An Improved Genetic-Based Link Clustering for Overlapping Community Detection

Yong Zhou and Guibin Sun(✉)

School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221008, China
yzhou@cumt.edu.cn, sunguibinbest@qq.com

Abstract. The problem of community detection in complex networks has been intensively investigated in recent years. And it was found that the communities of complex networks often overlap with each other. So in this paper, we propose an improved genetic-based link clustering for overlapping community detection. The first, the algorithm changes the node graph into the link graph. The second, the algorithm adopts the genetic algorithm to detect the link communities. The Third, the algorithm transforms the link communities into the node communities. Automatically, the nodes, which are linked with edges belonged to different link communities, will be the overlapping nodes. The last, in order to improve the quality of community detection, we define an effective method to solve the "excessive overlap" problem. The experimental results shows that the proposed algorithm is effective and efficient on both simulate networks and real networks.

Keywords: Genetic-based · Link clustering · Overlapping communities · Community detection

1 Introduction

Many complex systems in nature and society can be described in terms of networks or graphs. The study of networks is crucial to understanding both the structure and the function of these complex systems. Researchers found that a common feature which is called community structure exists in many complex networks. Community structures are always expressed as clusters of nodes with dense connections within cluster and sparse connections with the other clusters. The community structure plays an important role in the complex network which can help people to understand the function of the complex network and find the potential law in the complex network. Take the World Wide Web as an example, close hyperlink web pages form a community and they often talk about related topics.

The identification of community structure has attracted much attention from various scientific fields. A lot of algorithms have been proposed for detecting communities in complex networks. The traditional community detection algorithm is to divide the complex network into several disconnected communities (or clusters,

Z. Shi et al. (Eds.): IIP 2016, IFIP AICT 486, pp. 142–151, 2016.
DOI: 10.1007/978-3-319-48390-0_15

groups, etc.), and each node must be affiliated with one community. The representative algorithms include the modularity optimization algorithm [1, 2], spectral clustering method [3, 4], and so on. However, there are many overlapping networks in real world. That is to say, in the complex networks, some nodes can't belong to only one community, they can belong to multiple communities at the same time. For example, in a social network, each person can belong to more than one social group at the same time (e.g., school, family, friends, etc.).

Recently, the overlapping community structure has been widely studied. Some algorithms use the clique percolation to detect the overlapping community, such as the well-known CPM [5], SCP [6] and EAGLE [7]. Some algorithms utilize the local expansion by optimizing a local benefit function, such as LFM [8], MONC [9], CIS [10] and OSLOM [11]. Some label propagation based algorithms allow multiple labels for each node to detect overlapping structure, such as COPRA [12], SLPA [13], etc. Some algorithms are Based on the link clustering, such as LINK [14], Link Maximum Likelihood [15] and Link-Comm [16]. Although the overlapping community detection has obtained significant achievements, with the network structure increasingly complex, the community detection is more difficult. how to more accurately and effectively detect the overlapping community structure is still a great challenge.

In this paper, we propose an improved genetic-based link clustering for overlapping community detection. Firstly, the algorithm changed the node graph into the link graph. Secondly, the algorithm adopted the genetic algorithm to detect the link communities. Thirdly, the algorithm transformed the link communities into the node communities. Automatically, the nodes, which are linked with edges belonged to different link communities, will be the overlapping nodes. Last, in order to improve the quality of community detection, we defined an effective method to solve the "excessive overlap" problem. The effectiveness of the proposed algorithm is demonstrated by extensive tests on both simulate networks and real networks with a known community structure. Through experimental comparison, the proposed algorithm is effective and efficient in overlapping community detection.

2 Related Work

The genetic algorithm for overlapping community detection (GaoCD) [17] was newly proposed in 2013. In this paper, they proposed a genetic algorithm for overlapping community detection based on the link clustering. Different from those node-based overlapping community detection algorithms, the GaoCD algorithm applies a novel genetic algorithm to cluster on the edge set of network. The genetic representation and the corresponding operators effectively represent the link communities and make the number of the communities determined automatically. In the GaoCD algorithm, it mainly includes three components: objective function, genetic representation and genetic operators.

