



# A community detection algorithm based on multi-similarity method

Li Ni<sup>1,2</sup> · Pen ManMan<sup>1</sup> · Jiang Wenjun<sup>1</sup> · Li Kenli<sup>1</sup>

Received: 22 October 2017 / Revised: 3 December 2017 / Accepted: 22 December 2017 / Published online: 13 January 2018  
© Springer Science+Business Media, LLC, part of Springer Nature 2018

## Abstract

Social network detection and identification constitute an important topic in the field of sociology. Previous graph similarity has focus on either the topological structure of graph or the feature value of vertex. In this work, a multi-similarity measure method for community is described. The approach devised by using multi-similarity properties based on vertex features, relationship density and topology structure, and therefore is can be formulated and extended to practical implementation. The framework of community detection combines K-means clustering, spectral clustering and modularity algorithm-making it an effective basis for the realization of a social network interpretation. With this scheme, three evaluation criteria are proposed for methodology determination. The experimental results show a better working performance of the recommended method than traditional algorithms via statistical analysis.

**Keywords** Community detection · Multi-similarity · Vertex feature · Social network · K-means clustering

## 1 Introduction

Sociology believes that the formation of social phenomena on the basic of community rather than the individual [1]. Over the past decades, the study of social community occupies a mainly important space in social computing. Social community, which refers to a set of vertexes, results in inner interaction closer than that with the outer sets. Particularly, recently, with the increasing popularity of social networking application, the development is most pronounced in the structure complexity of social network [2]. Current technology promotes social network methodology from data collection, to network construction, to network dynamics, to network analysis, to results interpretation, etc., [3]. To better understand social network, a large number of approaches are proposed to capture network properties [4–6]. Similarity measure is one such method, with previous publications exploring the similarity of two vertexes [7–9] and vertex feature values [10,11]. The analysis of inner-vertices relationship is, however, still limited, from the point of view of structure or property. It is hard to extract a proper description

for the community scope. Few articles have been published on this aspect. Similarity measure will still have the challenges owing to the increasing complicated social network. For this reason, many current applications use similarity measure as a secondary approach or as a tool to study network structure via similarity analysis. Vertex similarities, however, still have the potential to be beneficial for network analysis, such as the nodes affected by each other.

To solve the above issue, this study proposes a multi-similarity method (MSM) for social network analysis, depicts the similarity-based method in combination with clustering algorithms for data collection, interprets the detection into community properties, and obtains a better working performance by providing the processing steps via laboratory setup.

The research contributions of this paper can be summarized as follows:

- Multi-similarity method (MSM): this study devises a multi-similarity method for graph topological structure and vertex feature capture which is based on six types of similarity properties.
- Similarity measure evaluation: define the properties of multi-similarity vertex in the network, and qualitatively identify the methodology design on similarity detection.
- Community detection: an optimized architecture combining MSM with K-means clustering algorithm is defined. The proposed method is utilized in a manner similar to the spectral clustering algorithm.

✉ Pen ManMan  
pengmanman@hnu.edu.cn

<sup>1</sup> College of Computer Science and Electric Engineering, Hunan University, Changsha 410082, Hunan, China

<sup>2</sup> College of Information and Electronic Engineering, Hunan City College, Yiyang 413000, Hunan, China

This paper presents similarity measure-related background knowledge and clustering algorithms in Sect. 2; illustrates the overall community structure in Sect. 3; shows the multi-similarity design concept and assessing indexes in Sect. 4; describes the community detection method devise process in Sect. 5; experimental results achieved in this study and the analysis in Sect. 6, and finally presents the research significances and future expectation and planning in Sect. 7.

## 2 Background and relate work

This section introduces the basic knowledge of similarity measure and clustering algorithms, so as to facilitate the depiction of subsequent methodology.

### 2.1 Similarity measure

Traditional similarity measure is characterized by Jaccard similarity based on neighbor vertexes [12]. Vertexes that share similar connection tend to settle in the same community. The more mutual friends of the two vertices, the more similar they appear to be. Jaccard similarity means that the similarity of two-pairs object is the common neighbor number, Jaccard similarity shows that  $v_j$  and  $v_i$  are more similar if they have more common neighbors, and  $Jaccard(v_i, v_j)$  increased with the increasing number of common friends. However, the Jaccard similarity equals 1 when the two have common friends; otherwise, the value of similarity is 0. This arbitrary detection method is certainly extreme and cannot correctly show the relationship between the two points of social networks.

The SimRank, which is based on neighbor over arbitrary graphs, plays an important role in similarity scoring [13]. The equation  $s(v_i, v_j)$  denotes that the similarity between  $v_i$  and  $v_j$  is the average value of the similarity between the neighbor vertexes of  $v_i$  and the neighbor vertexes of  $v_j$ . Just like Jaccard similarity, the SimRank metric value is 1 only for computing relation of the vertex and itself, otherwise is 0. Whereas, for SimRank, the more in-neighbors of vertex  $v_i$  and  $v_j$  is, the smaller the similarity between  $v_i$  and  $v_j$  [14] is. In other words, the SimRank score only accommodates the paths with equal length from a common source vertex, other paths for vertex-pair are fully ignored [15]. The optimized SimRank is Simrank++ [16], which improves the former approach in two ways: defining the evidence score as an increasing function of common and keeping the similarity score consistent with the weights on the graph [9].

