IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 13, 2020, accepted October 22, 2020, date of publication October 30, 2020, date of current version November 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3034997

TOSNet: A Topic-Based Optimal Subnetwork Identification in Academic Networks

HAYAT D. BEDRU[1], WENHONG ZHAO[2], (Member, IEEE), MUBARAK ALRASHOUD[3], AMR TOLBA[4,5], HE GUO[1], AND FENG XIA[6], (Senior Member, IEEE)

[1]School of Software, Dalian University of Technology, Dalian 116620, China
[2]Ultraprecison Machining Center, Zhejiang University of Technology, Hangzhou 310014, China
[3]Software Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 11437, Saudi Arabia
[4]Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia
[5]Mathematics and Computer Science Department, Faculty of Science, Menoufia University, Shebin-El-kom 32511, Egypt
[6]School of Engineering, IT and Physical Sciences, Federation University Australia, Ballarat, VIC 3353, Australia

Corresponding author: Wenhong Zhao (whzhao6666@outlook.com)

This work was supported in part by the International Scientific Partnership Program (ISPP) at King Saud University under Grant ISPP-78, and in part by the Zhejiang Provincial Fundamental Public Welfare Research Program under Grant LGG18E050025.

ABSTRACT Subnetwork identification plays a significant role in analyzing, managing, and comprehending the structure and functions in big networks. Numerous approaches have been proposed to solve the problem of subnetwork identification as well as community detection. Most of the methods focus on detecting communities by considering node attributes, edge information, or both. This study focuses on discovering subnetworks containing researchers with similar or related areas of interest or research topics. A topic-aware subnetwork identification is essential to discover potential researchers on particular research topics and provide quality work. Thus, we propose a topic-based optimal subnetwork identification approach (TOSNet). Based on some fundamental characteristics, this paper addresses the following problems: 1)How to discover topic-based subnetworks with a vigorous collaboration intensity? 2) How to rank the discovered subnetworks and single out one optimal subnetwork? We evaluate the performance of the proposed method against baseline methods by adopting the modularity measure, assess the accuracy based on the size of the identified subnetworks, and check the scalability for different sizes of benchmark networks. The experimental findings indicate that our approach shows excellent performance in identifying contextual subnetworks that maintain intensive collaboration amongst researchers for a particular research topic.

INDEX TERMS Academic social networks, collaboration intensity, network science, subnetwork identification, subnetwork ranking, topic modeling.

I. INTRODUCTION

Due to the fast-growing volume and variety of online scholarly data, researchers have shown tremendous interest in producing numerous techniques and applications to explore and analyze academic data. The extraction of academic networks aims at supplying comprehensive applications in the scientific research area. In an academic network, users are interested in searching for various information, such as researchers' profiles (e.g., h-index, number of co-authors, and citation counts), conferences, journals, venues, and publications [1]. Several problems in academic networks have been explored, and numerous systems have been developed, including CiteSeerX [2], DBLP,[1] Microsoft Academic Graph,[2] Google Scholar,[3] and ArnetMiner.[4] Since academic networks are bigger and uphold a highly temporal characteristic, analyzing them causes significant challenges to data mining techniques. Managing a big network is challenging; thus, dividing the big network into subnetworks provides significant insights towards its function and structure. Community structure detection is one of the fundamental applications that help analyze a big network [3], [4]. It helps to solve the discovery of subnetworks of nodes densely linked to one another than the rest of the network [5]. In recent years, there has

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Salehzadeh-Yazdi.

[1] https://dblp.org/
[2] http://research.microsoft.com/en-us/projects/mag/
[3] https://scholar.google.com/
[4] http://www.arnetminer.org

been growing attention in employing community detection on social networks to comprehend the network structure and exploit its outcomes extensively in different applications and intelligent systems [6].

Meanwhile, another approach taken as a crucial aspect of exploiting academic data is identifying influential nodes — the node can be an author, a paper, or a scientific team. For instance, the identification of influential authors intends to identify potential collaborators for particular research work.

Many existing studies have investigated specific aspects of community detection as well as subnetwork discovery [7]–[11]. However, despite the exerted efforts, existing approaches do not consider contextual limitations. Thus, they lack the feature of finding appropriate collaborators for a specific research topic. For instance, if a scholar wants to gather collaborators who have experiences in ‘‘Big Scholarly Data’’, there is no support in the existing methods that could help identify appropriate candidates for a given specific research topic. In essence, scholars would limit their research topics when they plan to look for collaborators. Thus, it is essential to build topic-oriented applications to identify authors as well as scientific teams.

The dearth of previous studies that could systematically handle community detection or subnetwork discovery concerning academic network applications has driven the present work. This paper studies the problem of discovering a subnetwork containing a collection of related scholars embedded in an extensive academic network. The discovered subnetworks contain relationships that are significant in the network. The relationships mainly depend on the research topic of scholars. Moreover, the authors in the subnetworks are supposed to have intensive collaboration with each other. To analyze the authors’ collaboration intensity level in a co-authorship network, we employ an index metric called Collaboration Intensity Index (CII) [12] that enables us to discover authors with an intense relationship. Identifying subnetworks enables us to uncover the complex structures of a big network by mining the interactions formed based on specific conditions or context. For instance, we expect authors within a subnetwork to have relatively similar research areas of interest. The splitting of networks in either slightly interconnected or disconnected subnetworks is essential when considering a scientific team’s formation in a specific research area.

We summarize the specific contributions of this paper as follows:

- Proposed a new method named TOSNet (Topic-based optimal subnetwork), which efficiently identifies an optimal subnetwork, which exists in big academic networks considering the research topic of authors.
- Proposed a method that employs the Latent Dirichlet Allocation method and Collaboration Intensity Index metric to discover more suitable and contextual subnetworks.
- Introduced subnetworks ranking metric, called $gnet - index$, which allows us to discover the optimal subnetwork from the extracted subnetworks.
- Detailed assessment demonstrating enhanced subnetwork detection prominence and accuracy in terms of a research topic.

This paper is organized as follows. Section II provides related works. Section III provides the problem formulation followed by the TOSNet method in Section IV containing explanations of the employed techniques and the subnetworks identification approach. Section V presents experiment settings, including dataset, baseline methods, and evaluation objectives, as well as the detailed analysis of the experimental results. In Section VI, we discuss some of the experimental results and findings. Finally, we provide the conclusion and future directions of our work in Section VII.

II. RELATED WORKS

In network science, discovering communities is one of the most crucial tasks. Communities provide a way to identify sets of connecting nodes and the interactions between them [4]. A ‘‘community’’ is a set of entities with dense connections amongst themselves but sparse connections between entities in different sets [13]. The term ‘‘community’’ has different interpretations in different disciplines. In the context of academic networks, communities might denote the grouping of entities such as authors, papers, institutions, and venues. Whereas in biological networks, a community can be a group of proteins and their interactions [14]. In academic social networks, terminologies like scientific teams/research teams, scientific collaborations, and subnetworks can be considered synonymous with the term ‘‘community’’ according to a research problem targeted to solve. In our work, a ‘‘community’’ refers to a subnetwork that is a collection of scientific collaborators or authors who have ‘‘previously’’ published one or more research works together, and they are included in the CiteSeerX [2]. Dividing a big network into smaller subnetworks provides a better method to comprehend and envision the type of interactions when some unknown features of networks exist. Numerous research has been conducted to solve the complexity of networks by proposing methods that detect the community structure in a network. Analyzing the community structure of academic networks is of particular interest due to the high volume of data and a network’s complicated inner structure. For instance, assume we have a small co-authorship network containing 100 authors and 50 papers and aim to detect communities. Each community contains a set of authors linked through co-authorship, i.e., two authors are linked if they have written one or more papers together. Thus, we can construct less than or equal to 50 number of communities from this given network. If authors A, B, and C co-authored five papers together, we do not need to construct five different communities. Instead, we will have a single community that includes authors A, B, and C. Thus, managing and analyzing the detected communities is much easier than managing the whole network at once. Besides, it reduces the computational time.

Most of the existing community detection methods have experimented on biological networks (i.e., Protein-Protein-Interaction networks [14]), social networks (i.e., communities in a Facebook and a community as similar posts in Twitter), citation networks, and co-authorship networks [10]. This section discusses approaches that considered node attributes, edge information, or both during detecting appropriate communities in a given network. In academic networks, node attributes may be the author's publications, citation counts, h-index, etc. Similarly, edge information refers to different types of connections amongst nodes in a network, for instance, a relationship between papers through citations and authorship collaborations between researchers.

