Community Detection Based on an Improved Modularity

Zhen Zhou, Wei Wang, and Liang Wang

National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing 100190, P.R.China

Abstract. Community detection is a very popular research topic in network science nowadays. Various categories of community detection algorithms have been proposed, such as graph partitioning, hierarchical clustering, partitional clustering, and so on. Among these algorithms, modularity-based approaches obtain more attention because modularity is a main criterion to evaluate community partitions. However, current modularity only measures the intra-links within communities and rarely considers the inter-links between them. In this paper, we encode both the intra-links and inter-links in an optimization framework to improve the modularity. The partitions can be computed by the greedy algorithm which utilizes the similar simulated annealing technique. The experimental results on four public datasets demonstrate that our improved modularity can reduce the links between communities, and achieve better performance than the original modularity.

Keywords: community detection, improved modularity, greedy algorithm.

1 Introduction

Many networks of interest, including social networks, computer networks, and metabolic and regulatory networks, are found to divide naturally into different communities or modules[19]. Such networks, in the abstract, contain two basic elements: individuals and relations. Individuals tend to form groups according to their relations, for instance, circles and teams in social networks. In general, a node or vertex is used to describe an individual and a link (or edge) between any two nodes represents their relations. Accordingly, groups are viewed as communities. The community, to some extent, can embody the latent rules of networks. Hence, the community becomes the entry point of researches of networks structure and functionality. Community detection is a fundamental research issue and attracts much interest over the last decade.

Community detection is to recognize the inherent structure of networks, i.e., dividing a network into several communities which have high density of edges within communities and low density between them. So, community detection is inextricably linked to graph partition and traditional clustering. Currently, the existing methods for community detection can be divided into several categories. One is *graph paritioning* and its typical algorithms contain the Kernighan-Lin algorithm[21] and the spectral bisection method[25]. *Hierarchical clustering* is also a technique especially for social networks[15,22]. *k-means*[14] and *fuzzy k-means*[16,5] are the most commonly-used algorithms of *partitional clustering*. *Spectral clustering* requires to compute the first k eigenvectors of a Laplacian matrix[24].

C.-L. Liu, C. Zhang, and L. Wang (Eds.): CCPR 2012, CCIS 321, pp. 638–645, 2012.

The concept of community detection was formally proposed in 2002 by Girvan and Newman, together with a *divisive algorithm*[10]. Two years later, they proposed the famous concept of modularity[20]. Modularity has been employed as a quality function in many algorithms, such as the divisive algorithms[7]. In addition, modularity optimization is itself a popular method for community detection. The main optimal techniques contain greedy techniques[3], simulated annealing[12], extremal optimization[6], spectral optimization[18], and so on. In this paper, we improve the modularity, as the original one only concerns edges within communities, which does not take the influence of inter edges into account. This will lead the modularity divide networks into large components, which is the so-called resolution problem[8]. The improved modularity considers both intra and inter edges, so it can more fully reflect the potential structure of networks. Then we take the greedy algorithm and similar simulated annealing technique to optimize this improved modularity. The experimental results on four public datasets available outperform those of using the original modularity.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the concept of modularity. In Section 3, we will detail our improved modularity and the analysis of parameter selection. Section 4 reports the experimental results. We conclude this work in Section 5.

2 Modularity

Modularity is based on the idea that a *null model* is expected to have no community structure, and the structure of the network consists of communities as long as it is sufficiently different from the null model. The null model is a copy of the original graph which keeps some of its structured properties, such as the degree sequence, but without community structure[7]. The most commonly-used null model is the *random graph*, though it is strictly not a null model. Modularity is defined as[3]

$$Q = \frac{1}{2m}\sum_{ij}^{n}(A_{ij} - \frac{k_i k_j}{2m})\delta(C_i, C_j). \tag{1}$$

where m and n stand for the number of edges and nodes, respectively. A is the adjacency matrix, if two nodes are connected, the corresponding A_{ij} represents the weight of connected edges between them, otherwise $A_{ij} = 0$. k_i and k_j are the degree of vertex i and j. C_i is the community vertex i belongs to. $\delta(C_i, C_j)$ is a sign function, i.e. if $C_i = C_j$, $\delta(C_i, C_j) = 1$, otherwise $\delta(C_i, C_j) = 0$. The fraction, $P_{ij} = \frac{k_i k_j}{2m}$, stands for the expected number of edges in the corresponding null model. The function δ only makes sense when vertex i and j are in the same community. The difference between $(A)_{ij}$ and $(P)_{ij}$ is the difference between the original graph and the corresponding null model. This discrepancy is expected to be as large as possible so that the graph far more likely consists of the community structure. Then Q can be seen as the discrepancy between the graph and null model. Therefore, the higher Q, the better. In this way, community detection becomes a procedure of finding the maximum Q. This is a NP-hard optimal problem[4]. Thus, heuristic optimization methods turn into the main research ideas. However, this modularity is far from perfection. In our method, we improve it and get a better performance.

