Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 109671, 8 pages
http://dx.doi.org/10.1155/2015/109671

Research Article

Effective Semisupervised Community Detection Using Negative Information

Dong Liu,[1] Dequan Duan,[1] Shikai Sui,[1] and Guojie Song[2]

[1]*School of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China*
[2]*School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China*

Correspondence should be addressed to Dong Liu; liudonghtu@gmail.com

Received 5 June 2014; Accepted 13 October 2014

Academic Editor: Qinggang Meng

The semisupervised community detection method, which can utilize prior information to guide the discovery process of community structure, has aroused considerable research interests in the past few years. Most of the former works assume that the exact labels of some nodes are known in advance and presented in the forms of individual labels and pairwise constraints. In this paper, we propose a novel type of prior information called negative information, which indicates whether a node does not belong to a specific community. Then the semisupervised community detection algorithm is presented based on negative information to efficiently make use of this type of information to assist the process of community detection. The proposed algorithm is evaluated on several artificial and real-world networks and shows high effectiveness in recovering communities.

1. Introduction

Many networked systems, including social and biological networks, are found to divide natural communities, that is, groups of vertices which are densely connected to each other while less connected to the vertices outside [1]. The community structure in real networks always has a specific function such as cycles or pathways in metabolic networks or collections of pages on the same or related topics on the web community [2]. To comprehensively understand the function of different networks, much research effort has been devoted to develop methods that can extract community structure from networks.

A lot of models and algorithms have been proposed for community detection, such as betweenness-based algorithms [1, 3], modularity-based methods [2, 4–6], spin model [7], and stochastic blockmodels [8]; see [9, 10] for a more comprehensive review. However, almost all existing approaches for community detection only make use of the network topology information, which completely ignore the background information of the network. However, in many real-world applications, we may know some prior information that could be useful in detecting the community structures. For instance, a few proteins have been known to belong to certain functional classes in protein-protein interaction networks [11]. Therefore, how to utilize prior information to guide the discovery process of community structure is an interesting question that is worthy of working on.

In recent years, a variety of semisupervised community detection algorithms have been proposed. Ma et al. [12] proposed a semisupervised method based on symmetric nonnegative matrix factorization, which incorporates pairwise constraints (via must-links and cannot-link) on the cluster assignments of nodes for identifying community structure in network. Eaton and Mansbach [13] presented a semisupervised algorithm based on spin-glass model, which can incorporate prior knowledge in the forms of individual labels (via known cluster assignments for a fraction of nodes) and pairwise constraints into the process of extracting community structure. Zhang et al. [14, 15] developed the methods that implicitly encode the pairwise constraints by modifying the adjacency matrix of the network, which can also be

regarded as the denoising process of the consensus matrix of the community structures. Liu et al. [16, 17] put forward two semisupervised algorithms based on discrete potential and label propagation, respectively. Both algorithms are especially suitable for the network with obscure community structure and exhibit almost linear complexity in time.

Although these approaches can improve accuracy and degree of noise resistant to community detection, they mostly focus on one kind of prior information; that is, the exact labels of a small portion of nodes are given. In some real application, it may not be easy to identify the exact community of a node, whereas we can easily point out the community that one node does not belong to. For a simplified example, assume that the web network can be grouped into some communities which represent pages on related topics. Further, supposing that the web page describes a female soccer game, it is hard to determine whether the web page belongs to sport community or feminism community. However, it does not belong to automobile community.

In machine learning, the negative information was first proposed by Hou et al. [18]. In their work, the negative information indicates whether a point does not belong to a specific category. They utilized the negative information to guide the process of semisupervised learning and made some experiments on image, digit, spoken letter, and text classification tasks. The experimental results showed the effectiveness of negative information. As far as we know, there is no community detection method concerning the negative information, although this information arises naturally in some applications.

In this paper, we propose a novel semisupervised community detection approach based on negative information. It has near-linear complexity in time and can incorporate the negative information into community detection. The algorithm has been evaluated on synthetic LFR benchmark networks [19] and on various real-world networks with community structure. The results show that negative information is helpful to improve the accuracy of identifying communities. Specifically, the algorithm exhibits almost linear complexity in time.

The rest of the paper is structured as follows. Section 2 includes reviews of the basic formulation and notations used in our approach. In Section 3, we describe our new semisupervised community detection algorithm 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 notations of network representation which will be used throughout this paper. Let $G = \{V, E\}$ denote an unweighted and undirected network, where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes and $E \subseteq V \times V$ is the set of edges. Multiple edges and self-connections are not allowed. The network structure is determined by $n \times n$ adjacency matrix A. Each element A_{ij} of A is equal to 1 if there is an edge connecting nodes i to j, and it is 0 otherwise. If there are k communities, a community-number (label) set $K = \{1, 2, \ldots, k\}$ is defined.