A. COMMUNITY IDENTIFICATION METHODS

Newman and Girvan [15] proposed an algorithm that extracts the structure of a network as communities. The authors used *edge betweenness* measures to discover the edges that need to be discarded from the communities. They defined *edge betweenness* as the number of shortest paths between a particular pair of nodes that comprise the edge. The method iteratively assesses *edge betweenness* until it is necessary to do so. Newman and Girvan discussed that the iteration is a critical step to discover communities successfully. The method proposed in [15] has two limitations: 1) the attributes of the nodes are not taken into considerations, and 2) the method might identify irrelevant communities wrongly. We argue that considering the characteristics of each node as well as each edge is crucial so that we may find nodes in a community having an intense relationship among themselves with similar features. Moreover, as described by Bhatt *et al.* [16], one can get a full meaning of the structures of communities in a complex network as far as both network topology and node attributes are considered in the process of community detection.

Numerous related works also focused on topic-modeling-based community detection methods. For instance, Le and Lauw [17] proposed a topic-based method called "Probabilistic LAtent Document Network Embedding (PLANE)" that integrates low-dimensional and topic distribution representations. The PLANE method employed *k-means* algorithm to the community detection process out of the embedded nodes. Bhatt *et al.* [16] proposed a context-oriented community detection method by applying a weighted knowledge graph. Bhatt *et al.* employed the Louvain community detection algorithm [18] and introduced a contextual similarity assessment method for describing node pair similarities to apprehend community contexts. The method iteratively updates the labels assigned to communities by community-context. Subsequently, it computes a context that better explains the nodes of each community. Besides, Bhatt *et al.* presented two main concerns while detecting contextual communities, such as "informativeness and purity." The former is "the specificity of a concept in a hierarchical knowledge graph." The latter is "the difference between the number of nodes subsumed by a concept of a given community and neighboring communities." In another work, Ma *et al.* [19] presented a network model that comprises two-layer such as collaboration network and paper similarity network layers. The latter is used to detect communities based on the similarity between each scholar's research topics in the collaboration network. According to the discussion in [19], the number of communities is equivalent to the number of research topics available in the network.

He *et al.* [20] proposed a model considering the detection of communities in a network as well as extracting their semantics simultaneously. Besides, the authors combined "a nested expectation-maximization and a belief propagation" algorithms and developed a learning process model that reveals subtle correlations between community detection and semantic extractions. The method proposed by He *et al.* identifies communities for particular semantics; however, in some cases, the method identifies more than one semantic in a community. One of the limitations of this method is that it requires a given number of communities to be detected *a priori*. In addition, Li *et al.* [21] proposed an author-topic-community (ATC) model on the basis of author-topic [22] and topic-link-LDA models [23]. ATC precisely models author profiles in the form of topic-author distribution and author community structure. According to Li *et al.* [21], authors grouped in a community if they write papers with joint research topics. Li *et al.* assumed that the presence of connections between authors depends on the similarity between the research interests of the authors as well as the community structure of the author in a particular academic social network.

Guedes *et al.* [24] proposed an algorithm called "Ranking Multiple Clustering Algorithm in Attributed Graphs (RM-CRAG)" for grouping graphs with attributes. For a given k value by the user, RM-CRAG generates the top-k groupings (possibly overlapping), in which these groupings are distinct from each other (not redundant). RM-CRAG does not handle edge attributes, unlike the TOSNet method. Besides, Sachan *et al.* [25] proposed a probabilistic model for the detection of communities, which mixes the relationship between vertices, the type of interaction, and information exchanged with members of other communities. This proposal does not detect overlapping communities. Besides, the user needs to inform previously existing communities and the number of topics (subjects) to be considered by the model, which is not always feasible in practice. In [26], the authors proposed a community detection algorithm called "Adjacent node Similarity Optimization Combination Connectivity Algorithm (ASOCCA)", that considers the similarities amongst nodes by measuring the clustering coefficient of each node in the given network. Similar to the previous work, their algorithm is mainly designed to detect non-overlapping communities in an unweighted and undirected network. ASOCCA constructs communities by dividing nodes with a higher degree of similarity. It also merges small size communities to their highly similar group of nodes. One of the limitations of ASOCCA is that it does not analyze the actual characteristics of the nodes; instead, it focuses on discovering neighboring nodes with a higher interaction or clustering coefficient value. Moreover, ignoring to investigate the characteristics of edges in a network is an essential limitation of existing community detection methods as many networks contain attributes linked to the edges that may be relevant in the community detection process.

B. COMMUNITY RANKING METHODS

In the process of detecting communities, it is also equally important to assess the quality of the communities identified. Therefore, modularity [15], normalized mutual information (NMI), and other similar techniques [27] have been designed to evaluate the detected communities. However, these models only measure the quality of each detected community. They do not include the ranking of communities, and it is scarce to find methods that can efficiently compare and output optimal communities. One of the few community ranking methods proposed recently is the one introduced by Li *et al.* [28], which is based on a model "k-influential community" that can capture an influential community in a network by adopting the idea of k-core. In another work, Lei and Wei [29] attempted to propose a method that detects an influential community. The method is a hierarchical agglomerative algorithm that detects communities from a complex network. The algorithm is initiated with the intuition that every community is a single node. A new re-normalization network (as it is named in [29]) is constructed by using a re-normalization technique. Also, Lei and Wei adopted a metric called "State of Critical Functionality (SCF)" to identify communities. When the SCF value of a node is bigger, the influence of its community becomes smaller. During the process of identifying an influential community, a single node is discarded from a network based on the node's SCF value. There are also more related works [30]–[32] that attempted to prevent the lack of community ranking applications.

Nonetheless, these methods do not identify influential communities for different domains. Moreover, they are computationally expensive to implement in a big network, a large-scale network with complicated inner structures. Due to these limitations, we are motivated to propose a topic-driven subnetwork identification algorithm and a mathematical model that can efficiently evaluate and rank subnetworks detected from a big network.

III. PROBLEM FORMULATION

We consider a heterogeneous authorship network containing authors and their corresponding papers/publications, that is denoted by $R = (A, P, L)$. Specifically, R consists of A, which is a set of authors, P, a set of publications, and L, a set of links ($L \subseteq (AP)$) between authors and papers, i.e., authorship links. A link $l = (a, p) \in L$, connects author a with a paper p. An author can have multiple papers, and a paper can be authored by one or more authors. We first apply Latent Dirichlet Allocation (LDA), a popular topic modeling algorithm, to the original dataset and extract authors by taking into account specific research topics. Consequently, we construct a new weighted-network using the extracted authors called an Author-Topic Network (ATN).

ATN: ATN is an edge-weighted network containing authors and their corresponding topics extracted using LDA from the network R, denoted by $ATN = (A, T, E, w)$. A is a set of authors, T is a set of topics, and $E \subseteq (A \times T)$ is a set of edges indicating the specific research topics of authors. The weight of edges is represented by w, which holds the topic probability distribution of each author in each topic. Note that we get the topics in T and w using LDA. Constructing this network aims to identify influential authors with a higher probability distribution on a particular topic. We first find the candidate authors for a specific topic and construct a weighted co-authorship network that will be used for the next step in our proposed approach.

A co-authorship network: This is an edge-weighted network denoted as $G = (A, E)$, where A is a set of authors, and E is a set of edges. An edge is a connection between authors if they co-author one or more papers on a specific research topic. We quantify the strength of the collaboration relationship of two authors by employing Collaboration Intensity Index (CII), which is considered as an edge weight [12]. Subsequently, we identify subnetworks from the newly created co-authorship network.

A subnetwork: A subnetwork is a small set of authors extracted from the network $G(A, E)$, which is denoted as $S = (A', E')$, where $A' \in A$ and $E' \in E$, $E' \subseteq (A' \times A') \cap E$. A' in S is a set of authors with a similar research topic t.

As the main objective of this work is to rank the subnetworks and identify the most prominent one, we propose a new formulation for the optimal subnetwork creation as follows.

Optimal subnetwork Let C be a collection of n number of subnetworks built over the same set of authors. $C = \{S_1, S_2, \ldots, S_n\}$, in which $S_i = (A'_i, E_i)$, $i = \{1, 2, \ldots, n\}$. Thus, the subnetworks preserve a positive parameter, i.e., $gnet - index$, which is proposed to rank the subnetworks. Therefore, the subnetwork with higher $gnet - index$ considered as an optimal subnet. In our work, an optimal subnet is assumed to be a scientific team formed based on a particular research topic. The overall architecture of our proposed approach (TOSNet) is presented in Fig. 1.

