COMMUNITY DETECTION ALGORITHM BASED ON LOCAL EXPANSION K-MEANS

L. Li[*], *K. Fan*[*], *Z. Zhang*[†], *Z. Xia*[‡]

Abstract: Community structure implies some features in various real-world networks, and these features can help us to analysis structural and functional properties in the complex system. It has been proved that the classic k-means algorithm can efficiently cluster nodes into communities. However, initial seeds decide the efficiency of the k-means, especially when detecting communities with different sizes. To solve this problem, we improve the classic community detection algorithm with Principal Component Analysis(PCA) mapping and local expansion k-means. Since PCA can preserve the distance information of every node pairs, the improved algorithm use PCA to map nodes in the complex network into lower dimension European space, and then detect initial seeds for k-means using the improved local expansion strategy. Based on the chosen initial seeds, the k-means algorithm can cluster nodes into communities. We apply the proposed algorithm in real-world and artificial networks, the results imply that the improved algorithm is efficient to detect communities and is robust to the initial seed of K-means.

Key words: *community structure, PCA, Local Expansion Strategy*

Received: April 13, 2016 **DOI:** 10.14311/NNW.2016.26.034
Revised and accepted: October 3, 2016

1. Introduction

Recently, many researchers notice that the complex network is a proper tools to describe variety of complex system in the real world [20–22], and thus the complex network has attracted the great attention in many fields such as physics, biology and social network et al. In complex network field, one of the important topology property is community structure which comprise of densely connected nodes, and some researchers have found that detecting community structure can reveal some valuable insights of the functional feature in the complex system [10, 17]. For example, communities in multimedia social network may imply people with the same hobby and trust relationship. Zhiyong Zhang et al. proposed an approach

[*] Lin Li, Kefeng Fan – Corresponding author, Research Center of Information Security, China Electronics Standardization Institute, Beijing, China, E-mail: lilincesi@126.com, kefengfan@163.com

[†] Zhiyong Zhang, Information Engineering College, Henan University of Science and Technology, Luoyang, China, E-mail: xidianzzy@126.com

[‡] Zhengmin Xia School of Electronic Information and Electrical Engineering, Shanghai Jiao tong University, Shanghai, China E-mail: zhengminxia@sjtu.edu.cn

to analyse and detect credible potential path based on community in multimedia social networks [25], the approach can effectively and accurately mine potential paths of copyrighted digital content sharing. Zhiyong Zhang et al. also proposed a trust model based on small world theory which shows the widely application of community struction [26]. The community structure in biology field may cluster proteins with the same function. So that many methods have been proposed to reveal this topological property in complex networks.

Many researchers have focused on community structure detection and create many works to detect reasonable communities in the complex network. For example, Newman and Girvan proposed the GN algorithm to divide the network into communities iteratively. The GN algorithm use the definition of betweenness to present the weight of each edge and the Modularity to measure the results of the community detection [16]. However, the high computing complexity of GN algorithm lead to be less efficient in large network with thousands of nodes. To solve this problem, many methods treat the Modularity as the object function and use optimization theory to detect communities in the complex network. Fang Wei et al. proposed a local expansion algorithm using improved Moduarity to uncover local communities [4]. Gong Maoguo et al. use multiobjective optimization theory to improve the community detection algorithm and use a multiobjective evolutionary algorithm to detect communities [7]. These kinds of algorithms can detect reasonable communities, but most of the optimizing Modularity algorithms own the problem of resolution limit. Based on local network topology, Lancichinetti et al. optimise a fitness function which expresses the statistical feature of communities to uncover reasonable community structure [13], and the proposed method named OSLOM. To detect communities in a near linear time, Raghavan et al. proposed the label propagation algorithm (LPA). In LPA, all nodes hold a single label and the method employ a label propagation to update the label iteratively [18]. However, both of the OSLOM and LPA uncover the community structure using the local topology information, thus some nodes can not be included into the corrected communities.

Since k-means algorithm is efficient to cluster nodes and fast convergence, some methods use k-means to detect community structure in the complex network. For instance, Jiang et al. proposed an algorithm based on page rank and k-means to reveal communities [9]. They map nodes of the complex network into European space and measure the node weight with Page Rank algorithm. Then the k-means algorithm cluster all nodes of the complex network into K groups. Since the sensitivity of the k-means algorithm to its initial seeds, any unreasonable initial seeds may lead k-means into the local optimum. Many improved k-means algorithms focus on choosing reasonable initial seeds to enhance the robustness of the k-means, but these improvements either need some empirical parameters or still heavily depend on the first chosen seed. Therefore, in this paper, we introduce an improved community detection algorithm based on PCA and local expansion k-means which can adaptively detect reasonable initial seeds. After PCA mapping, the dissimilarity of two nodes in the complex network can be represent by the space distance between points in low-dimension European space. Based on this feature, we map N nodes into low-dimensional Euclidean subspace with PCA and then propose a strategy to identify initial seeds. Using the chosen initial seeds, k-means algorithm clusters N nodes in the complex network into K communities. In this paper, we choose initial seeds of k-means with the propposed the loal expansion strategy, and then improved the community detection algorithm using the proposed strategy and k-means. The simulation results show that the proposed algorithm can detect communities more accurately in the complex network, like online social network and Internet.

2. Detecting communities based on local expansion k-means

In this paper, we define some follow notations to describe the algorithm. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected weight network. $\mathcal{V}$ is the node set with N nodes and $\mathcal{E}$ is the set of edge in the complex network. $\mathbf{A}$ is an weighted $N \times N$ adjacency matrix whose element a_{ij} means the weight of edge (i, j). If no edge exist between node i and j, the value of a_{ij} is 0.

The algorithm proposed in this paper includes three stages:first, calculate the European distance between nodes in the complex network. The distance of nodes belonging the same community is smaller than the nodes of different groups. And then we mapping N nodes into p-dimension space using PCA which can preserve the distance between nodes. Finally, we use the proposed the local expansion strategy to choose K initial seeds without any parameters. Finally, k-means algorithm is applied to uncover K communities in the complex network.

