Physica A 438 (2015) 321–334

Contents lists available at ScienceDirect

Physica A

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

A parameter-free community detection method based on centrality and dispersion of nodes in complex networks

Yafang Li, Caiyan Jia *, Jian Yu

Beijing Key Lab of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

HIGHLIGHTS

- Parameter-free to decide initial centers and the number of communities.
- A modified centrality measurement is presented.
- High accuracy of the proposed K-rank-D compared with other algorithms.

ARTICLE INFO

Article history:
Received 6 January 2015
Received in revised form 25 March 2015
Available online 9 July 2015

Keywords:
Clustering
Community detection
Rank centrality
Minimum distance
Complex network

ABSTRACT

K-means is a simple and efficient clustering algorithm to detect communities in networks. However, it may suffer from a bad choice of initial seeds (also called centers) that seriously affect the clustering accuracy and the convergence rate. Additionally, in K-means, the number of communities should be specified in advance. Till now, it is still an open problem on how to select initial seeds and how to determine the number of communities. In this study, a new parameter-free community detection method (named K-rank-D) was proposed. First, based on the fact that good initial seeds usually have high importance and are dispersedly located in a network, we proposed a modified PageRank centrality to evaluate the importance of a node, and drew a decision graph to depict the importance and the dispersion of nodes. Then, the initial seeds and the number of communities were selected from the decision graph actively and intuitively as the 'start' parameter of K-means. Experimental results on synthetic and real-world networks demonstrate the superior performance of our approach over competing methods for community detection.

1. Introduction

Networks are widely used to model complex relationships between individuals or organizations that are related to each other by various interdependencies like friendship, kinship, etc. To a network, community structure is an important character among its particular properties, e.g., small world [1], scale-free [2], and modularity [3,4]. A community indicates a set of nodes such that they are connected more closely than with nodes outside the community. Communities are also called clusters or modules, which are groups of vertices that probably share common properties and/or play similar roles within the network. Detecting communities is very important to understand the structure, function, and evolution of various complex networks [5]. It has attracted a lot of attention with widespread applications, such as identification of functional modules in biological networks [3,6–8], collection of pages dealing with the same or related topics on the web [9,10], grouping authors sharing similar research interest in co-author networks [11,12] and so on.

* Corresponding author.
E-mail address: cyjia@bjtu.edu.cn (C. Jia).

http://dx.doi.org/10.1016/j.physa.2015.06.043
0378-4371/

Numerous efforts have been made to extract community structure in complex networks, such as vertex and spectral clustering algorithms [13–17], modularity optimization algorithms [4,18–21], density-based algorithms [22,23], some other algorithms based on random walk [24–28] and statistical inference [29,30]. More detailed review can be referred to Ref. [31]. Among these algorithms, K-means [32] is a widely used clustering algorithm due to its efficiency and simplicity in practice. However, the results of K-means are seriously influenced by initial centers, which may lead to empty clusters or bad clustering results and seriously affect the rate of convergency. Additionally, it remains a problem to predefine a proper number of clusters.

To address the sensitivity issue, K-means++ [33] is introduced to extend K-means. It takes the first center from a data set randomly, it then chooses the other centers one by one with a probability. If a point is further from the selected centers, it has larger probability to be chosen as a new center. K-means++ has been shown to yield considerable improvement of convergence and computational time of K-means. However, K-means++ may choose outliers or data points located at low density area as cluster centers, which may lead to sub-optimal solutions. Besides, the first randomly chosen seed may not guarantee unique clustering result.

Recently, a novel density-based clustering algorithm is reported [34], which selects initial seeds based on local density peaks. The method is based on the idea that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with high density. We denote the method as Cluster-dp (the name comes from the Matlab code in Ref. [34]) in this paper for convenience. Cluster-dp allows us to select initial seeds and the number of clusters from a decision graph of a data set by hand tracking. But it requires us to set a cutoff distance d_c manually to calculate the local density of data points. Hence, a bad choice of d_c may mislead initial seeds. What is more, the algorithm cannot directly be applied to network data to incorporate its topological character in choice of initial seeds.

K-rank [35] is a representative vertex clustering algorithm for community detection, which takes node connectivity of a network into account to select initial seeds. The initializing procedure is comprised of two steps. First, it performs the PageRank algorithm to assign each node a rank value and sort them in descending order. It then tunes a parameter μ to make sure that the chosen seeds are located far from each other and have high PageRank values as well. However, it is hard to select a proper μ in a real application. Besides, we notice that if the PageRank values of nodes are almost equal, such as nodes in Girvan and Newman's artificial networks [3], initial seeds are likely to be chosen randomly. In this case, it could result in undesirable results.

Inspired by Cluster-dp which selects initial seeds that have anomalously large distance and higher local density by a decision graph, in this study, we propose an improved algorithm of K-rank, termed K-rank-D. It is a parameter-free method in deciding the number of communities and selecting initial centers that are expected to be dispersedly influential nodes in a network. It does not require parameters as Cluster-dp or K-rank does. The other difference between K-rank-D and Cluster-dp is that K-rank-D incorporates topological structure of nodes in a network to decide initial seeds. Experimental analysis on both real-world and synthetic data sets demonstrates that the initial seeds chosen by K-rank-D are reasonable. Besides, comparison results with other community detection algorithms indicate the good performance of K-rank-D.

The remainder of this paper is organized as follows. Section 2 presents our proposed K-rank-D algorithm. To verify the effectiveness of K-rank-D in initializing seeds and its performance to detect communities, several experiments on synthetic and real-world networks are carried out in Section 3. Section 4 draws the conclusion and gives further consideration.

2. Proposed algorithm

Our proposed algorithm assumes that initial key nodes (centers) have the following properties.

(1) *Outstanding*. The centers in a network are influential nodes, which are surrounded by low influential nodes. Similar with prototype-based clustering, in which each cluster has a prototype, each community is also regulated by a leader. The leader has high influence in its community, which can be reflected by its high centrality. Therefore, a node with high centrality is more likely to be chosen as initial seed.

(2) *Dispersedly-located*. The key initial seeds locate far from each other in a network. As each community is characterized by a center, the centers are expected to distribute evenly in the network. Thus, a center should have large distance from other centers.

To capture the initial key nodes characterized by the above two properties, a decision graph is drawn, in which one dimension evaluates the 'outstanding' of nodes and the other dimension characterizes the 'dispersedly-located' of nodes. Thus, K nodes distributed in the right upper part of the graph can be selected as initial centers. Then, vertex clustering algorithm Kmeans can be used to cluster nodes in the network into different communities. For executing the above process, the proposed K-rank-D algorithm needs to solve the following two problems. (1) Defining a centrality measurement to depict the centrality of nodes. (2) Giving a minimum distance measurement to describe the dispersion between nodes. In the study, we first transform a network into a geometrical structure of vectors, then a modified PageRank centrality was proposed to characterize the centrality of nodes. In addition, the minimum distance of each node with other nodes of higher centrality values was computed. The initial centers were then selected from the decision graph. With these carefully selected initial centers, nodes in the network were clustered by Kmeans algorithm.

2.1. Network transformation

Given a network $G = (V, E)$, in which $V = \{V_1, V_2, \ldots, V_n\}$ is a set of n nodes, E is a set of directed or undirected edges. The n nodes and their connections are interpreted by an adjacency matrix $A = [A_{ij}]_{n\times n}$, if node V_i and V_j are connected, $A_{ij} = 1$ for an unweighted network (or $A_{ij} = w_{ij} \neq 0$ for a weighted network, where w_{ij} is the weight between node V_i and V_j, indicating the connection strength between pairwise nodes), otherwise, $A_{ij} = 0$. In this paper, we use the term "node" and "vertex" interchangeably.

We first transform n nodes of a network into points of the same spatial space. One simple way is to calculate the pairwise similarities between nodes. Various similarity methods can be used [36], such as Jaccard similarity [37], cosine similarity [38] and so on. In this study, we adopt the signal similarity [39], which transfers the topological structure of a network into a geometrical structure of vectors in n-dimensional Euclidean space. To our knowledge and results of experimental tests, signal similarity has better performance than Jaccard, cosine similarity, etc.

The signal similarity is defined by a signaling propagating process. Each node is regarded as initial signal source to excite the whole network one time, all the other nodes record the amount of signals they have received. At each step, nodes send all their signals to their neighbors and themselves. After τ steps, the amount distribution of signals over the nodes can be taken as the influence of the source node on the whole network. General speaking, the source node influences its own community first and then affects other nodes by spreading signals. So obviously, nodes in the same community have similar effect on other nodes. Then the ith column of S represents the effect of node V_i to the whole network in τ steps. By normalizing S, we obtain n vectors $\bar{S}_1, \bar{S}_2, \ldots, \bar{S}_n$ in Euclidean space. The process can be described as

$$S = (A + I)^{\tau}, \qquad \bar{S}_{ij} = \frac{S_{ij}}{\sqrt{\sum_j S_{ij}^2}}, \tag{1}$$

where I is an n-dimensional identity matrix, τ is the total steps in signaling propagation ($\tau = 3$ in implementation). Since networks are usually sparse, the computation of S is not time-consuming. Also, we can choose other fast similarity transformation, like heat diffusion [40] or local random walks [28,41]. Alternatively, a network can be mapped into a low dimensional representative space by using MDS [28,42,43] or DeepWalk [44], etc.