IV. THE TOSNet METHOD

Herein, we give a detailed description of the process of TOSNet, including the adopted methods as well as the proposed algorithm.

A. TOPIC MODELING AND COLLABORATION INTENSITY MEASURE

For our optimal subnetwork identification approach, we first have to identify topics and collaboration networks. In the following subsections, we explain the methods we have employed to extract topics and collaborators from the CiteSeerX [2] dataset.

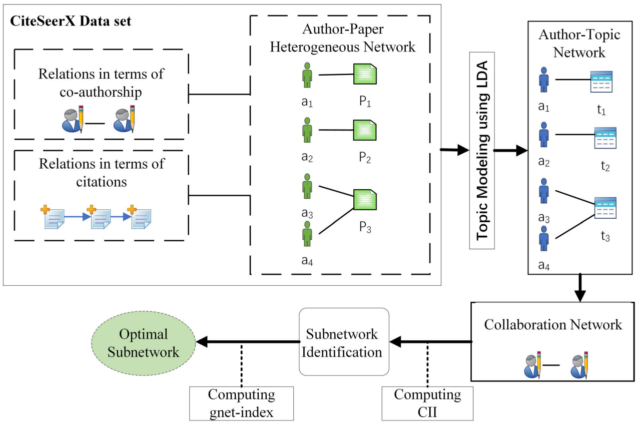


FIGURE 1. Structural Overview of TOSNet.

1) TOPIC EXTRACTION USING LDA

To correctly identify topic-based subnetworks from a big network, we opt to use topic extraction methods to organize networks based on topic relationships. For our experiment, we have examined a large set of unstructured data compiled from research papers. To extract research topics from this dataset, we have devised a mechanism to comprehend, organize, manage, and label this data accordingly. Topic modeling is applied in several scenarios like news article selection, searching queries on Quora related to each other, and recommendation methods, to mention some. In this paper, to address the issue of research topic extraction, we employ a probabilistic modeling approach known as Latent Dirichlet Allocation (LDA). We employ LDA to gather hidden variables and choose the total number of topics we want to discover in a given document collection. Giving the total number of topics *a priori* helps to select sensible and on-point research topics from the discovered ones. Hence, we could form subnetworks that contain authors with related research topics.

Furthermore, LDA can discover a set of topics from vast collections in an unsupervised way. LDA represents each topic as a probability distribution over words and each document as a distribution over topics. For our work, we assume that each document represents a collection of abstracts of a single author. For example, if an author X has published five papers, we merge the abstract of each paper and put them as one document. Therefore, the number of documents is equivalent to the number of authors in the dataset. So, for each author, we will have a topic vector t_a, which represents the probability distribution of an author over topics K. For example, t_a^i denotes the probability distribution of papers that the author has written related to topic i, $i = \{1, 2, \ldots, K\}$. We compute a corpus perplexity [33] to determine the appropriate number of topics to train the LDA model. The perplexity demonstrates how well the model characterizes a set of available documents. The lower the perplexity a certain number of topics preserves, the better fit it would be to train the model. The main objective of computing perplexity is to determine k that reduces the perplexity compared to other k values. The perplexity is mathematically computed as depicted in Eq. (1) [33].

$$perplexity(\mathcal{D}) = \exp\left\{-\frac{\sum_{d=1}^{\mathcal{M}} \log p\left(\mathbf{w}_d\right)}{\sum_{d=1}^{\mathcal{M}} \mathcal{N}_d}\right\} \tag{1}$$

where $\mathcal{D}$ is a corpus with a collection of $\mathcal{M}$ number of documents, d indicates a document, $\mathbf{w}_d$ is a word in a document d, and $\mathcal{N}_d$ represents the number of words in each document d.

Furthermore, computing the topic probability distribution (TDP) helps determine how experienced and skillful the researchers are on the given topics and analyze how many of their works are done on similar topics.

2) CO-AUTHORS RELATIONSHIP INTENSITY

One of the metrics that quantify the relationship between authors is the number of collaborations they have. It can be assumed that two authors have strong collaboration if they have more papers together than the rest of the network. Moreover, when two or more authors have more collaborations, their mutual understanding also gets more concrete. This enables them to produce effective scientific research works as well as increases their productivity [34]. However, investigating and analyzing the intensity degree of collaborators, considering only the collaboration frequency, might not give a convincing result. Besides, it might also lack to identify accurate influential authors. Thus, to compute the authors' intensity relationship degree, we employ an effective metric called Collaboration Intensity Index (*CII*). *CII* is a time-oriented metric that assesses the collaboration strength of connected authors by taking into account collaboration frequency that occurred between two authors and the number of papers that they have published individually in a certain period. The *CII* of two authors x and y who have collaboration relations from year t_1 to year t_2 can be mathematically computed as shown in (2).

$$CII = \frac{\Delta_{t_2-t_1} k_{xy}^2}{\Delta_{t_2-t_1} k_x \Delta_{t_2-t_1} k_y} \tag{2}$$

where $\Delta_{t_2-t_1} k_{xy}$ depicts the number of publications authors x and y have together. $\Delta_{t_2-t_1} k_x$ and $\Delta_{t_2-t_1} k_y$ depict the number of publications authors x and y individually published from year t_1 to t_2, respectively.

B. SUBNETWORK IDENTIFICATION APPROACH

1) COLLABORATION SUBNETWORKS

After constructing a co-authorship network $G(V, E)$, we compute the *CII* of each edge and assign it as a weight of the connection between two authors. To construct a subnetwork, we select authors whose *edge_weight* has a *CII* value greater than or equal to the average CII (*AvgCII*), i.e., if two authors have $CII \geq AvgCII$ value, then they form a subnetwork. The size of a subnetwork increases depending on the frequency of the collaboration of authors.

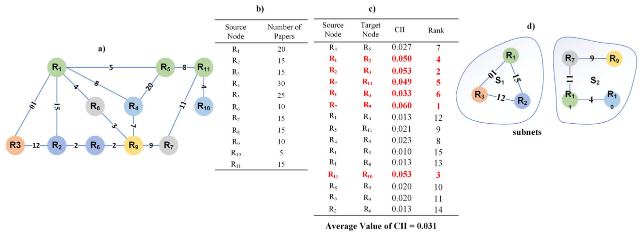
2) IDENTIFYING AN OPTIMAL SUBNETWORK

After finding the subnetworks of the given network, the next step is discovering a prominent subnetwork called an optimal subnet. An optimal subnet is a core subnetwork consists of nodes (i.e., authors) with higher and intense collaboration within a subnetwork. Whether or not there is a connection between subnetworks is not an issue for our case; instead, higher and intense collaboration within a subnetwork is the primary concern for identifying an optimal subnetwork.

Each subnetwork consists of authors and collaborations among them as nodes and edges, respectively. The collaboration between two researchers labeled as e_i, where $i = \{1, 2, \ldots., n\}$. Hence, to identify the optimal subnet, we consider the total number of collaborations in each subnetwork and edge weight. The proposed metric that is used to measure the weight of each subnetwork is called $gnet-index$. $gnet-index$ holds a value that is used to measure and rank the subnetworks. The optimal subnet is the one that has a larger number of collaborations as well as an edge weight greater than or equal to $all\_avg\_snet\_edge\_weight$. The $gnet-index$ of a subnetwork is calculated as:

$$gnet-index = \sum_{i=1}^{n} e_i \tag{3}$$

where e_i composes edge i labeled by the number of times two nodes get connected.



Source Node	Number of Papers
R_1	20
R_2	15
R_3	15
R_4	30
R_5	25
R_6	10
R_7	15
R_8	15
R_9	10
R_{10}	5
R_{11}	15

Source Node	Target Node	CII	Rank
R_4	R_5	0.027	7
R_1	**R_2**	**0.050**	**4**
R_2	**R_3**	**0.053**	**2**
R_7	**R_{11}**	**0.049**	**5**
R_1	**R_3**	**0.033**	**6**
R_7	**R_9**	**0.060**	**1**
R_1	R_4	0.013	12
R_5	R_{11}	0.021	9
R_4	R_9	0.023	8
R_1	R_5	0.010	15
R_1	R_8	0.013	13
R_{11}	**R_{10}**	**0.053**	**3**
R_8	R_9	0.020	10
R_6	R_9	0.020	11
R_2	R_6	0.013	14

Average Value of CII = 0.031

FIGURE 2. The connection between authors and their CII value. a) a toy example of a co-authorship network. b) the number of paper(s) each author has in the given network. c) the connection between authors and their CII value. The average CII value for the given network is 0.031. d) subnetworks extracted from the toy network.

