Semi-supervised Community Detection Framework Based on Non-negative Factorization Using Individual Labels

Zhaoxian Wang[1], Wenjun Wang[1], Guixiang Xue[2], Pengfei Jiao[1], and Xuewei Li[1(✉)]

[1] Tianjin Key Laboratory of Cognitive Computing and Application, School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
{zhaoxian_wang,wjwang,pjiao,lixuewei}@tju.edu.cn
[2] School of Computer Science and Software, Hebei University of Technology, Tianjin 300130, China

Abstract. Community structure is one of the most significant properties of complex networks and is a foundational concept in exploring and analyzing networks. Researchers have concentrated partially on the topology information for community detection before, ignoring the prior information of the complex networks. However, background information can be obtained from the domain knowledge in many applications in advance. Especially, the labels of some nodes are already known, which indicates that a point exactly belongs to a specific category or does not belong to a certain one. Then, how to encode these individual labels into community detection becomes a challenging and interesting problem. In this paper, we present a semi-supervised framework based on non-negative matrix factorization, which can effectively incorporate the individual labels into the process of community detection. Promising experimental results on synthetic and real networks are provided to improve the accuracy of community detection.

Keywords: Community detection · Semi-supervised framework · Non-negative Matrix Factorization (NMF) · Individual label

1 Introduction

Many systems take the form of networks, such as social and biological networks. An important property of the network is community structure which is first proposed by Girvan and Newman [1]. Community is a subgraph in which the vertices are more tightly connected with each other than with the vertices outside the subgraphs [2]. The nodes in the same community have similar features. Detecting the community can help us understand and analyze the network more deeply.

In the past few years, a large number of methods have been proposed to detect communities in the complex networks, including GN algorithm proposed by Girvan and Newman [3], modularity-based methods [4], stochastic blockmodels [5] and so on. Most of these approaches only take the topology information into consideration,

Y. Tan et al. (Eds.): ICSI-CCI 2015, Part II, LNCS 9141, pp. 349–359, 2015.
DOI: 10.1007/978-3-319-20472-7_38

little considering the background information. However, in the real world, some prior information can be learned from the network, which should be useful for us to identify the community structure.

Recently, many semi-supervised community detection algorithms have been proposed [6]-[9]. Ma el al proposed a semi-supervised method based on symmetric non-negative matrix, which incorporates the pairwise constraints into the adjacency matrix for finding the community structure [6]. A semi-supervised method based on the spin-glass model from statistical physics can integrate the prior information in forms of individual labels and pairwise constrains into community detection proposed by Eaton and Mansbach [7].Zhang [8] studied a semi-supervised learning framework which encodes pairwise constraints by modifying the adjacency matrix of network, which can also be regarded as de-noising the consensus matrix of community structures. Later, Zhang [9] added a logical inference step to utilize the must-link and cannot-link constraints fully. These algorithms use the prior information by transferring and modifying the adjacency matrix directly. After reconstructing the adjacency matrix, the semi-supervised problem is transformed into an unsupervised one [10].

Will the important nodes in the priors affect the result of community detection? We first extract the individual labels randomly, and later select the nodes in prior information concerning its importance. The centrality of nodes in networks can be assessed by degree, betweenness, closeness, eigenvector and so on [11-12].

In this paper, we propose a semi-supervised framework for community detection based on the NMF. One contribution of our framework is that it constructs a matrix by the positive and negative labels to more fully utilize the prior information. Another contribution is that we research the effect of important nodes in the priors on the community detection. This framework is applied to the artificial and real networks. The experimental results show that the framework can significantly improve the detection performance.

The remainder of this paper is organized as follows. Section 2 includes the review of basic problem formulation and notations used in our framework. In Section 3, we describe our semi-supervised community detection framework in detail. Experimental results on artificial and real-world networks are given in Section 4. Finally, a conclusion is presented in Section 5.