2.1 Objective Function

In the GaoCD algorithm, the partition density D is utilized to evaluate the link density within communities. The partition density D is proposed in the LINK algorithm [10], which emphasizes the community density and ignores the connection among communities. the partition density D is defined as follows.

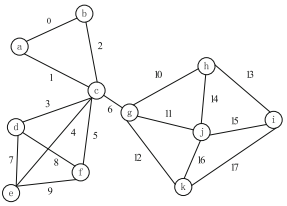
For a network with M links and N nodes, $P = \{P_1, P_2,\ldots, P_c\}$ is a partition of the links into C subsets. The number of links in subset P_c is m_c. The number of induced nodes, all nodes that those links touch, is n_c.

$$D = \frac{2}{M}\sum_{c} m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)} \tag{1}$$

2.2 Objective Function

In the GaoCD algorithm, a gene represents a link. An individual gene sequence in the population is represented as a gene type $[g_0, g_1,\ldots, g_i, \ldots, g_{m-1}]$. Among them, the m is the number of the edges in the network, $i \in [0, m)$ is the identifier of edges in the network, and each g_i is a random adjacent edge of edge i.

For example, in Fig. 1(a), e_0 has two adjacent edges e_1 and e_2. So the e_1 is the possible value of g_0. The encoding schema guarantees that every community partition can be encoded into a corresponding gene type and every gene type can be decode into an valid community partition. What' s more, the encoding schema can automatically determine the number of the communities, without any prior information.



(a) The example network

Index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
GeneType	2	6	1	7	5	8	11	9	7	5	14	12	17	15	16	17	12	13

(b) The genetic representation

Fig. 1. Illustration of the genetic representation

2.3 Genetic Operation

According to the genetic representation, The GaoCD algorithm adopts the corresponding genetic operators.

In the crossover operation, They randomly select two individuals from the current population. The exchanging positions are randomly generated and then exchange the genes in these positions between these two individuals. Since the g_i is always the identity

of the adjacent edges of e_i, the exchanged individuals also follow the genetic representation rule: g_i is an adjacent edge of e_i.

In the mutation operation, an individual is randomly selected from the current population and the positions are randomly generated. Then they reassign the gene values on these positions with a random adjacent edge.

3 An Improved Genetic-Based Link Clustering for Overlapping Community Detection

The GaoCD algorithm can effectively reveal overlapping structure. However, the GaoCD algorithm is also easy to appear the "excessive overlap" problem.

For example, the Fig. 2. is two kinds schematic diagrams of the "excessive overlap" problem. In the Fig. 2(a), all nodes should be divided into only one community, however, they are divided into two communities, making the node e and node b become the overlapping nodes. In the Fig. 2(b), the node e should only belong to the right community, however, it belongs to the both right community and left community.

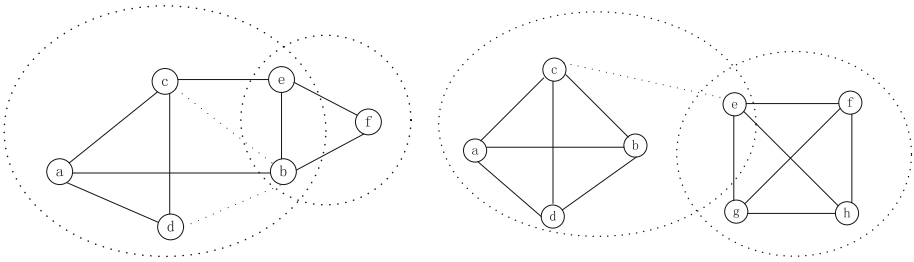


Fig. 2. Illustration of the "excessive overlap" problem

In order to avoid the "excessive overlap" problem and improve the community detection performance. we proposed an improved genetic-based link clustering for overlapping community detection.

3.1 Community Similarity

To solve the "excessive overlap" problem as shown in Fig. 2(a), we define a community similarity to measure the contact ratio of communities.