In Fig. 2(d), $s1$ and $s2$ have $edge\_weight \geq all\_avg\_snet\_edge\_weight$, in which $s1$ has three edges (i.e., $e_1 = 15, e_2 = 10, e_3 = 12$) with a $gnet - index$ value of 37 and $s2$ also has three edges (i.e., $e_1 = 11, e_2 = 9$, $e_3 = 4$) with a $gnet-index$ value of 24, therefore, the optimal core-subnet of the given network is $s1$. Algorithm 1 illustrates how the TOSNet method detects subnetworks and identify an optimal subnetwork in a big academic network. As shown in Algorithm 1, initially, authors with papers on specific topics are searched, with a linear run time, which is $O(n)$. Then, construction of a co-authorship network takes $O(n^2)$, where n is the number of collaborations between authors. Afterwards, collaboration intensity at each pair of nodes is calculated, that takes $O(n)$. Finally subnetworks detection takes $O(n)$. The total complexity of Algorithm 1 is $O(n^2)$.

Algorithm 1: Subnetworks Identification

```
Input: An Author-Topic Network $ATN(A, T, E, w)$,
       a topic $t \in T$, and $k$, i.e., the required number of
       subnetworks
Output: Top-$k$ subnetworks with their $gnet - index$
        value
Identify list of authors as $A$ from ATN who have paper
  on topic $t$
Construct a co-authorship network $G(A, E)$ // $A$
  (authors) $= \{a_1, \ldots, a_n\}$, $E$ (edges) $= \{e_1, \ldots, e_m\}$
for $e = 1$ to $m$ do
    Compute collaboration intensity index ($CII$) // using
      Eq. (2)
    $edge\_weight[m - 1] = CII$ value of each edge
    Compute the average $CII$,
      $AvgCII = \frac{\sum_j^{m-1} edge\_weight[j]}{m}$
if $CII_{e_i} \geq AvgCII$ then
    Insert authors linked by $e_i$ into $candidate\_authors$
      list
    $i++$
Construct subnetworks $S$ for $candidate\_authors$,
  $S = \{s_1, s_2, \ldots, s_n\}$
for $j = 1$ to $n$ do
    Compute average $edge\_weight$
      ($avg\_snet\_edge\_weight$)
    Compute the average of $avg\_snet\_edge\_weight$
      ($all\_snet\_avg\_edge\_weight$ of all subnetworks)
    if $avg\_snet\_edge\_weight\_s_j \geq$
      $all\_snet\_avg\_edge\_weight$ then
        Compute the $gnet - index$ of the subnetwork //
          using Eq. (3)
        $j++$
return top-$k$ subnetworks
```

V. EXPERIMENT

Herein, we present the details of how we conduct the experiments to validate the effectiveness and efficiency of our proposed approach. Subsequently, we discuss the experimental settings, including dataset, evaluation objectives, metrics, and analysis of the results.

A. EXPERIMENTAL SETTINGS

1) DATASET

A big dataset from CiteSeerX [2] is adopted in our experiment, which provides adequate training samples to validate our proposed approach. Specifically, a total of 4,625,758 research papers are collected from 2000–2019. After a necessary cleaning and preprocessing (i.e., removing solo-authored papers, stemming, lemmatization, and removing stop words) steps, a total of 1,699,965 abstracts are selected, each of which is then treated as a unique document. Subsequently, a total of 2,783,765 scholars who have partici-

TABLE 1. CiteSeerX Dataset Information After Data Preprocessing.

Parameters	Size
Number of Authors	2,783,764
Number of Edges	13,216,679
Average # of Papers per Author	3
Average # of Co-authors per Paper	4
Minimum Co-author in the Datases	2
Maximum # of Co-author in the Dataset	141

pated in the selected papers are identified for the experiment (See Table 1).

2) BASELINE METHODS

To evaluate TOSNet's performance in discovering key subnetworks, we employed the following representative baseline methods for comparison. The baseline methods have relatively similar objectives with our proposed method. They are mainly designed to discover core teams in a social network.

a: CORE SUBNET ABSTRACTING METHOD

Zhao *et al.* [35] introduced the subnetworks identification approach specifically for weighted networks by taking into account the strength of links in the network. The authors proposed a network metric called "h-strength", which measures a weighted network considering "link strength". "The h-strength of a network is equal to h_s, if h_s is the largest natural number such that there are h_S links each with strength at least equal to h_s in the network." The metric "h-strength" mainly abstracts the links' strength in the given weighted network as well as uncovers the core structure of a network [35]. If pairs of nodes have strong interactions and paths, they will have a high "h-strength". In addition, Zhao *et al.* [35] proposed a core subnetwork structure identification metric called "h-subnet", which is defined as "a subnetwork that includes the links whose strengths are larger than or equal to the h-strength of the network, and the nodes that are connected by these links." The core subnet contains links with high "h-strength" and nodes linked by those links [35].

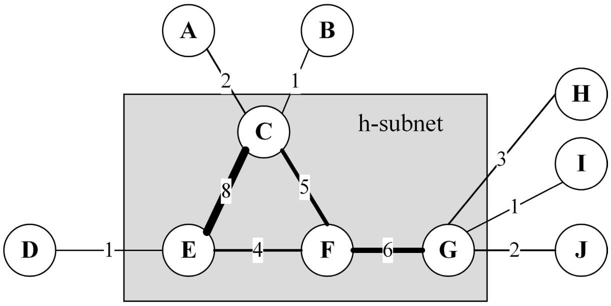


FIGURE 3. A toy network that contains its "h-strength" and the core subnet. The h-strength of this figure is 4, i.e., there are 4 strength of links greater than or equal to 4 [35]. The shaded rectangle depicts the "h-subnet" abstracted from the toy network. Table 2 shows the links, their strengths in a descending order, as well as the h_s of the network.

TABLE 2. List of edges in a descending order of their link strengths, extracted from the network in Fig. 3 [35]. The h_s of the network is highlighted in bold.

Source	Target	Link strength
C	E	8
F	G	6
C	F	5
E	F	**4**
G	H	3
A	C	2
G	J	2
B	C	1
D	E	1
G	I	1

b: H-BACKBONE AS CORE SUBNETWORK

Zhang *et al.* [36] introduced an approach to extract the core structure of weighted networks. They have proposed three metrics considering the edge betweenness of a network such as "bridge", "h-bridge", and "h-backbone". The first step to obtaining the "h-bridge" of a given weighted network is by calculating each edge's edge betweenness in the network. The "bridge" value of each edge is calculated by dividing the edge betweenness of an edge with the number of nodes in the network. Zhang *et al.* [36] defined "h-bridge" as "h-bridge (h_b) of a network is equal to h_b, if h_b is the largest natural number such that there are h_b links, each with bridge at least equal to h_b in the network". The metric "h-bridge" is used to rank the "bridge" of all edges. To identify the core subnetwork of a given network, "h-strength" [35] is considered along with "h-bridge". Hence, Zhang *et al.* [36] defined "h-backbone" as "a core subnetwork consisting of all edges with strengths larger than equal to the h-bridge or the h-strength in the network, together with their adjacent nodes".

3) EVALUATION OBJECTIVES

Experiments are performed for the following main objectives:

1) Topic-oriented collaborator identification, which evaluates the effectiveness of the proposed method on identifying researchers for a particular research topic;
2) Collaboration intensity, which assesses the level of intensity the candidate researchers have to each other in terms of collaboration;
3) Accuracy of key subnetwork identification, in terms of identifying collaborators from similar research areas of interest in a subnetwork.

For the first two objectives, LDA and CII are adopted to make the proposed method effective regarding identifying topic-based subnetworks and strengthen collaborations amongst authors in the identified optimal subnetwork, respectively.

B. RESULTS AND ANALYSIS

In this section, we present a detailed analysis of the experimental results and discuss the evaluation of our method in comparison with the baseline methods.

1) ANALYSIS OF THE TOSNet METHOD

a: THE PATTERN OF THE SUBNETWORKS

We trained the LDA model on the dataset and extracted topics by setting a varying number of topics. We determined the most appropriate number of topics by computing the perplexity of the corpus using Eq. 1) which enables us to measure and analyze the ''goodness-of-fit'' of the topic model fitting with different k values, i.e., k as 5, 10, 20, and 100. The perplexity computation finding suggests that fitting the LDA model with k at 20 is a relatively better choice than alternative k values that preserve high perplexity. Among the top-20 topics, we selected the 14th topic, which we labeled it as ''Social Behaviour''. Subsequently, we identified authors who have a higher topic probability distribution. With the application of TOSNet, we have constructed a co-authorship network of 3000 authors on the basis of the topic t and discovered 522 subnetworks. Fig. 4 unveils the structure of the network containing the identified subnetworks.