2.1 Nodes Mapping using PCA

Since nodes within the same community are densely connected, nodes belong to same community share more nearest neighbors. This knowledge implies that node i and j belonging to the same community own larger inner product $\langle \mathbf{a}_{i\cdot}, \mathbf{a}_{\cdot j} \rangle$, where vectors $\mathbf{a}_{i\cdot}$, $\mathbf{a}_{\cdot j}$ are the rows of the weighted adjacency matrix $\mathbf{A}$. So in this paper, we use the inner product $\langle \mathbf{a}_{i\cdot}, \mathbf{a}_{\cdot j} \rangle$ to represent the similarity between two nodes and construct similarity matrix $\mathbf{S} \in \mathbb{R}^{N\times N}$ whose element $s_{ij} = \langle \mathbf{a}_{i\cdot}, \mathbf{a}_{\cdot j} \rangle$. Notice that s_{ii}, $i = 1, \ldots, N$ is the degree of node i and $s_{ii} \geq s_{ij}, \forall i, j$. Then we define distance matrix $\mathbf{D} \in \mathbb{R}^{N\times N}$ whose elements are distances between two nodes:

$$d_{ij} = (s_{ii} + s_{jj} - 2s_{ij})^{1/2}. \tag{1}$$

We can notice that $d_{ii} = 0$ and $d_{ij} = d_{ji}$, so matrix $\mathbf{D}$ is a distance matrix of the complex network. The more similar two nodes are, the less distance between the two nodes in $\mathbf{D}$. According to D. J. Hand [8], PCA can be used to map each node into a lower dimension subspace while the Euclidean distance of two nodes in the lower dimension space is preserved as original length. Thus, in the lower dimension space, the nodes belonging to the same community are densely clustered after PCA mapping. Given the distance matrix $\mathbf{D}$ using (1), we use spectral decomposition to get the eigenvalue and eigenvector of $\mathbf{D}$: $\mathbf{D} = \mathbf{U\Lambda U}^{-1}$, where $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ is a diagonal matrix, λ_i is the i-th eigenvalue of $\mathbf{D}$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$, and each column of $\mathbf{U}$ is an eigenvector of $\mathbf{D}$. If the top p eigenvalues of $\mathbf{D}$ are larger than 0, $p = 1, \ldots, N$, we can create a matrix $\mathbf{P}$ whose columns are the top p eigenvectors $\alpha_1, \alpha_2, \ldots, \alpha_p$ corresponding to $\lambda_1, \lambda_2, \ldots, \lambda_p$ of $\mathbf{D}$. Then, the row vector of $\mathbf{P}$ is a p-dimensional coordinate of nodes in the Euclidean space. Using this PCA, we map N nodes into p-dimensional space and the distances between nodes are approximately preserved in the lower dimensional space.

2.2 Initial seeds detection with local expansion strategy

As mentioned before, k-means algorithm is efficient to cluster nodes and fast convergence. However, unreasonable initial seeds can make k-means converge to a local optimum. Many improved k-means algorithms focus on solving this problem. Two classic improvements are proposed to choose reasonable initial seeds: one is the k-means++ algorithm [1] and the other one is based on the given diameter [9]. k-means++ chooses the seeds as far away as possible and this rule leads to assigning K initial seeds into K distinct classes with great probability. But the unreasonable first initial seeds will still lead to a local optimum. Meanwhile, computational complexity is also a defect of k-means++. The second approach chooses initial seeds using a given parameter $\mathbf{R}$. One initial seed is chosen and then another point is chosen if the distance between them is larger than the given $\mathbf{R}$. This improvement assume that the diameter of all classes is $\mathbf{R}$. However, this assumption is unsuitable in community detection because the sizes of all communities are unknown and different from each other. According to the topology feature of community structure, we use the following three rules to detect the K initial seeds:

1. The Power Laws of the nodes degree causes the nodes with large degree is a small fraction. However, these nodes are usually the cores of the communities, this implies that these nodes with high degree should be the initial seeds;

2. Complete subgraph is a typical community structure in the complex network, thus choosing complete subgraph as the initial seeds is better than choosing a single node. The average weight of the complete subgraph g is calculated as
$$W_g = \frac{\sum_{i,j \in g} d_{ij}}{2 \times |g|}, \tag{2}$$
where $|g|$ is the count of edges whose two nodes belong to the same complete subgraph g. Smaller W_g reveals the subgraph with densely connected edges;

3. We involve nodes densely connected with the chosen core into the group until the local densely connections reach peak. This can keep the chosen initial seeds in different clusters. In this paper, we use the dense fitness function to measure the local dense connection:
$$F_{\mathcal{C}} = \frac{k_{\text{in}}^{\mathcal{C}}}{(k_{\text{in}}^{\mathcal{C}})^{\alpha} + (k_{\text{out}}^{\mathcal{C}})^{\beta}}. \tag{3}$$
Here, $\mathcal{C}$ is a node subset, $k_{\text{in}}^{\mathcal{C}}$ is the number of edges whose nodes are in the $\mathcal{C}$ and $k_{\text{out}}^{\mathcal{C}}$ is the count of edges that have one node in $\mathcal{C}$. α and β are parameters that control the weight of $k_{\text{in}}^{\mathcal{C}}$ and $k_{\text{out}}^{\mathcal{C}}$. Since we try to make local densely

connected nodes be a litter larger than a community which avoids having two initial seeds fall into one community, we define $\alpha = 0.9$ and $\beta = 1.1$. After many simulations, we notice that α and β will not affect the performance of the results if only α is a litter smaller than 0 and β is a litter larger than 0 — they do not need be changed according to different networks.

