Research on semi-supervised community discovery algorithm based on new annealing

Jinghong Wang1,2,3, Jiateng Yang1,2, Yichao He4
1College of Computer and Cyber Security, Hebei Normal University, Shijiazhuang City, Hebei Province, People’s Republic of China
2Hebei Key Lab Network and Information Security, Shijiazhuang City, Hebei Province, People’s Republic of China
3University of Illinois at Champaign, Urbana, Illinois 61801, USA
4School of Information Engineering, Hebei GEO University, Shijiazhuang Hebei 050031, People’s Republic of China

e-mail: 18331137621@163.com

Abstract: Based on the similarity of the community detection methods, the Givern-Newman (GN) algorithm is fast and accurate but has a higher running time. In order to improve the efficiency of GN Algorithm, this study presents a semi-supervised GN algorithm based on node similarity. By making full use of the constraint set of the prior knowledge must-link and cannot-link, the prior information is extended by the derived rules, and the extended information is verified by the method of distance measurement. Using new annealing maximisation algorithm to calculate node similarity iteratively, and validated using artificial and real networks. It proves that the proposed algorithm reduces the GN algorithm's running time and improves efficiency.

1 Introduction

With the development of the Internet, many social networking sites have sprung up. When identifying information sources, there are many unpredictable factors that affect the validity, authenticity, and reliability of the data. The accuracy of information acquired by physical instruments is restricted by the accuracy of the instrument, and will often contain some noise from the data collection process. Moreover, in the network transmission process (especially wireless network transmission), the accuracy of the information is affected by factors such as bandwidth, transmission delay, and energy [1]. These sites have evolved from simple dating sites to information dissemination sites, thus attracting more and more users to join. Community discovery is a key method for complex network analysis. It can identify closely linked subgraphs in the network, the nodes in subgraphs are closely linked, and the links between subgraphs are sparse. Researchers have come up with a number of community discovery methods, these methods can be divided into graph segmentation-based dichotomy, hierarchical clustering-based community discovery, heuristic-based community discovery, overlapping community discovery, etc.

The so-called complex network is nature of each entity abstraction for the nodes in the network, the relationship between entities abstract [2] of edges in the network, for social networks is a network in each user abstraction for the nodes in the network, between users’ contact abstraction for the edge of the network. The so-called community is the cluster structure in the network, which has the characteristics of a tight connection between the community and the connection of the community. Community detection is a popular automatic identification technology in network analysis technology [3]. So far, a large number of methods have been developed for community detection tasks, including hierarchical clustering, segmentation clustering, modular-based methods etc. They only focus on detecting closely connected subgraphs. However, complex networks may have many other types of structures, including core–periphery, hierarchical, multi-split structures, or a mixture thereof [4], the GN algorithm is a hierarchical clustering algorithm, is the classical network community discovery method cannot determine in advance the number of network communities, but with higher running time and is not applicable to large networks.

Most of the community discovery methods today belong to the category of unsupervised and cannot deal with the semi-supervised information (prior knowledge) given in advance. Using unsupervised learning to mark unlabelled, label samples will consume a lot of manpower and material resources and take a long time, and the final marking result is less accurate. The prior knowledge used in semi-supervised learning is noise-free data. Classless label samples are labelled according to the prior knowledge, which greatly reduced the labelling cost and improves operation efficiency [5].

To solve the above problems, this study proposes a semi-supervised community discovery algorithm based on node similarity [6]. This study analyses and compares the performance differences of known prior knowledge in community discovery, and combines semi-supervised and similarity calculation methods to replace the influence of boundary value calculation on running time in the traditional GN algorithm, thus reducing the running time of the GN algorithm and improving the operation efficiency of the algorithm. [7, 8]

2 Construct similarity

Community discovery methods based on network similarity are divided into edge-based similarity and node-based similarity [9]. This study uses the similarity of nodes and considers the relationship between nodes and their neighbour nodes. The similarity of nodes in the community is high, and the similarity of nodes between communities is low.

Definition 1: If the two nodes in the network have the same or similar neighbour nodes, the two nodes are considered to be similar. Neighbourhood information of a node needs to be considered while adding parameters. The perturbation of parameters solves the problem that network node clustering is easily affected by noise links and unbalanced clustering is formed [10].