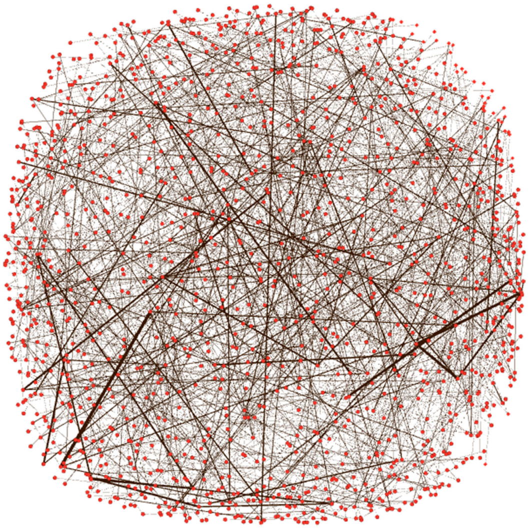


FIGURE 4. **A total of 3000 authors constructed the network. All the authors are colored in red to show their similarity in terms of a research topic. The thickness of the links represents the collaboration strength between pairs of authors.**

Further, Fig. 5 shows the co-authorship network structure constructed for 500 candidate authors comprised of 134 authors with intense collaborations and 60 subnetworks. As we can see from Figs. 4 and 5, as the number of nodes increases in a network, it becomes more challenging to manage, mining, analyze, and visualize the network. Hence, we can tell that dividing big networks with complicated inner structure into subnetworks is the best way to manage and analyze. Another interesting finding is that the larger size of the subnetwork does not guarantee the prominence of a subnetwork. Instead, a subnetwork can be identified as an optimal based on its authors' number of collaborations.

Having discovered subnetworks for the 14th topic, we have computed the $gnet - index$ value of each subnetwork and ranked them accordingly. Consequently, we have computed the average edge weight of each subnetwork and identified

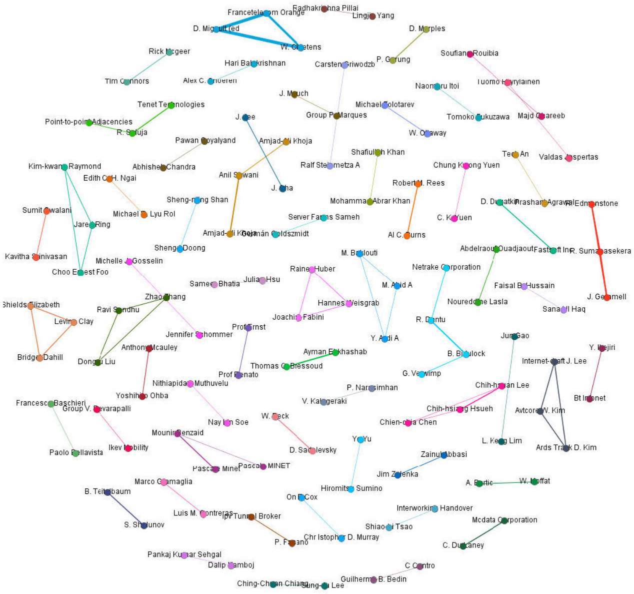


FIGURE 5. **A total of 134 authors selected from 500 candidates constructed the network. Different colors represent the sixty subnets identified from the network. The thicknesses of the links illustrate the intensity level between the pairs of authors in the subnets.**

subnetworks with an average edge weight higher than all subnetworks average edge weight. As a result, we have identified an optimal subnetwork with a group of 5 authors and a higher $gnet - index$ of 48. Fig. 6 shows the structure of the optimal subnet discovered using our method. Furthermore, Fig. 7 depicts the top five optimal subnetworks according to the $gnet - index$ value each of them obtained.

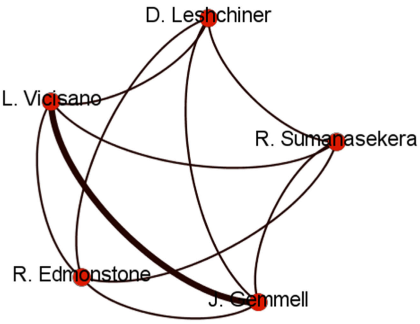


FIGURE 6. **The optimal subnet extracted from Fig. 4.**

Fig. 8a shows the size distribution of subnetworks versus the number of authors in a subnetwork. We found that amongst the 522 subnetworks, the minimum and maximum subnetwork sizes are 2 and 12 authors, respectively. On top of that, more than 50% of the subnetworks are collections of 2 authors, and less than 5% of the subnetworks have a size between 5-12. Moreover, we noticed that most subnetworks with lower $gnet - index$ are small in size, i.e., 2 authors in a subnetwork. Interestingly, the TOSNet approach allows us to combine subnetworks with smaller sizes and form a new subnetwork composed of any number of authors required by the researcher who plans to form a team because all the authors

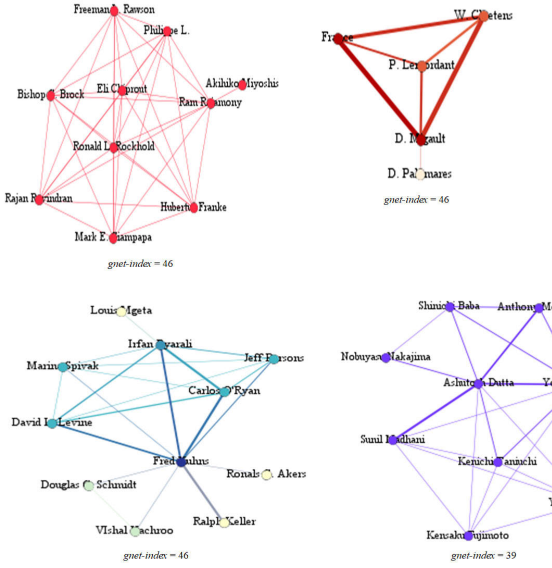
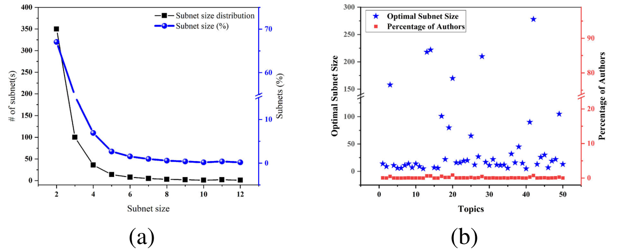


FIGURE 7. The optimal subnets in top five (2 – 5) according to *gnet – index* value each subnet obtains.

in the network are extracted based on a particular topic. From Fig. 8a, we can learn that the TOSNet approach optimizes the subnetwork identification process in identifying more manageable and feasible subnetworks with appropriate subnet size.



(a)

(b)

FIGURE 8. a) The distribution of subnets according to their size, i.e., the number of authors they contain. b) The percentage distribution of authors in the optimal subnets.

Besides, to further evaluate the optimally of the proposed method, we have detected subnetworks for 50 different topics. For each topic, we have identified 500 authors and constructed a co-authorship network. A different number of subnetworks with different sizes are detected for each topic. Fig. 8b illustrates the percentage of authors (out of the 500 authors) identified as members of the optimal subnets for 50 different topics. As can be seen from Fig. 8b, the higher number of the identified optimal subnets contain less than 10% of the total number of authors of each topic. As one of the main objectives of detecting subnetworks is to optimize the analysis of complex networks, we can tell that the TOSNet method is effective in detecting subnetworks and optimal subnet from a big academic network. Table 3 depicts five selected topics generated using the LDA model. We have assigned labels for each of the topics based on the words detected for each topic. Further, we have shown the optimal subnets discovered for each of the topics in Fig. 9.

TABLE 3. Selected topics generated by the LDA model. Note that the words included in each topic are 10 selected words that are meaningful.