3 The Proposed Algorithm

3.1 Improved Modularity

In Equation (1), $\delta(C_i, C_j)$ only works when $C_i = C_j$, which means that edges within communities would be computed. A good partition will make edge density within communities be high and low between them. So, there is obviously a puzzle for the original modularity: how the inter edges influence the the procedure of community detection? Maybe one will argue whether it matters. However, the modularity without taking the inter edges into account tends to split large communities as the large ones make more C_i equal to C_j so that Q has a higher value. Hence, one drawback of the original modularity is that it tends to 'eat' relatively small communities. To address such problem, we modify the modularity as

$$Q = \frac{1}{2m}\sum_{ij}^{n}\Big[(A_{ij} - \frac{k_i k_j}{2m})\delta(C_i, C_j) - \beta\big(A_{ij} - \frac{k_i k_j}{2m}\big)^{\alpha}\big(1 - \delta(C_i, C_j)\big)\Big], \quad (2)$$

where β and α are the undetermined parameters, other variables are the same as Equation (1). We call $(A_{ij} - \frac{k_i k_j}{2m})\delta(C_i, C_j)$ an *intra factor*, $\beta\big(A_{ij} - \frac{k_i k_j}{2m}\big)^{\alpha}\big(1 - \delta(C_i, C_j)\big)$ an *inter factor*. So this improved modularity considers both intra and inter edges of networks. The value of the inter factor depends on the parameters β and α. If large communities are expected, we increase β and reduce α, and vice versa. Ideally, β and α are expected to be automatically assigned according to different networks. Although a clear expression is very hard, the parameter analysis is a necessary task in our method. Our method adapts the greedy technique ultilizing the simulated annealing technique to find the optimal solution.

3.2 Optimization

The greedy technique[3], just as its name, always makes the best choice or the maximum value of modularity in each step. It may not reach the global optimal solution in some sense. However, simulated annealing is a probabilistic process for global optimization, and this is the reason why we take it. By combining these two methods, we hope to obtain a fast and probabilistic global optimal modularity algorithm. The procedure of our algorithm is summarised as follows.

1. Initially, each vertex is a community with label $C_i = i, i = 1, ..., n_c$, n_c is the number of communities. The adjacency matrix is $A_{n_c \times n_c}$. M is symmetric and the diagonal is zero. In addition, J is the index set of C_i's connected communities. First, $i = 1$.
2. We assume to merge the community C_i and $C_{j \in J}$. Compute the corresponding change of our modularity, ΔQ. After j sweeps all of C_i's connected communities, we find the maximum change, ΔQ_{max}. At this time, the other corresponding community is C_j.
3. If $\Delta Q_{max} > 0$, accept this change and rewrite $C_j = i$. Otherwise we give a probability to accept this.
4. If $i + 1 \leq n_c$, $i = i + 1$, and go to step 2, else go to step 5.

5. If $n_c = 1$, then stop the process and output results, else do the following. After i runs over all communities, we get a new partition. Then we merge communities with the same label, reset the label and update the number of communities, n_c, which is generally reduced. Update the adjacency matrix $A_{n_c \times n_c}$, where A_{ij} is the total weight between current C_i and C_j. Also, each index set will be updated. Go to step 2.

It should be noted that, in step 3, if $\Delta Q_{max} \leq 0$, then we give a probability to accept this result. This is what we called similar stimulated annealing technique. In order to simplify this problem, we just randomly make a real-valued number between 0 and 1, and the given probability 0.3 is the comparison criterion.