Three kinds of node similarity definitions follow formulae (1), (2), (3), and (4), where \( r_i \) represents the set neighbours of node \( i \), \( |r| \) represents the number of the set of potential is an element of the collection, \( r_i \cap r_j \) represent node \( i \) and node \( j \) shared neighbour numbers.

\[
S_j(i,j) = \frac{|r_i \cap r_j|}{|r_i \cup r_j|} \tag{1}
\]
\[
S(i, j) = \frac{|i \cap j|}{\sqrt{|i||j|}} \quad (2)
\]
\[
S_{\delta}(i, j) = \frac{|i \cap j|}{\min(|i|, |j|)} \quad (3)
\]
\[
S_{\alpha}(i, j) = \begin{cases} 
S(i, j), & i, j \in E \\
S(i, j) - \sigma, & i, j \notin E 
\end{cases} \quad (4)
\]

In a complex network, the more common neighbour nodes the two nodes have, the more similar the two nodes are, i.e. they belong to one community. Liu et al. [11] mainly used in the literature index network. Max et al. [12] show that if there is no link between two nodes in the net, a new similarity construction method can be constructed by subtracting a penalty term \(\sigma\) on the basis of \(S_{\alpha}\) similarity reading construction.

In this study, the similarity is calculated by using each data point in the data set as a node in the network. Adding an edge between two data points, \(i\) and \(j\), which are close or similar. To make full use of active node prior learning algorithm, combine the priori information with the similarity nodes, through the deterministic anti-annealing expectation-maximisation (EM) algorithm of the hybrid model, the node similarity is calculated iteratively and constructed the similarity matrix of the network nodes. Then, the clustering problem is transformed into the community division problem, and the corresponding clustering results can be obtained by using any kind of community partition method to partition the network.

### 3 New annealing EM (NAEM) algorithm for hybrid models

In this section, we prove that the EM algorithm with a mixed model always converges to a poor local maximum [13, 14]. Meanwhile, the NAEM algorithm is used to evaluate the parameters of the mixed model [3].

We run the EM algorithm 80 times in multiple small networks to analyse the convergence of the EM algorithm. The EM algorithm runs 80 times. Keep iterating the EM algorithm. When the log probability difference between two continuous iterations is \(<10^{-10}\), the algorithm stops. The results are shown in Tables 1–3. The results show that when the network is very complex, the EM algorithm always converges to the local maximum.

The NAEM algorithm is a maximum likelihood estimation algorithm. This algorithm improves the local optimal problem of the traditional EM algorithm and is used to estimate parameters of the mixed model. The convergence speed of the algorithm is greatly improved. The new posteriori parameterised \(\beta\) of the NAEM algorithm can be expressed as

\[
p_{\theta} = \frac{\left(\pi \prod \theta_{ij}^{\theta_{ij}}\right)^{\beta}}{\sum \left(\pi \prod \theta_{ij}^{\theta_{ij}}\right)^{\beta}} \quad (5)
\]

\(1/\beta\) represents the corresponding temperature.

If the posterior probability is known, then we use the same parameters as the EM algorithm. The NAEM algorithm process is as Algorithm 1.

**Algorithm 1**: A newly annealed EM algorithm is based on the mixed model.

1: Initialise

Initialise \(\beta = \beta_0 (0 < \beta_0 \ll 1)\) and initial parameters \(\theta^{(0)} = (\pi^{(0)}, \theta^{(0)})\)

2: set \(T, t = 0\)

3: Iteration to convergence

- E-step: calculate the posterior probability by

\[
p_{\theta} = \frac{\left(\pi \prod \theta_{ij}^{\theta_{ij}}\right)^{\beta}}{\sum \left(\pi \prod \theta_{ij}^{\theta_{ij}}\right)^{\beta}}
\]

- M-step: estimate \(\Theta^{(new)}\) by

\[
\sum \theta_{ij} = \frac{\sum A_i \theta_{ij}}{\sum \theta_{ij}}
\]

4: Increase \(\beta\)

5: If \(\beta \leq \beta_{\text{max}}, t = t + 1,\) return to step 3

6: Decrease \(\beta\)

7: Iteration to convergence

- E-step: the posterior probability is calculated by (5).