2.2. Initial seeds selection

2.2.1. Proposed centrality measurement

The PageRank centrality was originally developed by Brin and Page [45]. Assuming that a random walker follows the structure of a network by the transition matrix P and sometimes randomly jumps to another vertex in the network with the probability of $\frac{1}{n}$. Then the n-dimensional PageRank vector v can be calculated by power method [46,47] to iteratively update

$$v^{t+1} = \left((1 - \beta)P + e\frac{\beta}{n}\right) v^t, \tag{2}$$

until

$$E = |v^{t+1} - v^t| = \sum_{i=1}^{n} (v_i^{t+1} - v_i^t)^2 < \epsilon, \tag{3}$$

where the PageRank values of all nodes are initialized to $\frac{1}{n}$ in the first step, e is an unit vector, β is the re-start probability, and P is the transition matrix, which is defined as

$$P_{ij} = \frac{A_{ij}}{\sum_j A_{ij}}. \tag{4}$$

Large PageRank value indicates high centrality of a node, thus, it is more important in a network and more likely to be chosen as a seed. However, if all nodes demonstrate little diversity in PageRank value, it is difficult to find prominent nodes as initial seeds. To address this issue, we give a new centrality measurement, PageRank-D, which is modified by node distance. It is defined as

$$\bar{P}_i = P_i + \sum_j exp\left(-\frac{d(i,j)^2}{P(i)}\right), \tag{5}$$

where $d(i, j)$ is the Euclidean distance between node V_i and node V_j. It is easy to notice that PageRank value of each node is modified by the distance of the node from other nodes in a network. This suggests that nodes locating close with more nodes will get larger rank value, which indicates that they are more important compared with surrounded nodes. By this modification, the PageRank-D values of nodes vary distinctly between nodes and influential nodes will stand out by this rank value in a network.

2.2.2. Minimum distance

We use the minimum distance δ_i ($i = 1, 2, \ldots, n$) to measure the degree of dispersion among centers, it is calculated by computing the distance between node V_i and other nodes with higher centrality,

$$\delta_i = \min_{j:\bar{P}_j > \bar{P}_i} (d_{ij}). \tag{6}$$

Specifically, if there exist some nodes with the same centrality values, nodes with smaller node ID are ranked higher. For node V_k with maximal rank value, $\bar{P}_k = max\,(\bar{P}_i)$, obviously, it is more outstanding compared with its neighbors, and it has the largest likelihood to be chosen as an initial center. Hence, we assign its minimum distance δ_k as

$$\delta_k = max\,(\delta_i), \quad i \neq k. \tag{7}$$

Based on the assumptions of our algorithm, the initial seeds are those nodes with high centrality values and locate dispersedly in a network. Therefore, we draw a decision graph in 2-dimensional space to decide the initial seeds, where one dimension is PageRank-D values of nodes, the other is the minimum distance of nodes defined above. Through the decision graph, nodes that are located right upper in the decision graph are figured out as the seeds.

In case that community structure of a network is fuzzy, it is very difficult to identify initial seeds positioned distinctly in the decision graph, we estimate the number of communities K by F statistics [35,39,48] to find out the top K nodes with high comprehensive value as initial seeds. Comprehensive value (CV) of node V_i is defined as follows,

$$CV(i) = \frac{\bar{P}_i \delta_i}{\max\limits_{1 \le i \le n} (\bar{P}_i) \max\limits_{1 \le i \le n} (\delta_i)}. \tag{8}$$

2.3. Community detection by K-means

With obtained initial K centers as above, we adopt K-means to implement the task of community detection. The K-means algorithm searches for a partition of n nodes into K clusters that minimizes the following within-cluster sum of squares,

$$J(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{K} U_{ij} \|\bar{S}_i - Z_j\|^2, \tag{9}$$

subject to

$$\sum_{j=1}^{K} U_{ij} = 1, \quad i = 1, 2, \ldots, n, \; U_{ij} \in \{0, 1\},$$

where U is an $n \times K$ partition matrix, indicating the cluster membership of n nodes on K clusters. $\bar{S}_i$ is the presentation of ith node in n-dimensional space as revealed in Section 2.1. If node $\bar{S}_i$ is allocated to the jth cluster, $U_{ij} = 1$, otherwise, the entry is 0. $Z = \{Z_1, Z_2, \ldots, Z_K\}$ is the set of centroids of K clusters. With initial centers obtained by Section 2.2, the indicator matrix is calculated by

$$U_{ij} = \begin{cases} 1, & \forall k \neq j, \|\bar{S}_i - Z_j\|^2 \le \|\bar{S}_i - Z_k\|^2, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

$U_{ij} = 1$ means that the ith node is assigned to the jth cluster. If the distance between a node and two cluster centers are equal, the node is arbitrarily assigned to the cluster with smaller cluster index number.

With partition matrix U fixed, the centroid of each cluster is updated as follows,

$$Z_k = \frac{\sum\limits_{i \in c_k} \bar{S}_i}{n_k}, \quad k = 1, 2, \ldots, K, \tag{11}$$

where c_k represents the set of nodes in the kth cluster, n_k is the size of the kth cluster.

The objective function J can be minimized by iterative optimization of partition matrix U and the cluster centers Z. It finally converges to a local optimum when the assignments of nodes no longer change [49].

2.4. K-rank-D algorithm and complexity analysis

Algorithm 1 outlines our proposed K-rank-D algorithm. It is based on PageRank-D centrality and minimum distance to find initial seeds. Taking the detected K initial seeds as input, the standard K-means algorithm is used to extract communities of a network.

Algorithm 1 K-rank-D algorithm

Input: the adjacency matrix A of a network G.

1: Transform the network into a geometrical structure of vectors in n-dimensional space by Eq. 1.
2: Calculate the PageRank-D value of each node by Eq. 1.
3: Compute minimum distance of each node by Eq. 6 and Eq. 7.
4: Draw a decision graph to spot initial seeds of the network G and obtain the number of communities K.
5: Perform K-means algorithm until the assignment of each node remains unchanged.

Output: K communities of the network G.

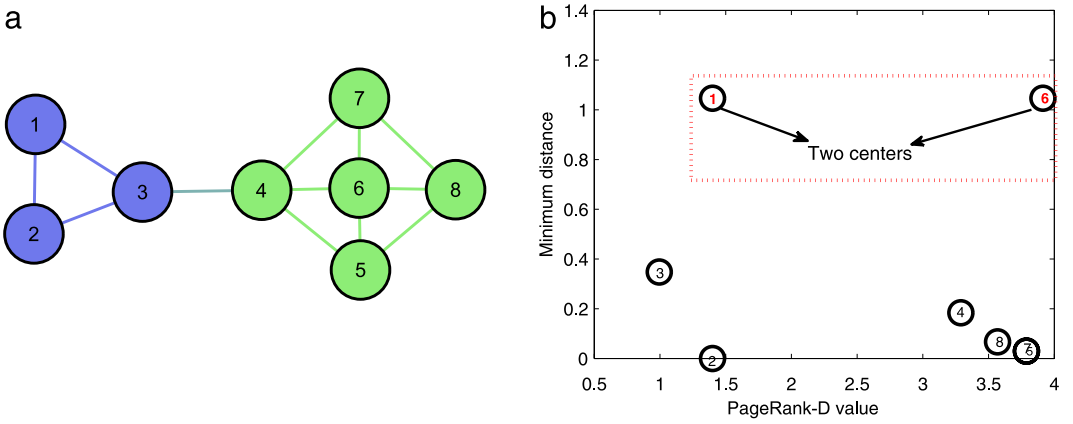

Fig. 1. Discovering communities in a network by K-rank-D. (a) The network consists of 8 nodes and 12 edges. (b) A decision graph is drawn according to centrality and minimum distance of nodes, where the horizontal axis indicates the modified PageRank value of nodes, and the vertical axis is the minimum distance of each node. By the decision graph, node 1 and node 6 are sought out as initial centers. With found initial centers, nodes are grouped into two clusters, where node 1, 2, and 3 are in one group, the remaining nodes form another group. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We provide an example in Fig. 1 to illustrate our proposed K-rank-D algorithm. Bearing in mind that initial centers should have high importance and locate dispersedly in a network, the PageRank-D value evaluating the importance of a node and the minimum distance characterizing the dispersion of nodes are used to draw the decision graph (Fig. 1(b)) of a network (Fig. 1(a)). By Fig. 1(b), nodes 1 and 6 "stand out" in the upper of the figure, indicating that these two nodes have large minimum distance and relatively large centrality values. Hence, they are selected as initial centers. After this operation, the initial centers and the number of communities ($K = 2$) were automatically determined as the 'start' parameter of K-means. Then, the network were separated into two communities (see Fig. 1(a), blue nodes form one community, green nodes form the other).