4 Experiments

4.1 Metrics

Some measures are essential to judge the partition qualities. An intuitive way is visual analysis, but this is always hard especially when a network is large. Two quantitative indicators are used here. One is *purity*[2], and the other is *normalized mutual information(NMI)*[13]. To compute purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned and being divided dividing by the number of vertices, n. Formally:

$$purity(\Omega, \mathbb{C}) = \frac{1}{n} \sum_k \max_j \mid \omega_k \cap c_j \mid, \tag{3}$$

where $\Omega = \{\omega_1, \omega_2, ..., \omega_K\}$ is the set of clusters and $\mathbb{C} = \{c_1, c_2, ..., c_J\}$ is the set of classes. Bad clustering has a purity value close to 0, while a perfect clustering has a purity of 1. High purity is easy to achieve when the number of clusters is large. In particular, the purity will be 1 if each vertex gets its own cluster. Thus, we can not simply use 'purity' to trade off the quality of the clustering against the number of clusters. A measure that allows us to make this tradeoff is normalized mutual information:

$$NMI(\Omega, \mathbb{C}) = \frac{I(\Omega; \mathbb{C})}{[H(\Omega) + H(\mathbb{C})]/2}, \tag{4}$$

where I is mutual information[9] and H is self information[11].

4.2 Datasets

The following four networks with groundtruth are obtained from Gephi Datasets[1]. The networks in these datasets are undirected and edges have constant weight 1.

1) *karate club*[26]: Zachary observed 34 members of a karate club for 2 years. The vertex 1 and 34 represent the instructor of this club and the administrator, respectively. During a period of study, the instructor and administrator had a disagreement. Finally,

[1] `https://wiki.gephi.org/index.php?title=Datasets`

the instructor left and established a new club, which took away about half of the original members. This is the network of karate club.

2) *dolphin community*[1]: This is the biological classification of dolphins proposed by Lusseau. The groundtruth has two clusters: one has 20 vertices and the other 42.

3) *college football*[10]: This is the world of the United States college football for the 2000 season: vertices in the graph represent teams (labeled by their college names) and edges represent regular-season games between two teams. Each conference has 8-12 teams.

4) *school*[23]: This dataset has 11 clusters, ten classes and a set of 10 teachers.

4.3 Results and Analysis

Table.1 summaries the experimental result. It shows that our modularity make a contribution to improve the clustering performance, especially for the dolphin community. Here, P and N represent purity and NMI, B and C represent the original and our modularity, respectively.

Table 1. The quantitative indicators of experiments

dataset	nodes	edges	P_B	P_C	N_B	N_C	β	α
karate club	34	78	0.8235	0.8235	0.6995	$\mathbf{0.8041}$	0.37	2
dolphin community	62	159	0.1613	0.9839	0.4647	$\mathbf{0.8888}$	0.45	2
college football	115	613	0.8783	0.9130	0.8890	$\mathbf{0.9242}$	0.45	2
school	238	5539	0.6555	0.7563	0.8731	$\mathbf{0.9135}$	0.20	2

The visualization of networks derives from Gephi[17]. Here shows the result of the first dataset in Figure.1. From Fig.1.b) we can see that the vertex 10 belongs to vertex 1. However, it should belong to vertex 34 with respect to groundtruth. Of course, vertex 10 is misclassified by the original modularity, which is assigned correctly by our modularity. In the dolphin community using our modularity, only the dolphin PL is misclassified, while the original modularity seperates the network into seven parts. In college football, both modularities maintain the conference structure, but the members of each conference are not the same as groundtruth. For example, the conference named Big Twelve has 12 members. Using the original modularity, its half members are misidentified while it is recognized correctly by ours. So the original modularity separates the Big Twelve into pieces while our method maintains the inherent structure of this conference. School dataset confirms what we have said above: the original modularity tends to cluster large communities. And this drawback is avoided in our modularity. Teacher set is misclassified by both modularities as it, to some extent, is semantic, which can not be reflected by using the degree.

The next problem is how to choose the parameters α and β. Empirically summarized from experiments, α is not sensitive, so it is set to a constant 2. Then we change β and get corresponding results for each β. We just need to find the value of β corresponding to the maximum of NMI. As Fig.2 shows, this job is not hard. Reduce the interval to 0.01 and we find the optimal values β may take on. In Fig.2, the left two datasets can simultaneously reach the optimal value of NMI , and so do the right two.