### 2.2 Clustering algorithm

Previous work highlights the function of K-means algorithm in community discovery [17]. K-means is a general used clus-

**Table 1** Symbols definition and their descriptions

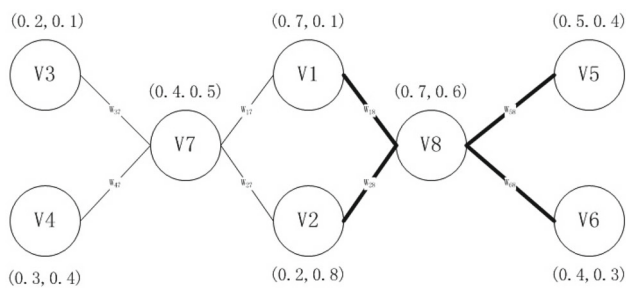
Definition	Description
G	A multi-feature, undirected, weighted social networks
V	Vertexes set
E	Connected link between two vertexes
F	Feature vector set of vertex
W	Connected weight set of vertexes
$C_i$	$i$ th cluster
$c_i$	The center of $C_i$
$d(v_i)$	Degree of vertex $v_i$
$N(v_i)$	Direct neighbors of vertex $v_i$
$EN(v_i)$	Exclusive neighbors of vertexes $v_i$ which is direct neighbors of vertex $v_i$ , but not common neighbor of $v_i$ and $v_j$
$d(v_i, v_j)$	A function about distance between $v_i$ and $v_j$
$w'(v_i, v_j)$	A weight function based on $v_i$ and $v_j$ common neighbors and mutual feature
$g(v_i)$	A function of vertex $v_i$ for common friends' feature and relation density
$g'(v_i)$	A function of vertex $v_i$ for exclusive neighbors feature and relation density
$s_l$	Similarity based on direct linked edge
$s_c$	Similarity based on direct common neighbors
$s_e$	Similarity based on exclusive neighbors
$s(v_i, v_j)$	Multi-similarity value of $v_i$ and $v_j$

tering algorithm for expectation maximization. The K-means clustering is employed for determining the initial cluster center, measuring the similarity between object and cluster center and resolving the objection function. On the foundation of expectation maximization, algorithms with Jaccard similarity, Cosine or SimRank, are developed to conduct community discovery [13].

Spectral clustering algorithm, in line with adjacency matrix or Laplacian matrix feature vector, is established for searching the minimum cut in network. The parameter involved is merely similarity matrix, which modifies sparse data processing effectively. For this reason, many current applications use K-means clustering as a secondary tool to complement similarity measure. The realization of spectral clustering algorithm is affected by the specific similarity matrix. Different matrix will definitely lead to different clusters. Thus, using an appropriate matrix is the key to address the problem [18].

## 3 Overview of network structure

This section will completely introduce the overall structure of the network, as well as features and relations of internal vertexes. As shown in Table 1, we describe the symbols definition and their descriptions about network structure.



**Fig. 1** An example of network architecture

In general, a number of parameters are required to characterize the network structure. An example of the social network is defined as a undirected graph  $G = (V, F, E, W)$  where  $V$  is the set of all the vertices,  $F$  is the feature set of these vertices,  $E$  is the set of undirect links and  $W$  is the relationship density for every community with the number of vertexes  $n \geq 1$ . Thus, a community containing  $M$  features can be defined as:

$$\mathbf{F}(\mathbf{v}) = (f_1, f_2, \dots, f_M), f_i \in [0, 1], i = 1, 2, \dots, M$$

As shown in Fig. 1, the network we adopted has all the feature probability values ranging from 0 to 1. According to the aforementioned principle, the greater the value is, the stronger the correlation is. The value of vertex v3 on features 1 and features 2 are 0.2 and 0.1, respectively, which is shown as (0.2, 0.1).

The value of the similarity between vertexes in Fig. 1 depends on the situation between their features and their relationships. Node v1 is directly connected to v7 and v8, so does v2, thereby v1 and v2 are named as common neighbor nodes for v7 and v8. Correspondingly, v2 and v7 have non-common neighbor nodes or exclusive neighbor nodes where v8 is exclusive neighbor for v2, as well as v1, v3 and v4 for v7. Thus, a path between exclusive vertexes v2 and v7 is formed as v2-v8-v1-v7. Therefore, three types of relationships exist within the structure, which are direct connection, common neighbor connection and exclusive vertexes connection.

Supposing that we divide the social network into  $k$  unconnected community, where  $C_i = (V_i, F_i, E_i, W_i)$   $i = 1, 2, \dots, k$ ,  $V = \cup_{i=1}^k V_i$ , and  $V_i \cap V_j = \phi$  when  $i \neq j$ , the purpose of the function is to address high-quality and efficient community results.

## 4 Multi-similarity method (MSM)

This section depicts the multi-similarity method design process on the foundation of vertex similarities and relations.

The method constructing together with the effectiveness evaluation is facilitated by proposed similarity properties.

### 4.1 Similarity properties index

In order to understand the similarity measure design with social network, it is required to study the natural characteristics of similarity. On the foundation of research [13–16, 19–21], a generalized similarity properties of the vertex are summarized as follows.

- p1 The similarity between the vertexes should be related to a commonneighbor, and increases with the number of common neighbors.
- p2 A larger number of similar features leads to increased similarity between vertexes. Granovetter [22] proposed that the more common the feature are between vertexes are, the more similarities in the vertexes cluster exist.
- p3 Similarity should be normalized to the scale of [0,1], while the upper and lower bounds of the value should make practical sense.
- p4 The value of the upper and lower bounds of similarity must be reasonable.
- p5 The similarity is consistent with the weights on the graph.
- p6 The value of similarity is related to the path.

Accordingly, the aforementioned properties are used for similarity measure method design. However, when distributed vertex have multiple features, the multi-similarity based method with this principle is required.