The complexity of K-rank-D involves three parts of this algorithm. To transform nodes in a network into Euclidean space (step 1), we simulate the signal propagating process directly in a network. The complexity of this step is $O(\tau(d+1)n^2)$, where d is the average degree of the network. In the initial seeds selection step, PageRank value can be easily calculated by power iteration method with complexity $O(mr)$, where m and r are respectively the number of edges and needed iterations. The modified PageRank value introduces node distance, which further induces the minimum distance. The complexity of distance computation is $O(n^2)$. Thus, the overall complexity of seeds initialization (step 2–4) is $O(mr + n^2)$. In the last step (step 5), we apply K-means to partition n nodes to K communities. The complexity of this step is $O(Kn^2t)$, where t is the number of iterations. With well chosen initial seeds, K-means can converge quickly and reduce required iterations largely in partitioning nodes of a network.

3. Experimental results and analysis

In this paper, to verify the effectiveness of our proposed K-rank-D algorithm, we compared K-rank-D against the state-of-the-art clustering algorithms K-rank, K-means, K-means++, and Cluster-dp. We also compared these vertex clustering algorithms with three well-known community detection methods, infomap [26], BGLL [50], and OSLOM [51]. In implementation, we fixed re-start probability $\beta = 0.15$ and $\epsilon = 10^{-8}$ to utilize PageRank algorithm. In deciding initial seeds, our proposed K-rank-D is parameter-free, but there are some parameters to be set for K-rank and Cluster-dp. Specifically, we tested μ in $[-1, -0.6]$ with step 0.1 for K-rank. For Cluster-dp, we ranged d_c such that the average number of neighbors was from 1% to 2% of the total nodes as revealed in Ref. [34]. We reported the best result of K-rank from different μ and that of Cluster-dp at certain d_c. The average results of K-means and K-means++ with 20 trials random initialization were recorded.

3.1. Evaluation measurements

In this study, Accuracy (ACC) [52] and Normalized Mutual Information (NMI) [52] are used to measure the performance of an algorithm. They are commonly used measurements defined as follows.

(1) **Accuracy (ACC)**. Given node V_i, l_{pi} is the assigned label by an algorithm, and l_{ti} is the true label. The accuracy is defined as the fraction of all nodes whose predicted labels are the same with the true labels. The ACC of a particular division of a network is defined as follows.

$$ACC = \frac{\sum_{i=1}^{n} \delta(l_{ti}, p_{map}(l_{pi}))}{n},$$

where $\delta(x, y)$ is a Kronecker function that the value is 1 if $x = y$, otherwise, 0. $p_{map}(l_{pi})$ is a permutation mapping function that maps the label l_{pi} of node V_i to the corresponding label in the ground-truth. n is the overall number of nodes in a network.

(2) **Normalized Mutual Information (NMI)**. The NMI is defined by

$$NMI(C, C') = \frac{-2\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij} \log \frac{n_{ij} n}{n_i^C n_j^{C'}}}{\left(\sum_{i=1}^{K} n_i^C \log \frac{n_i^C}{n}\right) + \left(\sum_{j=1}^{K'} n_j^{C'} \log \frac{n_j^{C'}}{n}\right)},$$

where C is the ground-truth cluster label, C' is the computed cluster label, K is the number of communities, n_i^C is the number of nodes in the ground-truth community i, $n_j^{C'}$ is the number of nodes in the computed community j, n_{ij} is the number of nodes in the ground-truth community i that are assigned to the computed community j.

The larger values of ACC and NMI, the better result an algorithm gets.

3.2. Data description

We used both synthetic and real-world networks to test the effectiveness of our proposed algorithm, K-rank-D. The details of these data sets are as follows.

(1) LFR networks [53]: Lancichinetti et al. present LFR benchmark networks, which are claimed to possess some basic statistical properties found in many real-world networks, such as power law distribution of the degree and community size. To specify the generated networks, several parameters are involved including N (number of vertices), μ_0 (mixing parameter), $\langle k \rangle$ (average degree), k_{max} (maximum degree of vertices), C_{min} (minimum community size), C_{max} (maximum community size), t_1 and t_2 (exponent of power-law distribution of nodes degree and community size, respectively). The mixing parameter μ_0 is defined such that every node connects a fraction of μ_0 links with other nodes outside its community, and shares $1 - \mu_0$ edges with nodes in its own community. That is to say, μ_0 determines how clear the community structure is. The smaller μ_0 is, the clearer community structure is. To evaluate the effectiveness of our proposed K-rank-D, we generate three groups of LFR benchmarks, of which the network sizes are set to 1000 (denoted as LFR1), 5000 (denoted as LFR2), and 10,000 (denoted as LFR3). Other parameters are set as follows: average degree $\langle k \rangle = 20$, maximum degree $k_{max} = 50$, $t_1 = 2, t_2 = 1$, community size ($[C_{min}, C_{max}]$) are respectively [10, 50], [20, 100], and [20, 100].

(2) GN networks [3]: The Girvan and Newman (GN) network has 128 vertices which are divided into four non-overlapping communities with 32 vertices respectively. The average degree of each node is $Z_{in} + Z_{out} = 16$. Namely, each node averagely has exactly 16 edges which randomly connect Z_{in} nodes in its own community and Z_{out} nodes in other communities. With the increase of Z_{out}, community structures become less clear and it is more challenging to partition the nodes into 4 groups.

(3) Real-world networks: Eight real-world networks are used in this section as summarized in Table 1. Zachary's karate club [54] is a social network representing personal relationship among 34 members at an American university in the 1970's. The club split into two communities since a conflict between the administrator and one of the teachers. Risk [55] is map of popular board games in 6 communities with 42 areas. Dolphins network [56] describes the frequent associations between 62 dolphins in a community living off Doubtful Sound and New Zealand. Lesmis [57] is a weighted network represents co-appearances of roles in Victor Hugo's novel "Les Miserables". US political books network [58] is a network of books about US politics published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com. Each vertex in the network is given one of labels "l", "n", or "c" to indicate its political attitude, "liberal", "neutral", or "conservative". Football network [3] contains the network of American football games (not soccer) between Division IA colleges during regular season Fall 2000. There are 115 nodes representing the football teams, which are divided into 12 conferences. An edge means there was a game between these two teams. Political

Table 1
The real-world networks for evaluation.

No.	Networks	n	m	K
1	Zachary's club [54]	34	78	2
2	Risk [55]	42	83	6
3	Dolphins [56]	62	159	2
4	Lesmis [57]	77	254	11
5	Political books [58]	105	441	3
6	Football [3]	115	613	12
7	Blogs [59]	1490	19515	2
8	PPI [60]	1628	11893	408

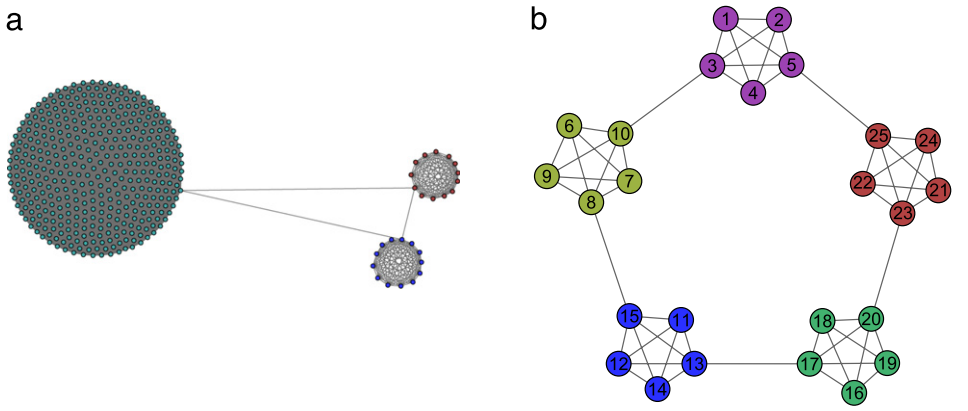

Fig. 2. Two special networks. (a) A big clique and two small cliques are linked by three single links. (b) An example of five cliques of the same size (five nodes) are linked by single links. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

blogs [59] is a directed network of hyperlinks between Weblogs on US politics. Each Weblog in the data set is associated with an attribute describing the political leaning of the Weblog, labeled as either "liberal" or "conservative". The PPI network [60] is built from 408 S. *cerevisiae* protein complexes.