- M-step: estimate \(\Theta^{(new)}\) by (6).

8: If \(\beta \geq 1, t = t + 1,\) return to step 6

9: Termination of the algorithm

---

**Table 1** Among them, ‘L-convergence’ means the log-likelihood of convergence of EM algorithm, ‘iteration-average’ means the average of restarts of convergence value, ‘R-number’ indicates the number of all 80 restart log-likelihood values. Run the results on the Karate network.

| L-convergence | Iterative-average | R-number |
|---------------|-------------------|----------|
| -499.6573     | 62                | 40       |
| -491.1758     | 66                | 14       |
| -501.9945     | 36                | 20       |

**Table 2** Result running on the dolphin network

| L-convergence | Iterative-average | R-number |
|---------------|-------------------|----------|
| -1128.3569    | 62                | 40       |
| -1185.7134    | 66                | 2        |
| -1205.5075    | 38                | 2        |

**Table 3** Result running on the adjnoun network

| L-convergence | Iterative-average | R-number |
|---------------|-------------------|----------|
| -3619.9127    | 70                | 3        |
| -3621.3281    | 70                | 9        |
| -3624.8334    | 136               | 4        |
The results show that when the network is very complex, the EM algorithm always converges to the local maximum. This algorithm is used for iterative calculation of similarity.

## 4 Givenr-Newman based on similarity and semi-supervised (SSGN) algorithm based on similarity

With the transition of supervised learning and unsupervised learning to semi-supervised learning, making full use of prior information to guide the clustering process has become the dominant part of semi-supervised clustering [15, 16]. Both traditional GN and Givenr-Newman based on semi-supervised (SGN) belong to the category of unsupervised learning and cannot deal with the pre-given semi-supervised knowledge [17]. In this study, a new semi-supervised clustering method based on a similarity degree is proposed, which is called SSGN. The semi-supervised learning method is used to process the information. By adding the must-link node and cannot-link node, it makes the network more obvious and improves the accuracy of the algorithm. [18–21]

![Flow chart](image)

**Definition 2:** Given a data set $X = \{x_1, x_2, \ldots, x_n\}$ and a point pair constraint set $C = C_\text{ML} \cup C_\text{CL} \neq \emptyset$, where $C_\text{ML}$ is a set of mandatory constraints, and its element $c = (x_i, x_j)$ means that $x_i$ and $x_j$ belong to the same cluster; $C_\text{CL}$ is a set of unconnected constraints, and its element $C_\text{CL} \neq (x_i, x_j)$ indicates that $x_i$ and $x_j$ belong to different clusters [22, 23].

**Definition 3:** Network structure graph $G(V, e) = (V, E, V, u \in E)$ is a collection of all nodes in the network, $e$ is the set of all edges in the network, provided the complex parallel community (CPC) network with $n$ nodes and $m$ edges.

The community in the network is represented by $C = \{C_1, C_2, \ldots, C_K\}$, $K$, $C_j = V$ and $C_i$ fields ($i = 1$) up to $C_j = i$ and $j, i = 1, 2, \ldots K$.

**Definition 4:** Divided communities meet sparsely populated communities and close links within communities, expressed as

$$
K = \sum_{j=1}^{K} |\{((u, v)) \in E, u \in C_i, v \in C_j\}|
$$

**Definition 5:** Must-link constraint, $C_{\text{ML}}$: $\forall v_i, v_j \in V, (v_i, v_j) \in C_{\text{ML}} \Rightarrow v_i, v_j$ which belong to the same community.

Cannot-link constraint, $C_{\text{CL}}$: $\forall v_i, v_j \in V, (v_i, v_j) \in C_{\text{CL}} \Rightarrow v_i, v_j$ that do not belong to the same community.

The test network in this study is a network with a priori information, and the must-link constraint node pairs and cannot-link constraint pairs are determined according to the division of attributes in a priori information. For example, in a network that contains a priori information, the creator of certain information and the user with a large fan are defined as a must-link constraint node, indicating that in a set, users without the information are defined as cannot-link constraint nodes. Also, expand the node tag according to the derivative rule, and divide the community by the similarity of nodes.