### 4.2 Multi-similarity method (MSM)

On the basic of para.3, the direct connection, common neighbor connection and exclusive vertexes connection are represented by  $s_d$ ,  $s_c$  and  $s_e$ , respectively. In the process of calculating, we emphasize the influence of the feature metric between vertexes and the relationship density of the user. The aforementioned properties, which are reflected in the three-similarity computing, are utilized for working performance evaluation.

#### 4.2.1 Similarity based on direct connection

A similarity exists between two directly connected vertexes, which is related to their own features and the strength of connections. We use a continuous function to depict the processing of vertex feature and relationship density. Generally, the relationship density is the main factor for analysis while the vertex feature is the supplement factor, since the feature of social network users is often cross domain. If their features are alike, the effect of vertex features on the relationship density is positive and vice versa. Thus, a combination of

the Gauss function is added to the feature metric of vertexes as well as the density to describe the correlation. Similarity based on direct linked edge is calculated as follows:

$$s_l(v_i, v_j) = \frac{w_{ij} e^{-\gamma \|\mathbf{F}(v_i) - \mathbf{F}(v_j)\|}}{w'(v_i, v_j)} \tag{1}$$

together with

$$w'(v_i, v_j) = \sum_{a_k \in N(v_i)} w(a_k, v_i) + \sum_{b_k \in N(v_j)} w(b_k, v_j) - w(v_i, v_j)$$

where  $\gamma$  is a parameter that determines the influenced speed of vertex feature to the vertex relationships,  $N(v_i)$  is directed neighbor of vertex  $v_i$ ,  $w(a_k, v_i)$  is weight value expressed as the relation number of the two vertexes and  $\mathbf{F}(v_i)$  is feature vector of vertex  $v_i$ . The similarity of the vertex feature will inevitably influence the relationship between them.

### 4.2.2 Similarity based on direct common neighbor

Influenced by social factors, in a social group, individual relation usually varies on the common neighbor as well as their features. If the features between two vertexes do not change (i.e., the two vertexes are the same), they will be more influenced by the relation density between the two vertexes. A continuous function is applied to depict the process of features and relations. The function in vertex  $v_i$  is described as:

$$g(v_i) = \sum_{a_k \in N(v_i) \cap N(v_j)} w(a_k, v_i) e^{-\gamma \|\mathbf{F}(a_k) - \mathbf{F}(v_i)\|} \tag{2}$$

The Jaccard similarity between  $v_i$  and  $v_j$  is

$$s(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{|N(v_i) \cup N(v_j)|} \tag{3}$$

Then similarity based on direct common neighbors is

$$s_c(v_i, v_j) = \frac{g(v_i, v_j)}{w'(v_i, v_j)} \tag{4}$$

The connection relation factor between the vertexes and the features factors of the vertexes are considered. For common neighbors we have  $g(v_i, v_j) = g(v_i) + g(v_j)$  where  $(v_i, v_j)$  is weight value sum of direct neighbor vertexes  $v_i$  and  $v_j$ . For exclusive neighbors,  $g'(v_i, v_j) = \sum_{a_k \in N(v_i)} w(a_k, v_i) + \sum_{b_k \in N(v_j)} w(b_k, v_j) - g(v_i, v_j)$  of  $(v_i, v_j)$  where  $(v_i, v_j)$  is the value sum of non-common directed neighbor vertexes  $v_i$  and  $v_j$ .

### 4.2.3 Similarity based on exclusive vertexes

The influence of similarity from neighbors is of maximum weight. On the other hand, the effect of exclusive neighbors is studied by using the ideas of Simrank through three steps of random walk strategy, whose similarity is

$$s_e(v_i, v_j) = \frac{\sum_{ev_i \in EN(v_i), ev_j \in EN(v_j)} s(v_i, ev_i) s(ev_i, ev_j) s(ev_j, v_j)}{|EN(v_i)| \times |EN(v_j)|} \tag{5}$$

in line with  $EN(v_i) = N(v_i) - N(v_i) \cap N(v_j)$ .

### 4.2.4 Similarity evaluation

The similarity value represents the homogeneity between specific vertexes within the social network. According to the preceding analysis, the formula is defined as follows:

$$s(v_i, v_j) = a \cdot s_l(v_i, v_j) + b \cdot s_c(v_i, v_j) + (1 - a - b) \cdot s_e(v_i, v_j) \tag{6}$$

$a \in [0, 1], b \in [0, 1]$

According to Eq. (6), the outcome of multi-similarity method caters to the requirements of the evaluation indexes. Since Due to the utilization of Jaccard method, values  $s_l$  and  $s_c$  are consistent with the definition of common neighbor in P1. The less the value of  $\|\mathbf{F}(v_i) - \mathbf{F}(v_j)\|$  is, the more the values of  $s_l$  and  $s_c$  reach. Meanwhile,  $s_e$  is the similarity value of SimRank based on Jaccard similarity, which is also satisfied with P1 and P2. Further, for each  $s(v_i, v_j) \in [0, 1]$ , the similarity between the vertex and itself is either 1 and no or 0, which results in the  $s(v_i, v_j)$  meets the qualification of P3 and P4. Likewise, based on Eqs. (1)–(6), the  $s(v_i, v_j)$  is proportional to the weight value, which is exactly in accordance with P5. In addition, P6 refers to the path of computing, in line with the random walk for the similarity measure algorithm implementation in this research.

Including some derivation details to justify the design theory on multi-similarities, positive results at this stage indicated that the strategy could indeed be effective in network interpretation. Specifically, the evaluation indexes are capable of generating measurable vertex responses of community detection algorithm design.