(4) Two special networks. The first special network (denoted as Case 1) consists of three communities, one of which is very large, a clique of 400 nodes. The other two communities are much smaller, each of them are a clique of 13 nodes. These three cliques are linked together by three single links between any two communities (see Fig. 2(a)). The second special network looks like a ring which connects 100 small cliques of 5 nodes by single links. Fig. 2(b) give an example of five cliques. We denote this network as Case 2. This two networks are commonly used to verify the resolution limit of a community detection algorithm [61,62].

3.3. Experimental results and analysis

3.3.1. Initial seeds analysis

In this section, we analyze how K-rank-D identifies initial seeds based on the distribution of nodes on a 2-dimensional decision graph and demonstrate its effectiveness. Fig. 3 is an illustrative example on Zachary's Karate Club network. As depicted in Fig. 3(a), node 1 and node 34 circled with red dash "stand out" in the decision graph. According to the assumptions: initial centers should be "outstanding" and "dispersedly-located", these two nodes will be selected as initial seeds. Fig. 3(b) gives the final membership of nodes, where nodes colored the same are grouped into the same community. The results obtained by K-rank-D are totally identical to the ground-truth. In addition, node 1 and node 34 locate centrally in each community shown in Fig. 3(b). This indicates the importance (high centrality) of these two nodes in their own communities. Further, it proves that the centers found are reasonable.

In this paper, we proposed a new method to measure node centrality, PageRank-D, through which we obtain a decision graph to decide initial centers. Here is an example why we modify the original PageRank to PageRank-D. Taking a GN-type network with $Z_{out} = 1$ as an example. Fig. 4(a) is the original PageRank value of nodes, the PageRank value of each node is nearly the same. This adds more difficulty to identify initial seeds. Fig. 4(b) shows the modified PageRank value of each node. In Fig. 4(b), important nodes exhibit large rank values. Utilizing the PageRank-D value of each node, we draw the decision graph in Fig. 4(c). Four nodes stand distinctly in this graph. They are respectively node 20, 52, 60, and 101. According to the assumptions in this paper, these four nodes are chosen as initial seeds. From Fig. 4(d), we can also observe that they distribute evenly in four communities. Besides, the final membership of all nodes agrees with the ground-truth.

Cluster-dp is able to identify initial seeds by a decision graph. However, Cluster-dp imposes difficulty to cope with network data sets in practice. This may happen by two reasons: (1) A parameter d_c is needed to be tuned so as to calculate

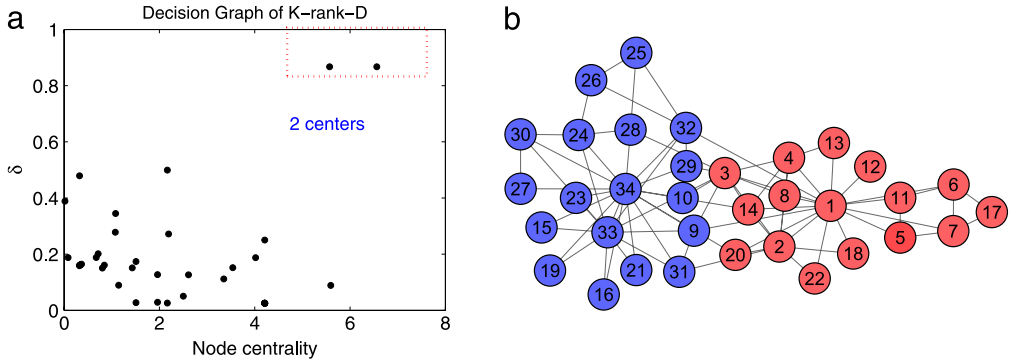

Fig. 3. An illustrative example on Zachary's Karate Club network. (a) The Decision graph of K-rank-D, node 1 and node 34 are circled as initial centers. (b) Results obtained by K-rank-D, the nodes in the same color are grouped into the same community, the membership of nodes is identical to the ground-truth. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

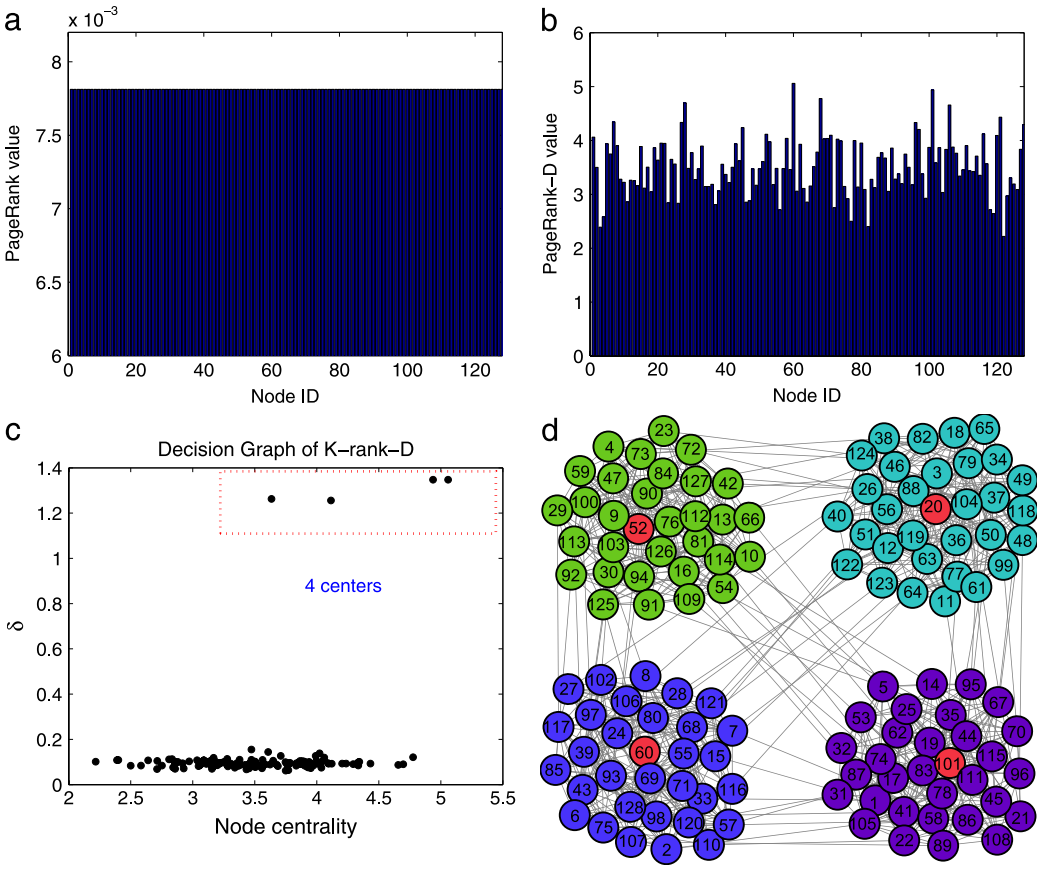

Fig. 4. A case study on a GN network with $Z_{out} = 1$. (a) PageRank value of each node, x axis indicates node ID, y axis is PageRank value. The PageRank value of all nodes are nearly the same. (b) PageRank-D value of each node obtained by our proposed method. (c) Decision graph. Nodes in red dash rectangle are initial centers identified by K-rank-D. (d) Distribution of four centers in the network. The nodes colored red are initial centers in four communities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the local density of nodes. (2) Cluster-dp does not take the topological property of nodes into account. Fig. 5 gives an example of an LFR network in case that μ_0 is 0.5. Cluster-dp found 43 initial seeds shown in Fig. 5(a) in contrast to the true 44 communities in this network. Our proposed K-rank-D discovered 44 initial centers demonstrated in Fig. 5(b) and the extracted communities exactly met the ground-truth.

We also give a case study on Political books data set. As introduced above, the Political books network is created according to the purchase history in Amazon.com. By Fig. 6, the ground-truth has three communities, which respectively represent the categories of books are conservative, neutral, and liberal. However, our proposed K-rank-D seeks out two distinctly

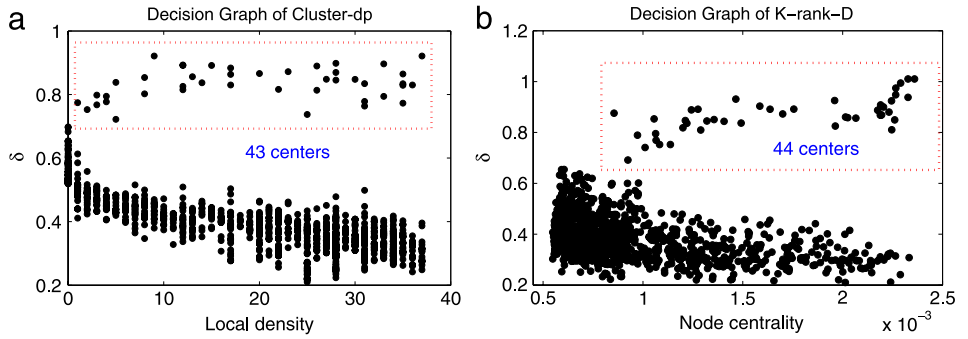

Fig. 5. Comparison of decision graphs on LFR1 ($n = 1000$, $mu = 0.5$) network. The ground-truth is 44 communities. (a) Decision graph by Cluster-dp. 43 initial centers in red rectangle are discovered. (b) Decision graph by our proposed K-rank-D, which identifies 44 initial centers and extracts 44 communities the same as the ground-truth.