The SSGN algorithm is a hierarchical splitting algorithm based on the idea of node similarity. The basic process is to continuously delete the edges with the smallest similarity to all source nodes in the network, and then recalculate the similarity between the remaining edge endpoints in the network. Repeat this process until all edges are deleted. Wherein full use of prior knowledge must-link constraints and cannot-link constraint sets, adopt breadth-first search, combine prior information, and combine similarity between nodes to reconstruct similarity matrix of network nodes.

Algorithm flow is as follows:

1. **Input:** $G(V, E)$, known must-link constraint set, cannot-link constraint set for community finding.

2. According to the rules of the derived relationship, the other unlabelled nodes are marked. The rules are as follows: $(v_i, v_j) \in C_{\text{CL}}$ and $(v_j, v_k) \in C_{\text{ML}} \Rightarrow (v_i, v_k) \in C_{\text{ML}}$; $(v_i, v_j) \in C_{\text{ML}}$ and $(v_j, v_k) \in C_{\text{CL}} \Rightarrow (v_i, v_k) \in C_{\text{ML}}$; to no node can be used as a marker.

3. Select the node in the $C_{\text{CL}}$ collection, remove the edge of the two node connections, and all the tag nodes are deleted.

4. Calculate the similarity between the two endpoints of each side of the network (add), select the smallest edge to delete.

5. Repeat step 4, calculate the similarity between the remaining edge endpoints to all edges that are deleted.

The flow chart is shown in Fig. 1.
5 Experimental results and analysis

The experimental operating environment is the 2 GHz computer Inter processor with 2 GB memory, the operating system is Windows 7, a programming environment for 7.9.0.592 Matlab.

5.1 Artificial data set

Artificial data set use the Lancichinetti-Fortunato-Radicchi (LFR) standard formation [24, 25], the rules are as follows: generation of network 1: 132 nodes, 262 edges, and four communities and finally community division results as shown in Fig. 1; network node 2: network 1: 132 nodes, 262 edges, and four communities and finally community division results as shown in Fig. 1; network node 2:

![](image)

**Fig. 2 Accuracy comparison of different similarity construction methods**

| Network | GN | SGN | SSGN |
|---------|----|-----|------|
| NMI     | 1  | 1   | 1    |
| time, s | 1997.86 | 200.77 | 140.79 |

**Table 4** Running time comparison between the SSGN algorithm and other algorithms in artificial neural network 1

| Network | GN | SGN | SSGN |
|---------|----|-----|------|
| NMI     | 1  | 1   | 1    |
| time, s | 21,347.8 | 2687.6 | 1678.9 |

**Table 5** Running time comparison between the SSGN algorithm and other algorithms in artificial neural network 2

| Network | GN | SGN | SSGN |
|---------|----|-----|------|
| NMI     | 1  | 1   | 1    |
| time, s | 1037.79 | 140.79 | 589.3 |

**Table 6** Comparison of running time of SSGN algorithm with different number of CML and CCL sets on artificial network 1

| Network | N = 8 | N = 10 | N = 20 |
|---------|-------|-------|-------|
| NMI     | 1     | 1     | 1     |
| time, s | 134.32 | 55.68 | 8.23  |

**Table 7** Comparison of running time of SSGN algorithm with a different number of CML and CCL sets on artificial network 2

| Network | N = 8 | N = 10 | N = 20 |
|---------|-------|-------|-------|
| NMI     | 1     | 1     | 1     |
| time, s | 1543.4 | 1037.79 | 589.3 |

**Table 8** Comparison of NMI between SSGN and other algorithms in real networks

| Dolphins | Karate | Football | Pol-books |
|----------|--------|----------|-----------|
| GN       | 1.0000 | 0.9601   | 0.9001    | 0.7960    |
| SGN      | 1.0000 | 1.0000   | 0.9878    | 0.8792    |
| SSGN     | 1.0000 | 1.0000   | 1.0000    | 0.9323    |

**Table 9** Comparison of running time of SSGN algorithm

| Dolphins | Karate | Football | Pol-books |
|----------|--------|----------|-----------|
| N = 8    | 1.0000 | 1.0000   | 1.0000    | 0.9323    |
| N = 10   | 1.0000 | 1.0000   | 1.0000    | 0.9567    |
| N = 20   | 1.0000 | 1.0000   | 1.0000    | 1.0000    |

similarity calculation method and the CANNOT-LINK set, MUST-LINK set of the selection of the direct impact of the accuracy of the algorithm.