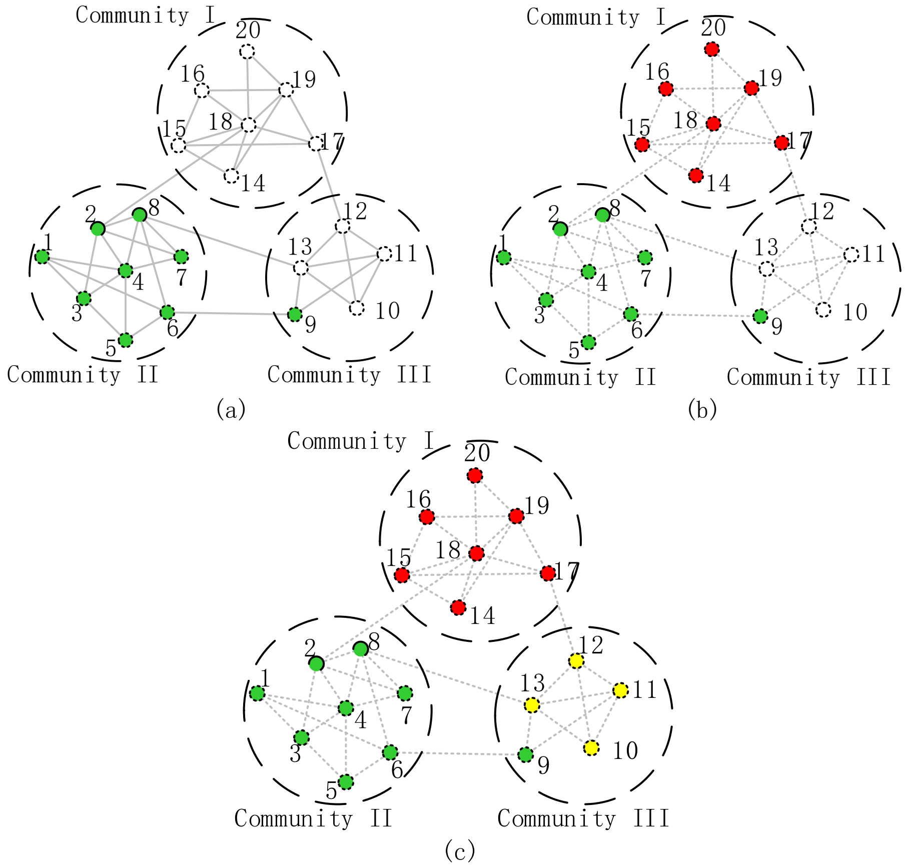
Based on the aforementioned three rules, we propose the local expansion strategy to detect initial seeds of k-means in Algorithm 1.

Algorithm 1 Local expansion strategy.

```
Ensure: $K$ initial seeds
  Set each node with the unselected label.
  Detect all complete subgraphs in the complex network as the candidate initial set.
  Calculate the average weight of all detected subgraph using (2).
  Sort these subgraph according to the average weight in descending order.
  $M \leftarrow 0$
  repeat
    Find the node with maximum degree in unselected nodes, and identify all complete subgraphs to which the chosen node belongs.
    From the identified complete subgraphs, choose the unmarked complete subgraph whose weight is maximum.
    Calculate the centrality of the chosen complete subgraph as one of the initial seeds of $k$-means.
    Initiate the group with the chosen complete subgraph, and use (3) to expand the group by involving the nodes local densely connected with the group into until none nodes can increase the $F_{\mathcal{C}}$ from Eq. (3).
    Mark all the selected nodes, and these nodes cannot be found in the following initial seed selection.
    $M \leftarrow M + 1$
  until $K$ initial seeds have been selected or no unmarked completed subgraph left
  if $M < K$ then
    Choose the left $K - M$ initial seeds as far away from the chosen $M$ seeds as possible.
  end if
  Output the $K$ initial seeds.
```

Because the local densely connected nodes is a litter larger than a community, maybe only $M \leq K$ initial seeds are chosen. In this paper, we use the strategy in Algorithm 1 to find the left $K - M$ initial seeds. The chosen M seeds make the left $K - M$ seeds be better than the strategy used in k-means++. The difference between the proposed local expansion strategy and the classic local expansion community detection algorithm is that we do not need find the groups of nodes with the largest fitness, thus the computing complexity is much less than classic local expansion algorithm. Based on the forementioned local expansion strategy, we can choose K reasonable initial seeds for k-means without changing any parameters.

Fig. 1 shows the workflow of the local expansion strategy in a toy network with three clusters and the chosen initial seeds is labeled by dashed line.



(a)

(b)

(c)

Fig. 1 *Local expansion strategy in a toy network with three communities. The expanded nodes are labeled with the same color. Notice that the expanded nodes is larger than a stand community.*

In Fig. 1(a), the first chosen initial seed is $(1, 3, 4)$ and then the initial seed is expanded using the local densely connected nodes. The expanded nodes are colored in green. We can notice that the expansion scope is a little larger than the community II, such that other initial seeds can not be the nodes in the community II. The second detected initial seed is $(18, 19, 20)$. Then as is shown in Fig. 1(b), the seed $(18, 19, 20)$ expands red color to all the nodes within community I. Finally, in Fig. 1(c), the third initial seed is $(10, 11, 13)$ and all the left nodes are colored by yellow.

2.3 Workflow of the local expansion k-means

Based on the forementioned concepts and formulas in Subsections 2.1 and 2.2, we proposed the community detection algorithm based on Local Expansion k-means as follows:

Algorithm 2 Workflow of the local expansion k-means.

```
Require: The weighted adjacency matrix $\mathbf{A}$
Ensure: Community set $C_{\max}$
  Calculate the similarity matrix $\mathbf{S}$ using the weighted adjacency matrix $\mathbf{A}$.
  Calculate the distance matrix $\mathbf{D}$ using (1) and matrix $\mathbf{S}$.
  Mapping all nodes of the complex network into $p$-dimensional Euclidean space
  with PCA.
  $\mathcal{C}_{\max} \leftarrow \emptyset$, $Q_{\max} \leftarrow 0$, $K \leftarrow K_{\min}$
  repeat
    Select $K$ initial seeds using the local expansion strategy (Algorithm 1).
    Identify $K$ communities $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$ with $k$-means algorithm.
    Calculate the the similarity-based modularity $Q_{\mathrm{s}}$.
    if $Q_{\mathrm{s}} > Q_{\max}$ then
      $Q_{\max} = Q_{\mathrm{s}}$.
      $\mathcal{C}_{\max} \leftarrow \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$.
    end if
  until $K > K_{\max}$
  Output the $\mathcal{C}_{\max}$.
```

In this paper, we use similarity-based modularity function Q_{s} to estimate the proper K. In general, the possible value of K is from 2 to $N-1$. Considering the proposed local expansion strategy, we notice that the maximum possible value of $K_{\max}$ is the number of the non-overlapping complete subgraphs and the minimum possible value $K_{\min}$ is $M-1$ where M is defined in the subsection 2.2. In Tab. I, we study the value of $K_{\min}$ and $K_{\max}$ in five real-world networks. We can notice that the scope of $[K_{\min}, K_{\max}]$ is much smaller than $[2, N-1]$.