Definition 1 ***Community Similarity.*** Given two communities $C1$ and $C2$, the community similarity is define as

$$S(C_1, C_2) = \frac{|C_1 \cap C_2|}{\min(|C_1|, |C_2|)} \tag{2}$$

Given a set of communities *CS* and a community *C*, we can define the near-duplicates of *C* to be all communities in *CS* that are within a contact ratio Δ, where Δ is the maximum community similarity threshold. When the community similarity of the two communities is beyond threshold Δ, the two communities will be merged. In the experiment, we found that it is very reasonable that this threshold Δ be set at about 0.66. The algorithm of calculating community similarity is shown in algorithm 1.

```
Algorithm1. CommSim
01 Input: a set of communities CS, the threshold Δ.
02 Output: updated CS.
03 for each i from 0 to CS.size() do
04   for each j from i +1 to CS.size() do
05      temp = S(C_i, C_j); // use the formula(2).
06      if temp > Δ
07         C_i = merge(C_i, C_j);
08         delete C_j from CS;
09      end
10   end
11 end
12 return CS;
```

3.2 Belonging Coefficients

To solve the "excessive overlap" problem as shown in Fig. 2 (b), we define the belonging coefficients to decide that the overlapping nodes belong to multiple communities or only a single community.

Definition 2 ***Belonging Coefficients.*** Given a community *C* and an overlapping node *v*, belonging coefficients is defined as fallow:

$$BC(v, C) = \frac{|E(v)||E(v)|}{|E(C)||K(v)|} \tag{3}$$

Among them, the node *v* denotes an overlapping node which belongs to community *C*. $E(v)$ denotes the edges which connect node *v* to the community *C*. $E(C)$ denotes the edges in the community *C*. $K(v)$ denotes the degree of node *v*.

For the nodes with multiple memberships, we use the belonging coefficients to determine whether nodes are excessive overlap nodes. In order to facilitate comparison, we introduce a threshold τ, where τ is the maximum difference between two belonging coefficients in different communities. In the experiment, we found that it is very reasonable that this threshold τ be set at about 0.25. The algorithm about the belonging coefficients is shown in algorithm 2.

```
Algorithm2. BelongCoefficient
01 Input: communities CS, overlapping nodes NS,
   threshold τ.
02 Output: updated CS.
03 for each i from 0 to NS.size() do
04   get the communities C related to the node i;
05   T = ∅; //The set T is used to store the belonging
     coefficients related to the node i.
06   for each j from 0 to C.size() do
07      T_j = BC(i, C_j); // use the formula(4).
08   end
09   sort(T); //sorting the set T and corresponding
     communities C.
10   max = T_0; //maximum belonging coefficient.
11   for each k from 1 to T.size() do
12        if T_0 - T_k > τ
13          delete the overlapping node i from relating
     community C_k, C_k+1, ..., C_T.size()-1;
14          break;
15        end
16   end
17 end
18 return CS;
```

4 Experiments

In this section, the IGLC algorithm is tested on the simulated data sets and real data sets, respectively. Experimental environment: Processor Inter (R) Core (TM) i5 3.1 GHz PC, memory 4G, the operating system is Windows 7, programming environment Matlab R2009a.

4.1 Experimental Data Sets

The Simulated Data Sets. Currently, the LFR benchmark network [18, 19] is the most commonly used data set in community detection. We generate two LFR benchmark networks, whose detail information are shown in Table 1. Some important parameters of the benchmark networks are as follow:

Table 1. The LFR benchmark networks

Num	*N*	*k*	*maxk*	*minc*	*maxc*	*on*	*mu*	*om*
S1	1000	20	50	10	50	100	0.1	2 ~ 8
S2	1000	20	50	10	50	100	0.3	2 ~ 8

N: the number of nodes; *k*: the average degree; *maxk*: the maximum degree; *minc*: the minimum for the community sizes; *maxc*: the maximum for the community sizes; *on*: the number of overlapping nodes; *mu*: mixing degree; *om*: the number of communities that each node can belong to;

The Real Data Sets. We make experiments on five well known social networks, whose real community structure have been given. Their specific information is shown in Table 2.

Table 2. The real network

Name	*Nodes*	*Edges*	*Source*
Karate	34	78	[20]
Dolphins	62	159	[20]
Political Books	105	441	[20]
Football	115	613	[20]
Netscience	379	914	[20]

4.2 Evaluation Criteria

In the experiments, we use the evaluation criteria normalized mutual information (*NMI*) [12] and extended modularity (*EQ*) [7] to evaluate the communities.

Normalized Mutual Information (NMI). The *NMI* is used to measure similarity between the results of algorithm with true class values.