## 5 Community detection algorithm

This section describes the community detection algorithm devised based on MSM. The algorithm is conducted on similarity matrix denoting the relative similarity of vertexes in social network. To accurately identify the social network, the

**Table 2** Similarity value among vertexes given Fig. 1 using Eq. (6)

s	v1	v2	v3	v4	v5	v6	v7	v8
v1	1	0.3760	0.0078	0.0107	0.3696	0.3804	0.0046	0.0764
v2	0.3760	1	0.0076	0.0106	0.3819	0.3886	0.0045	0.0815
v3	0.0078	0.0076	1	0.3715	0	0	0.0833	0
v4	0.0107	0.0106	0.3715	1	0	0	0.1276	0
v5	0.3696	0.3819	0	0	1	0.3844	0	0.0782
v6	0.3804	0.3886	0	0	0.3844	1	0	0.1389
v7	0.0046	0.0045	0.0833	0.1276	0	0	1	0.0784
v8	0.0764	0.0815	0	0	0.0782	0.1389	0.0784	1

input data is formulated through spectral clustering, the community is detected by K-means clustering and the outcomes optimized via modularity algorithm.

### 5.1 Detection based on K-means clustering

The architecture of community detection contains two main steps:

1. Establishing the clustering center based on graph clustering with parameter determination and initialization.
2. Constructing the objective function with the similarity method to detect community.

#### 5.1.1 Clustering center initialization

To start with, spectral clustering algorithm is utilized to perform graph representation on similarity matrix eigenvalues of the social network. For the deployment of K-means clustering, the number of clusters  $K$  and the controlling parameter  $\gamma$  are selected. At this stage, the clustering center has significant impact on clustering outcome. Generally, the center is selected randomly thus the outcome is inevitably unstable. For the purpose of community detection, the user multi-similarity value, based on PageRank [23] algorithm, is set as the initial center. Probability value of vertexes  $(v_i, v_j)$  as well as the influence rank value of vertex  $v_i$  is showed as

$$\begin{aligned}
 P_{v_i v_j} &= \frac{s(v_i, v_j)}{\sum_{v_j \in N(v_u)} s(v_u, v_j)} \\
 I_{v_i} &= \sum_{v_u \in N(v_i)} I_{v_u} P_{v_u v_i}
 \end{aligned}
 \tag{7}$$

where  $k$  seeds are the  $k$  maximum of  $I$ .

#### 5.1.2 Objection function

In our algorithm, the distance value is objection function in terms of the topological structure and feature attributes of vertexes. In particular, the distance between vertexes is

#### Algorithm 1 Community detection algorithm of k-means

**Require:** An undirected, weighted or unweighted multi-feature social network, number of clusters  $k$ , weight factor  $a, b$ , and  $\gamma$ , the maximized iteration number

**Ensure:**  $k$  clusters  $C_1, C_2, \dots, C_k$

- 1: Initial similarity value is  $s[i, j] = s(v_i, v_j)$ , according to equation (6)
- 2: Initial center vertexes are  $k$  vertices with respect to top  $k$   $I$  using equation (7)
 
$$c = TopK(I)$$
 {vertexes  $c$  having  $k$  maximized influence rank}
- 3: **for** each vertex  $v_i$  in  $V$  **do**
- 4:  $cluster[i] = argmin_{i,j} d(v_i, c_j)$  for all centers  $j=1, 2, \dots, k$
- 5: **end for**
- 6: Evaluate  $F_{objection}$  and the clusters quality through Modularity, SSE, and density
- 7: **if**  $F_{objection}$  || iteration number is maximized **then**
- 8: **return**  $k$  clusters  $C_1, C_2, \dots, C_k$
- 9: **end if**
- 10: Update  $c[j]$  for each cluster  $j=1, 2, \dots, k$ , for which the sum of distance values is minimum, using equation (9)
- 11: repeat steps 3-10 until convergence

related to the proposed similarity method, which is illustrated as:

$$F_{objection} = \sum_{i=1}^k \sum_{p \in C_i} d(p, c_i)^2
 \tag{8}$$

where initial distance  $d(i, j) = 1 - s(i, j)$  while the distance value is updated by

$$d(p, c_i) = 1 - \frac{\sum_{p \in C_i} s(i, p)}{|m_i|}
 \tag{9}$$

where  $|m_i|$  is the number of vertexes in  $C_i$ .  $c_i$  is the center of cluster  $C_i$ . Referring to the vertexes in Fig. 1, the relations among nodes are presented in Table 2. For the purpose of clustering, the similarities between each node and itself is 1 while for exclusive one is 0. Based on K-means clustering algorithm, a larger value denotes a closer relation of the vertexes.

The time complexity of the similarity method, which makes the three-step distance strategy is  $O(NET)$ . The mem-



ory allocation is mainly related to the number of edges. The clustering outcomes are optimized according to minimum modularity value, which results in a high inner-community similarity and a low external one. This approaches capable of dealing with medium and large-scale graphs. The pseudo code of proposed method is presented in Algorithm 1.

## 5.2 Evaluation of working performance

We tend to use the state-of-art technologies for working performance evaluation. However, current processing has to be adjusted to fit the proposed graph. For this reason, modularity, SSE (sum of the squared error) and density are recommended as criteria.

### 5.2.1 Modularity

Modularity has been proved highly effective in practice for community evaluation [24]. For clustering result estimation, an improved module is defined as:

$$Q = \frac{1}{\|W\|} \sum_{c=1}^{c=k} \sum_{i,j \in C_c} (w_{ij} e^{-\gamma \|\mathbf{F}(v_i) - \mathbf{F}(v_j)\|} - \frac{p_i p_j}{\|W\|}) \quad (10)$$

where  $\|W\| = \sum_{i=1}^{i=N} \sum_{j=1}^{j=N} w_{ij}$ ,  $p_i = \sum_{j=1}^{j=N} w_{ij} e^{-\gamma \|\mathbf{F}(v_i) - \mathbf{F}(v_j)\|}$ ,  $N$  is the number of vertexes, considering the difference between the community density and expectation value, the modularity characterizes the network. The greater the modularity value is, the more concentrated the vertexes within the community are.