2 Problem Formulation and Notations

We first give the notation of network which will be used throughout the paper. A network can be modeled as a graph G= (V, E), where V is the node set and E is the edge set. The network structure is defined by a N×N adjacency matrix A. N is the number of nodes. If there is an edge between node i and j, we set the element A_{ij} to 1, otherwise to 0. We assume G is an undirected and unweighted graph. Self-connections are not allowed.

NMF was first introduced by Lee and Seung as a method for study the substructure of data matrix [13]. It was defined as the factorization of a non-negative matrix A into the multiplication of two other non-negative matrices U and V, where A is a $N \times N$

matrix, U and V are $N \times K$ matrices, where K is the target number of communities to be detected in the network. In other words, NMF was aimed at mining the Euclidian distance between A and UV^T. The community structure can be inferred from V: node i belongs to the community k if V_{ik} is the largest element in the i-th row. We use the next objective (loss) function to quantify the quality of the factorization result. This function is based on the square loss function [14], which is equivalent to the square of the Frobenius norm of the difference between two matrices and is presented as follows.

$$L_{LSE}\left(A, UV^T\right) = ||A - UV^T||_F^2 \ . \tag{1}$$

Lee and Seung [15] presented iterative updating algorithms to minimize to the objective function L_{LSE} as follows.

$$(U)_{ij} \longleftarrow (U)_{ij} \frac{(AV)_{ij}}{\left(UV^TV\right)_{ij}}, (V)_{ij} \longleftarrow (V)_{ij} \frac{\left(A^TU\right)_{ij}}{\left(VU^TU\right)_{ij}}. \tag{2}$$

The prior information contains individual labels and pairwise constraints and we use the former one in this paper. There are positive and negative labels in the individual labels. If a node does belong to a community, we call this positive label (PL), while if a node does not belong to a community, we regard it as the negative label (NL). The matrix O of size N×K is constructed from the background information, where N is the number of nodes and K is the target number of communities in the complex network. For any node i, we define i-th row of O as follows.

1. If node i has the PL, and i belongs to the j-th community, then

$$O_{ik} = \begin{cases} 1, \ if \ k = j \\ 0, \ if \ k \neq j \end{cases} \text{, for k=1, 2,…K.} \tag{3}$$

2. If node i has the NL, and i does not belongs to the j-th community , then

$$O_{ik} = \begin{cases} 0, \ if \ k = j \\ \frac{1}{K-1}, \ if \ k \neq j \end{cases} \text{, for k=1, 2,…K.} \tag{4}$$

3. If node i has no priors, then

$$O_{ik} = \frac{1}{K} \text{, for k=1, 2,…K.} \tag{5}$$

In this paper, we use the normalized mutual information (NMI) to evaluate the performance of our framework on detecting the community structure [16]. This value can be formulated as follows. In Eq.6, R is the real community result and B is the found community result. In general, the larger the value of NMI, the better the partition of the network will be.

$$\mathrm{NMI(R, B)} = \frac{-2\sum_{i=1}^{c_R}\sum_{j=1}^{c_B} N_{ij} \log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{c_R} N_{i.} \log\left(\frac{N_{i.}}{N}\right) + \sum_{j=1}^{c_B} N_{.j} \log\left(\frac{N_{.j}}{N}\right)}. \tag{6}$$

3 The Semi-supervised Community Detection Framework Based on NMF Using Individual Label

In this section, based on the individual labels discussed above, we first present the semi-supervised community detection framework which incorporates the prior information into the NMF objective function. Then we will see how the important nodes in prior information affect the result of community detection.

3.1 Description of Our Framework Based on NMF with Individual Labels

In this section, we propose the semi-supervised framework based on NMF which can make use of the individual labels to improve the performance of community detection. NMF can factorize a non-negative matrix A into the multiplication of two other non-negative matrices U and V, where A is an N×N matrix, both U and V are N×K matrices. We can infer the community structure in the network from V. In the i-th row, it is easy to known, if V_{ik} is the largest element, then node i belongs to the community k. If we have known that node i belongs to the community k, then we can enhance the value of V_{ik}, however, if that node i does not belong to the community k, then the value of $\mathrm{V_{ik}}$ will be punished.

