Neurocomputing 367 (2019) 46–54

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Link communities detection: an embedding method on the line hypergraph

Haicheng Tao[a], Zhe Li[b,*], Zhiang Wu[c], Jie Cao[a]

[a] School of Computer Science and Engineering, Nanjing University of Science & Technology, China
[b] College of Economics and Management, Hubei Engineering University, China
[c] School of Information Engineering, Nanjing University of Finance & Economics, China

ARTICLE INFO

Article history:
Received 16 May 2019
Revised 5 July 2019
Accepted 6 July 2019
Available online 9 July 2019

Communicated by Prof. Zidong Wang

Keywords:
Link communities
Overlapping communities
Line hypergraph
Graph embedding

ABSTRACT

Recent advances have verified ground-truth communities perceive several characteristics. That is, communities are overlapped and densely connected. Not only that, the organization of communities, in a general sense, is hierarchical. To capture all of these characteristics, we propose a framework based on link embedding method. Firstly, we define close-knit link groups which preserve the hierarchical structures and carefully transform the problem of mining close-knit link groups as mining cosine patterns which can be implemented efficiently. Secondly, we construct the weighted line hypergraph and embed each link into a low dimension vector. Finally, we simply employ *K*-means algorithm to obtain the link communities. Overlapping structures are naturally obtained by interpreting the link communities as nodes communities. Experimental results on three real-world networks demonstrate the proposed approach is able to identify much higher-quality overlapping communities in terms of four external measures, compared with six classical overlapping community detection methods.

1. Introduction

Community structures within complex networks play a vital role in many areas, ranging from physics [1] and bioinformatics [2] to social sciences [3] and computer sciences [4]. Numerous methods exist for discovering both crisp and fuzzy (i.e., overlapping) communities. Almost all of these methods are constructed upon an underlying assumption: nodes in the same community are connected more densely than those between different communities [5].

In fact, the design of various approaches is driven by imaginary features of the community. For instance, one of the most influential methods, based on the modularity (Q) maximization (e.g., Fast Newman [5], Louvain [6]), is driven by the above underlying community assumptions. Furthermore, the Infomap algorithm [7] makes the assumption that nodes in the same community are connected by some paths upon which information flows more quickly and easily. Hence, characterizing distinctive features of communities within real-world networks largely contributes to the success of community detection methods.

To be specific, firstly, real-life communities are *overlapping*, since there could be nodes belong to multiple communities simultaneously. For instance, a person often has connections to a number of social groups simultaneously, such as scientific activities, family, friends, and hobbies.

Secondly, the true closely-knit community is usually very small, though it defies the modularity based optimization and evaluation. In [8], the community size is proven to obey the pow-law distribution on six real-world networks, which implies a vast majority of communities are in small-size. For example, approximately 62% communities contain less than 10 users, and 80% communities contain at most 20 users in the `Friendster` network.

Thirdly, the organization of communities is *nested* and *hierarchical*. Small communities build larger ones which group together to form much larger ones [9]. Fig. 1 shows an ego-network as well as its labeled communities extracted from Facebook. The student marked in red has different circles or communities with respect to his social relationships. More in detail, that student is in a research group while he/she may also share the similar music interest with others. Meanwhile, they are also in the same department and thus form a bigger community containing those small ones.

Last but not least, the overlapping nodes between communities are densely connected in real world networks which is different from conventional views. In [8], authors study 6 large social networks with ground-truth communities and point out that overlaps

* Corresponding author.
E-mail address: lizhe_hbeu@vip.163.com (Z. Li).

https://doi.org/10.1016/j.neucom.2019.07.003
0925-2312/

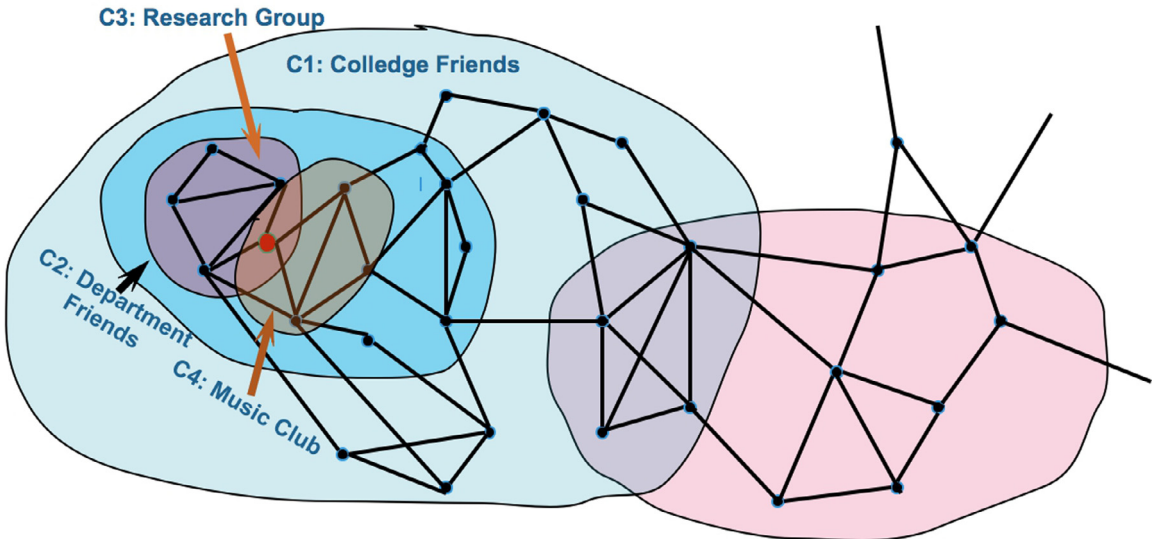

Fig. 1. An ego-network from Facebook with complex community structures.

of communities are more densely connected than non-overlapping parts of communities. Meanwhile, existing methods either identify those overlaps as sole communities or merge the overlapping communities into a single one.

In this situation, it calls for a flexible framework which is able to identify hierarchical communities at different levels. That is, it is expected to reveal the smallest but closely-connected communities; but sometimes the macro-view or medium-view of structures is also needed to characterize a skeleton structure of the whole network. Meanwhile, at any level, the multiple membership of every node should be expressed.

One of the most popular method to detect overlapping community is the Clique Percolation Method (CPM) [10], where the union of adjacent k-1 common nodes of k-cliques are defined as communities. Since one node can participate in multiple k-cliques simultaneously, overlaps arise naturally between these communities. However, once the k is specified, CPM-like methods can not capture the multiple hierarchical structures at different levels. Another method is based on links rather than nodes [9], where links are partitioned via hierarchical clustering. But this method is also built on the assumption that nodes are more sparsely connected between communities than those within communities which contradicts with the cases in the highly dense networks [8].