Assume that there are three kinds of nodes, that is, traditional label (TL) nodes V^{TL}, negative label (NL) nodes V^{NL}, and unlabeled (UL) nodes V^{UL}. Define the set of TL nodes $V^{\mathrm{TL}} = \{v_1, v_2, \ldots, v_t\}$ with cardinality $|V^{\mathrm{TL}}| = t$, the set of NL nodes $V^{\mathrm{NL}} = \{v_{t+1}, v_{t+2}, \ldots, v_{t+l}\}$ with cardinality $|V^{\mathrm{NL}}| = l$, and the set of UL nodes $V^{\mathrm{UL}} = \{v_{t+l+1}, v_{t+l+2}, \ldots, v_{t+l+u}\}$ with $|V^{\mathrm{NL}}| = u$, where typically $t \ll u$, $l \ll u$, and $t + l + u = n$. Further, suppose that we are given a set of nodes $V = V^{\mathrm{TL}} \cup V^{\mathrm{NL}} \cup V^{\mathrm{UL}}$. The label indicator matrix of V^{TL} is defined as follows: $L^{\mathrm{TL}}_{ij} = 1$ if and only if v_i belongs to the jth community; otherwise $L^{\mathrm{TL}}_{ij} = 0$. We define the label indicator matrix of V^{NL} as $L^{\mathrm{NL}}_{ij} = 1$ if and only if v_i does not belong to the jth community; otherwise $L^{\mathrm{NL}}_{ij} = 0$. Note that, different from L^{TL}, the row vectors of L^{NL} may have more than one element which is equal to 1. The goal of our approach is to infer the exact labels for nodes in $V^{\mathrm{NL}} \cup V^{\mathrm{UL}}$.

In this paper, label propagation task is to propagate the TL under the guidance of NL information to all of the nodes in $V^{\mathrm{NL}} \cup V^{\mathrm{UL}}$, accomplishing label prediction of nodes without TL. The result of label propagation for community detection depends on the weights of the edges of network, so how to construct the weight matrix W plays a decisive role. In this work, the simple weight matrix can be defined as

$$W_{ij} = \frac{A_{ij}}{\deg(i)}, \tag{1}$$

where $\deg(i)$ represents the degree of node i. Obviously, in label propagation process, the labeled nodes propagate seed labels to their neighbours with uniform probability.

3. The Proposed Algorithm

In this section, the details of our proposed algorithm based on negative information are presented, and then the time complexity and the convergence property of the algorithm are analyzed. There are mainly two steps of the algorithm. The first is to determine the particular parameter matrices, and the second is to propagate labels via an iterative process.

3.1. Parameter Matrices Construction. Using the idea of the work by Hou et al. [18], we introduce two matrices, that is, the initial label matrix $L \in R^{(t+l+u)\times K}$ and the parameter matrix $P \in R^{(t+l+u)\times K}$, where K is number of communities. L_{ij} represents the probability that v_i belongs to the jth community, and P is a matrix that shows the role of each node and indicates when an NL node can be regarded as a TL node and when it is considered as an unlabeled node. We also define two parameters $0 < \alpha_l < 1$ and $0 < \alpha_u < 1$, which take different values for labeled nodes (including TL and NL) and unlabeled nodes.

For any node v_i, the $L_{i\cdot}$ and $P_{i\cdot}$ are defined as follows.

(1) If v_i Has the TL. Based on the indictor matrix L^{TL}, if v_i belongs to the jth community, then

$$L_{ix} = \begin{cases} 1, & x = j; \\ 0, & x \neq j, \end{cases} \tag{2}$$

$$P_{ix} = \alpha_l \quad \text{for } x = 1, 2, \ldots, k. \tag{3}$$

(2) If v_i Has the NLs. According to the L^{NL}, we can define an index set $I_i = (i_1, i_2, \ldots, i_p)$, which contains the sets that v_i does not belong to; then

$$L_{ix} = 0 \quad \text{for } x = 1, 2, \ldots, k. \tag{4}$$

$$P_{ix} = \begin{cases} \alpha_l, & x \in I_i; \\ \alpha_u, & x \notin I_i. \end{cases} \tag{5}$$

(3) If v_i Is an Unlabeled Node. Consider

$$L_{ix} = 0 \quad \text{for } x = 1, 2, \ldots, k, \tag{6}$$

$$P_{ix} = \alpha_u \quad \text{for } x = 1, 2, \ldots, k. \tag{7}$$

How to make use of these two matrices in the proposed algorithm will be explained in the next subsection. Note that α_l is close to 0 and α_u is close to 1.