To use the individual labels to improve the result, we denote the new representation of V where the matrix O summarized from the individual labels are used as a multiplication factor in Eq.7. In this paper, the sign of operation ⊗ indicates the dot product.

$$\mathrm{d(O,V)} = \mathrm{O} \otimes \mathrm{V}\,. \tag{7}$$

For NMF there is an interesting fact that the estimates are always scale invariant. For example, we add a multiplication factor c to U and the other factor $\frac{1}{c}$ to V to get different U and V. The product UV^T will not change. Although there is no explicit control for NMF, standard NMF tends to estimate sparse components. The factorized matrices are obtained through minimizing an objective function defined in Eq.8.

$$\min_{U\geq 0, V\geq 0} ||A - UV^T||^2 + \lambda_s \sum_{k=1}^{K} ||V_k||_1\,. \tag{8}$$

In the formula (5), the parameter $\lambda_s \geq 0$.Adding penalties to NMF is a common strategy since they not only improve the interpretability, but also improve numerical stability of the estimation by making the NMF optimization less under constrained. The assessment algorithm for the penalized NMF is studied in many papers. The main iteration rule in our work is presented as follows. And the parameter $ë_s$ is set to 1 in the experiments.

The algorithm of Sparse NMF with individual labels is described.

```
program Inflation
  const λ_s = 1;
  begin
    construct O;
    initialize {U, V}, positive random matrices;
    repeat
      set (U)_ij ← (U)_ij (AV)_ij / (UV^T V)_ij;
      set (V)_ij ← (V)_ij (A^T U)_ij / (VU^T U)_ij + λ_s;
      V=O⊗V;
      normalization of U,V;
    until convergence
end
```

3.2 The Evaluation of the Important Nodes

In the research of the social network, many methods have such a hypothesis, namely the importance of node is equivalent to its connection with other nodes, which makes the node significant. The basic idea of these methods is the importance of difference between different nodes in the network is obtained by some useful information, such as the degree of node, the shortest path, the weight of nodes and edges.

The proposed indexes of important nodes mainly can be divided into centrality and prestige. Measurement methods mainly include the node degree, betweenness, closeness, eigenvector and so on. In this paper, we use the degree and betweenness of nodes. A brief introduction about these two methods will be presented in the following.

The degree of nodes refers to the number of edges connected to the node in the network. The size of degree can reflect the importance of nodes to a certain extent. The larger the degree of node, the more important the node may be, because it may be located in the center of the network.

Betweenness was first put forward for measuring the individual's social status in the study of social network in 1977 by Freeman [17]. The betweenness of node u refers to all the shortest paths in the network through the node u. We define the set of shortest path between nodes i and j as s_{ij}, and the betweenness formula of u after normalization is presented as follows.

$$B_u = \sum_{i,j} \frac{\sum_{l \in s_{ij}} \delta_l^u}{|s_{ij}|}. \tag{9}$$

The symbol $\sum_{l \in s_{ij}} \delta_l^u$ is the number of shortest paths through node u.

The size of betweenness can reflect the importance of nodes in a way. The larger betweenness of the node, the more important the node is. The betweenness is useful for us to find the important nodes with large flow.

According to the importance of nodes, we select some prior information on purpose in the following experiments. The detail description about experiments will be presented in section 4.

4 Experiment and Discussion

In this section, we design a set of experiments, whose data set are LFR artificial networks [18] and real-world networks including Amazon's network of political books [19], the network of blogs about US politics [20] and adjacency network of common adjectives and nouns in the novel David Copperfield by Charles Dickens [21]. The normalized mutual information (NMI) is used to evaluate the performance of detecting communities with our framework, which is discussed above. The closer to 1 the NMI, the better the partition of the network will be.

4.1 Artificial Networks

In this subsection, the experiments include evaluating the performance of the framework with different percentage of priors and measuring the ability of the framework to detect communities with different important nodes in the background information.