To address these limitations, we propose an embedding method based on links, which can preserve the hierarchical structures at different levels, and construct a weighted hypergraph using these structures. After that, we can get a low dimensional representation using the spectral clustering method. Consequently, a highly dense and overlapping communities can be obtained by simply running K-means algorithm on them only once. The hierarchical structures are constructed by cosine patterns which will be explained in details below. Experiments on real networks with "ground truth communities" information validate the proposed method.

2. Related work

In the literature, algorithms for overlapping community detection can be roughly classified into five categories [11], i.e., clique percolation methods(CPM), local expansion methods, fuzzy detection methods, agent based methods and line graph and link partitioning methods. The clique percolation method is based on the k-clique which is fully connected subgraph of k nodes. The communities identified by CPM are the union of k-cliques which share $k-1$ common nodes. Since nodes can participate in multiple k-cliques, the overlaps between communities naturally occur. However, CPM can not capture the underlying structure on the circumstance that the network is highly connected [12]. The methods based on local expansion try to maximize the fit function by expanding a defined seed. Similar to CPM, Greedy Clique Expansion method (GCE) [13] chooses the clique as seed. As for fuzzy detection methods, nodes are represented as soft membership vectors by various network embedding methods, e.g., spectral clustering methods and model based methods [14,15], which can detect highly overlapping communities. Most agent based methods rely on the label propagation algorithms in which nodes are grouped together by the propagation of the same property [16,17]. In order to get overlaps, multiple labels are assigned to nodes. The line graph and link partitioning methods are the most relevant to our methods. Two strategies can be applied to this type of methods, i.e. bottom-up strategy and top-down strategy. For bottom-up strategy, link communities can be obtained via hierarchical clustering based on link similarities. In [18], authors measure the similarity between links using Jaccard index and build a link dendrogram which then can be partitioned by the maximum partition density. Meanwhile, by transforming the node graph into line graph, existing node partitioning algorithms can be applied to find crisp link communities so as to find overlapping nodes communities [19,20]. This line of methods employ the top-down strategy. Since there could be much more links after transforming the node graph into line graph, authors in [21] reduce the number of links through random sampling and then map the link into a 2-dimensional space using geometric embedding method. Lastly, *DBSCAN* is used for clustering. However, these methods can not capture either hierarchical structures or highly dense overlapping structures in real world networks. While our proposed method can preserve the hierarchical structures due to the link representation of the network and detect highly overlapping communities using the embedding methods borrowed from fuzzy detection methods.

3. Framework overview

In this section, we start with an analysis of the fundamental techniques for link community detection, and then describe the central idea of our approach as well as the overview of our detection framework.

3.1. Motivations

Recently, link communities are deemed to be more intuitive than node communities on revealing structures of the real-world network, due to the fact that each link usually has a unique position whereas the node naturally occupies multiple positions owing to its links [18]. Thus, the overlapping community detection problem could be treated as a natural byproduct of link communities, [22] just by transforming the crisp link communities to fuzzy node communities. Most of the existing methods for detecting link communities are based on the so-called *link clustering* [18–24]. The pioneering work [18] used a Jaccard-type similarity score for a pair of links to generate the hierarchical link clustering and split clusters according to the newly-proposed measure partition density. This is a bottom-up strategy by constantly merging the closest pair of link clusters. However, to compute the similarity matrix of all link pairs are very expensive, since the number of edges is usually far more than that of nodes.

By contrast, another research stream performs link clustering by exploiting the top-down strategy. They transform the network into the corresponding line graph and then detect link communities by various algorithms for node partitioning on this generated line graph, such as modularity optimization [19,20], matrix factorization [22], local density optimization [23] and geometric embedding [21]. These top-down methods have reduced computational cost remarkably, but encountering another problem: the line graph is very noisy, i.e., containing excessive weak-ties. In fact, Lim et al. [21,24] have noticed this problem and proposed the naive link sampling strategy to alleviate it.

3.2. Central idea

The heart of our framework is to combine the strengths of both bottom-up and top-down strategies. First of all, we extend the similarity of *a pair of* links, proposed by Ahn et al. [18], to that of *a group of* links. We prove that if suitable thresholds are set, several types of weak-ties can be excluded from link groups. Also, by virtue of our previous work on cosine pattern mining [25,26], we can utilize an efficient algorithm for mining link groups satisfying the threshold constraints. Secondly, we model the groups of links as a *hypergraph* for the reason that each group usually contains more than two links. Compared with the traditional line graph used in [19–22,24], our hypergraph model can significantly reduce the number of nodes in the line-space, since links with high similarity have been placed in a group and connected by hyperedge. Meanwhile, a host of weak-ties have been filtered, which makes the hypergraph contain less noise. Finally, hypergraph embedding technique is employed to partition the link groups into a number of link communities.

At the end of this subsection, we give the notations that will be used hereafter. Assume the original undirected unweighted network is $\mathcal{G}(V,E)$ with n nodes and m edges. Moreover, $v_i(1 \le i \le n)$ and e_{ij} are introduced to represent the ith vertex as well as the edge connecting v_i and v_j respectively. The link community to be detected in this paper is formally defined in Definition 1.

Definition 1. Given the network $\mathcal{G}(V,E)$, the collection of link communities is $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$, where $L_k \subset E, 1 \le k \le K$. The elements of $\mathcal{L}$ is pairwise disjoint: $L_k \cap L_{k'} = \emptyset,\ \forall k \ne k'$. The union of the elements of $\mathcal{L}$ is the subset of E: $L_1 \cup \cdots \cup L_k \subseteq E$.

4. Link community: definition and properties

In this section, we define the link community that is derived by extending the similarity of a pair of links, proposed by Ahn et al. [18], to that of a group of links. Then, we discuss some important properties especially about some type of weak-ties can be excluded from the proposed link community.