3.2. Description of the Algorithm. The algorithm is motivated by the fact that the nodes having the same traditional label are grouped together as one community through labels propagation process. We initialize a small number of nodes with user-defined labels based on prior information (including TLs and NLs) and let the TLs propagate through the network. As the labels propagate, the exact labels of the NL and unlabeled nodes can be achieved. Then we will show how to iteratively propagate the TL under the guidance of NL information and unlabeled nodes.

This process is iteratively performed, where, at every step, each node absorbs some label information from its neighbors and retains some label information of its initial state. Let $\mathscr{F}$ denote a set of computed label matrices, for all $F \in \mathscr{F} \subset R^{(t+l+u)\times K}$; its row vector corresponds to the possibilities of a specific node belonging to all the communities. The exact label of one node can be determined by the index of the largest element of the corresponding row vector of F. The iterative formula is defined as follows:

$$F_{ij}^{h+1} = P_{ij} \sum_{x=1}^{n} W_{ix} F_{xj}^{h} + \left(1 - P_{ij}\right) L_{ij}, \tag{8}$$

where h denotes the times of iterations. The first term shows the label information that v_i absorbs from its neighbors and the second term represents the label information retained from its initial label.

Specifically, if v_i has a TL which indicates it belongs to the jth community, then $L_{ij} = 1$ and $L_{ix} = 0$ $(x \neq j)$. In this case $L_{ix} = \alpha_l$ for $x \in K$ and α_l is close to 0. Thus, the second term in (8) plays a major role in each iteration; that is, the predicted label is consistent with the given TL. If v_i has an NL indicating that v_i belongs to the jth community, that is, $j \in I_i$, then $P_{ij} = \alpha_l$ and $F_{ij}^{h+1} \approx (1 - P_{ij}) L_{ij} \approx L_{ij} \approx 0$. On the contrary, if $j \notin I_i$, no much prior information can help to determine whether v_i belongs to the jth community or not. Therefore, we regard it as an unlabeled node and $P_{ix} = \alpha_u$, $L_{ix} = 0$. In this case, the first term in (8) plays a major role in each iteration. If v_i is unlabeled, there is no prior information about its label and $P_{ix} = \alpha_u$, $L_{ix} = 0$. Thus (8) is dominated by its second term. In summary, the iteration equation can be rewritten as

$$F^{h+1} = PWF^{h} + (I - P) L, \tag{9}$$

where I denotes an $n \times n$ identity matrix.

To summarize, the main procedure of the method is presented in Algorithm 1.

3.3. Analysis of the Algorithm. In this subsection we will analyze our method theoretically. First we will discuss the time complexity of the algorithm. Second we will analyze the convergence property of the iteration of the algorithm.

The algorithm mainly contains three computational parts: constructing the weight matrix W, constructing the label and parameter matrices L, P, and iterating (9) until convergence. In the first part, $O(m)$ time is required to construct the weight matrices, where m denotes the number of edges. In the second part, the label and parameter matrices can be derived with computational complexity $O(n)$, where n denotes the number of nodes. In the last part, the time complexity of each iteration is $O(m)$. Assuming (9) is converged at h iterations, the last part of the algorithm requires $O(hm)$ time. Since the time complexity of the algorithm depends on the highest complexity of the three parts involved in it, the overall time complexity is $O(hm)$ for the proposed algorithm.

The convergence of the algorithm is analyzed as follows. According to the initial condition that $F^0 = L$, (9) can be rewritten as

$$F^{h} = (PW)^{h} L + (I - P) \sum_{i=1}^{h-1} (PW)^{i} L. \tag{10}$$

Since $w_{ij} \geq 0$ and $\sum_i w_{ij} = 1$, from the theorem of Perron-Frobenius [20], the spectral radius of W satisfies $\rho(W) \leq 1$. Recall that the elements of P are either α_l or α_u, where $0 < \alpha_l < 1$ and $0 < \alpha_u < 1$; thus

$$\begin{aligned} &\lim_{t \to \infty} (PW)^{t} = 0 \\ &\lim_{t \to \infty} \sum_{i=0}^{t-1} (PW)^{i} = (I - PW)^{-1}. \end{aligned} \tag{11}$$

input: adjacency matrix A, the initial label matrix L^{TL}, L^{NL}, the constants α_l and α_u
output: The TL of all the nodes
(1) construct the weight matrix W by (1).
(2) for $i = 1 : n$
(3) If v_i has the TL
(4) construct L and P by (2) and (3), respectively.
(5) If v_i has the NLs
(6) construct L and P by (4) and (5), respectively.
(7) If v_i is unlabeled node
(8) construct L and P by (6) and (7), respectively.
(9) iterate (9) until convergence.
(10) output the labels of each node v_i by $f_i = \mathrm{argmax}_{i \ll k} F_{ij}$.