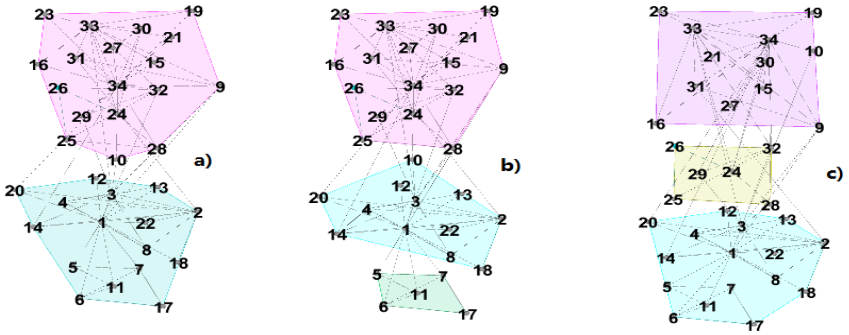

Fig. 1. a) Zachary's karate club. Social network of friendships between 34 members of a karate club at a US university in the 1970[26]. b) The result of community detection with original modularity. c) The result of community detection with our modularity.

As you can see, β varies according to different networks. We hope to find out what is the primary determinant of β? Unfortunately, there is not an available analytic solution of β. Here are some suggestions.

1) It is not hard to think that β should be associated with the network degree distribution, just as shown in Fig.3. The left two figures can be seen as a power law distribution and the right two as normal distribution. So we can simply assign different networks with different β. In a word, the degree distribution of a network determines the value of β.

2) An interesting phenomenon in Fig.2 is that the NMI curve reverses the normalized number of clustering curve in the first two panels and synchronizes in the last two. In our opinion, this is another evidence of our suggestion 1), but this is not yet confirmed.

5 Conclusion

In this paper, we take the inter edges of networks into account and improve the modularity. By using the greedy algorithm and similar simulated annealing technique, we optimize our modularity. Tests on four public datasets show that our modularity outperforms the original one both in purity and NMI.

Community detection is popular in social network analysis in recent years. However, what this field lacks the most is a uniform and precise definition of community. Another outstanding problem is defining a benchmark which implies the natural partition of a network, the one that all algorithms can compare with. This also contains how to evaluate a partition. For the future work, we aim to investigate the above two problems and come up with effective algorithms to solve them.

Acknowledgement. This work is jointly supported by National Natural Science Foundation of China (61175003), Hundred Talents Program of CAS, The strategic Priority Research Program of CAS (XDA06030300), and National Basic Research Program of China (2012CB316300). I am indebted to thanks the SocioPatterns collaboration[2] and all the people who give me useful suggestions and advice to improve my job at various stages.

[2] `http://www.sociopatterns.org`

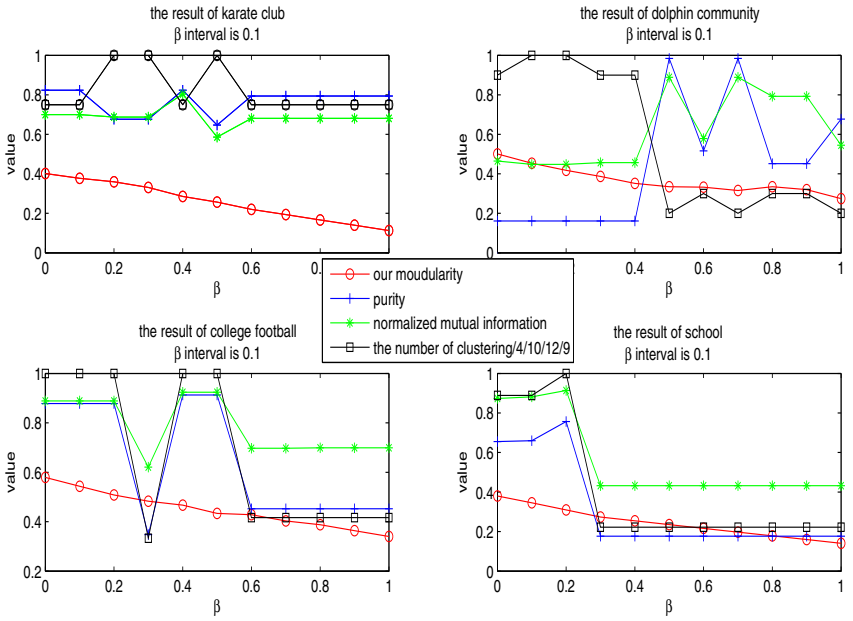

Fig. 2. The curves of improved modularity, purity, NMI and the normalized number of clustering. The horizontal axis represents the value of β. The vertical axis is the corresponding values.

