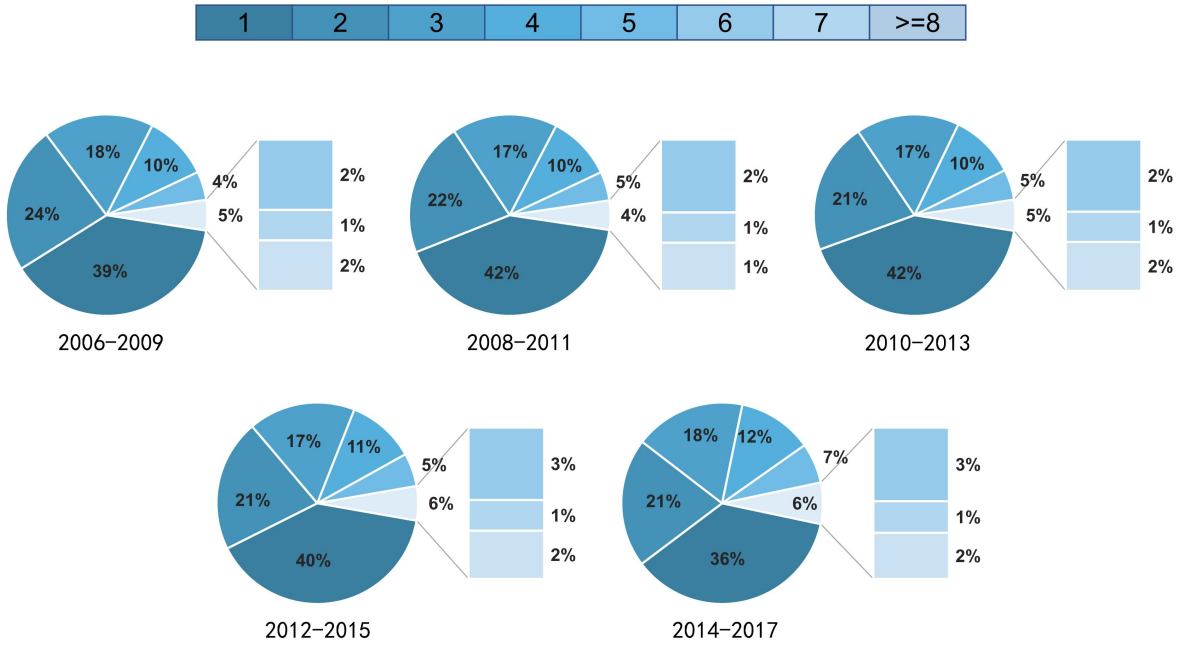

Figure 11: The number of authors who collaborate in one paper.

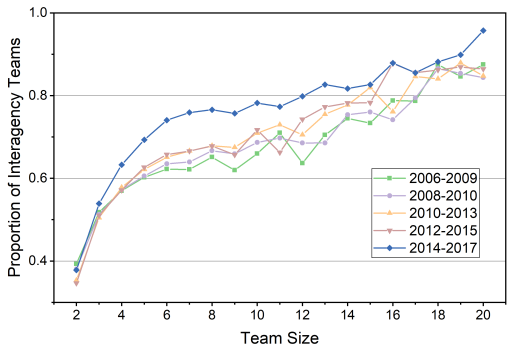

Figure 12: The proportion of interagency teams in academic teams.

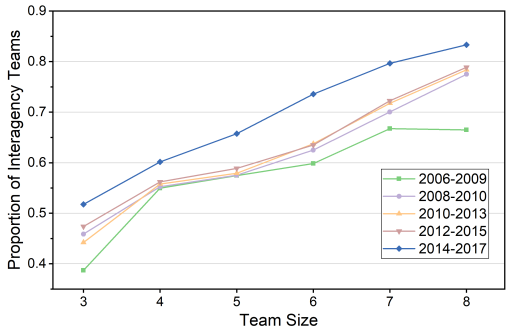

Figure 13: The proportion of interagency teams in academic teams with top-20% cited.

teams should be firstly recognized to support continuous studies. In this work, we employ pairwise familiarity and higher-order familiarity to recognize collaborative teams in academia. Our proposed approach MOTO significantly outperforms baseline methods in a real-world, large-scale network. Teamwork patterns are also analyzed. Teams with members from different institutions widely exist in academia and generally achieve better performance. The number of teams also has an influence on team outputs. Our work provides a way to mine a large number of collaborative teams, which considers both collaboration behaviors and preferences. The proposed method MOTO can also be applied in other disciplines that feature abundant collaboration relationships. Considering the mobility in academia, the definition of familiarity will be optimized based on more data such as subjective consciousness or dynamic collaboration relations in our future work. We will also mine the recognized teams in-depth to identify and investigate new research patterns.

REFERENCES

[1] Nesreen K Ahmed, Jennifer Neville, Ryan A Rossi, and Nick Duffield. 2015. Efficient graphlet counting for large networks. In *2015 IEEE International Conference on Data Mining*. IEEE, 1–10.

[2] Austin R Benson, David F Gleich, and Jure Leskovec. 2016. Higher-order organization of complex networks. *Science* 353, 6295 (2016), 163–166.