The LFR benchmark network is an artificial network for community detection, which is claimed to possess some basic statistical properties found in real networks. The generator of LFR allows us to specify the number of nodes (N), average degree (k), maximum degree (maxk), exponent for the degree sequence (t1), exponent for the community size distribution (t2), minimum for the community sizes (minc), maximum for the community sizes (maxc) and the mixing parameter (u). In LFR, both community size and degree distributions are power laws, where vertices and communities are generated by sampling. With the increase of u, the structure of network becomes vague, and the detection of communities becomes more challenging.

In this paper, we set the number of nodes to 1000, the minimum community size to 10 and 20, the maximum community size to five times the minimum community size, the average degree to 20, the exponent of the vertex degree and community size to -2 and -1, respectively, and the mixing parameter to different values 0.7 and 0.8.

The percentage of the labeled nodes in the network is an important factor in the experiments. To fully use the individual labels, we construct the matrix O and incorporate it into the updating process of NMF, The average performance of our framework based on different percentage of the used priors of half positive labels and half negative labels is displayed in Fig. 1. There is a positive correlation between NMI and the used priors. There are abnormal points in the first row and the second column picture, where the value of NMI decreases when the used priors is 1%. The NMI of the standard NMF is the NMI where the used priors are 0. Compared with standard NMF, the NMIs of our framework are higher. Obviously, when u becomes 0.8, the result of community detection by our framework turns to be weak. There is a question that the value of NMI is below to 1 when the used priors are 100% in Fig.1. In Fig.1, the priors are half positive labels and half negative labels, so when the used priors are 100%, there are some fuzzy labels in the prior information, so the result of community detection is not exactly the same as the real one.

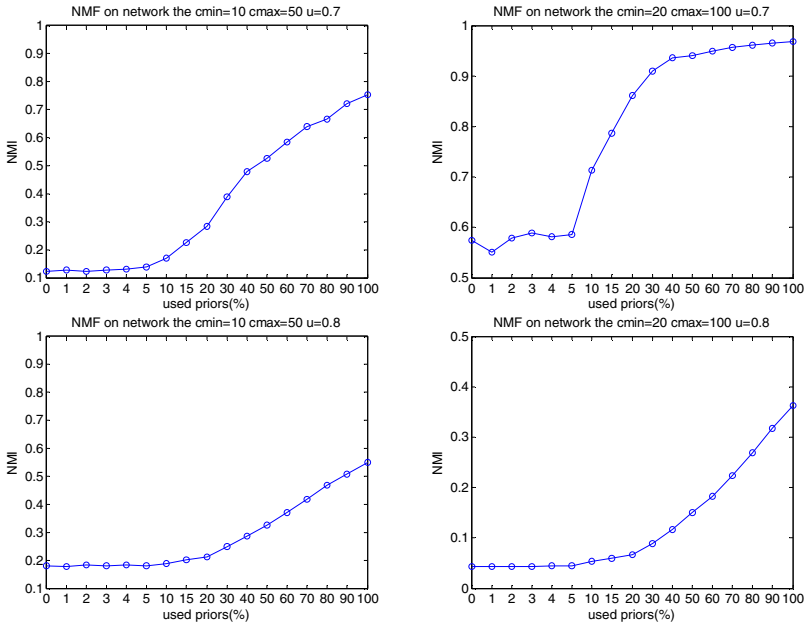

Fig. 1. Performance of our framework in the terms of NMI as a function of the percentage of priors with half positive and negative labels added on LFR networks