4.1. The definition

Given a pair of links, e.g., e_{ir} and e_{jr} incident on a node r, the similarity can be computed via *cosine* instead of Jaccard index in the [18] defined as

$$\cos(e_{ir}, e_{jr}) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|} \tag{1}$$

where N_i is the neighborhood of node i. The reason we choose *cosine* as our similarity measure will be explained later. Moreover, we can easily extend the link pairs into link groups as follows.

$$\cos(e_{1r}, e_{2r}, \ldots, e_{qr}) = \frac{|N_1 \cap \cdots \cap N_q|}{|N_1 \cup \cdots \cup N_q|} \tag{2}$$

The cosine similarity lies between 0 and 1. 1 is for the equivalent structure which is rare and hard to find. Whereas in this paper, we aim to find structures(link groups) which are looser than equivalent structures by setting a threshold for *cosine* similarity appropriately. Meanwhile, it can be observed that the *cosine* similarity of link groups is derived from node r's neighbors, i.e., impost nodes, which is irrelevant to the node r, i.e., keystone in the Eq. (2). If we represent the network as the *transaction* mode, where each line corresponds to a node(keystone) and items(impost nodes) in this line are its neighbors. To find these link groups in social network is equivalent to mine items associated with the transaction in the realm of frequent pattern mining. Moreover, due to the *cosine* measure holds Ordered Anti-Monotone Property (OAMP), we can use an efficient algorithm proposed for mining cosine patterns in our previous work [25,26] to find link groups. So from the perspective of pattern mining, we define the close-knit link groups as follows.

Definition 2. (Close-knit Link Groups) Given an unweighted network $\mathcal{G}(V,E)$, t_s^* for the minimum support threshold and t_c^* for the minimum cosine threshold, close-knit link groups are defined by $\mathcal{F}(\mathcal{G}, t_s^*, t_c^*) = \{X \subseteq subset(E) | supp(X) \ge t_s^*, \cos(X) \ge t_c^*\}$.

Note that the close-knit link groups are hierarchical since subset of cosine patterns could also be included within a threshold for t_s^* and t_c^*.

4.2. The properties

In [27], Mark Granovetter presents that the set of nodes made up of weakly connected nodes(weak tie nodes) comprises a low-density network whereas the set consisting of the same nodes but with nodes which are strong connected(strong tie nodes) will be densely knit, i.e., close-knit groups. Notice that our method can well handle the weak-ties by using the threshold t_s^*, especially for two special types of weak ties, i.e., *bridge* containing only one path between two endpoints and *local bridge* attaching two endpoints v_a and v_b which have no friends in common, that is, the length of path connecting v_a and v_b will be more than two if deleting the local bridge. Given a link, e.g., e_{ij}, node i and j could be (I) both impost nodes or (II) one of them is the keystone. Then, we have the following properties:

Property 1. *The bridges are excluded in close knit link groups with* $t_s^* \ge 2/n$.

Proof. Assume e_{ij} is the bridge.

For (I), according to the definition of bridge, we have $|N_i \cup N_j| = 0$ and thus e_{ij} can not be in cosine patterns.

For (II), let node i be the keystone in the cosine pattern P and node j and node p are impost nodes. Since $t_s^* \ge 2/n$, there should

be at least another keystone for cosine pattern P, e.g. node q. So, we have two paths between node i and node j, i.e., $i \to j$ and $i \to p \to q \to j$ which contradicts with the definition of bridge. □

Property 2. *The local bridges with span more than 3 are excluded in close knit link groups with* $t_s^* \geq 2/n$.

Proof. Assume e_{ij} is the local bridge.

For (*I*), this property is also true since $|N_i \cup N_j| = 0$.

For (*II*), we can also get two paths between node i and node j, i.e., $i \to j$ and $i \to p \to q \to j$, and the length of path will be 3 if excluding the imediate path, i.e., $i \to j$, which contradicts with the definition of local bridge. □

5. Line hypergraph embedding method

This section discusses the embedding method on line hypergraph to discover link communities of the original network. Recall that a set of close-knit link groups have been mined, denoted as $\mathcal{H}$ in Section 4.1. Hence, we introduce the representation of $\mathcal{H}$ as a hypergraph at first.

5.1. Hypergraph transformation

A hypergraph is a graph in which an edge can connect more than two vertices [28], and it is widely used for representing a set of complex relational objects in many real-world problems. Here, we employ the *weighted hypergraph* to represent the close-knit relations among multiple edges.

Definition 3. Given the original network $\mathcal{G}(V, E)$ and the $\mathcal{H}$ set of close-knit link groups, the weighted hypergraph is denoted as $\Gamma(V^+, E^+, w)$ which satisfies the following conditions.

- A node $v_{e_{ij}} \in V^+$ corresponds to the link e_{ij} between v_i and v_j in $\mathcal{G}$, and $V^+ = \{v_{e_{ij}} | \exists h_z \in \mathcal{H}, e_{ij} \in h_z\} \subseteq E$.
- A hyperedge $e_z \in E^+$ corresponds to the link group $h_z \in \mathcal{H}$, and it connects the node $v_{e_{ij}}$ if $e_{ij} \in h_z$. Thus, $|E^+| = |\mathcal{H}|$.
- The weight $w(e_z) = \cos(h_z)$, i.e., the cosine similarity of the corresponding link group.

The traditional line graph proposed in [19,20] (or so-called link-space graph in [21,24]) simply treats every link as the node of line graph and adds an edge of line graph if two links of original graph have a common node. Compared with this model, our hypergraph has far less edges, since (i) a hyperedge is able to connect multiple nodes, and (ii) only links of original graph with strong similarity are connected by a hyperedge. In essence, this concise hypergraph model, which transforms the original network into the link space, benefits from the definition of the local close-knit link group, that is, the bottom-up handling strategy.

5.2. The embedding method

Similar to the approach often used in the literature (e.g., [21,22]), to find link communities and thereby overlapping communities of nodes on the original graph $\mathcal{G}$ can be achieved by using a *crisp* clustering algorithm on the hypergraph Γ. So, our focus here is on the hypergraph clustering. As shown by Yan et al. [29], the graph embedding paradigm is able to represent vertices of a graph in a low-dimensional vector space while preserving the structure of the original graph. With proper graph embedding, we can process graph data more flexibly and efficiently, by using various classic vector-based clustering algorithms.

Intuitively, the *incidence matrix* of the hypergraph Γ is a $|V^+| \times |E^+|$ matrix denoted as $\mathbf{C}$, where each element $c(v, e) = 1$ if $v \in e$ and 0 otherwise. Let $d(v) = \sum_{e \in E^+} w(e)c(v, e)$ denote degree of every node, and $\mathbf{D}_v$ be a $|V^+| \times |V^+|$ diagonal matrix containing the vertex degree. Then, the *adjacency matrix* [28] of the hypergraph Γ is

$$\mathbf{A}_\Gamma = \mathbf{C}\mathbf{W}\mathbf{C}^\top - \mathbf{D}_v, \tag{3}$$

where $\mathbf{W}$ is the $|E^+| \times |E^+|$ diagonal matrix containing the weights of hyperedges. Once the adjacency matrix of hypergraph is available, the graph embedding is generally viewed as the factorization on a proximity matrix among nodes [29–31]. Formally, let $\mathbf{S}$ denote a $|V^+| \times |V^+|$ proximity matrix, and $\mathbf{Z}$ denote a $|V^+| \times Z$ embedding matrix, where Z is the embedding dimensions (i.e., the dimension of latent space). Then, we adopt the L2-norm below as the loss function which need to be minimized:

$$\min ||\mathbf{S} - \mathbf{Z} \cdot \mathbf{Z}^\top||_F^2. \tag{4}$$

The objective of Eq. (4) is to find an optimal rank-Z approximation of the proximity matrix $\mathbf{S}$. The solution is equivalent to perform the Singular Value Decomposition (SVD) on $\mathbf{S}$, and select the largest Z singular value with corresponding singular vectors to construct the optimal embedding vectors.