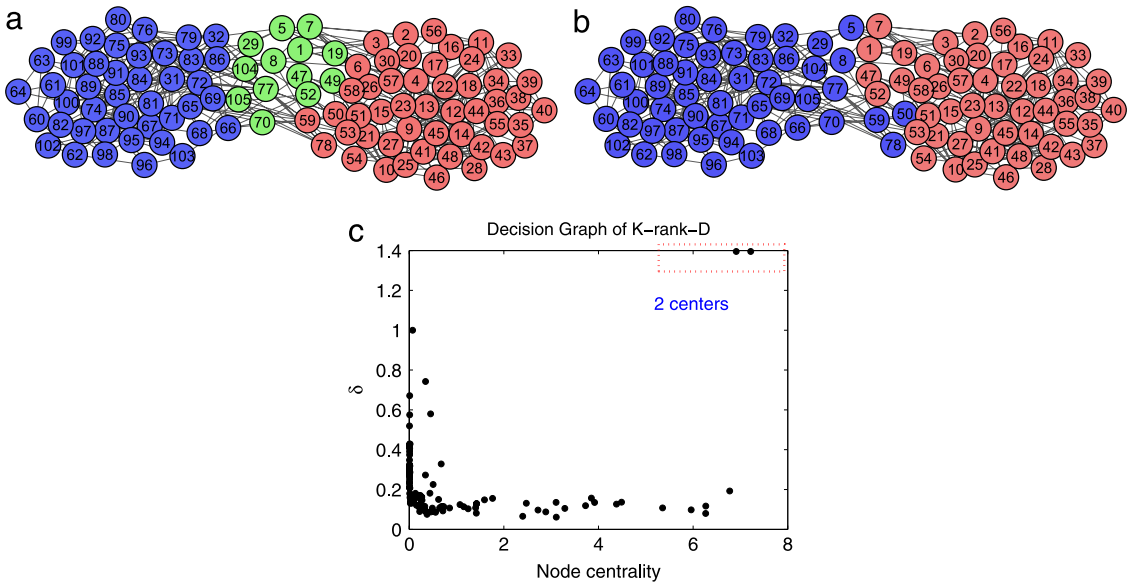

Fig. 6. Case study on the Political book network. (a) and (b) illustrate the ground-truth and communities discovered by our proposed K-rank-D, respectively. Nodes in the same color are clustered into the same group. (c) is the decision graph drawn by K-rank-D. Obviously, two nodes are spotted as initial seeds. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

positioned nodes, node 9 and node 73, to be initial seeds. Hence, it finally captures two cohesive connected clusters. Though it is not consistent with the ground-truth, according to the density of links in the 'neutral' community compared with the other two communities, nodes in the neutral may not be clustered into a single community. We also observe that the modularity of the ground-truth is 0.4149, but K-rank-D achieves 0.4454. This demonstrates that K-rank-D is able to discover more closely connected communities, and it further proves the number of clusters found by K-rank-D is reasonable.

3.3.2. Results on synthetic and real-world networks

To empirically evaluate the performance of our proposed K-rank-D, first, we compared K-rank-D with these algorithms on GN and LFR networks as shown in Figs. 7–10. By these figures, it is easy to notice that when community structure is clear, almost all methods perform well. As community structure gets fuzzier, their performance drops with different degree. What is more, we have the following conclusions from these figures. (1) K-means++ and K-means exhibit similar performance on GN and LFR networks, but K-means++ is more robust in some cases. The reason may be that K-means++ chooses centers far away from each other and K-means initializes seeds randomly. (2) K-rank shows good performance when community structure is clear, such as on GN networks when $Z_{out} \leq 6$ and on LFR networks when $\mu_0 \leq 0.5$, but the quality drops when $\mu_0 \geq 0.6$ on LFR networks. For Cluster-dp, its performance falls when $\mu_0 \geq 0.5$ on LFR2 networks and LFR3 networks. This may be owing to its sensitiveness to the cutoff distance d_c and it cannot be directly applied to deal with network data sets. Of all algorithms, the performance of BGLL drops when community structure is still clear (when $\mu_0 = 0.3$ on LFR networks), this may explain that the largest modularity will not always lead to the best results to some extent. (3) Infomap and OSLOM detect the exact communities when community structure is clear, but their performance drop dramatically

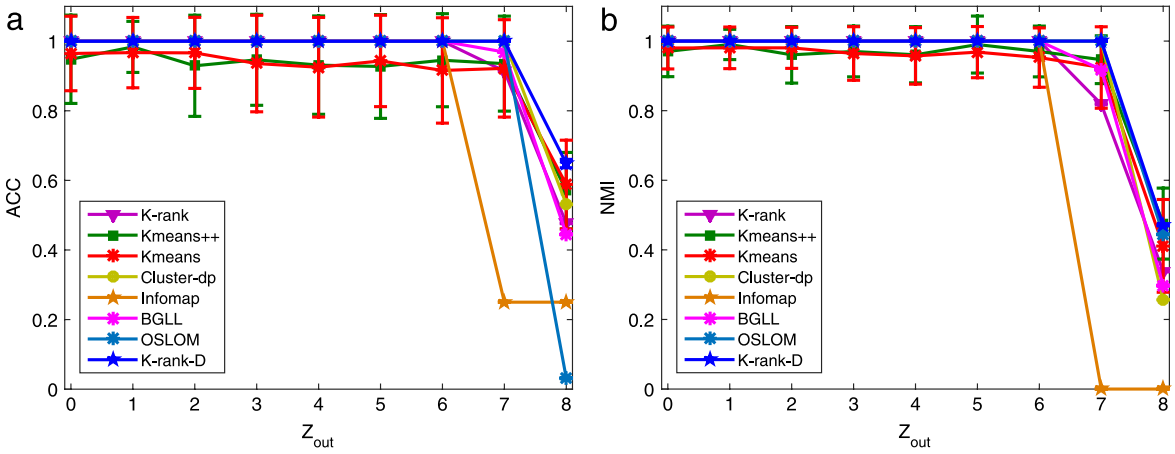

Fig. 7. ACC and NMI of different algorithms on GN networks. The horizontal axis Z_{out} indicates the number of edges a node connects with nodes outside its own community. With the increasing of Z_{out}, community structure becomes fuzzier.

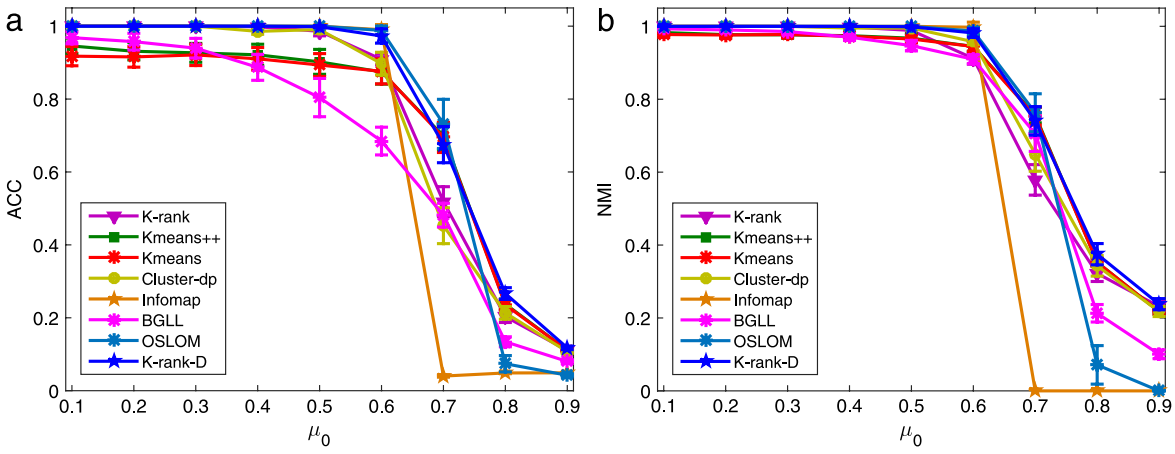

Fig. 8. Results on LFR1 ($n = 1000$) data sets. The horizontal axis is μ_0 ranging from 0.1 to 0.9. The vertical axis is the performance of different algorithms, where (a) is accuracy and (b) is NMI.

Table 2
Comparison results of different algorithms on special networks.