### 5.2.2 SSE (sum of the squared error)

SSE uses Euclidean distance to assess the cluster quality from the perspective of data feature. A smaller value of SSE indicates a better-selected center of the cluster. The sum of the squared error based on the data feature between vertexes is

$$SSE = \sum_{i=1}^k \sum_{v_j \in C_i} \|F(v_j) - F(c_i)\|_2^2 \quad (11)$$

### 5.2.3 Density

The density function is the most significant specification representing the clustering. Destiny stands for the ratio between edges in the cluster and total edges in the graph, which lies in the interval of [0,1]. According to Eq. (12), the greater the density value, the stronger the connection of the edge in the clusters.

**Table 3** Data sets information

Data sets	DBLP	pblog	Systhetic data
The number of vertexes	5000	1490	100
The number of edges	12,796	33,433	1000
Average weighted degree	13	20	20
Unweighted density	0.001	0.191	0.151
Average path length	2.185	2.237	2.217

$$Density = \frac{\sum_{i=1}^{i=k} \sum_{v_j, v_c \in C_i} \|E(v_j, v_c)\|}{\|E\|} \quad (12)$$

## 6 Experiments

This section shows the implemented methodology for community detection. The laboratory establishing integrates K-means [25], spectral clustering [26] and modularity [24] algorithm for clustering. In order to demonstrate method effectiveness, three evaluation parameters modularity [24], sum of squared error (SSE), and density [20] are employed.

### 6.1 Dataset description

We conducted experiments on two real datasets and a synthetic dataset. The DBLP network dataset has 5000 users with 12,796 links between these vertexes having two attributes. The other one is political blog network of 1490 web blogs with 19,090 links containing one attribute, which is named as Pblog in the following parts. Details are given in Table 3.

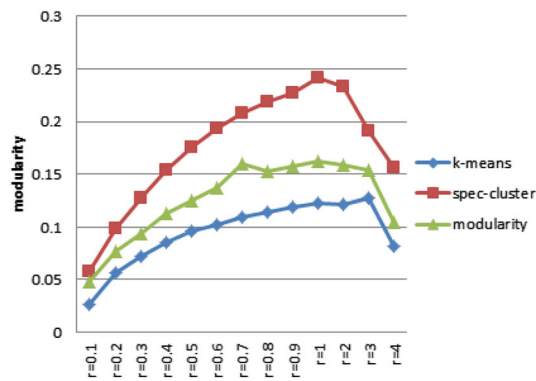
### 6.2 Experimental procedures

Experiments were performed on the experimental targets with different topological densities. Before the compiling of the algorithm, typical parameters  $\gamma$ ,  $a$  and  $b$  are quantified based on the natural characteristics of target community, which are set as 0.8, 0.4 and 0.4, respectively. To quantify the clustering, the aforementioned indexes are presented for estimation.

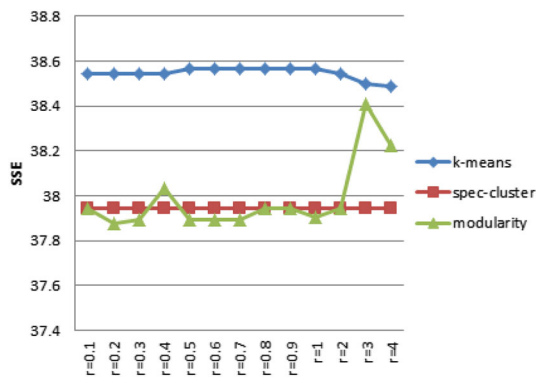
Aiming at evaluating the working performances of the proposed method, experiments with traditional approaches are carried out subsequently: Sa-clustering [22], W-clustering [27] and KSNAP [28] algorithm are employed to deal with the same targets.

### 6.3 Results and analysis

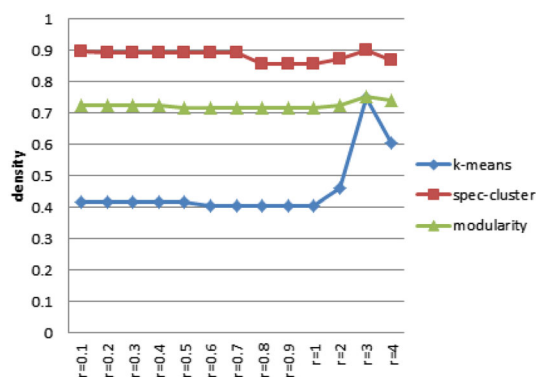
The evaluation indexes on synthetic dataset are exhibited in Figs. 2, 3, 4, and 5. In general, a lower value of variable  $\gamma$  results in a lower modularity. The curve of  $\gamma$  appears to be



**Fig. 2** Curve graphs show feature weight factor  $\gamma$  and cluster number  $k$  versus modularity is analysed on synthetic dataset



**Fig. 3** Curve graphs show feature weight factor  $\gamma$  and cluster number  $k$  versus SSE is analysed on synthetic dataset



**Fig. 4** Curve graphs show feature weight factor  $\gamma$  and cluster number  $k$  versus density is analysed on synthetic dataset

smoother with the variation of SSE and density. According to Fig. 3, the k-means and spec-cluster combined algorithm brings improved density values.

In Figs. 6, 7, and 8 we can see that our proposed similarity method combined with the k-means and spec-cluster algorithm yields improved density values.