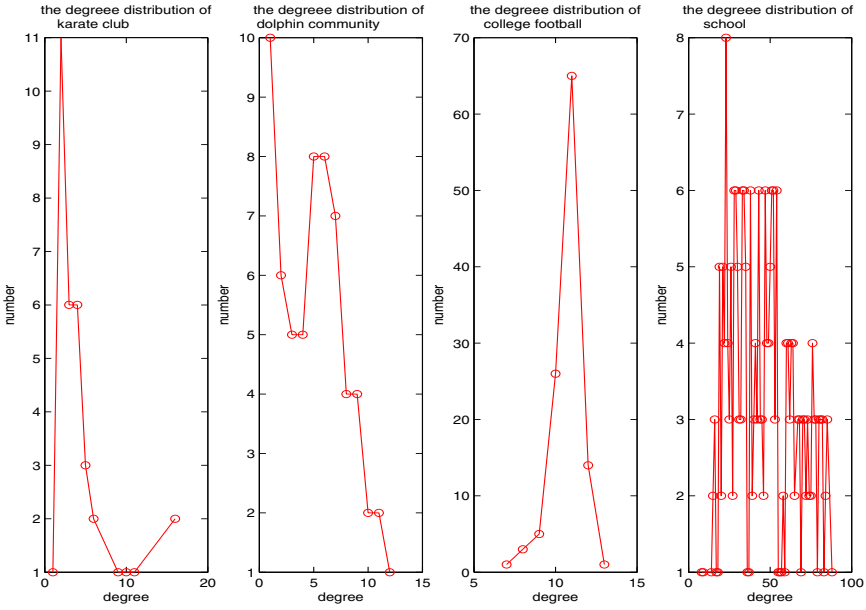

Fig. 3. The degree distribution of four datasets. The horizontal axis represents the value of degree and the vertical axis is the number of nodes with the given degree.

References

1. Areans, A., Feránndez, A., Gómez, S.: Analysis of the structure of complex networks at different resolution levels. New Journal of Physics (2008)
2. Artiles, J., Gonzalo, J., Sekine, S.: Weps 2 evaluation campaign: overview of the web people search clustering task. In: World Wide Web (2009)
3. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics (2008)
4. Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. IEEE Trans. Knowl. Data Eng. (2008)
5. de, F., de Carvalho, A.T., Tenório, C.P.: Fuzzy k-means clustering algorithms for interval-valued data based on adaptive quadratic distances. Fuzzy Sets and Systems (2010)
6. Duch, J., Arenas, A.: Community detection in complex networks using extremal optimization. Physical Review E (2005)
7. Fortunato, S.: Community detection in graphs. Physics Reports (2010)
8. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. Proceedings of the National Academy of Sciences (2007)
9. Gierlichs, B., Batina, L., Tuyls, P., Preneel, B.: Mutual Information Analysis: a Comprehensive Study. Journal of Cryptology (2010)
10. Girvan, M., Newman, M.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences (2002)
11. Gray, R.M.: Entropy and information theory. Springer (2010)
12. Guimerà, R., Amaral, L.A.N.: Functional cartography of complex metabolic networks. Nature (2005)
13. Hu, X., Zhang, X., Lu, C., Park, E., Zhou, X.: Exploiting wikipedia as external knowledge for document clustering. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2009)
14. Laurent, G., Olivier, M., Pierre, C., Alfred, H.O.: Graph based k-means clustering. Signal Processing (2012)
15. Legendre, P., Birks, H.J.B.: Clustering and Partitioning. Springer (2012)
16. Li, M.: Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters. IEEE Transactions on Knowledge and Data Engineering (2008)
17. Bastian, M., Heymann, S., Jacomy, M.: Gephi: An Open Source Software for Exploring and Manipulating Networks. In: International AAAI Conference on Weblogs and Social Media (2009)
18. Newman, M.: Evaluation of clustering. Physical Review E (2006)
19. Newman, M.: Modularity and community structure in networks. Proceedings of the National Academy of Science (2006)
20. Newman, M., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E (2004)
21. Porter, M.A., Onnela, J.-P., Mucha, P.J.: Communities in networks. Notices of the AMS (2009)
22. Shepitsen, A., Gemmell, J., Mobasher, B., Burke, R.: Personalized recommendation in social tagging systems using hierarchical clustering. In: ACM Conference on Recommender Systems (2008)
23. Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L.: High-resolution measurements of face-to-face contact patterns in a primary school. PLoS One (2011)
24. von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing (2007)
25. Yan-ping, Z., Yang, W., Shu, Z.: Detecting communities using spectral bisection method based on normal matrix. Computer Engineering and Applications (2010)
26. Zachary, W.W.: An information flow model for conflict and fission in small groups. Journal of Anthropological Research (1977)