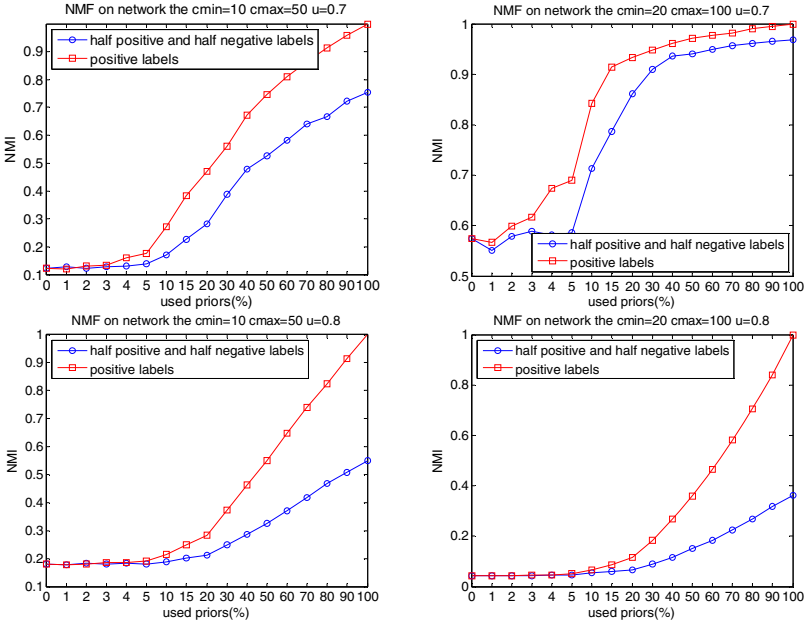

Fig. 2. Performance of our framework in the terms of NMI as a function of the percentage of priors with positive labels added on LFR networks

The individual labels contain positive and negative labels. We know that the positive labels can more accurately describe the community than the negative labels. In Fig. 2, we randomly extract the priors with all positive labels in the network. Compared with the priors with half positive and half negative labels, we can know that the positive labels are more useful than the negative labels in the process of detecting communities. From the Fig.2, we can clearly see that the NMIs increase consistently as the used priors except some nodes and it is faster than the Fig 1 in the growth trend. Compared with the result in Fig1, the value of NMI is up to 1, when the used priors are 100%, for the labels of the nodes in the priors are positive.

The important nodes in the network can be measured by degree and betweenness. To evaluating the effect of important nodes on the performance of the community detection by our semi-supervised framework, we reset the nodes according to the degree of nodes and the betweenness of nodes in descending order respectively. Different percentage of priors with nodes from top to down is obtained to be combined to the NMF's updating process. The labels in the priors are positive. The result is showed in Fig.3. Obviously, the effect of degree and betweenness is not stable. At least in our framework, their influence is not obvious. However, there is an interesting thing that when the u is 0.8, the influence of degree and betweenness is especially obvious.

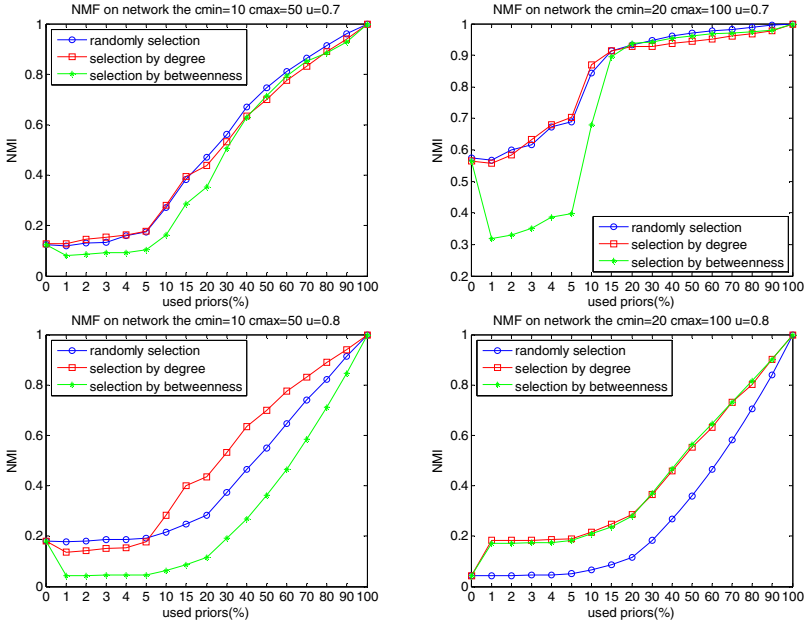

Fig. 3. Performance of our framework in the terms of NMI as a function of the percentage of priors with top degree and betweenness added on LFR networks

In summary, incorporating the individual labels into the process of NMF updating process can effectively improve the performance of the community detection in the

LFR network, especially the positive labels. However, the effect of important nodes is not obvious.