Algorithm 1: Semisupervised community detection using negative information.

Obviously, (9) will converge to

$$\lim_{t \to \infty} F^t = (I - P)(I - PW)^{-1} L. \tag{12}$$

4. Experiments and Discussion

In this section, we give a set of experiments to show the effectiveness of the proposed algorithm. The relevant data sets involving the experiments are LFR artificial networks [19] and real-world networks including the Zachary's network of karate club [21] and the Lusseau's network of bottlenose dolphins [22]. In all the experiments of this section, α_l and α_u are set to 0.05 and 0.95, respectively.

4.1. Artificial Networks. In this subsection, the ability of the algorithm to identify communities is tested in LFR benchmark networks. Our experiments include evaluating the performance of the algorithm with various amounts of NL nodes, measuring the ability of the algorithm to recover communities with different parameter μ in benchmark networks, comparing the accuracy of our algorithm with label propagation algorithm (LPA) [23] and Infomap algorithm [24] and analyzing the relationship between the percentage of NL nodes and the percentage of TL nodes in the proposed algorithm. In the following experiments, the choice of NL and TL nodes is random, and the number of NLs of each NL node is set to 20 percent of the number of communities. Note that the NLs of each NL node are selected randomly.

The LFR benchmark network is an artificial network for community detection, which is claimed to possess some basic statistical properties found in real networks, such as heterogeneous distributions of degree and community size. Many parameters are involved to specify properties of generated networks in this benchmark: N (number of nodes), $\langle k \rangle$ (average degree), $k_{\max}$ (maximum degree), $c_{\min}$ and $c_{\max}$ (minimum and maximum community size), τ_1 (exponent of power-law distribution of nodes degree), τ_2 (exponent of power-law distribution of community sizes), and μ (mixing parameter).

To evaluate the performance of our algorithm on discovering community structures, the normalized mutual information (NMI) measure [25] is used to quantitatively compare the known partition with the partition found by the algorithm:

$$\mathrm{NMI}(A, B) = \frac{-2 \sum_{i=1}^{c_A} \sum_{j=1}^{c_B} N_{ij} \log\left(N_{ij} N / N_{i.} N_{.j}\right)}{\sum_{i=1}^{c_A} N_{i.} \log\left(N_{i.}/N\right) + \sum_{j=1}^{c_B} N_{.j} \log\left(N_{.j}/N\right)}, \tag{13}$$

where c_A is the real number of community and c_B denotes the number of found community. The matrix N presents the confusion matrix, where N_{ij} is simply the number of nodes in the real community i that appears in the found community j. $N_{i.}$ and $N_{.j}$ are the sum over row i and column j of confusion matrix, respectively. N is obviously the number of nodes. If the found partition is identical to the real communities, then NMI takes its maximum value, 1. However, if the found partition is entirely independent of the real partition, NMI = 0 corresponds to the situation that the entire network is found to be one community. The closer to 1 of the NMI, the better partition of the network will be.

The number of NL nodes in each community is an important factor that affects the ability of the algorithm to identify communities. In order to quantify the relationship between the accuracy of the algorithm and the number of NL nodes in each community, the experiment is performed on the LFR benchmark networks (see Figure 1). The following parameters were employed: $N = 1000$, $\langle k \rangle = 15$, $\mu = \{0.7, 0.8\}$, $\tau_1 = 2$, $\tau_2 = 1$, $c_{\min} = 20$, and $c_{\max} = 50$. We fix the percentage of TL nodes 20% in each community. The result of the experiment suggests that the partition accuracy of the algorithm increases with increase of the percentage of NL nodes in each community.