As noted already, the graph embedding is technically feasible, if the proximity matrix $\mathbf{S}$ is determined on the hypergraph. In the literature of complex network analysis [32], a vast number of proximity measures between nodes have been presented. Here, we introduce two feasible choices for defining $\mathbf{S}$: one for low-order proximity and the other for high-order proximity.

1. *Low-order proximity matrix.* The Laplacian eigenmaps [33] is a classic low-order proximity. According to [28], the Laplacian matrix of the hypergraph is defined as

$$\mathbf{S}_l = \frac{1}{2}(\mathbf{I} - \mathbf{D}_v^{-1/2}\mathbf{A}_\Gamma\mathbf{D}_v^{-1/2}), \tag{5}$$

   where $\mathbf{I}$ is an identity matrix.
2. *High-order proximity matrix.* There are many high-order proximity measures in graph [32], such as Katz index and adamic-adar, etc. Here, we choose the rooted pagerank value as the measure, which is indeed the probability that a random walk from a node and will locate another node in the steady state. Let $\mathbf{P}$ denote the probability transition matrix. Then the high-order proximity matrix is

$$\mathbf{S}_h = (1 - \alpha) \cdot (\mathbf{I} - \alpha\mathbf{P})^{-1}, \tag{6}$$

   where $\alpha \in [0, 1)$ is the probability to randomly walk to a neighbor. According to [28], the transition probability matrix of the hypergraph is

$$\mathbf{P} = \mathbf{D}_v^{-1}\mathbf{C}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{C}^\top,$$

   where $\mathbf{D}_e$ is a diagonal matrix containing hyperedge degrees, where the degree of an hyperedge is $\delta(e) = \sum_{v \in V^+} c(v, e)$.

After embedding the link groups into low dimension vectors, we run K-means on them and then we get link communities. The memberships of each node can be interpreted as the communities of links which connect to that node. In [21], authors point out that the membership can not be correctly assigned for highly-mixed community structures. To deal with this, we count the fraction of each node's incident edges that belong to communities as their *belonging coefficient*. Finally, the node is assigned to the communities whose *belonging coefficient* is larger than its average *belonging coefficient*.

6. Experimental validation

In this section, we present the comparison results of our method with six representative overlapping community detection algorithms. The algorithms to be compared are listed below.

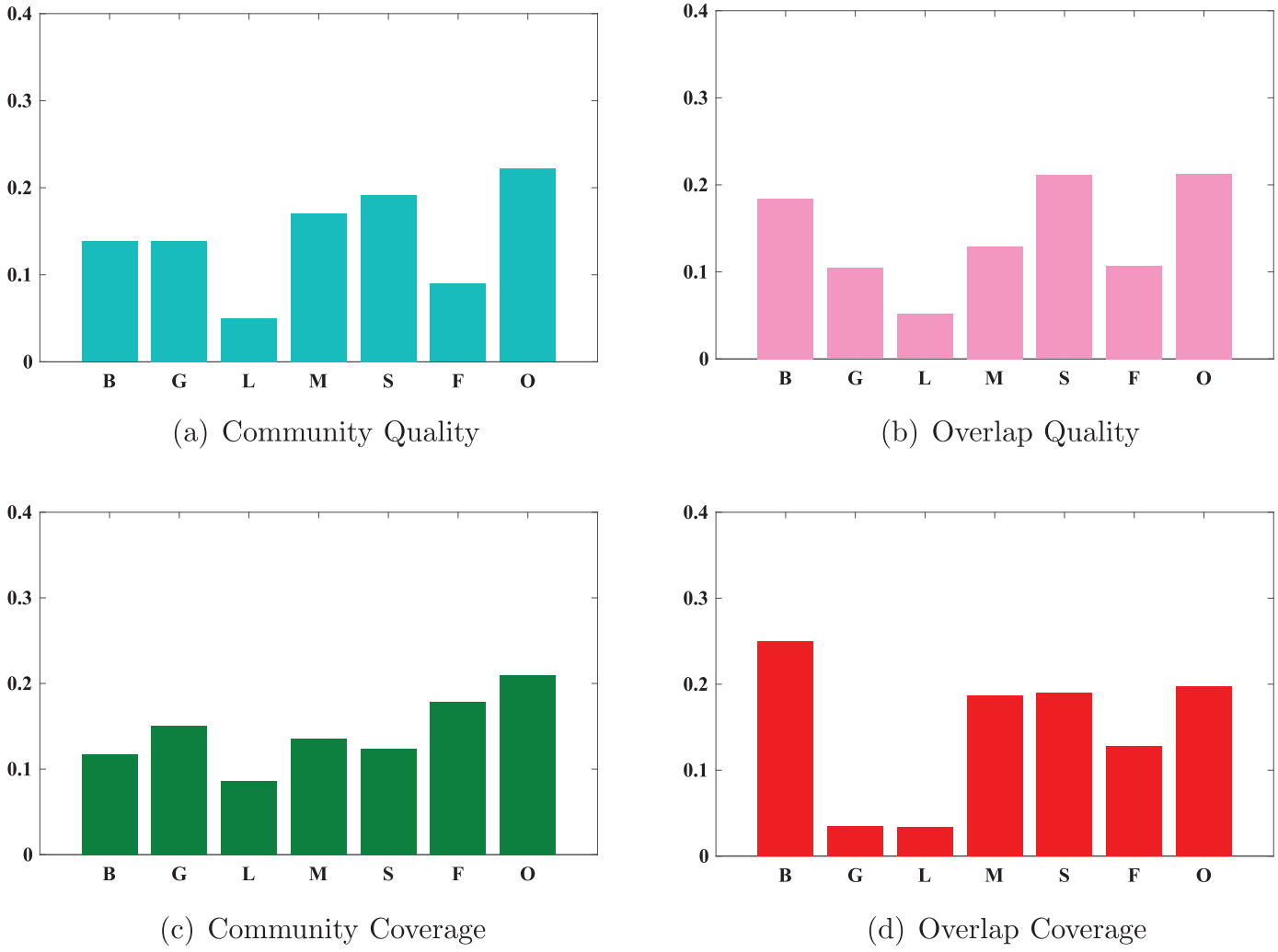

Fig. 2. Performance results on the word association network.