4.2 Real-World Networks

In this subsection, we test our framework with real-world networks, Amazon's network of political books, the network of blogs about US politics and the adjacency network of common adjectives and nouns in the novel David Copperfield by Charles Dickens. Firstly, we will give the data description, and then the experiments' performance will be presented.

The network of Amazon's political books contains 105 books on US politics and 441 edges. The nodes are books sold by bookseller Amazon, which have been manually given labels as "liberal", "neutral", or "conservative". Edges represent co-purchasing of books. The network of blogs about US politics consists of 1490 nodes and is treated as an undirected network in this paper. The nodes in the network are divided into "liberal" and "conservation" according to the content in the blogs, which represent the blogs and the edges in the network represent that a URL presented on the page of one blog references another political blog. If there is reference relationship between two blogs, the edge between blogs forms. There are 112 vertices and 850 edges in the adjacency network of common adjectives and nouns in the novel David Copperfield by Charles Dickens where the vertices represent common adjectives and nouns and the edges connect any two words that appear adjacent to one another at any point in the book.

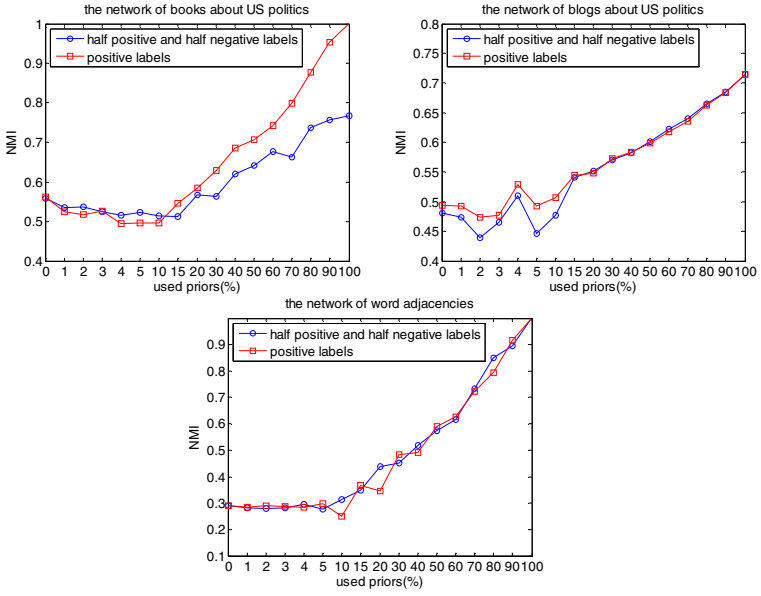

Fig. 4. Performance of our framework in the terms of NMI as a function of the percentage of priors with top degree and betweenness added on the real-world networks

Applying our proposed framework to these real-world networks, the result of the community detection is shown in Fig.4. The performance of our framework on the real-world networks is consistent with that on the LFR network. However, the result of the network of blogs about US politics is abnormal and when the used priors is set 100%, the NMI is less than 1. In Fig.4b and Fig.4c the two lines overlap. There are two communities in these two real networks, which is known previously, based on the construction of O discussed above, we can find that if node i does not belong to one community, it must belong to the other one, then there is no difference between the positive labels and negative labels because we can construct the same O at node i. Further, the percentage of priors is important to the result of community detection.