Topic 4	Topic 12	Topic 15	Topic 16	Topic 20
Speech Recognition	**Communication Channel**	**Network Security**	**Astrophysics**	**Controlling System**
configuration	transmitter	internet	planet	controller
speech	antenna	architecture	atmosphere	nonlinear
recognition	forward	traffic	stars	coordination
information	receiver	address	giant	parameter
mutual	channels	domain	density	control
beamforming	signals	provider	galaxy	model
estimation	access	security	temperature	robust
parameter	decoder	networks	observation	robot
assumption	multiple	offload	orbit	system
Gaussian	communication	user	earth	track

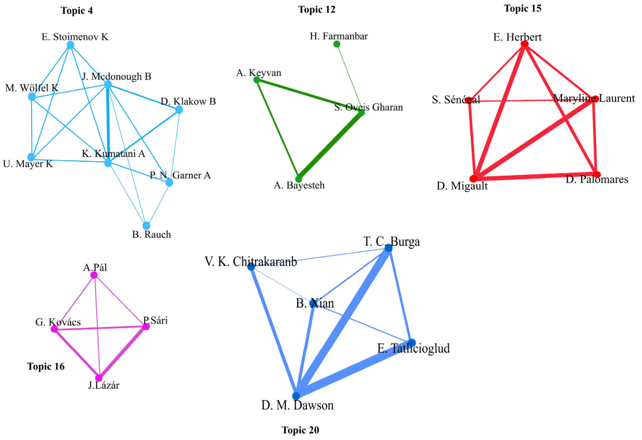


FIGURE 9. The optimal subnets of the topics depicted in Table 3.

b: SUBNET VERSUS GNET – INDEX DISTRIBUTION

In this work, as we discussed previously, we proposed *gnet – index* for ranking the discovered subnetworks and single out an optimal subnetwork. Thus, we computed the *gnet – index* value for the 522 subnetworks and found that more than 50% of the subnetworks have a *gnet – index* value of 1. Fig. 10 shows subnetworks vs *gnet – index* distribution. As we can see from Fig. 10, the number of subnetworks with a higher *gnet – index* is small compared to subnetworks with lower *gnet – index*. The distribution graph (i.e., Fig. 10) includes all the 522 subnetworks despite the average edge weight value (*avg_snet_edge_weight*) of each subnetwork. Figure 10 subtly explains the characteristics of the TOSNet method by showing how many of the detected subnetworks lies under which range of *gnet – index* value.

2) COMPARISON WITH BASELINE METHODS

As a comparison, we implemented the methods proposed in [35] in the same network we used for the proposed approach. Fig. 11 shows the identified subnetworks using the baseline method. The optimal subnetwork, i.e., ''h-subnet'' discovered using the first baseline method, contains a total of 83 authors. Moreover, baseline [35] identifies a combination of disconnected subnetworks as an optimal subnet. Besides, the optimal subnet consists of subnetworks with different research areas of interest. This makes it unreliable

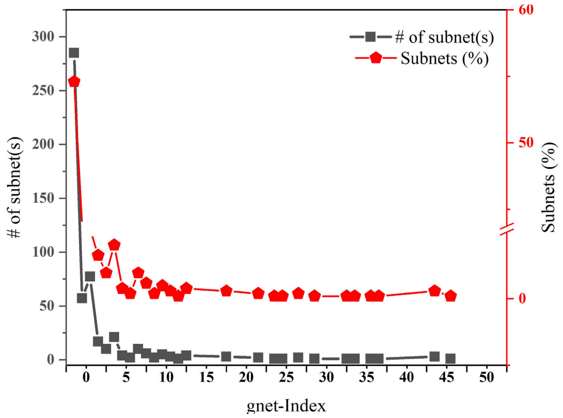


FIGURE 10. The distribution of subnets according to their ***gnet-index***.

to be chosen as a means to identify influential collaborators for specific research work.

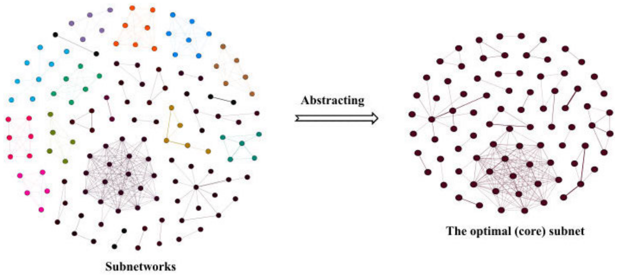


FIGURE 11. The identified subnetworks and the optimal subnet discovered after implementing the abstraction method as proposed in [35].

Subsequently, we have applied the baseline method [36] on the same CiteSeerX dataset we used for the TOSNet method and baseline [35]. According to the experimental findings conducted on this dataset, the TOSNet method efficiently identifies subnetworks that comprise authors with similar research topics. For baseline methods, there are probabilities the authors in the network might not preserve contextually related research areas of interest. Fig. 12 depicts the optimal subnet (''h-backbone'') identified using the baseline method proposed in [36]. The optimal subnetwork identified by the baseline method [36] contains 541 authors. Besides, inside the identified optimal subnet, there are ten disconnected sub-subnets. This shows as the baseline method [36] does not give promising results in identifying optimal and effective subnetworks from a big network. Moreover, both baseline methods are not efficient while implementing them in big networks; neither can they discover the required type of subnetworks in terms of a research topic as well as subnetwork size. On top of that, unlike the TOSNet approach, the baseline methods are limited to dealing exclusively with homogeneous networks rather than heterogeneous networks.

VI. DISCUSSION

Three fundamental characteristics differentiate TOSNet from existing approaches. First, our approach extracts subnetworks by considering a research topic similarity among authors in a network. We have adopted LDA topic modeling to mine topics from a given network and compute the topic probability distribution of each author. Identifying subnetworks by taking into account specific research topics helps unravel influential researchers on those specific research areas of interest and produce quality work. Furthermore, scientific researchers usually seek collaborators with specific topics; hence, it is necessary to design contextual subnetwork identification for scientific collaborator recommendations. The second feature that makes our approach different is that it focuses on the computation of collaboration intensity between authors when detecting subnetworks from a co-authorship network. The more intense the collaboration between authors is, the more productive they can be when they get a chance to work together again. Moreover, studying the intensity level of collaborations in a co-authorship network helps to discover effective scientific teams. Third, our approach ranks the identified subnetworks based on their total number of collaborations they maintain and single out an optimal subnetwork. The proposed method effectively generates the optimal subnetwork with a reasonable subnet size, i.e., the number of authors in a subnet.

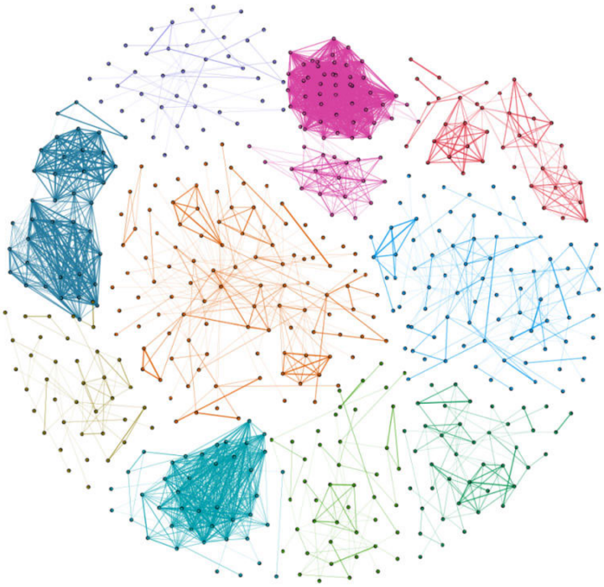


FIGURE 12. The identified optimal subnet (''h-backbone'') containing 541 authors, using the baseline method proposed in [36].

According to the comparative analysis results, our method outperforms the baseline methods in terms of discovering contextually related subnetworks and generating manageable and reasonable subnetworks from big heterogeneous networks. The identified subnetworks can be considered as scientific teams without further computational process. To further check the similarity of subnetworks, we have employed the ''network density'' measurement metric. The network density measures how close the network is to complete. A complete subnetwork has all possible collaborations, and its density equals 1. The higher density indicates that the authors in the network have intense and dense collaborations with each other. In contrast, lower density indicates the sparsity nature of a network. The subnetworks identified by our proposed method have density values of 0.5 and more. Hence, our proposed method has an average density of 0.725; in contrast, the two baselines [35] and [36] have average density values of 0.056 and 0.253, respectively.