As can be seen from the results as Figs. 6, 7, and 8, the proposed method dose have a better implementation compared to other models. In the analysis of modularity, the similarity

measure-based algorithm gains a larger value in clustering (Fig. 6), which is because of the optimization of vertexes distance during pre-processing. Meanwhile, since the initial cluster vertex and vertex features are formulated, the SSE value for K-means tends to be the best outcome (Fig. 7). Further, with the increasing of  $k$ , the value of density reduces, especially for K-means algorithm (Fig. 8), which generate a more desirable clustering outcome.

Besides, on the basic of the same vertex feature and topology structure, the resolution of different processing on the density are tested. The proposed detection algorithm obtained the best value of density by selecting four different K values, and the results are shown in Figs. 8, and 9, which is stronger than that of other methods. Under the condition of  $K = 3$ , the density reaches its best performance. Results in other cases are slightly better than the traditional technique results when processing separately.

## 7 Concluding remarks

In this paper, a multi-similarity-based method for community detection is described. The approach is basically obtained via the optimization of similarity measure as well as the analysis of social network. The method presented is improved with the combination of clustering algorithms. Specific properties are put forward for methodology design and interpretation. Results indicate that the working performance of a structured community detection is estimated by three variables.

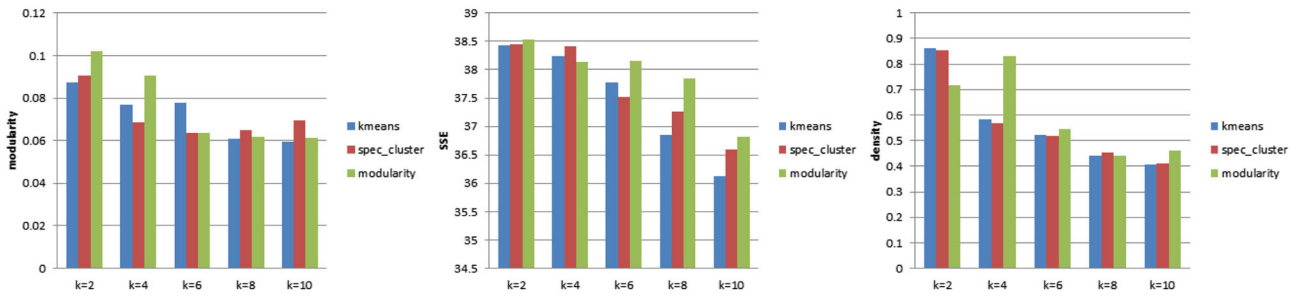
Firstly, by investigating the characteristic of social network, the community structure is depicted as the detecting target. This analysis also illustrates the similarity relation of inner and external vertexes. The application of similarity measure is provided with the property indexes for similarity-based method evaluation.

Secondly, the devise of the similarity-based detection algorithm is provided. Based on the address of three different relations, the algorithm is formulated. The multi-similarity vertexes can be effectively detected due to the satisfactory of detection algorithm to the evaluation indexes.

Thirdly, the community detection method is optimized with the integration of K-means clustering, spectral clustering and modularity algorithm. The input data is formulated as well as the initialization of clustering is revised, which offers an efficient center selecting procedure.

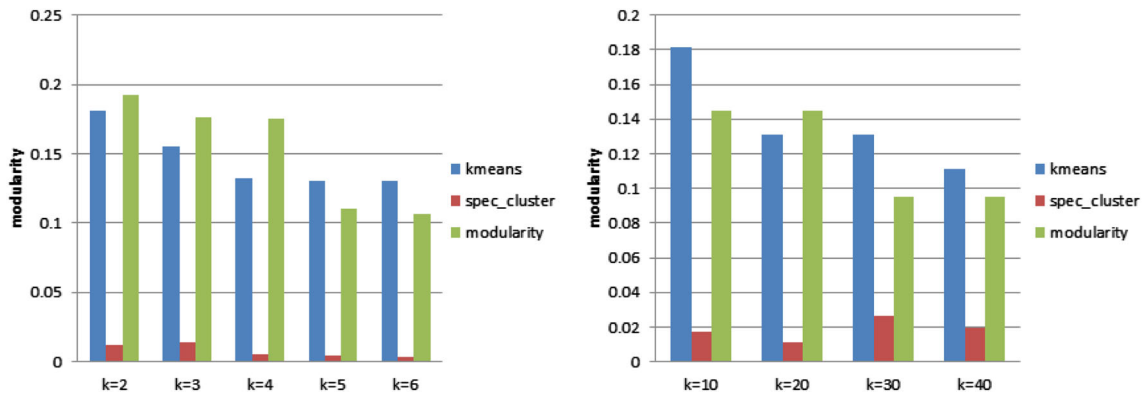
The paper also presented a laboratory architecture for effectiveness certification. A better working performance is obtained on community datasets by highlighting the results of the three criteria. In addition, all experiments are conducted on traditional clustering algorithms while results are carefully analyzed with density function.

This study introduces a new social network detection opportunity to communities of multi-similarities, which also