- CFinder (F): the famous Clique Percolation Method [10], where the size of k-cliques is set to 4.
- Link (L): the link clustering method [18].
- MOSES (M): the stochastic block model based local optimization scheme [15].
- GCE (G): the local greedy optimization strategy using k-cliques as seed [13], where the size of k-cliques is also set to 4.
- SVI (S): the mean-field variational inference model based on link sampling [34].
- BigClam (B): the cluster affiliation model by using the nonnegative matrix factorization [35].
- Ours (O): the proposed method where $\tau_s^* = 2$, $\tau_s^* = 0.5$ and 5 for embedding dimensions for fast computation.

In the experiments, we used the software package with its default settings provided by the authors for every compared algorithm. We implemented our method in Python.

6.1. Evaluation measures

When the ground-truth communities are unknown, modularity is the most widely used *internal* measure for evaluating both crisp and overlapping communities in the literature [36]. However, the performance evaluation based on modularity is likely to be misleading, because (i) communities with high modularity are usually obtained by merging small communities, that is, communities with small scales are often covered up, i.e., the resolution limit of modularity [37]; and (ii) communities with high modularity usually contain nodes that are connected more densely than those between different communities, which is not true in networks with pervasive and dense overlaps [8,18]. Hence, the modularity is not suitable for evaluating the results on highly dense overlapping and hierarchical networks that this paper works on.

Alternatively, we attempt to use the *external* measures for evaluations. We select three networks including some descriptive metadata associated with every node, and extract the implicit "ground-truth communities" information from the descriptive metadata. For example, an author in the scientific collaboration network is usually associated with several keywords such as social network analysis, data mining, information fusion, and so on. Intuitively, similarity between keywords of authors indicates whether they have common interest or not. So it is expected that authors with high similarity are in the same community. Note that this evaluation strategy is first used in [18] and subsequently used in [35]. In general, we construct four evaluation measures [18], of which the computational details will be introduced along with different datasets.

- *Community Quality:* It measures the similarity of nodes in pairs within each community compared to a null model. The similarity between nodes is carefully defined by the metadata of

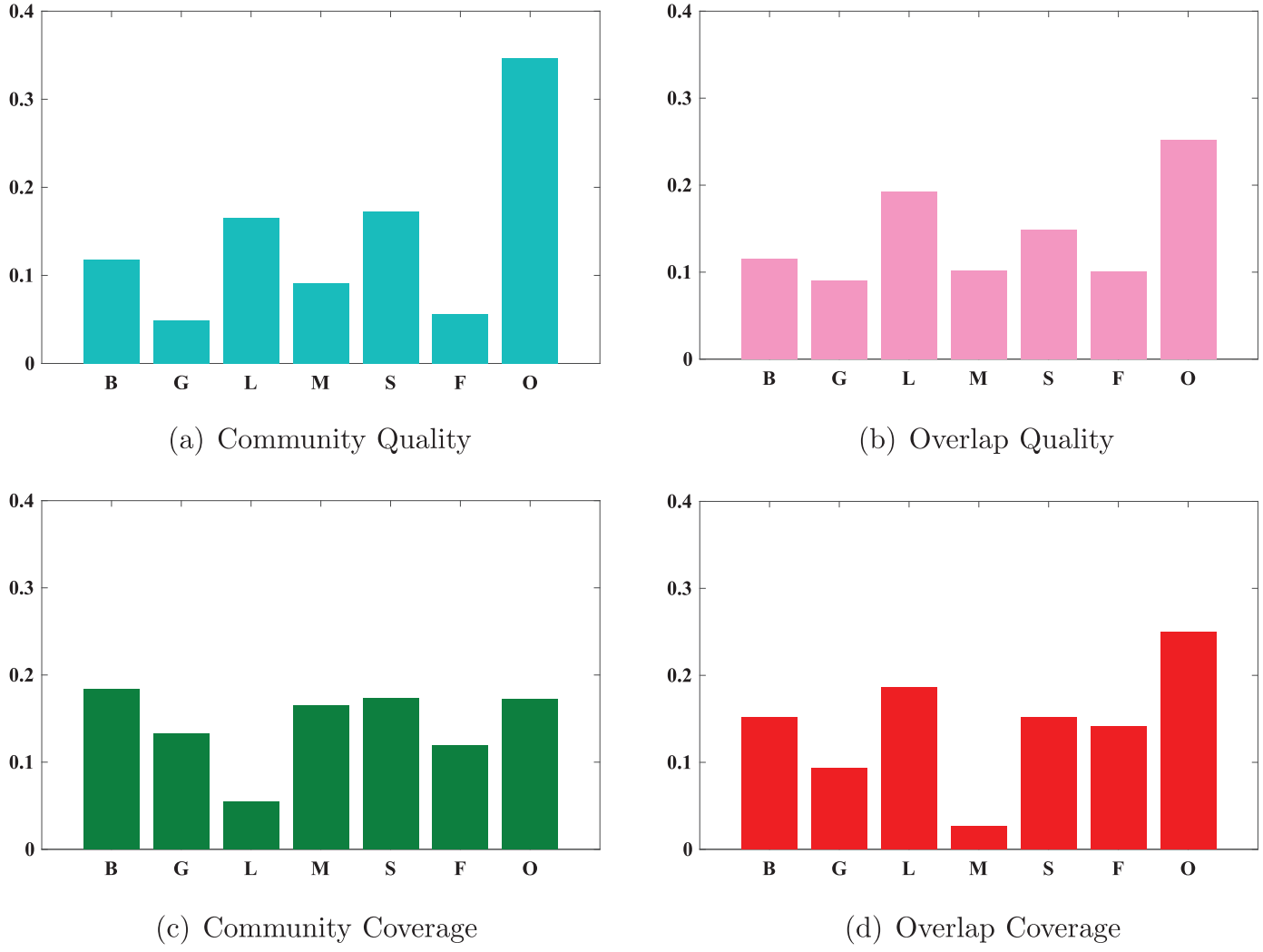

Fig. 3. Performance results on the scientific collaboration network.

different networks. The general form of the community quality is

$$\text{Comm\_Quality} = \frac{\sum_{v_i,v_j \text{within same comm.}} \mu_{ij}}{\sum_{v_i,v_j} \mu_{ij}}, \tag{7}$$

where μ_{ij} is the similarity between node v_i and v_j based on the metadata of the network.