Special	Algorithms							
	K-rank	K-means++	K-means	Cluster-dp	infomap	BGLL	OSLOM	K-rank-D
	Accuracy ± std							
Case 1	$0.99_{\pm 0.00}$	$0.99_{\pm 0.00}$	$0.99_{\pm 0.01}$	$0.99_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.96_{\pm 0.00}$	$0.96_{\pm 0.00}$	$0.99_{\pm 0.00}$
Case 2	$1.00_{\pm 0.00}$	$0.89_{\pm 0.02}$	$0.96_{\pm 0.01}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.25_{\pm 0.00}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$
	NMI ± std							
Case 1	$0.91_{\pm 0.00}$	$0.91_{\pm 0.00}$	$0.90_{\pm 0.01}$	$0.91_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.86_{\pm 0.00}$	$0.66_{\pm 0.00}$	$0.91_{\pm 0.00}$
Case 2	$1.00_{\pm 0.00}$	$0.98_{\pm 0.00}$	$0.99_{\pm 0.00}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.82_{\pm 0.00}$	$1.00_{\pm 0.00}$	$1.00_{\pm 0.00}$

when community structure gets fuzzier. To be specific, when $Z_{out} \geq 7$ on GN networks and when $\mu_0 \geq 0.8$ on LFR networks (when $\mu_0 \geq 0.7$ on LFR1 networks), infomap is unable to discover communities and assign all nodes into one cluster. Similarly, OSLOM declines greatly when $\mu_0 \geq 0.8$. However, the performance of K-rank-D decreases slightly when $\mu_0 \geq 0.7$. One reason is that when community structure is not clear, K-rank-D encounters trouble to decide the exact number of communities. Though K-rank-D does not perform as well as infomap and OSLOM on LFR2 and LFR3 networks when $\mu_0 = 0.7$, K-rank-D is the best when $\mu_0 \geq 0.8$ in most cases, this may due to the good seeds initialization in K-rank-D. To conclude, by comparing with clustering algorithms (K-rank, Kmeans, Kmeans++, and Cluster-dp), infomap, BGLL, and OSLOM, K-rank-D has good performance to extract community structure and it is fit for the networks no matter whether they have clear community structure or not.

Table 2 gives results of different algorithms on two special networks. On these networks, our proposed K-rank-D is also capable of inferring the true number of communities correctly and shows good performance. This chart has two

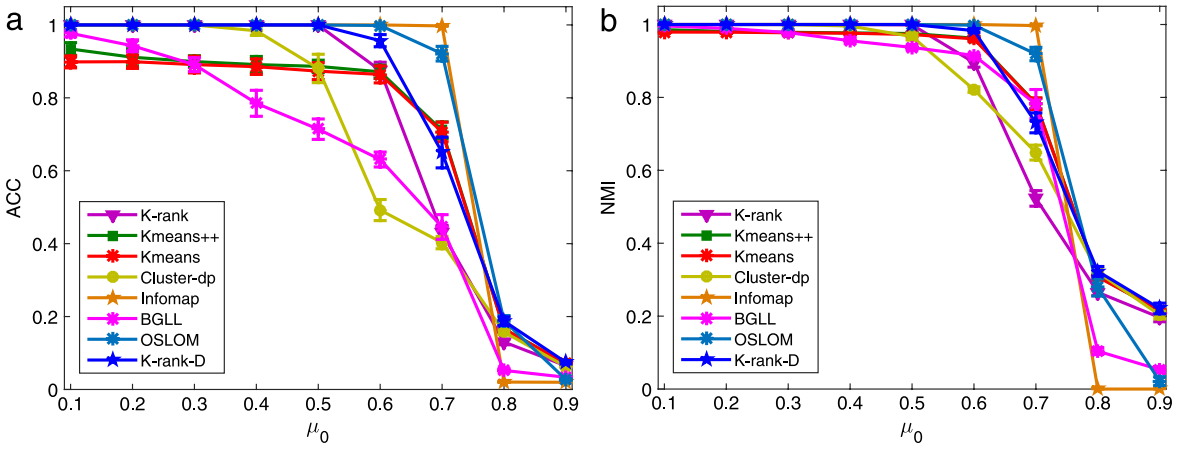

Fig. 9. Results on LFR2 ($n = 5000$) data sets. The horizontal axis is μ_0 ranging from 0.1 to 0.9. The vertical axis is the performance of different algorithms, where (a) is accuracy and (b) is NMI.

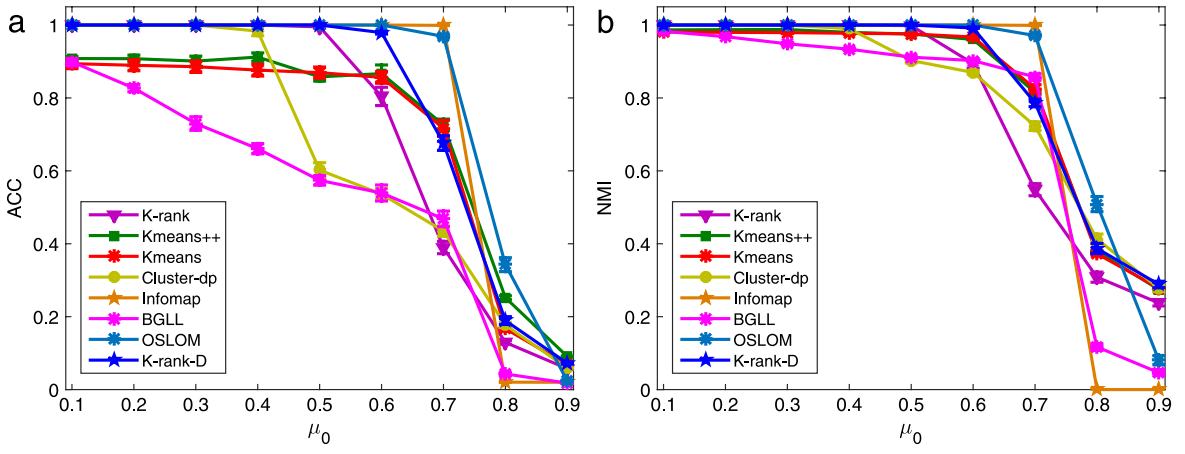

Fig. 10. Results on LFR3 ($n = 10{,}000$) data sets. The horizontal axis is μ_0 ranging from 0.1 to 0.9. The vertical axis is the performance of different algorithms, where (a) is accuracy and (b) is NMI.

interesting observations. Firstly, algorithms based on vertex clustering are not influenced by resolution limit as modularity maximization algorithms do [63]. Secondly, it is further verified that to acquire good clustering results, both *outstanding* and *dispersedly-located* should be considered to select initial seeds, as K-rank-D, K-rank, and Cluster-dp do.

We further conduct experiments on eight real-world networks to compare K-rank-D with K-means, K-means++, K-rank, Cluster-dp, infomap, BGLL, and OSLOM. The results with standard deviation (std) are reported in Table 3. From the tables, we have the following observations. (1) By comparing with K-means, it is necessary and effective to select initial seeds carefully. We also notice that if a network is of clear community structure, like Zachary's club, Dolphins, Political books and PPI networks, random initialized centers have high probability to get stable structure and almost no influence on the final results, so the deviations tend to be zero (the significant figures after decimal point is set to 2). For example, the standard deviations on accuracy and NMI are respectively 3.376e−16 and 2.688e−16 on Political books network, the standard deviations are respectively 0.0026 and 0.000025 on PPI network. (2) To decide initial seeds, "*outstanding*" and "*dispersedly-located*" should be jointly considered. As K-means++ only takes "*dispersedly-located*" (distance between centers) into account, it does not perform as well as K-rank-D, Cluster-dp, and K-rank, which consider these two aspects together. Besides, the randomness of the first selected seed of K-means++ may be the main reason that it is worse than K-means on Risk, Lesmis, and Political Blogs data sets. (3) Our proposed K-rank-D algorithm exhibits better performance than K-means, K-means++, K-rank, and Cluster-dp. This demonstrates that the strategy of seeds initialization introduced in this study is more effectiveness. Additionally, K-rank-D outperforms infomap, BGLL, and OSLOM in most cases though BGLL and infomap perform better on the Lesmis data set.

In this study, we adopted signal propagation to transform a network into Euclidean space. Different similarity methods may lead to different spatial structure. In our method, the transformation method of a network is not limited to signal propagation (Signal for short). Other faster similarity methods can be used, such as local random walk [28,41] (LRW for short, its complexity is $O(nd^l)$, where d is the average degree of nodes, l is the steps of random walks) and heat diffusion [40] (Heat-

Table 3
Comparison results of different algorithms on real world networks.