Network	Nodes	$K_{\min}$	$K_{\max}$	Best K
Karate [23]	34	2	22	2
Dolphins [14]	61	3	40	105
Lesmis [11]	76	3	47	101
Football [16]	115	8	45	104
Collaboration network [5]	126	13	58	107

Tab. I *Scope of possible K in five real-world networks.*

3. Applications

In this section, we test the performance of the proposed algorithm in some real-world networks with known community partitions, such as Karate club network, the National Collegiate Athletic Association (NCAA) College-Football network. To validate the robust and efficient of the algorithm, we also apply the proposed algorithm into some artificial networks: Girvan-Newman artificial network [16] (GN artificial network) and LFR benchmark networks [13].

3.1 Karate network

The Zarchary's karate network is a well-known society network data. In the 1970s, Zarchary observed a karate club and described the members of the club as a network with 34 nodes and 78 edges. Nodes in the network are members of the club and the edges represent the social interactions of people. According to the observation, the club was divided into two smaller communities after a dispute between administrator and instructor.

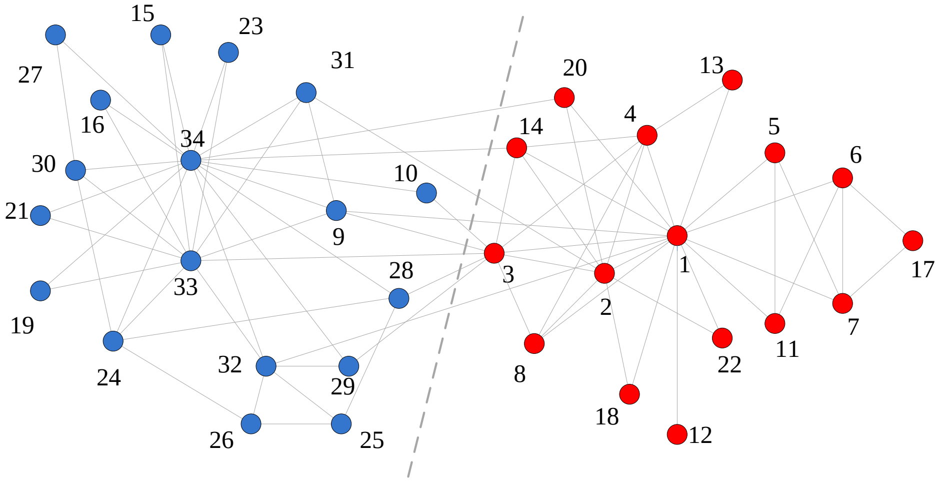


Fig. 2 *The testing result of the proposed method in the Zachary's Karate network.*

As shown in Fig. 2, the Zarchary's karate network is divided into two communities by the proposed algorithm. The blue community represent administrator group while the red one is instructor group. This partition exactly corresponds to the situation in the real-world and many works support this partition [6, 19, 24]. Comparing with the GN algorithm [16], the communities of administrator and instructor are successful detection. However, the GN algorithm also identifies the third community which include nodes 24, 25, 26, 28, 29, 32. We can notice that the third group is unreasonable, because node 24 more densely connects with the community of instructor.

3.2 NCAA College-football network

We use the NCAA College-football network as the second benchmark network to test the performance of the proposed algorithm. In [16], Girvan and Newman proposed the NCAA College-football network with 115 nodes and 613 edges. In this network, each node represents one college football team in NCAA, and edges represent the regular season game between the teams.