[3] Clara Calero, Renald Buter, Cecilia Cabello Valdés, and ED Noyons. 2006. How to identify research groups using publication analysis: an example in the field of nanotechnology. *Scientometrics* 66, 2 (2006), 365–376.

[4] Tanmoy Chakraborty, Sikhar Patranabis, Pawan Goyal, and Animesh Mukherjee. 2015. On the formation of circles in co-authorship networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 109–118.

[5] Yang Chen, Cong Ding, Jiyao Hu, Ruichuan Chen, Pan Hui, and Xiaoming Fu. 2017. Building and analyzing a global co-authorship network using google scholar data. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 1219–1224.

[6] Petr Chunaev. 2020. Community detection in node-attributed social networks: A survey. *Computer Science Review* 37 (2020), 100286.

[7] Nancy J Cooke, Margaret L Hilton, et al. 2015. *Enhancing the effectiveness of team science*. National Academies Press Washington, DC.

[8] Dongqi Fu, Dawei Zhou, and Jingrui He. 2020. Local Motif Clustering on Time-Evolving Graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 390–400.

[9] Weiwei Gu, Jifan Liu, et al. 2019. Exploring small-world network with an elite-clique: Bringing embeddedness theory into the dynamic evolution of a venture capital network. *Social Networks* 57 (2019), 70–81.

[10] Kara L Hall, Amanda L Vogel, Grace C Huang, Katrina J Serrano, Elise L Rice, Sophia P Tsakraklides, and Stephen M Fiore. 2018. The science of team science: A review of the empirical evidence and research gaps on collaboration in science. *American Psychologist* 73, 4 (2018), 532.

[11] Shin-Yuan Hung, Hui-Min Lai, David C Yen, and Chun-Yi Chen. 2017. Exploring the Effects of Team Collaborative Norms and Team Identification on the Quality of Individuals' Knowledge Contribution in Teams. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* 48, 4 (2017), 80–106.

[12] Muhammad Aqib Javed, Muhammad Shahzad Younis, Siddique Latif, Junaid Qadir, and Adeel Baig. 2018. Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications* 108 (2018), 87–111.

[13] Isaac Jones, Ran Wang, Jiawei Han, and Huan Liu. 2016. Community cores: Removing size bias from community detection. In *Tenth International AAAI Conference on Web and Social Media*.

[14] Julio Juárez, Cipriano Santos, and Carlos A Brizuela. 2021. A Comprehensive Review and a Taxonomy Proposal of Team Formation Problems. *ACM Computing Surveys (CSUR)* 54, 7 (2021), 1–33.

[15] Young Ji Kim, David Engel, Anita Williams Woolley, Jeffrey Yu-Ting Lin, Naomi McArthur, and Thomas W Malone. 2017. What makes a strong team? Using collective intelligence to predict team performance in League of Legends. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2316–2329.

[16] Xiangjie Kong, Yajie Shi, Shuo Yu, Jiaying Liu, and Feng Xia. 2019. Academic social networks: Modeling, analysis, mining and applications. *Journal of Network and Computer Applications* 132 (2019), 86–103.

[17] Leah Kulp, Aleksandra Sarcevic, Megan Cheng, Yinan Zheng, and Randall S Burd. 2019. Comparing the effects of paper and digital checklists on team performance in time-critical work. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[18] Chenhao Ma, Reynold Cheng, Laks VS Lakshmanan, Tobias Grubenmann, Yixiang Fang, and Xiaodong Li. 2019. LINC: A motif counting algorithm for uncertain graphs. *Proceedings of the VLDB Endowment* 13, 2 (2019), 155–168.

[19] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. 2004. Superfamilies of evolved and designed networks. *Science* 303, 5663 (2004), 1538–1542.

[20] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. 2002. Network motifs: Simple building blocks of complex networks. *Science* 298, 5594 (2002), 824–827.

[21] Roland Molontay and Marcell Nagy. 2019. Two decades of network science: as seen through the co-authorship network of network scientists. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 578–583.

[22] Ashwin Paranjape, Austin R Benson, and Jure Leskovec. 2017. Motifs in temporal networks. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*. 601–610.

[23] Leonardo Costa Ribeiro, Márcia Siqueira Rapini, Leandro Alves Silva, and Eduardo Motta Albuquerque. 2018. Growth patterns of the network of international collaboration in science. *Scientometrics* 114, 1 (2018), 159–179.

[24] Alex Rodriguez and Alessandro Laio. 2014. Clustering by fast search and find of density peaks. *Science* 344, 6191 (2014), 1492–1496.

[25] Miloš Savić, Mirjana Ivanović, and Bojana Dimić Surla. 2016. A community detection technique for research collaboration networks based on frequent collaborators cores. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. 1090–1095.

[26] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. 2016. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering* 29, 1 (2016), 17–37.

[27] Filipi N Silva, Diego R Amancio, Maria Bardosova, Luciano da F Costa, and Osvaldo N Oliveira Jr. 2016. Using network science and text analytics to produce surveys in a scientific topic. *Journal of Informetrics* 10, 2 (2016), 487–502.