- *Overlap Quality:* To measure the quality of overlapping nodes, we extract the true overlap for each node from metadata which corresponds to its real number of communities. Then, the *Mutual Information* which measures the information that two variables share is employed to relate true overlaps and the detected overlaps. Given the detected overlaps by a particular method, this measure indicates how much information about the true overlaps are gained.
- *Community Coverage:* It measures the fraction of nodes that belong to at least two communities.
- *Overlap Coverage:* This measure counts the average memberships of nodes.

Except the community coverage, the range of other three measures does not always lie in 0 and 1. Then for convenient of comparison, we normalize these measures into [0,1] by using the min-max normalization schema.

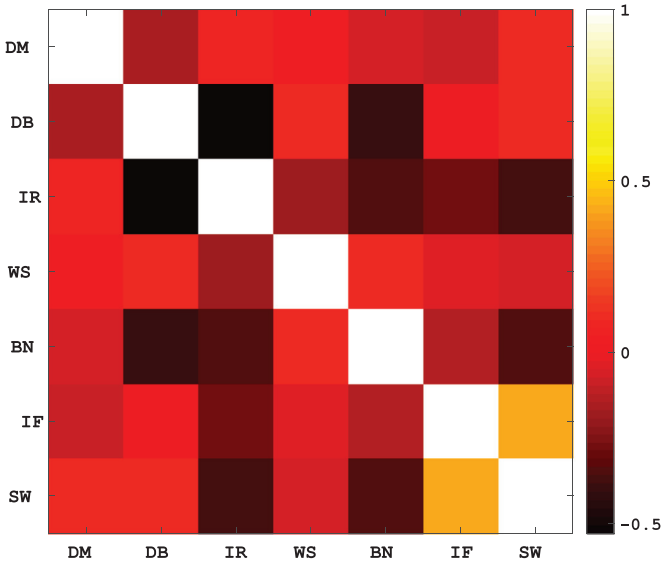

Fig. 4. Correlations between Topics.

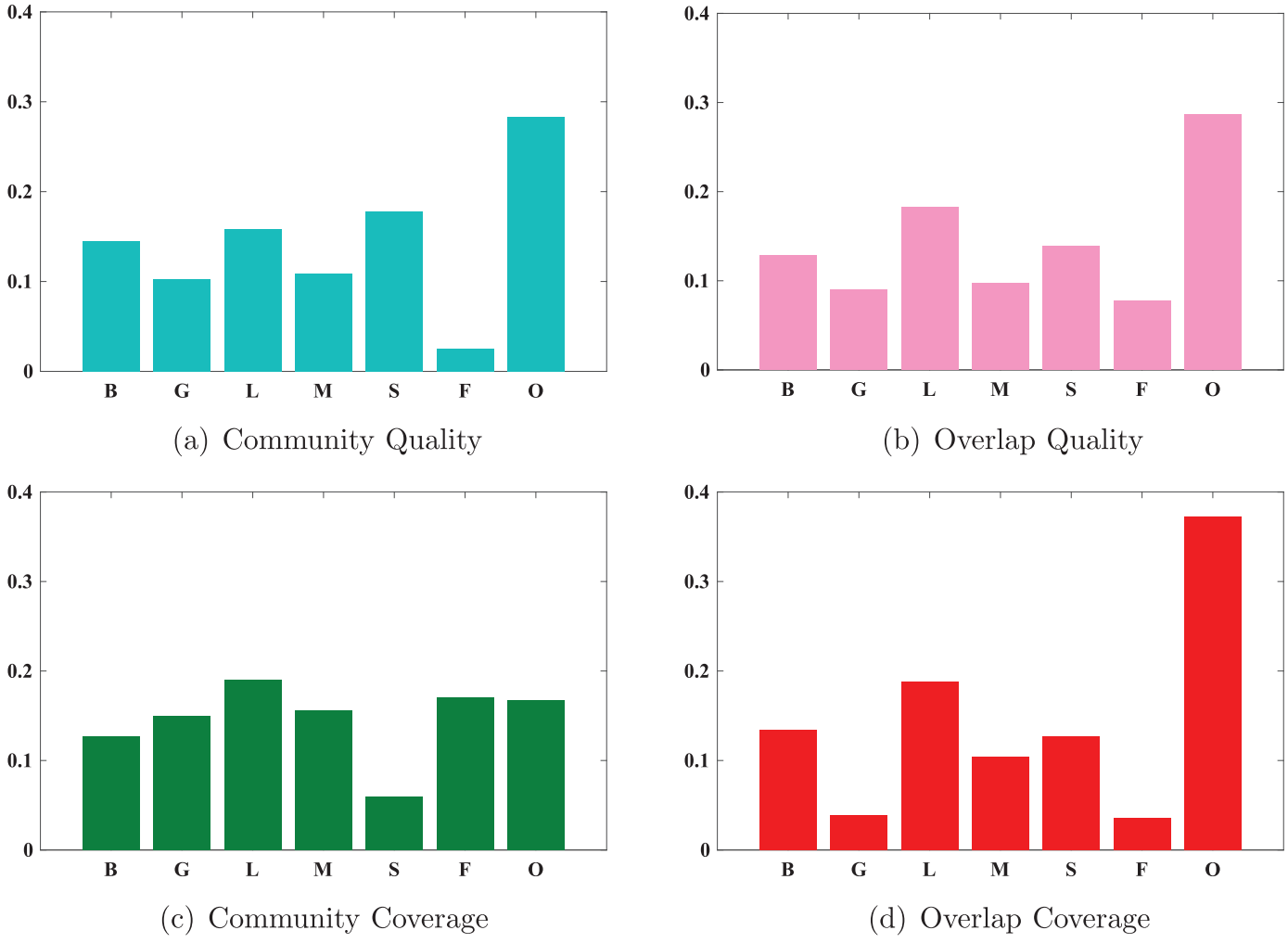

Fig. 5. Performance results on the user check-in network.

6.2. Word association network

Word Association network [38] is collected by participants who write down the word that first comes to their mind when presented a word. This network contains 7,094 nodes (words) and 31,771 edges (associations). As a lexical database for English, wordnet[1] is used as the metadata source for this network. We compute the similarity of words in pairs based on the shortest path that connects the senses in the is-a(hypernym/hypnoym) taxonomy. Then the community quality can be computed using Eq. (7). Since one word usually has multiple meanings, we use mutual information between the number of meanings for each word and its number of memberships detected by particular methods to quantify the overlap quality.