Assuming that the true class values of the data sets are $C = \{C_1, C_2, \ldots, C_k\}$, and the class labels obtained by the algorithm are $U = \{U_1, U_2, \ldots, U_l\}$, where k and l denote the number of clusters in C and U. The number of nodes in the Ci $(1 \le i \le k)$ and Uj $(1 \le j \le l)$ are n_i and n_j respectively. The length of intersection of C_i and U_j is n_{ij}, so *NMI* is defined as Eqs. (4-5).

$$NMI = \frac{2 \times I(C, U)}{H(C) + H(U)} \tag{4}$$

$$NMI = \frac{-2 \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{n_{ij}}{n} \log \frac{n_{ij}}{n_i \times n_j}}{\sum_{i=1}^{k} n_i \log \frac{n_i}{n} + \sum_{j=1}^{l} n_j \log \frac{n_j}{n}} \tag{5}$$

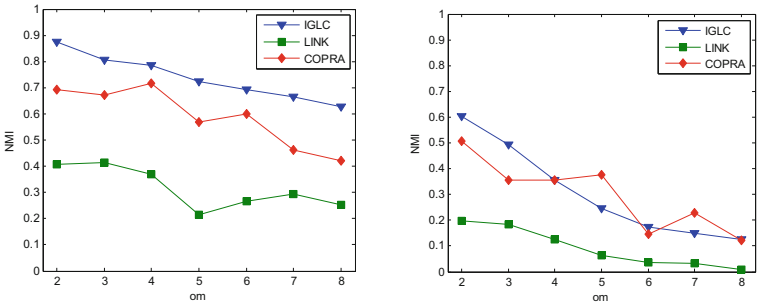
Extended Modularity (EQ). The *EQ* is a variant of the commonly used modularity (*Q*) metric [1], which is defined for overlapping communities by Shen. This extended modularity is defined as follow:

$$EQ = \frac{1}{2m} \sum_{C} \sum_{i,j \in C_k} \frac{1}{O_i O_j} [A_{ij} - \frac{k_i k_j}{2m}] \tag{6}$$

4.3 Experimental Results

In the experiments, we use two algorithms to compare with the proposed IGLC algorithm. The two algorithms are COPR [12] and LINK [14], respectively. The parameters of IGLC are set as follows: $size = 100$, $gens = 100$, $pc = 0.6$, $pm = 0.4$, $\Delta = 0.66$, $\tau = 0.25$. The parameter of COPRA is set as follows: $v = 4$. The LINK algorithm don't need parameters.

The Results on the Simulated Data Sets. The results on the two simulated data sets are shown in the Fig. 3. The abscissa is the *om* whose value ranges from 2 to 8, and the ordinate is the *NMI* value. The NMI value of per *om* is the average of 10 times.



(a) S1($mu = 0.1$) (b) S2($mu = 0.3$)

Fig. 3. The results on the simulated data sets

(1) *Compared with the LINK algorithm.*
In the two LFR benchmark networks, the results of IGLC are all better than the results of LINK. Because the LINK algorithm exists "excessive overlap" problem, which seriously reduces the quality of community detection. However, the proposed IGLC algorithm solves the "excessive overlap" problem very well. Therefore, the proposed IGLC algorithm effectively improves the quality of community detection.

(2) *Compared with the COPRA algorithm.*
In the low mixing degree LFR benchmark network($mu = 0.1$), the *NMI* values of IGLC are all better than the *NMI* values of COPRA. In the high mixed degree LFR benchmark network ($mu = 0.3$), the *NMI* values of IGLC are most better than the *NMI* values of COPRA, Only in a few cases, the COPRA has higher NMI value (e.g. $om = (5, 7)$ in the S1. In addition, with the increase of *om*, the community detection is becoming more and more difficult. The *NMI* value of COPRA present fluctuations, which demonstrates that the COPRA algorithm has poor robustness. However, the *NMI* value of proposed IGLC algorithm present the steady downward trend, which demonstrates that our algorithm has good robustness.

In conclusion, the proposed IGLC algorithm can obtain better quality of overlapping community detection compared with the LINK algorithm and COPRA algorithms in most cases.