In order to compare the SSGN algorithm has a different definition of the traditional community network running accurately and efficiently, the experiment is divided into two parts (Tables 4 and 5): first of all, the use of artificial data to train the SSGN algorithm, the algorithm is tested with the known solves the constraints, cannot link constraints quantity change accuracy and run-time variation. Verify the algorithm parameters definition: the number of nodes in the network hypothesis is n, the number of edges is m, clustering number is k, the iteration number of the algorithm for T cannot link constraint set, must link constraints set number of elements x and y, respectively, for a total of N, artificial experimental hypothesis number is equal, tabulated in Tables 6 and 7, SSGN algorithm default (N = 8).

The experimental results are as follows: by comparing the results table shows that all the algorithms in precision consistent, the GN algorithm's running time is the highest, the SGN algorithm's running time is shorter than the GN algorithm, the SSGN algorithm running time is significantly shorter in the GN algorithm, increased slightly compared to SGN algorithm.

Fig. 2 shows the SSGN algorithm based on the similarity of different construction methods of the accuracy of alignment graph selection in artificial networks in the prior information. Compared with the results, the results show that all algorithms in the same accuracy, with the increase of the number of cannot-link nodes to the must-link node, SGN algorithm running time will be shortened. It is worth noting that in the selection of must-link constraint and cannot-link constraint sets, the more data you select, the faster the algorithm will run, in practical applications, a number of identified attributes are selected, the more obvious.

5.2 Real network analysis

Real network data using the classical network (Tables 8 and 9): dolphins (dolphin network), karate (Zachary karate club network), football (American college football network) and books on politics (American political work network) data acquisition (reference http://www.orgnet.com/). The traditional network has a clear community structure, which is as follows: dolphins (the dolphin network [26]) consists of two communities, 62 nodes, and 382 edges, in which each node has individual dolphins, community, known as the dolphin population, between nodes even the edges between two dolphins are in frequent contact. Karate (Zachary karate club network) consists of 34 nodes and 192 sides, each node has members of a club, the edge between two nodes denotes the...
relationship of social interaction between members of the two, the club is divided into two communities, a club supervisor, a principal as the centre in two communities, initialise the structure diagram as shown in Fig. 3. Football (American college football network) is made up of 132 nodes and 1343 sides, each node represents a team, and between the two nodes the two teams connected to this side are playing. All teams are divided into 12 leagues, or 12 communities, the probability of playing against a team outside the league is greater than the probability of playing against a team outside the league. Books on politics (American political work network) consists of 42 nodes of 127 edges, each node is a book, each book in Amazon is to sell books about American politics, means that the edge between two nodes while the customer has to purchase the two books, the entire network according to the political insight into three factions (community), i.e. the ‘liberals’ and ‘centrist’ and ‘conservative’.

According to the analysis method of an artificial neural network, the accuracy of each algorithm is as follows:

Fig. 4 shows the NMI algorithm and other algorithms in the real network on the SSGN comparison.

Fig. 5 shows the SSGN algorithm has a different number of C-must-link (CML) and C-cannot-link (CCL) sets in the real network of NMI contrast.

In a word, SSGN algorithm is more efficient and more accurate than traditional GN and SGN Algorithm when the accuracy of SSGN algorithm is the same. In the selection of CML and the number of CCL sets, the more CML and CCL ensemble algorithms are selected, the lower the time complexity and the higher the accuracy (Fig. 6).

6 Summary and prospect

This study according to the similarity structure method, make full use of prior knowledge of must link constraints, cannot link constraint set, with a priori information and combined with the similarity between nodes, re-construct the network nodes of similar degree matrixes, the GN algorithm reduces the repeated computation of boundary value time caused by the high complexity of the problem and improve the accuracy of using the semi-supervised GN algorithm based on node similarity. In a real network, for must link constraints, cannot link constraints set elements of the quantity and quality of SSGN algorithm has the decisive effect of complex networks are defined more clearly, determine the nature of the relationship between the more scientific, will greatly enhance the running efficiency and the accuracy of the algorithm.