[28] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*. 243–246.

[29] Lovro Šubelj, Dalibor Fiala, Tadej Ciglarič, and Luka Kronegger. 2019. Convexity in scientific collaboration networks. *Journal of Informetrics* 13, 1 (2019), 10–31.

[30] Jurriaan Van Diggelen, Mark Neerincx, Marieke Peeters, and Jan Maarten Schraagen. 2018. Developing effective and resilient human-agent teamwork using team design patterns. *IEEE intelligent systems* 34, 2 (2018), 15–24.

[31] Sara Elena Garza Villarreal and Satu Elisa Schaeffer. 2016. Local bilateral clustering for identifying research topics and groups from bibliographical data. *Knowledge and Information Systems* 48, 1 (2016), 179–199.

[32] Wei Wang, Jing Ren, Mubarak Alrashoud, Feng Xia, Mengyi Mao, and Amr Tolba. 2020. Early-stage reciprocity in sustainable scientific collaboration. *Journal of Informetrics* 14, 3 (2020), 101041.

[33] Xinyu Wang, Zhou Zhao, and Wilfred Ng. 2016. USTF: a unified system of team formation. *IEEE Transactions on Big Data* 2, 1 (2016), 70–84.

[34] Feng Xia, Ke Sun, Shuo Yu, Abdul Aziz, Liangtian Wan, Shirui Pan, and Huan Liu. 2021. Graph Learning: A Survey. *IEEE Transactions on Artificial Intelligence* 2 (2021), 109–127. https://doi.org/10.1109/TAI.2021.3076021

[35] Feng Xia, Haoran Wei, Shuo Yu, Da Zhang, and Bo Xu. 2019. A survey of measures for network motifs. *IEEE Access* 7 (2019), 106576–106587.

[36] Feng Xia, Shuo Yu, Chengfei Liu, Jianxin Li, and Ivan Lee. 2021. CHIEF: Clustering with Higher-order Motifs in Big Networks. *IEEE Transactions on Network Science and Engineering* (2021). https://doi.org/10.1109/TNSE.2021.3108974

[37] Jin Xu, Shuo Yu, Ke Sun, Jing Ren, Ivan Lee, Shirui Pan, and Feng Xia. 2020. Multivariate relations aggregation learning in social networks. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. 77–86.

[38] Jaewon Yang and Jure Leskovec. 2013. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 587–596.

[39] Shuo Yu, Hayat Dino Bedru, Ivan Lee, and Feng Xia. 2019. Science of scientific team science: A survey. *Computer Science Review* 31 (2019), 72–83.

[40] Shuo Yu, Yufan Feng, Da Zhang, Hayat Dino Bedru, Bo Xu, and Feng Xia. 2020. Motif discovery in networks: A survey. *Computer Science Review* 37 (2020), 100267.

[41] Shuo Yu, Jiaying Liu, Feng Xia, Haoran Wei, and Hanghang Tong. 2020. How to optimize an academic team when the outlier member is leaving? *IEEE Annals of the History of Computing* 01 (2020), 1–1.

[42] Shuo Yu, Feng Xia, and Huan Liu. 2019. Academic Team Formulation Based on Liebig's Barrel: Discovery of Anticask Effect. *IEEE Transactions on Computational Social Systems* 6, 5 (2019), 1083–1094. https://doi.org/10.1109/TCSS.2019.2913460

[43] Shuo Yu, Feng Xia, and Huan Liu. 2019. Academic team formulation based on Liebig's barrel: Discovery of anticask effect. *IEEE Transactions on Computational Social Systems* 6, 5 (2019), 1083–1094.

[44] Shuo Yu, Feng Xia, Kaiyuan Zhang, Zhaolong Ning, Jiaofei Zhong, and Chengfei Liu. 2017. Team recognition in big scholarly data: Exploring collaboration intensity. In *2017 IEEE 3rd Intl Conf on Big Data Intelligence and Computing (DataCom)*. IEEE, 925–932.

[45] Shuo Yu, Jin Xu, Chen Zhang, Feng Xia, Zafer Almakhadmeh, and Amr Tolba. 2019. Motifs in Big Networks: Methods and Applications. *IEEE Access* 7 (2019), 183322–183338.

[46] Yongsheng Yu, Cheng Dong, Hongqi Han, and Zhong Li. 2018. The Method of Research Teams Identification Based on Social Network Analysis: Identifying Research Team Leaders Based on Iterative Betweenness Centrality Rank Method. *Information studies: Theory & Application* 41, 7 (2018), 105–110.

[47] Fang Zhang and Shengli Wu. 2018. Ranking scientific papers and venues in heterogeneous academic networks by mutual reinforcement. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. 127–130.

[48] Zhihong Zhang, Dongdong Chen, Lu Bai, Jianjia Wang, and Edwin R Hancock. 2020. Graph Motif Entropy for Understanding Time-Evolving Networks. *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[49] Huan Zhao, Xiaogang Xu, Yangqiu Song, Dik Lun Lee, Zhao Chen, and Han Gao. 2019. Ranking Users in Social Networks with Motif-based PageRank. *IEEE Transactions on Knowledge and Data Engineering* (2019).