In the LFR networks, the mixing parameter μ represents the ratio between the external degree of each vertex with respect to its community and the total degree of the node. The larger the value μ of the network is, the harder its community structure is detected. Then the experiment is designed for testing the accuracy of our algorithm with the various parameter μ (see Figure 2). In the experiment, we randomly select 4 and 8 labeled nodes in each community, respectively. The following parameters about the LFR benchmark networks were employed: $N = 1000$, $\langle k \rangle = 15$, $\mu = \{0.1, 0.2, \ldots, 0.9\}$, $\tau_1 = 2$, $\tau_2 = 1$, $c_{\min} = 20$, and $c_{\max} = 50$. We fix the percentage

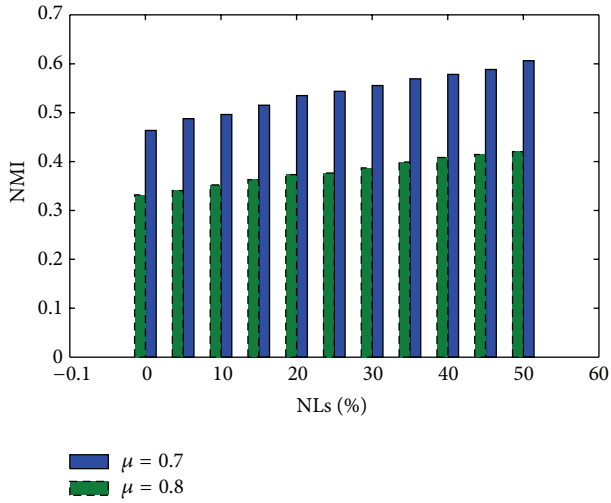

Figure 1: The relationship between the accuracy of the algorithm and the number of NL nodes in each community. In the LFR benchmark, $\mu = \{0.7, 0.8\}$, the partition accuracy of the algorithm increases with the increase of the percentage of NL nodes in each community. Each point in the figure represents the average over 100 runs on randomly generated LFR benchmark networks with the given parameters.

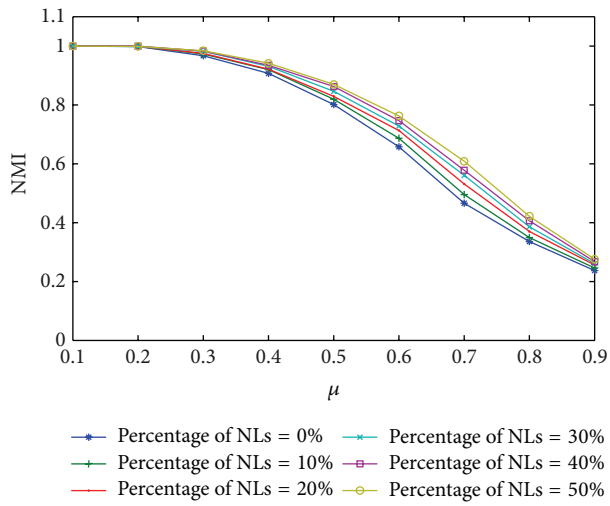

Figure 2: Testing the accuracy of our algorithm with the various parameter μ. Each point in the graph represents the average over 100 runs on randomly generated LFR benchmark networks with the given parameters.

of TL nodes 20% in each community and determine the percentage of NL nodes in each community by searching the grid $\{0\%, 10\%, \ldots, 50\%\}$.

To evaluate the effectiveness of our algorithm, we compare it to label propagation algorithm (LPA) and Infomap algorithm. Both algorithms can discover community structure without prior knowledge. We generate benchmark networks with the following parameters: $N = 1000$, $\langle k \rangle = 15$, $\mu = \{0.1, 0.2, \ldots, 0.9\}$, $\tau_1 = 2$, $\tau_2 = 1$, $c_{\min} = 20$, and $c_{\max} = 50$. Then, we randomly select 20% TL nodes and, respectively, select 0% and 20% NL nodes in each community (see Figure 3).

Figure 3 presents the advantages of our algorithm under a certain situation. Compared with the two comparative algorithms, our proposed algorithm gives almost the same results as the LPA and Infomap when the mix parameter ranges from 0.1 to 0.3. Our algorithm also presents better quality than LPA. Although our algorithm is no better than the Infomap algorithm during the mix parameter covers from 0.4 to 0.7, it outperforms under the condition that the mix parameter is high. In particular the mix parameter arrives at 0.8; the NMI value of the Infomap algorithm is almost 0, while it is more than 0.3 for our algorithm, as depicted in Figure 3. It means that our algorithm is particularly suitable for the community detection of high parameter μ. In other words, our algorithm is more favorable with obscure community structure in networks. On the other hand, the result of the experiment shows that the NL nodes can help increase the accuracy of community partition.