The Results on the Real Data Sets. Table 3 shows community detection results of three algorithms on real data sets, whose detailed information is shown in Table 2. The bold in each row is the optimal community detection result. The evaluation criterion in Table 3 is the extended modularity *EQ*.

Table 3. The results on the real network

Name	LINK	COPRA	IGLC
Karate	0.146	0.423	**0.514**
Dolphins	0.351	0.683	**0.729**
Political Books	0.254	**0.813**	0.804
Football	0.557	0.685	**0.689**
Netscience	0.457	0.812	**0.893**

In the Table 3, compared with the LINK and COPRA algorithm, the proposed IGLC algorithm can obtain better clustering results on most networks. Although the proposed IGLC algorithm don't have the best *EQ* values on the Political Books network, The *EQ* values are also the second best values.

In conclusion, the proposed IGLC algorithm can achieve acceptable results on real data sets, so the IGLC algorithm is reasonable and effective in overlapping community detection.

5 Conclusion

In this paper, we propose an improved genetic-based link clustering for overlapping community detection(IGLC). The IGLC algorithm mainly includes two parts. One part is adopting the GaoCD algorithm to detect the link communities. The other is transforming the link communities into the node communities and adopting the community similarity and the belonging coefficients to solve the "excessive overlap" problem. Through experimental comparison, the proposed algorithm is effective and efficient on both simulate networks and real networks.

References

1. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**(2), 026113 (2004)
2. Lee, J., Gross, S.P., Lee, J.: Modularity optimization by conformational space annealing. Phys. Rev. E **85**(5), 056702 (2012)
3. Shen, H.W., Cheng, X.Q.: Spectral methods for the detection of network community structure: a comparative analysis. J. Stat. Mech: Theory Exp. **10**, P10020 (2010)

4. Jiang, J.Q., Dress, A.W.M., Yang, G.: A spectral clustering-based framework for detecting community structures in complex networks. Appl. Math. Lett. **22**(9), 1479–1482 (2009)
5. Palla, G., Derényi, I., Farkas, I., et al.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**(7043), 814–818 (2005)
6. Kumpula, J.M., Kivelä, M., Kaski, K., et al.: Sequential algorithm for fast clique percolation. Phys. Rev. E **78**(2), 026109 (2008)
7. Shen, H., Cheng, X., Cai, K., et al.: Detect overlapping and hierarchical community structure in networks. Physica A **388**(8), 1706–1712 (2009)
8. Lancichinetti, A., Fortunato, S., Kertész, J.: Detecting the overlapping and hierarchical community structure in complex networks. New J. Phys. **11**(3), 033015 (2009)
9. Havemann, F., Heinz, M., Struck, A., et al.: Identification of overlapping communities and their hierarchy by locally calculating community-changing resolution levels. J. Stat. Mech: Theory Exp. **2011**(01), P01023 (2011)
10. Kelley S.: The existence and discovery of overlapping communities in large-scale networks. Rensselaer Polytechnic Institute (2009)
11. Lancichinetti, A., Radicchi, F., Ramasco, J.J., et al.: Finding statistically significant communities in networks. PLoS ONE **6**(4), e18961 (2011)
12. Gregory, S.: Finding overlapping communities in networks by label propagation. New J. Phys. **12**(10), 103018 (2010)
13. Xie, J., Szymanski, B.K., Liu, X.: SLPA: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In: IEEE 11th International Conference on Data Mining Workshops (ICDMW), pp. 344–349. IEEE (2011)
14. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**(7307), 761–764 (2010)
15. Ball, B., Karrer, B., Newman, M.E.J.: Efficient and principled method for detecting communities in networks. Phys. Rev. E **84**(3), 036103 (2011)
16. Kim, Y., Jeong, H.: Map equation for link communities. Phys. Rev. E **84**(2), 026110 (2011)
17. Shi, C., Cai, Y., Fu, D., et al.: A link clustering based overlapping community detection algorithm. Data Knowl. Eng. **87**, 394–404 (2013)
18. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E **78**(4), 046110 (2008)
19. Lancichinetti, A., Fortunato, S.: Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys. Rev. E **80**(1), 016118 (2009)
20. Newman. Network Data [EB/OL]. http://www-personal.umich.edu/~mejn/netdata/. 19 April 2013