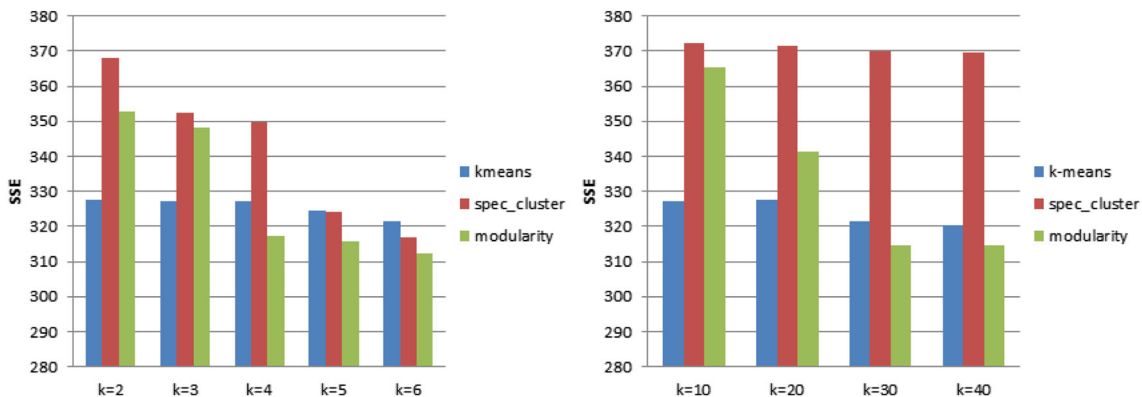
**Fig. 5** The histograms describe no. of cluster number  $k$  versus modularity, SSE, and density on synthetic dataset in three kinds of cluster algorithm:  $k$  means ( $k$  means community detection algorithm based on

MSM), spec-cluster (spectral clustering algorithm based on the MSM), and modularity (modularity maximization cluster algorithm)



**Fig. 6** No. of cluster number  $k$  versus modularity on pblog dataset (the first figure) and DBLP dataset (the second figure), using three kinds of cluster algorithm:  $k$  means ( $k$  means community detection algorithm

based on MSM), spec-cluster (spectral clustering algorithm based on the MSM), and modularity (modularity maximization cluster algorithm)



**Fig. 7** No. of cluster number  $k$  versus SSE on pblog dataset (the first figure) and DBLP dataset (the second figure), using three kinds of cluster algorithm:  $k$  means ( $k$  means community detection algorithm based on

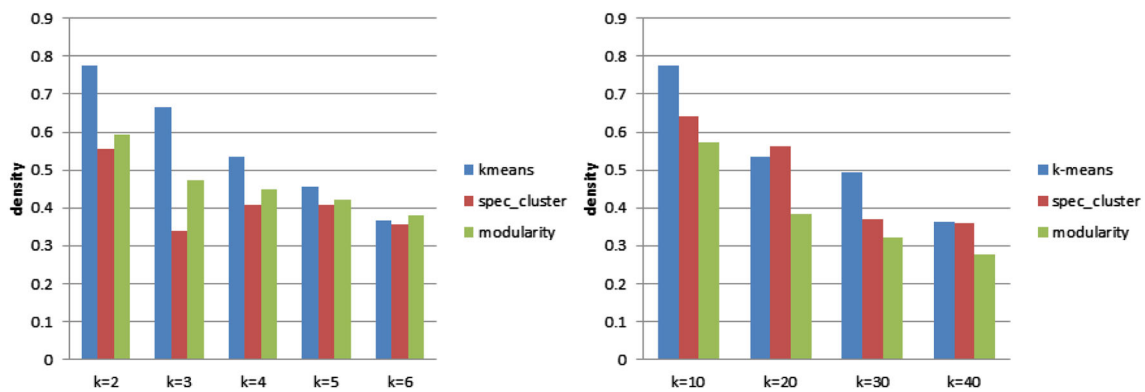
MSM), spec-cluster (spectral clustering algorithm based on the MSM), and modularity (modularity maximization cluster algorithm)

facilitate the design processes. A detection strategy like this, even if imperfect, may allow users to point in complex communities in social networks.

Future work should address more sophisticated situations where large-scale and multiple data are represented to explore

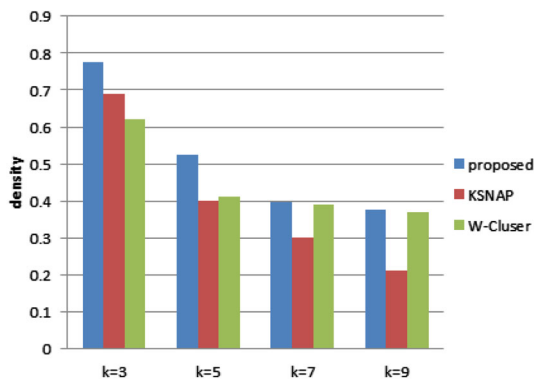
whether the current multi-similarity-based method can be extended to a more complex community.





**Fig. 8** No. of cluster number  $k$  versus density on pblog dataset (the first figure) and DBLP dataset (the second figure), using three kinds of cluster algorithm:  $k$  means ( $k$  means community detection algorithm

based on MSM), spec-cluster (spectral clustering algorithm based on the MSM), and modularity (modularity maximization cluster algorithm)



**Fig. 9** No of cluster number  $k$  versus density. Cluster density is analysed on pblog dataset using three kinds of cluster algorithm: proposed, KSNAP, and W-Cluster. “proposed” is the community detection algorithm of  $k$ means in Algorithm 1

**Acknowledgements** The research was partially funded by the Key Program of National Natural Science Foundation of China (Grant Nos. 61133005, 61432005), the National Natural Science Foundation of China (Grant Nos. 61370095, 61472124, 61572175), International Science & Technology Cooperation Program of China (2015DFA11240). The work was supported by research Project of the Education Department of Hunan Province (Grant No. 14c0210).