In the proposed algorithm, the percentage of NL nodes and the percentage of TL nodes are important factors that influence the accuracy of community partition. To analyze the relationship between NL and TL, we, respectively, set the percentage of TL nodes to $\{5\%, 10\%, \ldots, 50\%\}$ and the percentage of NL nodes to $\{0\%, 10\%, \ldots, 50\%\}$ (see Figure 4). We generate benchmark networks with the following parameters: $N = 1000$, $\langle k \rangle = 15$, $\mu = 0.5$, $\tau_1 = 2$, $\tau_2 = 1$, $c_{\min} = 20$, and $c_{\max} = 50$.

As can be seen from Figure 4, the proposed algorithm performs better with the increase of TL nodes. It is consistent with intuition, since there is more exact label information available. Moreover, with the increase of NL nodes, the algorithm can achieve higher accuracies. This means that NL is actually helpful to community detection and the algorithm can use this information effectively. In particular, NL is more beneficial when TL nodes are rare, since the increase of accuracy brought by NL will become smaller with the increase of TL nodes.

4.2. Real-World Networks. In this subsection, we verify our algorithm from empirical networks, the karate club network and the dolphins social network, which have been applied as benchmarks to evaluate many community detection algorithms since the true community structures are known in the two networks. In general, the karate club network can be split into two disjoined groups due to the disagreement between the administrator and the instructor of the club, and the dolphins social network can be separated into two groups due to the temporary disappearance of a dolphin. However, the NL node is equivalent to TL node provided that the networks are divided into two communities. In the following experiments, we assume that Donetti's result [26] is the true partition of the karate clue network and Pan's conclusion [27] is the true community of dolphin social network.

The karate club network is constructed by Zachary over a period of two years and is composed of 34 nodes corresponding to members of the club and 78 edges representing

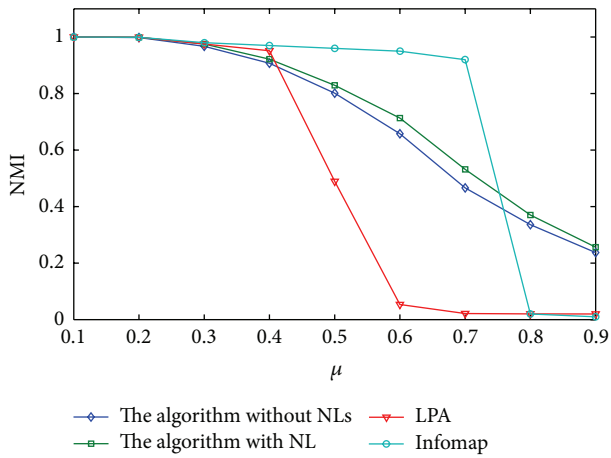

Figure 3: Results (NMI) of community detection algorithms in the LFR benchmark networks. The algorithms include LPA, Infomap, and the proposed algorithm with different parameters. Each point in the graph represents the average over 100 runs on randomly generated LFR benchmark networks with the given parameters.

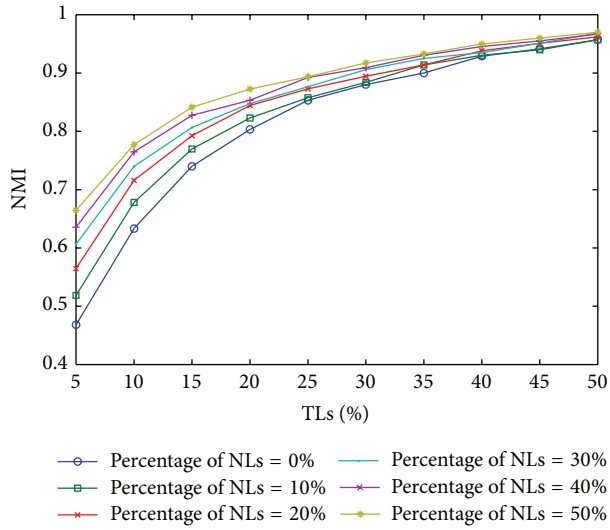

Figure 4: Results (NMI) of the proposed algorithm with different selections of NL and TL in the LFR benchmark networks. Each point in the graph represents the average over 100 runs on randomly generated LFR benchmark networks with the given parameters.

the connections of the individuals outside the activities of the club. In Donetti's result, the network is split into four communities. We select the nodes $\{25, 33, 17, 1\}$ as TL nodes and the nodes $\{28, 4, 3, 11\}$ as NL nodes for four different communities, respectively. Each NL node has one NL. The parameters a_u and a_l are set to 0.95 and 0.05, respectively. Applying the proposed algorithm, the results of community detection for karate club network are shown in Figure 5. It is clear that the result of our proposed method is in agreement with the partition of Donetti's method.