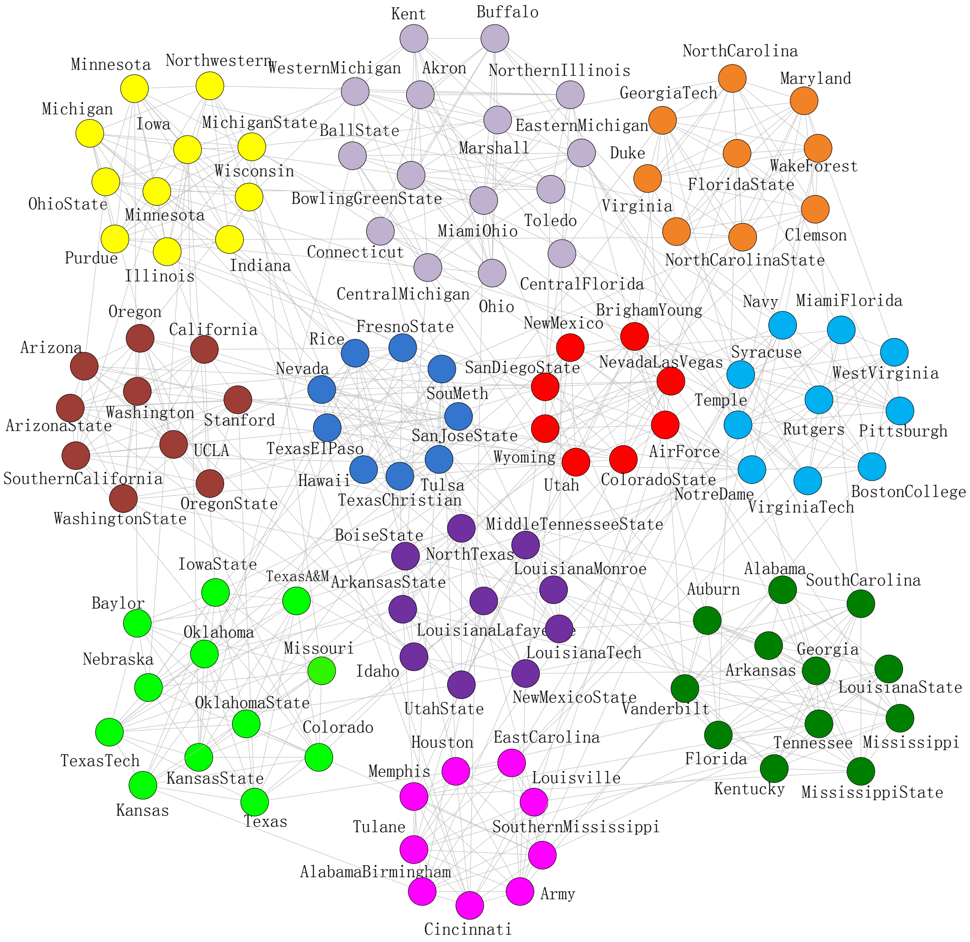


Fig. 3 *Clustering result of the proposed method in NCAA football network.*

As is shown in Fig. 3, the proposed algorithm divided the NCAA College-football network into eleven groups. The color of each node shows the community to which the node belongs, thus the nodes with the same color belong to the same community. The proposed algorithm correctly identifies five communities which are *Atlantic Coast*, *Big10*, *Big12*, *Pac10* and *Mountain West*. In the left six communities, no more than three nodes are misclassified in each group and total only 8 nodes are clustered into the incorrect groups. We can notice that some misclassified communities are also consistent with the definition of the community. For example, Fig. 3 shows that the dark blue community is a complete subgraph which implies Texas Christian is a reasonable member of the dark blue community, although the Texas Christian is not the member of the dark blue group in the real-world. Comparing with other algorithms, the result of the proposed method is better than GN algorithm in [16] and the classic LPA. For example, the GN algorithm identified 11 nodes into the incorrect groups and LPA misclassified 10 nodes.

3.3 Renren online social network

RenRen online social network is a famous online social network in China. It provides a platform for the public to submit their photo, blog and video. As most users of RenRen are students in the school, the classmate relationship is very common. Thus this character forms the communities which uncover the classmate relationship, pepple with the same hobby. Detecting communities in this social network can be used to advertising promotion or friends recommendation. Treating the first author of this paper as root node, we crawl the friend relationship in RenRen online social network from 2010 to 2013, and create a social network with 86357 nodes. Then we use the proposed algorithm to the RenRen online social network. The community detection result is shown in Fig. 4.

Due to the limited space, we only show a part of connected components in the social network. Outside the circle, the network is the backbone network of the RenRen social network, and each node is a detected community with different edge density. In the larger circle, we show five communities which are different classes in Shanghai Jiao tong university. In the backbone network, each node is a group people which comes from the same school or company. The communities in the backbone network show the classmate relationship of the first author of this paper when he studied in SDU, HUST and SJTU. In the larger circle, there are three SJTU labs which focus on three different academic fields, and the author of this paper works together with them. The A0803491 is the PhD. class, and the author Lin Li is also a member of the class. Community IV is a department of EE in SJTU, and the members in the community all focus on wireless communication. In Fig. 4, we can notice that the proposed algorithm can identify meaningful communities, and these commnities can be used to analyse social relationship in the social network or build trust model in the multimedia social network.

3.4 GN artificial network

The aforementioned analysis implies that the algorithm proposed in this paper can reveal reasonable communities in real-world networks. Although real-world networks can provide the reasonable explain for the simulation results, most of them do not have the unique correct partitions. In this subsection, we use two artificial networks to further test the performance of the proposed algorithm, One artificial network is GN artificial network which is proposed by Girvan and Newman, the other one is LFR benchmark networks proposed by Lancichinetti et al. Both of them can generate predetermined community structure, and the Normalized Mutual Information (NMI) [3] is used to evaluate the quality of the proposed algorithm.