The data used in this study is mainly non-overlapping complex networks. The algorithm for overlapping network algorithms is to re-compute the similarity of the relatively fuzzy nodes in the community according to the divided community results and to redefine the communities based on similarity values. The accuracy of dividing the boundary nodes of overlapping networks needs to be improved. The next step will be to further study the must-link and cannot-link constraint nodes in overlapping networks to improve the accuracy of community partitioning in overlapping networks.

7 Acknowledgments

This paper was presented at the 3rd Asian Conference on Artificial Intelligence Technology (ACAIT 2019). This work was supported by the HEIBEI Provincial Department of Education (ZD2018023); Science and Technology Fund Project of Hebei Normal University (L2019Z10); The Hebei Province Overseas Funding Project (C201816); The Natural Science Foundation of China (61672206); The Hebei Provincial Examination Institute Key Projects (KSZX201822); and The Hebei Natural Science Foundation (F2018205162).

8 References

[1] Xu, W., Yu, J.: ‘A novel approach to information fusion in multi-source datasets: a granular computing viewpoint’, Inf. Sci., 2017, 378, pp. 410–423
[2] Jiang, Y.: ‘Community detection in complex networks’, Beijing Jiaotong University, Beijing, 2014, pp. 32–33
[3] Kardak, S., Aggarwal, C.C., Jaideep, S., et al.: ‘Community detection with prior knowledge’. SIAM Data Mining, Dalian, Liaoning, China, 2013
[4] Xu, W., Guo, Y.: ‘Generalized multigranulation double-quantitative decision-theoretic rough set’, Knowl.-Based Syst., 2016, 105, (1), pp. 190–205
[5] Xu, W., Li, W.: ‘Granular computing approach to two-way learning based on formal concept analysis in fuzzy datasets’, IEEE Trans. Cybern., 2016, 46, (2), pp. 366–379
[6] Xu, W., Li, M., Wang, X.: ‘Information fusion based on information entropy in fuzzy multi-source incomplete information system’, Int. J. Fuzzy Syst., 2017, 19, (4), pp. 1200–1216
[7] Rosvall, M., Bergstrom, C.T.: ‘Maps of random walks on complex networks reveal community structure’, Proc. Natl. Acad. Sci., 2008, 105, (4), pp. 1118–1123
[8] Sang, B., Guo, Y., Shi, D., et al.: ‘Decision-theoretic rough set model of multi-source decision systems’, Int. J. Mach. Learn. Cybern., 2018, 9, pp. 1941–1954
Chai, B., Wang, J., Yu, J.: ‘A parameter selection method of the deterministic anti-annealing algorithm for network exploring’, *Neurocomputing*, 2017, **226**, pp. 192–199.

Girvan, M., Newman, M.E.J.: ‘Community structure in social and biological networks’, *Proc. Natl. Acad. Sci.*, 2002, **99** (12), pp. 7821–7826.

Liu, Z., Li, P., Zheng, Y., *et al.*: ‘Community detection by affinity propagation’. Technical report, 2008.

Ma, X., Gao, L., Yong, X., *et al.*: ‘Semi-supervised clustering algorithm for community detection in complex networks’, *Physica A*, 2010, **389**, (1), pp. 187–197.

Salton, G., Mcgill, M.J.: ‘Introduction to modern information retrieval’ (McGraw-Hill Book Company, Academic Press, New York, 1983), pp. 45–89.

Zhang, Z.Y., Sun, K.D., Wang, S.Q.: ‘Enhanced community structure detection in complex networks with partial background information’, *Sci. Rep.*, 2013, **3**, p. 3241.

Cheng, J., Leng, M., Li, L., *et al.*: ‘Active semi-supervised community detection based on must-link and cannot-link constraints’, *PLOS ONE*, 2014, **9**, (10), pp. 1–18.

Ng, A.Y., Jordan, M.I., Weiss, Y.: ‘On spectral clustering: analysis and an algorithm’. Proc. Advances in Neural Information Processing Systems, New York, USA, 2001, pp. 849–856.