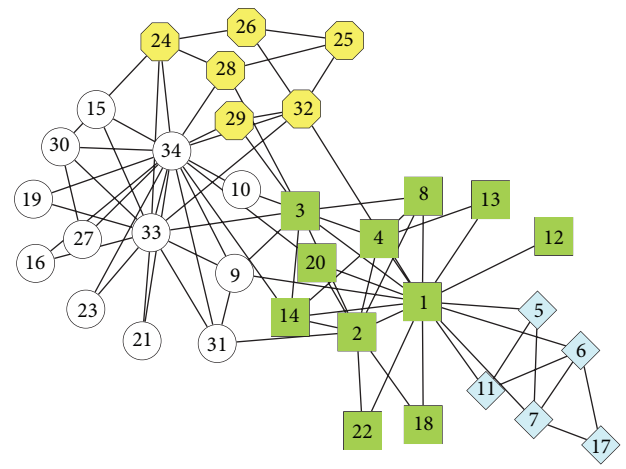

Figure 5: Detecting community structure in the karate club network via the proposed algorithm. Four communities are detected, which are denoted by the different shapes.

The dolphin social network, consisting of 62 nodes indicating bottlenose dolphins and 159 edges representing the associations between dolphin pairs occurring more often than expected by chance, is constructed by Lusseau over a period of seven years from 1994 to 2001. In Pan's conclusion, the network is divided into four communities. We select the nodes $\{SN63, Trigger, SN90, Oscar\}$ as TL nodes and the nodes $\{TSN103, MN105, SN89, Bumper\}$ as NL nodes for four different communities, respectively. Each NL node has one NL. The parameters a_u and a_l are 0.95 and 0.05, respectively. Applying the proposed algorithm, the results of community detection for dolphin social network are shown in Figure 6. It is obvious that the result of our proposed method is approximately consistent with the result of Pan's methods.

5. Conclusions

In this paper, a semisupervised community detection algorithm is proposed based on negative information, which indicates whether a node does not belong to a specific community. It has near-linear complexity in time and can incorporate the NL and TL into community detection. As seen from our experimental results on both real and artificial networks, incorporating NL into community detection procedure can significantly improve performance, especially in the situation where the traditional labels are rare. Moreover, the more TLs and NLs applied in our algorithm, the better the community partition result.

Unfortunately, it is an implicit restriction that the number of communities must be known in advance, since the selection of the TL nodes should cover all the communities. Our future work will concentrate on the issue of detecting communities without preknowing the community number. In other words, we will devote part of our energy on the research

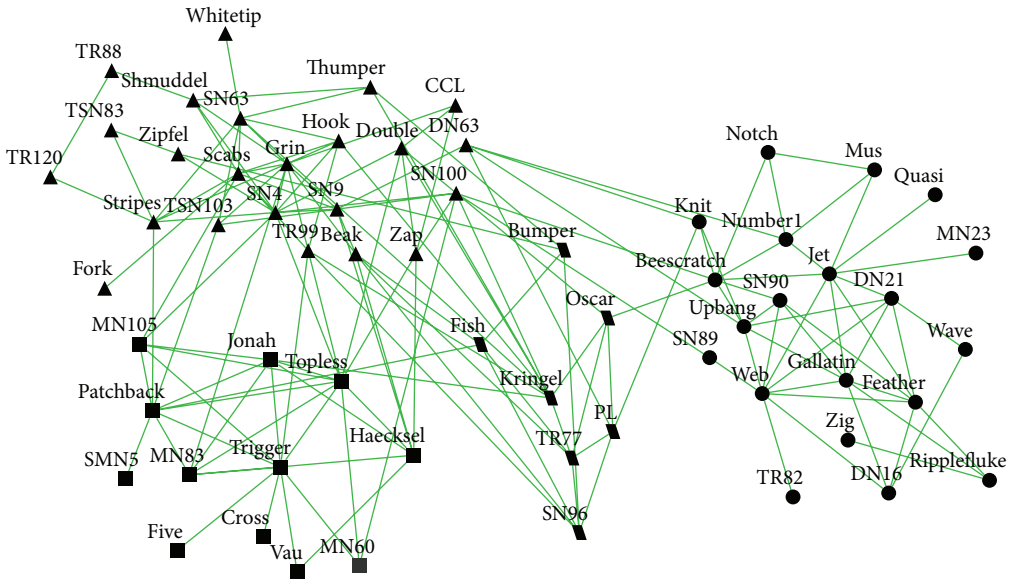

FIGURE 6: Detecting community structure in the dolphin social network via the proposed algorithm. Four communities are detected, which are denoted by the different shapes.

of an improved semisupervised community detection algorithm which is capable of identifying communities accurately without labeled nodes of any community in the future.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National High-Tech Research and Development Program (863 Program) (no. 2014AA015103), the Joint Funds of the National Natural Science Foundation of China (no. U1404604), the Natural Science Plan Project of the Education Department of Henan Province (nos. 2010B520014 and 2010A520024), and the Soft Science Research Program of Science and Technology Department of Henan Province (no. 112400450405).