Algorithms	No. of networks							
	1	2	3	4	5	6	7	8
	Accuracy ± std							
K-rank	**1.00**$_{\pm 0.00}$	**0.97**$_{\pm 0.00}$	**1.00**$_{\pm 0.00}$	0.68$_{\pm 0.00}$	0.84$_{\pm 0.00}$	**0.91**$_{\pm 0.00}$	0.93$_{\pm 0.00}$	0.91$_{\pm 0.00}$
K-means++	**1.00**$_{\pm 0.00}$	0.78$_{\pm 0.07}$	**1.00**$_{\pm 0.00}$	0.66$_{\pm 0.04}$	**0.85**$_{\pm 0.01}$	0.85$_{\pm 0.06}$	0.70$_{\pm 0.17}$	**0.92**$_{\pm 0.00}$
K-means	**1.00**$_{\pm 0.00}$	0.84$_{\pm 0.12}$	**1.00**$_{\pm 0.00}$	0.69$_{\pm 0.02}$	**0.85**$_{\pm 0.00}$	0.84$_{\pm 0.06}$	0.80$_{\pm 0.13}$	0.80$_{\pm 0.00}$
Cluster-dp	0.82$_{\pm 0.00}$	0.86$_{\pm 0.00}$	**1.00**$_{\pm 0.00}$	0.64$_{\pm 0.00}$	0.84$_{\pm 0.00}$	0.90$_{\pm 0.00}$	**0.95**$_{\pm 0.00}$	**0.92**$_{\pm 0.00}$
Infomap	0.82$_{\pm 0.00}$	0.85$_{\pm 0.00}$	0.54$_{\pm 0.00}$	**0.79**$_{\pm 0.00}$	0.79$_{\pm 0.00}$	0.90$_{\pm 0.00}$	0.69$_{\pm 0.00}$	0.81$_{\pm 0.00}$
BGLL	0.64$_{\pm 0.00}$	0.85$_{\pm 0.00}$	0.58$_{\pm 0.00}$	**0.79**$_{\pm 0.00}$	0.83$_{\pm 0.00}$	0.86$_{\pm 0.00}$	0.75$_{\pm 0.00}$	0.82$_{\pm 0.00}$
OSLOM	0.97$_{\pm 0.00}$	0.57$_{\pm 0.00}$	0.92$_{\pm 0.00}$	0.58$_{\pm 0.00}$	0.83$_{\pm 0.00}$	0.91$_{\pm 0.00}$	0.88$_{\pm 0.00}$	0.80$_{\pm 0.00}$
K-rank-D	**1.000**$_{\pm 0.000}$	**0.97**$_{\pm 0.00}$	**1.00**$_{\pm 0.00}$	0.70$_{\pm 0.00}$	**0.85**$_{\pm 0.00}$	**0.91**$_{\pm 0.00}$	**0.95**$_{\pm 0.00}$	**0.92**$_{\pm 0.00}$
	NMI ± std							
K-rank	**1.00**$_{\pm 0.00}$	**0.96**$_{\pm 0.00}$	**1.00**$_{\pm 0.00}$	0.75$_{\pm 0.00}$	0.56$_{\pm 0.00}$	0.91$_{\pm 0.00}$	0.67$_{\pm 0.00}$	**0.95**$_{\pm 0.00}$
K-means++	**1.00**$_{\pm 0.00}$	0.83$_{\pm 0.05}$	**1.00**$_{\pm 0.00}$	0.75$_{\pm 0.02}$	0.56$_{\pm 0.01}$	0.89$_{\pm 0.02}$	0.25$_{\pm 0.20}$	**0.95**$_{\pm 0.00}$
K-means	**1.00**$_{\pm 0.00}$	0.89$_{\pm 0.07}$	**1.00**$_{\pm 0.00}$	0.77$_{\pm 0.02}$	0.55$_{\pm 0.00}$	0.90$_{\pm 0.03}$	0.39$_{\pm 0.19}$	0.93$_{\pm 0.00}$
Cluster-dp	0.47$_{\pm 0.00}$	0.91$_{\pm 0.00}$	**1.00**$_{\pm 0.00}$	0.78$_{\pm 0.00}$	0.57$_{\pm 0.00}$	0.90$_{\pm 0.00}$	**0.72**$_{\pm 0.00}$	**0.95**$_{\pm 0.00}$
Infomap	0.69$_{\pm 0.00}$	0.94$_{\pm 0.00}$	0.46$_{\pm 0.00}$	**0.83**$_{\pm 0.00}$	0.54$_{\pm 0.00}$	0.91$_{\pm 0.00}$	0.37$_{\pm 0.00}$	0.93$_{\pm 0.00}$
BGLL	0.58$_{\pm 0.00}$	0.94$_{\pm 0.00}$	0.48$_{\pm 0.00}$	0.75$_{\pm 0.00}$	0.57$_{\pm 0.00}$	0.89$_{\pm 0.00}$	0.37$_{\pm 0.00}$	0.93$_{\pm 0.00}$
OSLOM	0.84$_{\pm 0.00}$	0.58$_{\pm 0.00}$	0.61$_{\pm 0.00}$	0.65$_{\pm 0.00}$	0.57$_{\pm 0.00}$	0.91$_{\pm 0.00}$	0.50$_{\pm 0.00}$	0.91$_{\pm 0.00}$
K-rank-D	**1.00**$_{\pm 0.00}$	**0.96**$_{\pm 0.00}$	**1.00**$_{\pm 0.00}$	0.80$_{\pm 0.00}$	**0.60**$_{\pm 0.00}$	**0.92**$_{\pm 0.00}$	**0.72**$_{\pm 0.00}$	**0.95**$_{\pm 0.00}$

Table 4
Comparison results of different transformation methods on real world networks.

Algorithms	No. of networks							
	1	2	3	4	5	6	7	8
	Accuracy							
Signal	**1.00**	**0.97**	**1.00**	0.70	0.85	0.91	0.95	0.92
LRW	**1.00**	**0.97**	**1.00**	0.54	**0.85**	**0.93**	**0.96**	0.90
Heat-D	**1.00**	**0.97**	**1.00**	**0.83**	0.84	0.75	**0.96**	0.85
	NMI							
Signal	**1.00**	**0.96**	**1.00**	0.80	**0.60**	0.92	0.72	**0.95**
LRW	**1.00**	**0.96**	**1.00**	0.63	**0.60**	**0.93**	0.74	0.94
Heat-D	**1.00**	**0.96**	**1.00**	0.69	**0.85**	0.84	0.80	0.80

Table 5
Comparison results of different transformation methods on LFR2 networks.

Algorithms	μ_0								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	Accuracy								
Signal	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.95**	**0.65**	**0.18**	**0.07**
LRW	**1.00**	**1.00**	**1.00**	**1.00**	0.98	0.79	0.38	0.12	0.06
Heat-D	**1.00**	**1.00**	**1.00**	**1.00**	0.96	0.62	0.37	0.15	**0.07**
	NMI								
Signal	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**	**0.73**	**0.32**	**0.22**
LRW	**1.00**	**1.00**	**1.00**	**1.00**	0.99	0.89	0.53	0.25	0.19
Heat-D	**1.00**	**1.00**	**1.00**	**1.00**	0.97	0.75	0.51	0.28	**0.22**

D for short, its complexity is $O(m\lambda)$, where λ is the number of iterations of heat diffusion process). Comparison results on real world networks and LFR2 ($n = 5000$) networks shown in Tables 4–5 (results on LFR1 ($n = 1000$) and LFR3 ($n = 10{,}000$) networks are similar with Table 5). These tables suggest that local random walk and heat diffusion also perform well by embedding into our method. To further speedup the clustering process, the dimension of each node can be reduced by using MDS [28,42,43] or DeepWalk [44], etc. But it will be at some cost of accuracy of the results.

4. Conclusion

In this study, we present a new clustering algorithm for detecting communities in networks, named K-rank-D. Compared with previous work on seeds initialization, our proposed method is parameter-free, which seeks out dispersedly influential nodes as initial centers. In addition, K-rank-D enables us to determine the number of communities in a network actively and intuitively. The method has been tested with a variety of networks. By analyzing initial seeds, we find that the initial seeds chosen by K-rank-D are reasonable, which are evenly and centrally located in each community. What is more, experimental results on real-world and synthetic data sets demonstrate the effectiveness and high accuracy of K-rank-D.

The initial seeds selection method in K-rank-D can be easily extended by using other dispersion measurements and nodes importance measurements. We hope that this new method will help to decide the number of communities in study of community detection in complex networks. However, for a network without very clear community structure, such as the GN networks when $Z_{out} = 8$ and LFR networks when $\mu_0 > 0.7$, it is also hard for K-rank-D to find exactly right number of communities. How to find the optimal number of communities in these cases needs further study.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China (No. 61473030 and No. 61370129), the Fundamental Research Funds for the Central Universities (No. 2014JBM031 and No. K15JB00070), the Program for Changjiang Scholar and Innovative Research Team in University (No. IRT201206) and the Opening Project of State Key Laboratory of Digital Publishing Technology. The authors would like to acknowledge the anonymous reviewers for their constructive comments.