For this network, we set 100 for the size of communities. The results in Fig. 2 show that our method performs best on three measure, i.e., community quality, overlap quality and community coverage. In terms of overlap coverage, the BigClam algorithm is superior than other methods since the number of identified communities is very small in the word association network. In detail, BigClam only identifies 24 communities, which leads to a large number of overlapping nodes.

6.3. Scientific collaboration network

This network is extracted from ArnetMiner[2], where a node represents an author and an edge between authors indicates they publish at least one common paper. This co-author network consists of 4270 nodes and 15,055 edges from seven relevant research topics in computer science. In order to get more information for authors, we associated this network with a knowledge graph dataset, where top 100,000 frequent tags are extracted from the publications in ArnetMiner [39]. Tags such as feature extraction, data mining and support vector machine etc. are assigned to scholars, research organizations and publication venues according to their publications. To measure the community quality, the Jaccard similarity between the sets of tags of authors is used. Meanwhile, authors who publish papers in different topics will more easily participate in multiple collaborations. Hence, we use mutual information between the number of topics each author works in and its number of memberships to quantify the overlap quality.

We obtain 50 communities using our method in this network. Fig. 3 shows the comparison results of six methods in terms of four measures. As can be seen, communities detected by using our method are more meaningful (see Fig. 3(a)). Meanwhile, our method is also the winner on other three evaluation measures.

[1] https://wordnet.princeton.edu/.

[2] http://resource.aminer.org/lab-datasets/soinf/.

This indicates our method is able to discover more overlapping nodes that correspond to authors who have truly participated in multiple research fields.

Furthermore, we dive into the link communities detected by our method. In this network, label information about seven topics are given for each link, i.e., Data Mining (DM), Database Systems (DB), Information Retrieval (IR), Web Services (WS), Bayesian Networks (BN), Information Fusion (IF) and Semantic Web (SW). Then, we count the fraction of links belonging to the same topic for each community. And we define the consistency between a pair of topics in terms of the similarity for their fractions. The similarity is measured by the Pearson correlation. Fig. 4 shows the correlation between topics. The correlation between *Semantic Web* and *Information Fusion* is much stronger than other pairs of topics which is consistent with the truth that *Semantic Web* and *Information Fusion* are definitely closely related topics in computer science. This also validates that our method can detect meaningful communities. However, we can see that correlations between *Database Systems* and other topics, especially *Information Retrieval*, are all weak since the majority of links (about 44%) are related with *Database Systems*.

6.4. User check-in network

This network is collected from a popular location-based social network, i.e., Gowalla, in which users can check in and share their locations. We extract one month data and the network is consisted of 2704 nodes and 17,796 edges. We define the home location of each user as the average positions of his/her check-ins, i.e., the average coordinates of check-in latitudes and longitudes. Manual inspection has shown that this inference of the home location could achieve 85% accuracy [40]. To compute the similarity between users, we use the euclidean distance between their home locations based on the assumption that the social relationship is closer for users who are geographically closer. Given that more places the user visits, more likely he/she will join in different social communities. We use mutual information between the number of locations each user checked in and its number of memberships to quantify the overlap quality.

The size of communities is set 30 for this network. Comparing with the six baselines, our method achieves the best results on community quality, overlap quality and overlap coverage in Fig. 5. However, the Link method performs best on community coverage. For the communities detected by Link method, we find that the average size of communities is only about 6.4 (23.1 for our method) and the number of communities is 89 (30 for our method). In generally, the larger the number of communities is, the higher possibility of nodes can participate in multiple communities. And due to the far less nodes in each community for Link method, the node distribution will be sparser. In consequence, we can see that the Link method gets the largest community coverage (see Fig. 5(c)) but the coverage for overlaps is relatively small (see Fig. 5(d)).

Above all, the proposed method is robust since it shows the best performance on at least three measures for all datasets while the six baselines do not.

7. Conclusion

This paper proposes a novel method for detecting communities that preserves the overlapping and yet hierarchical structures in real life networks. The method consists in finding the close knit link groups which shows the hierarchical structures at different levels by mining the cosine patterns. Then, we represent the link groups as line hypergraph and embed them into a low dimension space based on the spectral method. Finally, we can get link communities by simply running *K*-means on link groups. Thus, overlapping communities can be naturally obtained by transforming the links into nodes. With “ground-truth communities” information on three real life networks, our experiments on four *external* measures indicate that our method holds great promise for highly dense and hierarchical networks.

Disclosure of conflicts of interest

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript

Acknowledgment

This work was supported in part by National Key Research and Development Program of China under Grant 2016YFB1000901, the National Natural Science Foundation of China (NSFC) under Grant 71571093, Grant 91646204, Grant 71372188, the National Center for International Joint Research on E-Business Information Processing under Grant 2013B01035 and Natural Science Foundation of Hubei Provincial under Grant 2018CFB314.

References

[1] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3) (2010) 75–174.
[2] V. Spirin, L.A. Mirny, Protein complexes and functional modules in molecular networks, Proc. Natl. Acad. Sci. 100 (21) (2003) 12123–12128, doi:10.1073/pnas.2032324100.
[3] S.P. Borgatti, A. Mehra, D.J. Brass, G. Labianca, Network analysis in the social sciences, Science 323 (5916) (2009) 892–895, doi:10.1126/science.1165821.
[4] Y. Ruan, D. Fuhry, S. Parthasarathy, Efficient community detection in large networks using content and links, in: Proceedings of the 22nd International Conference on World Wide Web, in: WWW '13, ACM, New York, NY, USA, 2013, pp. 1089–1098, doi:10.1145/2488388.2488483.
[5] M. Newman, M. Girvan, Finding and evaluating community structure in networks., Phys. Rev. E Stat. Nonlinear Soft Matter Phys. 69 2 Pt 2 (2004) 026113.
[6] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exp. 2008 (10) (2008) P10008.
[7] M. Rosvall, C.T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proc. Natl. Acad. Sci. 105 (4) (2008) 1118–1123, doi:10.1073/pnas.0706851105.
[8] J. Yang, J. Leskovec, Structure and overlaps of ground-truth communities in networks, ACM Trans. Intell. Syst. Technol. (TIST) 5 (2) (2014) 26.
[9] A. Lancichinetti, S. Fortunato, J. KertAsz, Detecting the overlapping and hierarchical community structure of complex networks, New J. Phys. 11 (3) (2009) 19–44.
[10] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (7043) (2005) 814–818.
[11] J. Xie, S. Kelley, B.K. Szymanski, Overlapping community detection in networks: the state of the art and comparative study, ACM Comput. Surv. 45 (4) (2013).
[12] W. Chen, Z. Liu, X. Sun, Y. Wang, A game-theoretic framework to identify overlapping communities in social networks, Data Min. Knowl. Discov. 21 (2010) 224–240.
[13] C. Lee, F. Reid, A. McDaid, N. Hurley, Detecting highly overlapping community structure by greedy clique expansion, 2010 arXiv preprint arXiv: 1002.1827.
[14] S. Zhang, R.-S. Wang, X.-S. Zhang, Identification of overlapping community structure in complex networks using fuzzy c-means clustering, Physica A Stat. Mech. Appl. 374 (1) (2007) 483–490, doi:10.1016/j.physa.2006.07.023.
[15] A. McDaid, N. Hurley, Detecting highly overlapping communities with model-based overlapping seed expansion, in: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, 2010, pp. 112–119, doi:10.1109/ASONAM.2010.77.