We generate the GN artificial network which contains 128 nodes, and the 128

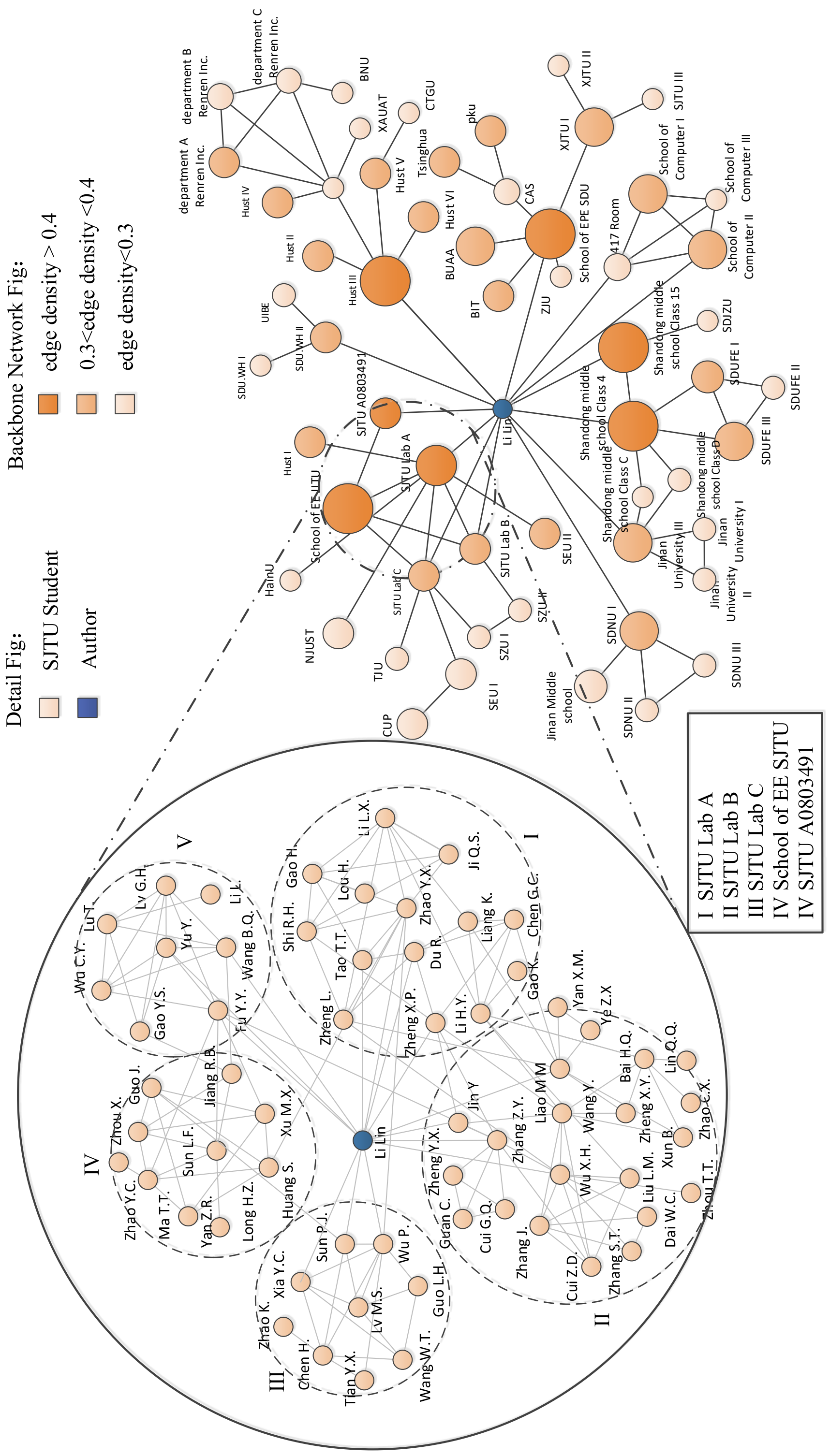


Fig. 4 *The testing result of the proposed algorithm in Renren online social network.*

nodes are divided into 4 equal communities. The node average degree $\bar{k}$ is 16 and the nodes are connected by edges with the probabilities depending on whether the two nodes belong to the same group or not. k_{in} shows the number of the edges on average connected with nodes belonging to the same community, and the k_{out} means the edges between communities, and $k_{\text{in}} + k_{\text{out}} = 16$.

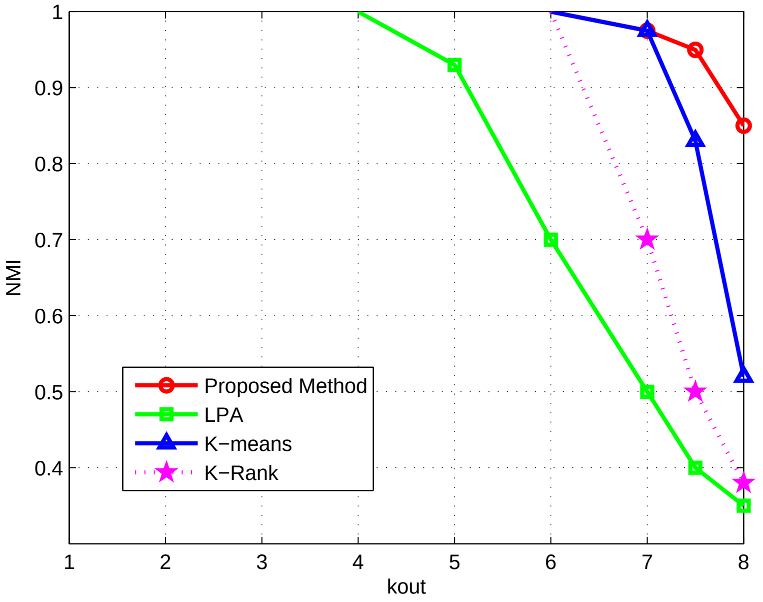


Fig. 5 *Clustering results of four testing methods in GN artificial network.*

In Fig. 5, the cluster results of four algorithms are shown. The four algorithms are: k-means [15], LPA [**?**], k-rank [**?**] and the proposed algorithm. As Fig. 5 shown, LPA performs worst when $k_{\text{out}} \geq 4$. This is because LPA uses random strategy to update the label of nodes, so the cluster results are not robust for different networks. When $k_{\text{out}} \geq 6$, the proposed algorithm is better than other two methods. This is because classic k-means algorithm uses global topology feature to cluster nodes, and the proposed local expansion strategy improves the robustness of k-means. When $k_{\text{out}} \geq 8$, each node has approximate intra-community edges and inter-community edges. This feature causes the unreasonable community definition, thus four of the algorithms fail to detect communities.

3.5 LFR benchmark network

The GN artificial network has a defect that all of its communities have equal size. To overcome this problem, Lancichinetti et al. [12] proposed the LFR benchmark network which is more realistic. The community size and node degree of the LFR benchmark network accord with heterogeneous distribution, and this feature reflects the real properties of the complex network. Thus in this paper, we further validate the proposed algorithm and compare the result with other four methods in LFR benchmark networks. The generated LFR benchmark network contains 1000 nodes and the community size follows two kinds: $[10, 50]$ and $[20, 100]$. The mixing parameter μ shows the fraction of the edges that connect with other communities. We still use NMI to verify the performance of the testing algorithms.

Fig. 6 shows the cluster results of the five algorithms in the LFR benchmark network with the community size ranging from 10 to 50. When $\mu \leq 0.5$, all the algorithms have similar performance. LPA cannot detect reasonable community structure when $\mu \geq 0.6$. When $\mu = 0.7$, the proposed algorithm works better than other algorithms except K-rank algorithm. We can notice that the proposed algorithm can also uncover communities when $\mu = 0.8$ which implies that the proposed algorithm is robustness for ambiguous communities.

Fig. 7 shows the simulation results of the five algorithms when community size ranges from 20 to 100. We notice that all of the five algorithms works well when $\mu \leq 0.4$. When mu changes from 0.5 to 0.7, GBLL [2], k-means and the proposed algorithms perform better than other two methods. The performance of the proposed algorithm is the best when $\mu = 0.8$.