Moreover, to assess efficiency, we have constructed co-authorship networks and discovered subnetworks with different numbers of authors using our approach and the baseline methods. The experimental findings show that the running time of TOSNet has linear scalability concerning the number of collaborations that exist in the co-authorship network. In contrast, the baseline method [35] is costly in terms of time while applying it in a big academic network. Also, the second baseline method [36] takes the time complexity of $O(n^2)$ for detecting subnetworks from a co-authorship network, where n is the number of edges in the co-authorship networks. Although the proposed method and the baseline method [36] have equivalent time complexity, our method is more efficient as we can get more information while using it to detect subnetworks, for example, research topics of each author in particular subnetworks. To show the computational time of the proposed method and the baselines, we have generated five artificial networks with different numbers of collaborations ranging from 10,000 to 160,000. Fig. 13 depicts the running time of our proposed algorithm and baseline methods. The TOSNet method shows linear running time to discover subnetworks and identify one as an optimal subnetwork from co-authorship networks with different sizes. Whereas the efficiency and scalability of the baseline method [36] decrease as the number of authors in the network increases.

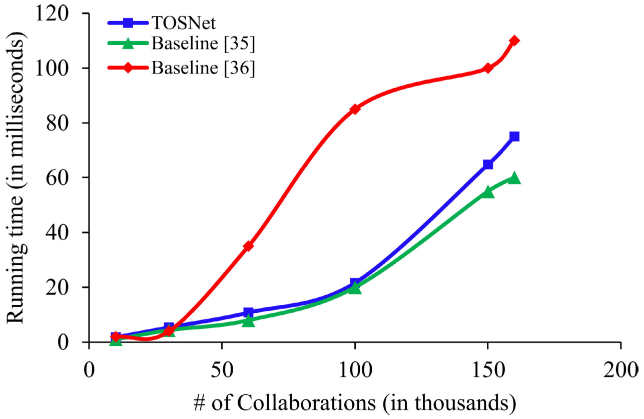


FIGURE 13. The running time of the TOSNet and baseline methods.

Moreover, we have evaluated the accuracy of the proposed method with respect to the size of the authors in the subnetwork. TOSNet identified subnetworks with a size range of 2 – 12. In contrast, the subnetworks identified by the baseline methods are large-scale, and some of the authors in the subnetwork do not have similar research topics. Hence, to do the optimal size distribution for the baseline methods, we have implemented them after identifying candidates for each selected topic. Fig. 14 illustrates the distribution of optimal subnets size identified using the TOSNet and baseline methods for 50 different topics. It shows that the baseline methods mostly identify optimal subnets containing 50 and more authors, unlike the TOSNet method. This finding indicates that the proposed approach is more effective and accurate in revealing reliable and manageable subnetworks that have included authors with related research areas of interest. Also, the proposed method optimizes the process of subnetwork identification in big heterogeneous networks.

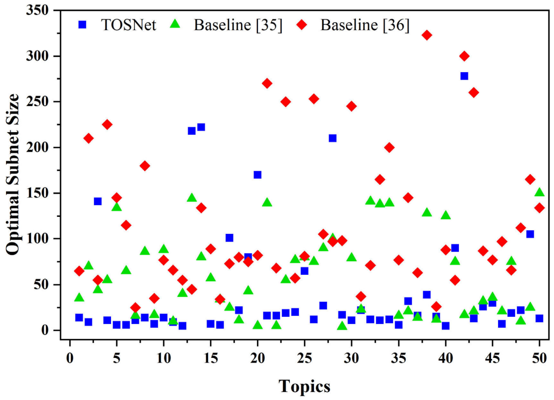


FIGURE 14. The optimal subnets size distribution in 50 different topics detected using the TOSNet and baseline methods.

Additionally, we have assessed the performance of TOSNet and baseline methods on a co-authorship network using modularity (Q) metric [37], since there exists no compelling metric to evaluate the efficacy of detected communities or subnetworks. We have employed modularity to generally evaluate the quality of the detected subnetworks using the TOSNet and baseline methods. We found that our proposed approach has a reasonable average modularity value of 0.897. While the baseline methods [35] and [36] have relatively smaller average modularity values of 0.656 and 0.584, respectively. As discussed by Newman and Clauset [37], higher Q indicates dense interconnections amongst authors, while lower Q indicates a random and unrelated collection of edges in the network. Besides, it is considered that higher modularity indicates the detected subnetworks are effective and good in quality. Fig. 15 illustrates the modularity (Q) value computed for groups of subnetworks detected for 50 different topics using the TOSNet and baseline methods. From Fig. 15, we can see that the subnetworks detected, using the TOSNet method, for more than 40% of the topics have modularity Q values higher than 0.6. In contrast, subnetworks detected, using the baseline methods, for less than 25% of the topics have measured lower modularity values. This finding shows how the TOSNet method detects quality subnetworks from a big academic network, that of heterogeneous.

According to the comprehensive studies and experimental results, we can conclude that TOSNet can accurately and

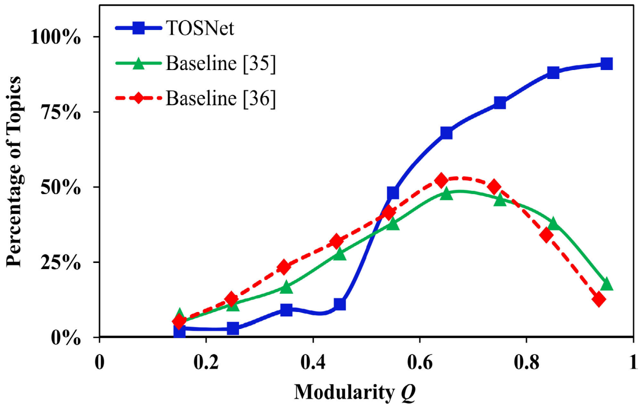


FIGURE 15. The modularity Q values distribution of the subnetworks detected using the TOSNet method.

effectively discover topic-based subnetworks from a network that has a complicated inner structure and a very big-scale in size. Moreover, it can identify the most relevant optimal subnetwork for a particular research topic according to the proposed metric, i.e., $gnet - index$. Also, the experimental findings have verified the effectiveness of TOSNet in terms of identifying significant subnetworks with a good modularity value.

Although our approach performs well in identifying contextual subnetworks against the baseline methods, it is necessary to evaluate the proposed approach by applying it on networks with known ground-truth community structures. As future work, we are interested in evaluating TOSNet against other well-known community detection or subnetwork identification methods using networks with a known number of communities.

VII. CONCLUSION

Although identifying subnetworks or community detection is a long-studied issue in different disciplines, especially in computer science and biology, it still lacks a satisfactory solution. For instance, if a researcher wants to form a scientific team for a particular research topic, s/he first needs to identify collaborators who have experiences on that specific topic. To do that, s/he needs to find a contextual community detection method that could enable her/him to discover relevant collaborators. Nevertheless, there are little to no methods that can accurately and efficiently reveal contextual subnetworks in terms of research topics or areas of interest. In this paper, we have introduced a new approach for classifying big networks into smaller groups, i.e., discovering subnetworks from big academic networks, which is known as a Topic-based Optimal Subnetwork identification (TOSNet). In TOSNet, we have first adopted a topic modeling method, i.e., LDA algorithm, to identify authors with a higher topic probability distribution. We have then employed an algorithm that computes the collaboration intensity index to accurately discover subnetworks that contain groups of authors with similar or relatable research topics. Moreover, we have ranked the detected subnetworks with a new metric introduced by us, called $gnet - index$. We have specifically considered a heterogeneous authorship network from the CiteSeerX dataset, which consists of authors and papers as nodes and connections between these two entities as edges. The experimental findings have demonstrated the effectiveness of TOSNet.

For future work, we plan to expand our approach to fit with temporal networks by considering time because some researchers, once influential and active on specific topics, might not be active another time.