References

[1] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.

[2] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 23, pp. 8577–8582, 2006.

[3] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, Article ID 066133, 2004.

[4] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 69, no. 2, Article ID 026113, 2004.

[5] J. Duch and A. Arenas, "Community detection in complex networks using extremal optimization," *Physical Review E*, vol. 72, no. 2, Article ID 027104, 2005.

[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, Article ID P10008, 2008.

[7] S.-W. Son, H. Jeong, and J. D. Noh, "Random field Ising model and community structure in complex networks," *European Physical Journal B*, vol. 50, no. 3, pp. 431–437, 2006.

[8] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, no. 1, Article ID 016107, 2011.

[9] S. Fortunato, "Community detection in graphs," *Physics Reports: A Review Section of Physics Letters*, vol. 486, no. 3–5, pp. 75–174, 2010.

[10] M. E. J. Newman, "Communities, modules and large-scale structure in networks," *Nature Physics*, vol. 8, no. 1, pp. 25–31, 2012.

[11] J. Weston, C. Leslie, E. Ie, D. Y. Zhou, A. Elisseeff, and W. S. Noble, "Semi-supervised protein classification using cluster kernels," *Bioinformatics*, vol. 21, no. 15, pp. 3241–3247, 2005.

[12] X. Ma, L. Gao, X. Yong, and L. Fu, "Semi-supervised clustering algorithm for community structure detection in complex networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 389, no. 1, pp. 187–197, 2010.

[13] E. Eaton and R. Mansbach, "A spin-glass model for semi-supervised community detection," in *Proceedings of the 26th AAAI Conference on Artificial Intelligence and the 24th Innovative Applications of Artificial Intelligence Conference (AAAI '12)*, pp. 900–906, July 2012.

[14] Z.-Y. Zhang, "Community structure detection in complex networks with partial background information," *Europhysics Letters*, vol. 101, no. 4, Article ID 48005, 2013.

[15] Z. Y. Zhang, K.-D. Sun, and S. Q. Wang, "Enhanced community structure detection in complex networks with partial background information," *Scientific Reports*, vol. 3, article 3241, 2013.

[16] D. Liu, X. Liu, W. Wang, and H. Bai, "Semi-supervised community detection based on discrete potential theory," *Physica A: Statistical Mechanics and its Applications*, vol. 416, pp. 173–182, 2014.

[17] D. Liu, H.-Y. Bai, H.-J. Li, and W.-J. Wang, "Semi-supervised community detection using label propagation," *International Journal of Modern Physics B*, vol. 28, Article ID 1450208, 2014.

[18] C. Hou, F. Nie, F. Wang, C. Zhang, and Y. Wu, "Semisupervised learning using negative labels," *IEEE Transactions on Neural Networks*, vol. 22, no. 3, pp. 420–432, 2011.

[19] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Physical Review E*, vol. 78, no. 4, Article ID 046110, 2008.

[20] G. H. Golub and C. F. van Loan, *Matrix Computations*, vol. 3, JHU Press, 2012.

[21] W. Zachary, "An information flow model for conflict and fission in small groups1," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

[22] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: can geographic isolation explain this unique trait?" *Behavioral Ecology and Sociobiology*, vol. 54, no. 4, pp. 396–405, 2003.

[23] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 76, no. 3, Article ID 036106, 2007.

[24] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 4, pp. 1118–1123, 2008.

[25] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment*, no. 9, Article ID P09008, pp. 219–228, 2005.

[26] L. Donetti and M. A. Muñoz, "Detecting network communities: a new systematic and efficient algorithm," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2004, Article ID P10012, 2004.

[27] Y. Pan, D.-H. Li, J.-G. Liu, and J.-Z. Liang, "Detecting community structure in complex networks via node similarity," *Physica A: Statistical Mechanics and Its Applications*, vol. 389, no. 14, pp. 2849–2857, 2010.