In Fig. 8, we test the proposed algorithm in the LFR benchmark network with the community size changing from 20 to 50. Since all of the community is smaller than 50 which means the edges between each pair of communities is less densely, many algorithms work better than the results with large communities. As is shown in Fig. 8, all of the five algorithms perform well when μ is less than 0.5. The LPA cannot detect reasonable community structure when $\mu \geq 0.6$. When $\mu = 0.7$,

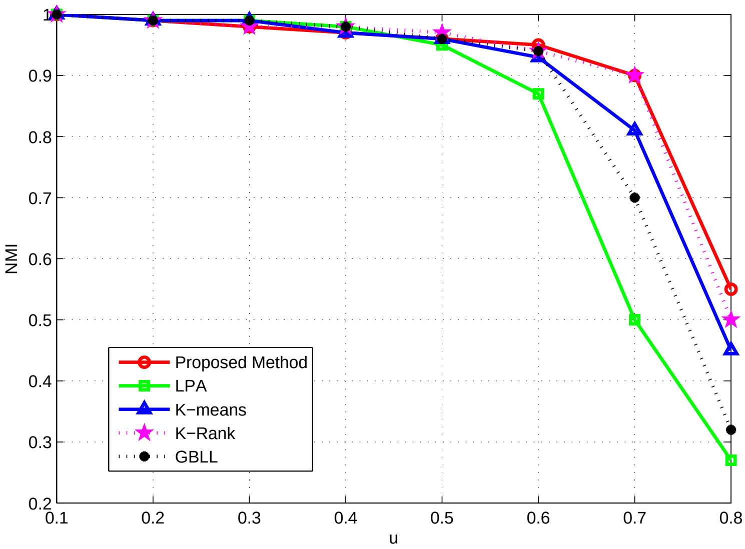


Fig. 6 *Clustering results of five testing methods in LFR artificial network. The community size ranges from 10 to 50.*

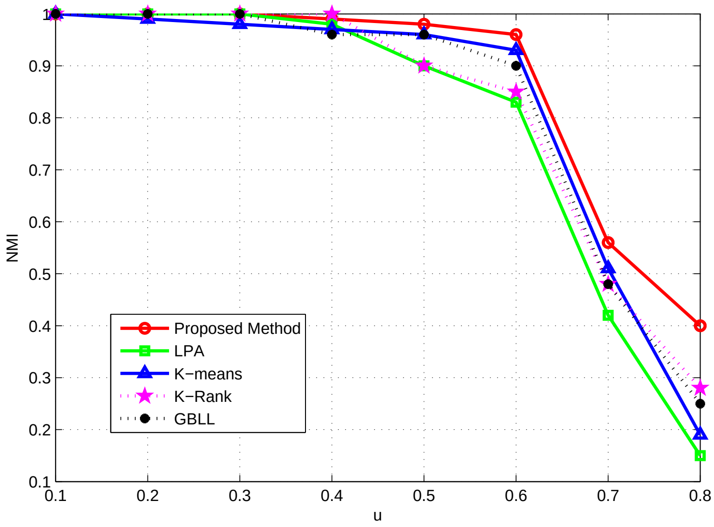


Fig. 7 *Clustering results of five testing methods in LFR artificial network. The community size ranges from 20 to 100.*

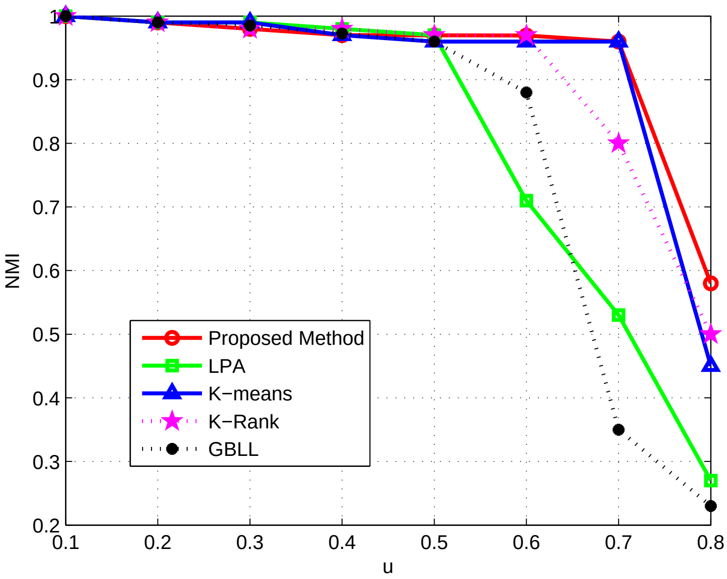


Fig. 8 *Clustering results of five testing methods in LFR artificial network. The community size ranges from 20 to 50.*

K-Rank algorithm works not better than k-means and the proposed algorithm. When $\mu = 0.8$, all of the algorithms fail to uncover reasonable community structure. Based on Figs. 6-7, we can know that when μ is small which means the community structure is clear, the proposed algorithm performs similar with four of other algorithms. When μ is large which implies the ambiguous community structure, the proposed algorithm can still reveal some reasonable communities. Since the proposed algorithm choose more reasonable initial seeds, the simulation results are better than other algorithms.

4. Conclusion

In this paper, we introduce an algorithm to detect communities using PCA and local expansion k-means. We map network nodes into low-dimension space with PCA, and then based on the topology feature of the community structure, the local expansion strategy is proposed to choose reasonable initial seeds. Finally, the chosen initial seeds can improve the robustness of k-means when community structure is detected. Applying the proposed algorithm in many networks, we notice that the algorithm proposed in this paper can detect communities efficiently. Since the computing complexity and function of k-means is not suitable for large-scaling media social network, in the future, we will improve the proposed algorithm to uncover overlapping community structure in the large-scale social network.