References

[1] D.J. Watts, S.H. Strogatz, Collective dynamics of small-worldnetworks, Nature 393 (6684) (1998) 440–442.
[2] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[3] M. Girvan, M.E. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. 99 (12) (2002) 7821–7826.
[4] M.E. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2) (2004) 026113.
[5] M.E. Newman, The structure and function of complex networks, SIAM Rev. 45 (2) (2003) 167–256.
[6] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.-L. Barabási, Hierarchical organization of modularity in metabolic networks, Science 297 (5586) (2002) 1551–1555.
[7] D.M. Wilkinson, B.A. Huberman, A method for finding communities of related genes, Proc. Natl. Acad. Sci. 101 (suppl 1) (2004) 5241–5248.
[8] R. Guimera, L.A.N. Amaral, Functional cartography of complex metabolic networks, Nature 433 (7028) (2005) 895–900.
[9] G.W. Flake, S. Lawrence, C.L. Giles, F.M. Coetzee, Self-organization and identification of web communities, Computer 35 (3) (2002) 66–70.
[10] Y. Dourisboure, F. Geraci, M. Pellegrini, Extraction and classification of dense communities in the web, in: Proceedings of the 16th International Conference on World Wide Web, ACM, 2007, pp. 461–470.
[11] A. Perianes-Rodríguez, C. Olmeda-Gómez, F. Moya-Anegón, Detecting, identifying and visualizing research groups in co-authorship networks, Scientometrics 82 (2) (2010) 307–319.
[12] B. He, Y. Ding, J. Tang, V. Reguramalingam, J. Bollen, Mining diversity subgraph in multidisciplinary scientific collaboration networks: A meso perspective, J. Informetrics 7 (1) (2013) 117–128.
[13] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (5814) (2007) 972–976.
[14] S. Zhang, R.-S. Wang, X.-S. Zhang, Identification of overlapping community structure in complex networks using fuzzy c-means clustering, Physica A 374 (1) (2007) 483–490.
[15] X. Ma, L. Gao, X. Yong, L. Fu, Semi-supervised clustering algorithm for community structure detection in complex networks, Physica A 389 (1) (2010) 187–197.
[16] H.-W. Shen, X.-Q. Cheng, Spectral methods for the detection of network community structure: a comparative analysis, J. Stat. Mech. Theory Exp. 2010 (10) (2010) P10020.
[17] M. Newman, Spectral methods for community detection and graph partitioning, Phys. Rev. E 88 (4) (2013) 042822.
[18] A. Medus, G. Acuna, C. Dorso, Detection of community structures in networks via global optimization, Physica A 358 (2) (2005) 593–604.
[19] A. Clauset, M.E. Newman, C. Moore, Finding community structure in very large networks, Phys. Rev. E 70 (6) (2004) 066111.
[20] H. Shen, X. Cheng, K. Cai, M.-B. Hu, Detect overlapping and hierarchical community structure in networks, Physica A 388 (8) (2009) 1706–1712.
[21] R. Shang, J. Bai, L. Jiao, C. Jin, Community detection based on modularity and an improved genetic algorithm, Physica A 392 (5) (2013) 1215–1231.
[22] H. Jin, S. Wang, C. Li, Community detection in complex networks by density-based clustering, Physica A 392 (19) (2013) 4606–4618.
[23] M. Gong, J. Liu, L. Ma, Q. Cai, L. Jiao, Novel heuristic density-based method for community detection in networks, Physica A 403 (2014) 71–84.
[24] H. Zhou, Distance, dissimilarity index, and network community structure, Phys. Rev. E 67 (6) (2003) 061901.
[25] H. Zhou, R. Lipowsky, Network brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities, in: Computational Science-ICCS 2004, Springer, 2004, pp. 1062–1069.
[26] M. Rosvall, C.T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proc. Natl. Acad. Sci. 105 (4) (2008) 1118–1123.
[27] D. Lai, H. Lu, C. Nardini, Finding communities in directed networks by pagerank random walk induced network embedding, Physica A 389 (12) (2010) 2443–2454.
[28] W. Wang, D. Liu, X. Liu, L. Pan, Fuzzy overlapping community detection based on local random walk and multidimensional scaling, Physica A 392 (24) (2013) 6578–6586.
[29] M.B. Hastings, Community detection as an inference problem, Phys. Rev. E 74 (3) (2006) 035102.
[30] M.E. Newman, E.A. Leicht, Mixture models and exploratory analysis in networks, Proc. Natl. Acad. Sci. 104 (23) (2007) 9564–9569.
[31] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (3) (2010) 75–174.
[32] J. MacQueen, et al. Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol.1, California, USA, 1967, pp. 281–297.
[33] D. Arthur, S. Vassilvitskii, k-means++: the advantages of careful seeding, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.
[34] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science 344 (6191) (2014) 1492–1496.
[35] Y. Jiang, C. Jia, J. Yu, An efficient community detection method based on rank centrality, Physica A 392 (9) (2013) 2182–2194.
[36] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A 390 (6) (2011) 1150–1170.
[37] P. Jaccard, Étude comparative de la distribution florale dans une portion des alpes et des jura, Bull. Soc. Vaud. Sci. Natur. 37 (1901) 547–579.
[38] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., New York, NY, USA, 1986.
[39] Y. Hu, M. Li, P. Zhang, Y. Fan, Z. Di, Community detection by signaling on complex networks, Phys. Rev. E 78 (1) (2008) 016115.
[40] H. Ma, I. King, M.R. Lyu, Mining web graphs for recommendations, IEEE Trans. Knowl. Data Eng. 24 (6) (2012) 1051–1064.
[41] W. Liu, L. Lü, Link prediction based on local random walk, Europhys. Lett. 89 (5) (2010) 58007.
[42] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: A review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37.
[43] J. Yang, D. Hubball, M.O. Ward, E.A. Rundensteiner, W. Ribarsky, Value and relation display: interactive visual exploration of large data sets with hundreds of dimensions, IEEE Trans. Vis. Comput. Graphics 13 (3) (2007) 494–507.
[44] B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: Online learning of social representations, in: KDD, ACM, 2014, pp. 701–710.
[45] L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking: bringing order to the web.

[46] A. Arasu, J. Novak, A. Tomkins, J. Tomlin, Pagerank computation and the structure of the web: Experiments and algorithms, in: Proceedings of the Eleventh International World Wide Web Conference, Poster Track, 2002, pp. 107–117.
[47] M. Franceschet, Pagerank: Standing on the shoulders of giants, Commun. ACM 54 (6) (2011) 92–101.
[48] A. Li, Z.H. Zhang, Y. Meng, Fuzzy mathematics and application, 2005.
[49] S.Z. Selim, M.A. Ismail, K-means-type algorithms: a generalized convergence theorem and characterization of local optimality, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1) (1984) 81–87.
[50] 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.
[51] A. Lancichinetti, F. Radicchi, J.J. Ramasco, S. Fortunato, Finding statistically significant communities in networks, PloS One 6 (4) (2011) e18961.
[52] A. Strehl, J. Ghosh, Cluster ensembles—a knowledge reuse framework for combining multiple partitions, J. Mach. Learn. Res. (2003) 583–617.
[53] A. Lancichinetti, S. Fortunato, F. Radicchi, Benchmark graphs for testing community detection algorithms, Phys. Rev. E 78 (4) (2008) 046110.
[54] W.W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. (1977) 452–473.
[55] K. Steinhaeuser, N.V. Chawla, Identifying and evaluating community structure in complex networks, Pattern Recognit. Lett. 31 (5) (2010) 413–421.
[56] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations, Behav. Ecol. Sociobiol. 54 (4) (2003) 396–405.
[57] D.E. Knuth, The Stanford GraphBase: A Platform for Combinatorial Computing, Vol.37, Addison-Wesley, Reading, 1993.
[58] V. Krebs, (unpublished), http://www.orgnet.com/.
[59] L.A. Adamic, N. Glance, The political blogosphere and the 2004 us election: divided they blog, in: Proceedings of the 3rd International Workshop on Link Discovery, ACM, 2005, pp. 36–43.
[60] J. Vlasblom, S.J. Wodak, Markov clustering versus affinity propagation for the partitioning of protein interaction graphs, BMC Bioinform. 10 (1) (2009) 99.
[61] R. Aldecoa, I. Marín, Surprise maximization reveals the community structure of complex networks, Sci. Rep. 3 (2013) 1060.
[62] Y. Jiang, C. Jia, J. Yu, An efficient community detection algorithm using greedy surprise maximization, J. Phys. A 47 (16) (2014) 165101.
[63] S. Fortunato, M. Barthelemy, Resolution limit in community detection, Proc. Natl. Acad. Sci. 104 (1) (2007) 36–41.