REFERENCES

[1] X. Kong, Y. Shi, S. Yu, J. Liu, and F. Xia, "Academic social networks: Modeling, analysis, mining and applications," *J. Netw. Comput. Appl.*, vol. 132, pp. 86–103, Apr. 2019.
[2] C. L. Giles, K. D. Bollacker, and S. Lawrence, "Citeseer: An automatic citation indexing system," in *Proc. ACM JCDL*, 1998, pp. 89–98.
[3] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Phys. Rep.*, vol. 659, pp. 1–44, Nov. 2016.
[4] B. Chatterjee and H. N. Saha, "Detection of communities in large scale networks," in *Proc. IEEE 10th Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Oct. 2019, pp. 1051–1060.
[5] G. M. Slota, J. W. Berry, S. D. Hammond, S. L. Olivier, C. A. Phillips, and S. Rajamanickam, "Scalable generation of graphs for benchmarking HPC community-detection algorithms," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, 2019, p. 73.
[6] P. Symeon, K. Yiannis, V. Athena, and S. Ploutarchos, "Community detection in social media, performance and application considerations," *J. Data Mining Knowl. Discovery*, vol. 24, no. 3, pp. 515–554, 2012.
[7] T. H. P. Silva, M. M. Moro, A. P. C. Silva, W. Meira, and A. H. F. Laender, "Community-based endogamy as an influence indicator," in *Proc. IEEE/ACM Joint Conf. Digit. Libraries*, Sep. 2014, pp. 67–76.
[8] C. Zhe, A. Sun, and X. Xiao, "Community detection on large complex attribute network," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2041–2049.
[9] G. Rossetti and R. Cazabet, "Community discovery in dynamic networks: A survey," *ACM Comput. Surv.*, vol. 51, no. 2, p. 35, 2018.
[10] M. A. Javed, M. S. Younis, S. Latif, J. Qadir, and A. Baig, "Community detection in networks: A multidisciplinary review," *J. Netw. Comput. Appl.*, vol. 108, pp. 87–111, Apr. 2018.
[11] H. Nguyen, S. Shrestha, D. Tran, A. Shafi, S. Draghici, and T. Nguyen, "A comprehensive survey of tools and software for active subnetwork identification," *Frontiers Genet.*, vol. 10, p. 155, Mar. 2019.
[12] S. Yu, F. Xia, K. Zhang, Z. Ning, J. Zhong, and C. Liu, "Team recognition in big scholarly data: Exploring collaboration intensity," in *Proc. IEEE 15th Int. Conf. Dependable, Autonomic Secure Comput., 15th Int. Conf. Pervas. Intell. Comput., 3rd Int. Conf. Big Data Intell. Comput. Cyber Sci. Technol. Congress(DASC/PiCom/DataCom/CyberSciTech)*, Nov. 2017, pp. 925–932.
[13] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.
[14] X. Hu, M. Ng, F.-X. Wu, and B. A. Sokhansanj, "Mining, modeling, and evaluation of subnetworks from large biomolecular networks and its comparison study," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 2, pp. 184–194, Mar. 2009.
[15] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 2, Feb. 2004, Art. no. 026113.
[16] S. Bhatt, S. Padhee, A. Sheth, K. Chen, V. Shalin, D. Doran, and B. Minnery, "Knowledge graph enhanced community detection and characterization," in *Proc. 12th ACM Int. Conf. Web Search Data Mining (WSDM)*, New York, NY, USA, 2019, pp. 51–59.
[17] T. M. V. Le and H. W. Lauw, "Probabilistic latent document network embedding," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 270–279.
[18] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "The Louvain method for community detection in large networks," *J. Stat. Mechanics: Theory Exp.*, vol. 10, Mar. 2011, Art. no. P10008.

[19] Y. Ma, Z. Ji, and L. Song, "A two-layer network model reveals the adhesion scientist career stage and research topic in China," *IEEE Access*, vol. 8, pp. 52726–52737, 2020.

[20] D. He, Z. Feng, D. Jin, X. Wang, and W. Zhang, "Joint identification of network communities and semantics via integrative modeling of network topologies and node contents," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 116–124.

[21] C. Li, W. K. Cheung, Y. Ye, X. Zhang, D. Chu, and X. Li, "The author-topic-community model for author interest profiling and community discovery," *Knowl. Inf. Syst.*, vol. 44, no. 2, pp. 359–383, Aug. 2015.

[22] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," 2012, *arXiv:1207.4169*. [Online]. Available: http://arxiv.org/abs/1207.4169

[23] Y. Liu, A. Niculescu-Mizil, and W. Gryc, "Topic-link LDA: Joint models of topic and author community," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 665–672.

[24] G. P. Guedes, E. Ogasawara, E. Bezerra, and G. Xexeo, "Discovering top- k non-redundant clusterings in attributed graphs," *Neurocomputing*, vol. 210, pp. 45–54, Oct. 2016.

[25] M. Sachan, D. Contractor, T. A. Faruquie, and L. V. Subramaniam, "Using content and interactions for discovering communities in social networks," in *Proc. 21st Int. Conf. World Wide Web*, New York, NY, USA, 2012, pp. 331–340.

[26] X. Pan, G. Xu, B. Wang, and T. Zhang, "A novel community detection algorithm based on local similarity of clustering coefficient in social networks," *IEEE Access*, vol. 7, pp. 121586–121598, 2019.

[27] T. Chakraborty, A. Dalmia, A. Mukherjee, and N. Ganguly, "Metrics for community analysis: A survey," *ACM Comput. Surv.*, vol. 50, no. 4, p. 54, 2017.

[28] R.-H. Li, L. Qin, J. X. Yu, and R. Mao, "Influential community search in large networks," *Proc. VLDB Endowment*, vol. 8, no. 5, pp. 509–520, Jan. 2015.

[29] M. Lei and D. Wei, "Identifying influence for community in complex networks," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Jun. 2018, pp. 5346–5349.

[30] N. Du, X. Jia, J. Gao, V. Gopalakrishnan, and A. Zhang, "Tracking temporal community strength in dynamic networks," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 3125–3137, Nov. 2015.

[31] J. Zhan, V. Guidibande, and S. P. K. Parsa, "Identification of top-K influential communities in big networks," *J. Big Data*, vol. 3, no. 1, p. 16, Dec. 2016.

[32] F. Bi, L. Chang, X. Lin, and W. Zhang, "An optimal and progressive approach to online search of top-k influential communities," *Proc. VLDB Endowment*, vol. 11, no. 9, pp. 1056–1068, May 2018.

[33] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[34] S. Wuchty, B. F. Jones, and B. Uzzi, "The increasing dominance of teams in production of knowledge," *Science*, vol. 316, no. 5827, pp. 1036–1039, May 2007.

[35] S. X. Zhao, P. L. Zhang, J. Li, A. M. Tan, and F. Y. Ye, "Abstracting the core subnet of weighted networks based on link strengths," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 5, pp. 984–994, May 2014.

[36] R. J. Zhang, H. E. Stanley, and F. Y. Ye, "Extracting h-backbone as a core structure in weighted networks," *Sci. Rep.*, vol. 8, no. 1, pp. 1–7, Dec. 2018.

[37] M. E. J. Newman and A. Clauset, "Structure and inference in annotated networks," *Nature Commun.*, vol. 7, no. 1, p. 11863, Sep. 2016.

HAYAT D. BEDRU received the bachelor's degree in computer science and information technology from Adama University, Adama, Ethiopia, and the M.Sc. degree in software engineering from the HiLCoE School of Computer Science and Technology, Addis Ababa, Ethiopia. She is currently pursuing the Ph.D. degree in software engineering with the Dalian University of Technology, Dalian, China. Her research interests include network science, data science, and computational social science.

WENHONG ZHAO (Member, IEEE) received the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2002. Since 1991, he has been working with the Zhejiang University of Technology, Hangzhou, China, where he is currently a Full Professor with the Ultraprecison Machining Center. His research interests include big data, embedded systems, intelligent systems, and precision machining.

MUBARAK ALRASHOUD received the Ph.D. degree in computer science from Ryerson University, Toronto, ON, Canada, in 2015. He is currently an Associate Professor and the Head of the Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. His research interests include social network analysis, vehicular ad hoc networks, social networks, and software engineering.

AMR TOLBA received the M.Sc. and Ph.D. degrees from the Department of Mathematics and Computer Science, Faculty of Science, Menoufia University, Egypt, in 2002 and 2006, respectively. He is currently an Associate Professor with the Faculty of Science, Menoufia University. He is also with the Department of Computer Science, Community College, King Saud University (KSU), Saudi Arabia. He has authored or coauthored over 60 articles. His main research interests include socially aware networks, vehicular ad hoc networks, the Internet of Things, intelligent systems, and cloud computing.

HE GUO received the B.Sc. degree from Jilin University, Changchun, China, in 1982, and the M.Sc. degree from the Dalian University of Technology, Dalian, China, in 1989. He is currently a Full Professor with the School of Software, Dalian University of Technology. His research interests include distributed computing, computer vision, artificial intelligence, and software engineering.

FENG XIA (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He is currently an Associate Professor and the Discipline Leader with the School of Engineering, IT and Physical Sciences, Federation University Australia. He has published two books and over 300 scientific papers in international journals and conferences. His research interests include data science, social computing, and systems engineering. He is a Senior Member of ACM.