Acknowledgement

Thank to Mark Newman's great help in finishing this paper. This work is supported by 2015 Intelligent Manufacturing Special Project: The Comprehensive Standardized Test, National Key Basic Research Program of China (2013CB329603). National Natural Science Foundation of China Grant No. 61370220, Program for Innovative Research Team (in Science and Technology) in University of Henan Province Grant No. 15IRTSTHN010, Plan For Scientific Innovation Talent of Henan Province Grant No. 134100510006.

References

[1] ARTHUR D., VASSILVITSKII S. k-means++: The advantages of careful seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms.* Society for Industrial and Applied Mathematics. 2007, pp. 1027–1035.

[2] BLONDEL V.D., GUILLAUME J.L., LAMBIOTTE R., ETIENNE L. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment.* 2008, 10, P10008, doi: 10.1088/1742-5468/2008/10/p10008.

[3] DANON L., GUILERA D.A., DUCH J., ALEX A. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment.* 2005, 2005(9), P09008, doi: 10.1088/1742-5468/2005/09/p09008.

[4] FANG W., CHEN W., MA L., ZHOU A.Y., Detecting Overlapping Community Structures in Networks with Global Partition and Local Expansion, Progress in WWW Research and Development. 2008, pp. 43–55, doi: 10.1007/978-3-540-78849-2_7.

[5] FRANK K.A. Mapping interactions within and between cohesive subgroups. *Social networks.* 1996, 18(2), pp. 93–119, doi: 10.1016/0378-8733(95)00257-x.

[6] FU X.H, LIU L.H, WANG C. Detection of community overlap according to belief propagation and conflict, *Physica A.* 2013, 392, pp. 941–952, doi: 10.1016/j.physa.2012.09.023.

[7] GONG M.G., MA L.J., ZHANG Q.F., JIAO L.C. Community detection in networks by using multiobjective evolutionary algorithm with decomposition, *Physica A.* 2012, 390, pp. 4050–4060, doi: 10.1109/cec.2011.5949886.

[8] HAND D.J., SMYTH P., MANNILA H. Principles of Data Mining, MIT Press, 2001. not complete!

[9] JIANG Y.W., JIA C., YU J. An efficient community detection method based on rank centrality. *Physica A: Statistical Mechanics and its Applications.* 2013, 392(9), pp. 2182–2194, doi: 10.1016/j.physa.2012.12.013.

[10] WU J., LU R., JIAO L., LIU F., YU X., WANG D., SUN B. Phase transition model for community detection, *Physica A.* 2013, 392, pp. 1287–1301, doi: 10.1016/j.physa.2012.11.032.

[11] KNUTH D.E. The Stanford GraphBase: A Platform for Combinatorial Computing, Addison-Wesley, Reading, MA. 1993, doi: 10.1109/cec.2011.5949876.

[12] LANCICHINETTI A., FORTUNATO S., RADICCHI F. Benchmark graphs for testing community detection algorithms. *Physical Review E.* 2008, 78(4), 046110, doi: 10.1103/physreve.78.046110.

[13] LANCICHINETTI A., RADICCHI F., RAMASCO J.J., FORTUNATO S. Finding statistically significant communities in networks. *PloS one.* 2011, 6(4), e18961, doi: 10.1371/journal.pone.0018961.

[14] LUSSEAU D., SCHNEIDER K., BOISSEAU O.J., PATTI H., ELISABETH S., STEVE M.D. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology.* 2003, 54(4), 396–405, doi: 10.1007/s00265-003-0651-y.

[15] MACQUEEN J. Some methods for classification and analysis of multivariate observations, *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.* 1967, 1(14), pp. 281–297.

[16] NEWMAN M.E.J., Girvan M. Finding and evaluating community structure in networks. *Physical review E.* 2004, 69(2), 026113, doi: 10.1360/zf2012-42-5-537.

[17] NGUYEN N.P., DINH T.N., XUAN Y., THAI M.T. Adaptive algorithms for detecting community structure in dynamic social networks, INFOCOM, *Proceedings IEEE.* 2011, pp. 2282–2290, doi: 10.1109/infcom.2011.5935045.

[18] RAGHAVAN U.N., ALBERT R., KUMARA S. Near linear time algorithm to detect community structures in large-scale networks, *Physical Review E.* 2007, 76(3), 036106, doi: 10.1103/physreve.76.036106.

[19] SHANG R., BAI J., JIAO L., JIN C. Community detection based on modularity and an improved genetic algorithm. *Physica A: Statistical Mechanics and its Applications.* 2013, 392(5), pp. 1215-1231, doi: 10.1016/j.physa.2012.11.003.

[20] STROGATZ S.H., Exploring complex networks, *Nature.* 2001, 410(268), doi: https://doi.org/10.1038/35065725.

[21] WATTS D.J., STROGATZ S.H. Collective dynamics of “small-world” networks. *Nature.* 1998, 393(6684), pp. 440–442, doi: 10.1038/30918.

[22] XIE J., KELLEY S., SZYMANSKI B.K. Overlapping community detection in networks: The state-of-the-art and comparative study.*ACM Computing Surveys (CSUR).* 2013, 45(4), 43, doi: 10.1145/2501654.2501657.

[23] ZACHARY W.W. An information flow model for conflict and fission in small groups, *Journal of Anthropological Research.* 1977, 33, pp. 452–473, doi: 10.1086/jar.33.4.3629752.

[24] ZHANG Z., JIANG X., MA L., TANG S.T., ZHENG Z.M. Detecting communities in clustered networks based on group action on set. *Physica A: Statistical Mechanics and its Applications.* 2011, 390(6), pp. 1171–1181, doi: 10.1016/j.physa.2010.11.029.

[25] ZHANG Z.Y., WANG K.L. A Formal Analytic Approach to Credible Potential Path and Mining Algorithms for Multimedia Social Networks. *The Computer Journal.* 2015, 58(4), pp. 668–678, doi: 10.1093/comjnl/bxu035.

[26] ZHANG Z.Y, WANG K.L. A Trust Model for Multimedia Social Networks. *Social Networks Analysis and Mining.* 2013, 3(4), pp. 969–979, doi: 10.1007/s13278-012-0078-4.