## References

- Hechter, M.: Principles of Group Solidarity. University of California Press, Berkeley (1988)
- Liu, C., Liu, J., Jiang, Z.: A multiobjective evolutionary algorithm based on similarity for community detection from signed social networks. *IEEE Transact. Cybern.* **44**(12), 2274–2287 (2014)
- Alhadj, R.: Introduction to the second issue of Social Network Analysis and Mining journal: scientific computing for social network analysis and dynamicity. *Soc. Netw. Anal. Mining* **1**, 73–74 (2011)
- Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. *Nature* **393**(6638), 440–442 (1998)
- Adamic, L.A., Huberman, B.A.: Power-law distribution of the world wide web. *Science* **287**(5461), 2115 (2000)
- Devinatz, V.G.: Introduction to social networking, management responsibilities, and employee rights: the evolving role of social networking in employment decisions. *Empl. Responsib. Rights J.* **27**, 305–306 (2015)
- Lee, S.-H., Kim, J.-M., Choi, Y.-K.: Similarity measure construction using fuzzy entropy and distance measure. *LNAI* **4114**, 952–958 (2006)
- Ronald, R.Y.: Monitored heavy fuzzy measures and their role in decision making under uncertainty. *Fuzzy Sets Syst.* **139**(3), 491–513 (2003)
- Rbill, Y.: Decision making over necessity measures through the Choquet integral criterion. *Fuzzy Sets Syst.* **157**(23), 3025–3039 (2006)
- Hsieh, C.H., Chen, S.H.: Similarity of generalized fuzzy numbers with graded mean integration representation. In: Proceedings of the Eighth International Fuzzy Systems Association World Congress, pp. 551–555 (1999)
- Chen, S.-M., Chen, J.-H.: Fuzzy risk analysis based on similarity measures of generalized fuzzy numbers. *IEEE Transact. Fuzzy Syst.* **11**(1), 450–56 (2003)
- Tan, P.N., Steinback, M., Kumar, V.: Introduction to data mining. *Data Anal. Cloud* **22**(6), 1–25 (2006)
- Jeh, G., Widom J.: SimRank: a measure of structural-context similarity. In: Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 538–543 (2002)
- Yu, W., Lin, X., Zhang, W., et al.: More is simpler: effectively and efficiently assessing vertexpair similarities based on hyperlinks. *Proc. Vldb Endow.* **7**(1), 13–24 (2013)
- Zhao, P., Han, J., Sun, Y.: P-Rank: a comprehensive structural similarity measure over information networks[C]. In: Proceedings of the ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November pp. 553–562 (2009)
- Antonellis, I., Molina, H.G., Chang, C.C.: Simrank++: Query rewriting through link analysis of the click graph. *Comput. Sci.* **1**, 1177–1178 (2007)
- MacKay, D.C.: Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge (2003)
- Ding, S., Jia, H., Zhang, L.: Research of semi-supervised spectral clustering algorithm based on pairwise constraints. *Neural Comput. Appl.* **24**(1), 211–219 (2014)

19. Nawaz, W., et al.: Intra graph clustering using collaborative similarity measure. *Distrib. Parallel Databases* **33**(4), 583–603 (2015)
20. Sun, Y., Han, J., Yan, X., et al.: PathSim: meta path-based top-K similarity search in heterogeneous information networks. *Proc. VLDB Endow.* **4**(11), 992–1003 (2011)
21. Qin, X., Dai, W., Jiao, P., et al.: A multi-similarity spectral clustering method for community detection in dynamic networks. *Sci. Rep.* **6**, 31454 (2016)
22. Niknam, T., Amiri, B., Olamaei, J., AREFI, A.: An efficient hybrid evolutionary optimization algorithm based on PSO and SA for clustering. *J. Zhejiang Univ.* **10**(4), 512–519 (2009)
23. Page, L.: The PageRank Citation Ranking: Bringing Order to the Web, pp. 1–14. Stanford InfoLab, Stanford (1998)
24. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. *Phys. Rev. E: Stat.* **69**(2), 026113 (2003)
25. Macqueen J.: Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297 (1967)
26. Shi, J., Malik, J.: Normalized cuts and image segmentation. *IEEE Transact. Pattern Anal. Mach. Intell.* **22**(8), 888–905 (2002)
27. Park, E.J., Dollinger, A., Huether, L., et al.: The nano-fractal structured tungsten oxides films with high thermal stability prepared by the deposition of size-selected W clusters. *Appl. Phys. A* **123**(6), 418 (2017)
28. Tian, Y., Hankins, R.A., Patel, J.M.: Efficient aggregation for graph summarization. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 567–580 (2008)



**Li Ni** received the Bachelor's degree in computer science from University of National Defence Technology, Changsha, China, in 2002, and the Master's degree in computer software and theory from XiangTan University, in 2009. She has been a Ph.D. candidate at Hunan University, Changsha, China, since September 2014. Her research interests include data mining, social community detection and so on.



**Pen ManMan** Ph.D., Professor, doctoral tutor. In 1985 graduated from the Department of computer science at Hunan University. In 1988 Master degree from Hunan University in Computer Science. In 2006, Ph.D. is graduated from Hunan University College of computer and communication. After graduating from school to teach master. Her research interests include data mining, computer architecture, big data computing and so on.



**Jiang Wenjun** received the Bachelor's degree in computer science from Hunan University, Changsha, China, in 2004, and the Master's degree in computer software and theory from Huazhong University of Science and Technology, Wuhan, China, in 2007. She has been a Ph.D. candidate at Central South University, Changsha, China, since September 2009. Currently, she is a visiting Ph.D. student at Temple University, Philadelphia, USA. Her research interests include trust and social influence evaluation models and algorithms in online social networks.



**Li Kenli** received the Ph.D. degree in computer science from HuaZhong University of Science and Technology, China, in 2003, and the M.S. degree in mathematics from Central South University, Hunan, China, in 2000. He has been a visiting scholar at University of Illinois at Champaign-Urbana and Urbana from 2004 to 2005. He is now a professor of computer science and technology at Hunan University, China. He has published more than 130 papers in international conferences and journals. He is currently served on the editorial boards of *IEEE Transactions on Computers*. He is a senior member of CCF. He is outstanding youth scientific researcher in China. His major research contains parallel computing, data mining and so on.