[16] J. Xie, B.K. Szymanski, X. Liu, Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process, in: Proceedings of the IEEE 11th International Conference on Data Mining Workshops, 2011, pp. 344–349.
[17] U.N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks., Phys. Rev. E Stat. Nonlinear Soft Matter Phys. 76 3 Pt 2 (2007) 036106.
[18] Y.-Y. Ahn, J.P. Bagrow, S. Lehmann, Link communities reveal multiscale complexity in networks, Nature 466 (7307) (2010) 761–764.
[19] T.S. Evans, R. Lambiotte, Line graphs, link partitions, and overlapping communities, Phys. Rev. E 80 (1) (2009) 016105.
[20] T.S. Evans, R. Lambiotte, Line graphs of weighted networks for overlapping communities, Eur. Phys. J. B 77 (2) (2010) 265–272.
[21] J. Kim, S. Lim, J. Lee, B.S. Lee, Linkblackhole*: robust overlapping community detection using link embedding, IEEE Trans. Knowl. Data Eng. (2018) 1.
[22] D. He, D. Jin, C. Baquero, D. Liu, Link community detection using generative model and nonnegative matrix factorization, PLOS One 9 (1) (2014).
[23] X. Bai, P. Yang, X. Shi, An overlapping community detection algorithm based on density peaks, Neurocomputing 226 (2017) 7–15.
[24] S. Lim, S. Ryu, S. Kwon, K. Jung, J.G. Lee, Linkscan*: overlapping community detection using the link-space transformation, in: Proceedings of the IEEE International Conference on Data Engineering, 2014, pp. 292–303.
[25] J. Cao, Z. Wu, J. Wu, Scaling up cosine interesting pattern discovery: a depth–first method, Inf. Sci. 266 (2014) 31–46.
[26] Z. Wu, J. Cao, J. Wu, Y. Wang, C. Liu, Detecting genuine communities from large-scale social networks: a pattern-based method, Comput. J. 57 (9) (2014) 1343–1357.
[27] M. Granovetter, The strength of weak ties: a network theory revisited, Sociol. Theory 1 (1983) 201–233.
[28] D. Zhou, J. Huang, B. Schölkopf, Learning with hypergraphs: clustering, classification, and embedding, in: Proceedings of the Advances in Neural Information Processing Systems 19, NIPS, 2006, pp. 1601–1608.
[29] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.
[30] P.D. Hoff, A.E. Raftery, M.S. Handcock, Latent space approaches to social network analysis, J. Am. Stat. Assoc. 97 (460) (2002) 1090–1098.
[31] M. Ou, P. Cui, J. Pei, Z. Zhang, W. Zhu, Asymmetric transitivity preserving graph embedding, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1105–1114. San Francisco, CA, USA, August 13–17, 2016
[32] L. Lu, T. Zhou, Link prediction in complex networks: a survey, Physica A Stat. Mech. Appl. 390 (6) (2011) 1150–1170.
[33] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: Proceedings of the Advances in Neural Information Processing Systems 14 NIPS, 2001, pp. 585–591.
[34] P.K. Gopalan, D.M. Blei, Efficient discovery of overlapping communities in massive networks, Proc. Natl. Acad. Sci. (2013) 201221839.
[35] J. Yang, J. Leskovec, Overlapping community detection at scale: a nonnegative matrix factorization approach, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, ACM, 2013, pp. 587–596.
[36] M.E.J. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. 103 (23) (2006) 8577–8582, doi:10.1073/pnas.0601602103.
[37] S. Fortunato, M. Barthélemy, Resolution limit in community detection, Proc. Natl. Acad. Sci. 104 (1) (2007) 36–41, doi:10.1073/pnas.0605965104.
[38] D.L. Nelson, C.L. McEvoy, T.A. Schreiber, The university of south florida free association, rhyme, and word fragment norms, Behav. Res. Methods Instrum. Comput. 36 (3) (2004) 402–407.
[39] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, Arnetminer: Extraction and mining of academic social networks, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '08, ACM, New York, NY, USA, 2008, pp. 990–998, doi:10.1145/1401890.1402008.
[40] E. Cho, S.A. Myers, J. Leskovec, Friendship and mobility: User movement in location-based social networks, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '11, ACM, New York, NY, USA, 2011, pp. 1082–1090, doi:10.1145/2020408.2020579.

Haicheng Tao received the M.S. degree in software engineering from University of Science and Technology of China in 2012. He is now a Ph.D. candidate of Nanjing University of Science and Technology. His current research interests include social network analysis and data mining.

Zhe Li is currently pursuing the Ph.D. degree with the School of Automatic Control, Northwestern Polytechnical University. He is currently a Lecturer with Hubei Engineering University. His current research interests include consistency control of multi-agents in complex environments, multirates with multi-agent systems, and distributing control systems.

Zhiang Wu received the Ph.D. degree in computer science from Southeast University, Nanjing, China, in 2009. He is currently a Professor with the Jiangsu Provincial Key Laboratory of E-Business, Nanjing University of Finance and Economics, Nanjing. His current research interests include distributed computing, social network analysis, and data mining. Dr. Wu is a member of ACM and CCF.

Jie Cao received the Ph.D. degree from Southeast University, Nanjing, China, in 2002. He is currently a Professor and the Dean of the School of Information Engineering, Nanjing University of Finance and Economics, Nanjing. His current research interests include cloud computing, business intelligence, and data mining. Prof. Cao was a recipient of the Young and Mid-Aged Expert with Outstanding Contribution in Jiangsu Province and was selected in the Program for New Century Excellent Talents in University.