5 Conclusions

In this paper, a semi-supervised community detection method based on NMF with individual labels is proposed. Unlike previous works which transfer the individual labels into the adjacency matrix, we formulate it into the objective function and incorporate it into the process of NMF updating. As can be seen from the extensive experiments on both artificial and real networks that using the individual labels can significantly improve performance, especially in the situation where the individual labels are positive. Moreover, we extract some important nodes with large degree and betweenness into the priors and the effect of these nodes on the community detection is not obvious.

A number of improvements of our framework may be possible. Firstly, we hope to apply the semi-supervised framework to other matrix-based community detection methods, such as spectral clustering and its variants. Secondly, it would be interesting to investigate the abnormal phenomenon in the experiments. With the increase of used priors, the performance of community detection is poor at some points. In this case the research about how to improve the result is meaningful. Finally, we will investigate how the priors guide the process of community detection.

Acknowledgments. This work was supported by the Major Project of National Social Science Fund (14ZDB153), the National Science and Technology Pillar Program (2013BAK02B06 and 2015BAL05B02), Tianjin Science and Technology Pillar Program (13ZCZDGX01099, 13ZCDZSF02700), National Science and Technology Program for Public Well-being (2012GS120302).

References

1. Strogatz, S.H.: Exploring complex networks. J. Nature **410**(6825), 268–276 (2001)
2. Newman, M.E.J.: Detecting community structure in networks. J. The European Physical Journal B-Condensed Matter and Complex Systems **38**(2), 321–330 (2004)
3. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. J Proceedings of the National Academy of Sciences **99**(12), 7821–7826 (2002)
4. Newman, M.E.J.: Modularity and community structure in networks. J Proceedings of the National Academy of Sciences **103**(23), 8577–8582 (2006)

5. Karrer, B., Newman, M.E.J.: Stochastic blockmodels and community structure in networks. J. Physical Review E **83**(1), 016107 (2011)
6. Ma, X., Gao, L., Yong, X., et al.: Semi-supervised clustering algorithm for community structure detection in complex networks. J Physica A: Statistical Mechanics and its Applications **389**(1), 187–197 (2010)
7. Eaton, E., Mansbach, R.: A Spin-Glass Model for Semi-Supervised Community Detection. AAAI (2012)
8. Zhang, Z.Y.: Community structure detection in complex networks with partial background information. J. EPL (Euro physics Letters) **101**(4), 48005 (2013)
9. Zhang, Z.Y., Sun, K.D., Wang, S.Q.: Enhanced community structure detection in complex networks with partial background information. J. Scientific reports (2013)
10. Yang, L., Cao, X., Jin, D., et al.: A Unified Semi-Supervised Community Detection Framework Using Latent Space Graph Regularization. J (2014)
11. Nan, H., Wen-Yan, G.: Evaluate nodes importance in the network using data field theory. In: International Conference on Convergence Information Technology, pp. 1225–1234. IEEE (2007)
12. Freeman, L.C.: Centrality in social networks conceptual clarification. J. Social networks **1**(3), 215–239 (1979)
13. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. J. Nature **401**(6755), 788–791 (1999)
14. Wang, R.S., Zhang, S., Wang, Y., et al.: Clustering complex networks and biological networks by non-negative matrix factorization with various similarity measures. J. Neurocomputing **72**(1), 134–141 (2008)
15. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. Advances in neural information processing systems, 556–562 (2001)
16. Zhong, S., Ghosh, J.: Generative model-based document clustering: a comparative study. J. Knowledge and Information Systems **8**(3), 374–384 (2005)
17. Freeman, L.C.: A set of measures of centrality based on betweeness. J. Sociometry, 35–41 (1977)
18. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. J. Physical review E **78**(4), 046110 (2008)
19. Newman, M.E.J.: Modularity and community structure in networks. J Proceedings of the National Academy of Sciences **103**(23), 8577–8582 (2006)
20. Adamic, L.A., Glance, N.: The political blogosphere and the 2004 US election: divided they blog. In: Proceedings of the 3rd international workshop on Link discovery, pp. 36–43. ACM (2005)
21. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. J. Physical review E **74**(3), 036104 (2006)