2010 IEEE International Conference on Data Mining Workshops

A Comparison of Objective Functions in Network Community Detection

Chuan Shi and Yanan Cai
Beijing University of Posts and Telecommunications, Beijing, China 100876.
Email: shichuan@bupt.edu.cn

Philip S. Yu
University of Illinois at Chicago, IL, USA, 60607-7053
Email: psyu@cs.uic.edu

Zhenyu Yan
Fair Isaac Corporation, San Rafeal, CA, USA, 94903.
Email: yan_zhen_yu@hotmail.com

Bin Wu
Beijing University of Posts and Telecommunications, Beijing, China 100876.
Email: wubin@bupt.edu.cn

***Abstract*—Community detection, as an important unsupervised learning problem in social network analysis, has attracted great interests in various research areas. Many objective functions for community detection that can capture the intuition of communities have been introduced from different research fields. Based on the classical single objective optimization framework, this paper compares a variety of these objective functions and explores the characteristics of communities they can identify. Experiments show most objective functions have the resolution limit and their communities structure have many different characteristics.**

***Keywords*-Community detection, objective functions, single-objective optimization, multi-objective optimization.**

I. INTRODUCTION

Detecting communities in real-world networks has received a great deal of attentions. The main reason is that communities are supposed to play special roles in the structure-function relationship, and thus detecting communities (or modules) can be a way to identify substructures which could correspond to important functions. For example, the communities in WWW are sets of web pages sharing the same topic [1]; the modular structures in biological networks are widely believed to play important roles in biological functions [2].

Loosely speaking, the communities are groups of nodes that are densely interconnected but only sparely connected with the rest of the network [3,4]. To extract such groups of nodes, one typically chooses an objective function that captures the intuition of a community as a group of nodes with better internal connectivity than external connectivity. The objective is usually NP-hard to optimize, so one usually employs heuristics (e.g., betweenness [5]) or approximation algorithms (e.g., spectral method [18], genetic algorithm [10,11,12]) to find sets of nodes that approximately optimize the objective function and can be understood as communities. As a consequence, the contemporary community detection problem (Ω, P) can be formally defined as a single-objective optimization problem: determine the partition C^* for which

$$P(C^*) = \min_{C \in \Omega} P(C) \tag{1}$$

where Ω is the set of feasible partitions, C is a community structure of a given network G and $P : \Omega \rightarrow R$ is the measure function. Without loss of generality, we assume P is to be minimized. Most contemporary algorithms for Community Detection (CD) are based on the classical single-objective optimization. Different algorithms vary in the optimization criterion P and optimization techniques, e.g., the modularity Q in GN [5], the "cut" function in spectral method [18] and the "description length" in the information-theoretic based method [15].

In order to capture the notion of community, many objective functions (see Section III) have been proposed based on graph theory, physical phenomenon, and even by intuition. In applications, many researchers have found that some objective functions are "biased". That is, the optimization on some objective functions returns results that might be substantially different from the real communities. For example, the modularity optimization (i.e., optimization based on Q) may fail to identify modules smaller than a certain size [16]. Moreover, different objective functions might lead to different partitions. Some might generate a huge number of communities with small sizes, whereas some might return a small number of large communities. Thus, it is interesting to compare these objectives on real networks and summarize the characteristics of communities they identify. Understanding structural properties of communities identified by different objective functions can guide the selection of the most appropriate objective in the context of a given network and application.

This paper comprehensively compares and analyzes 11 popular objectives that are already widely used or can potentially be used in community detection. To do so, we use a set of real networks as benchmarks

978-0-7695-4257-7/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDMW.2010.107

and propose some new quality measure metric. From extensive experiments, we find that the communities identified by these objectives have different characteristics.These findings reveal the trait of objective functions and their structural characteristics, which can guide a user to choose the suitable objective functions in applications.

The rest of the paper is organized as follows. Section II introduces the related work. In section III, we summarize the 11 objective functions that are either widely used or potentially can be used for community detection. Section IV does the extensive experiments on 12 real networks and observe the structure characteristics of communities identified by 11 objectives based on the single objective optimization framework. Finally, Section V concludes this paper.

II. Related Work

There have been many algorithms to analyze the community structures in complex networks. The algorithms use methods and principles of physics, artificial intelligence, graph theory and even electrical circuits. One of the most known algorithms proposed so far is the Girvan-Newman (GN) algorithm which uses a foundational measure criterion of community, modularity Q [5]. Generally speaking, the larger the value of Q is, the more accurate a partition into communities is. Thus, the community detection becomes a single-objective modularity optimization problem. Since the search for the optimal (largest) modularity value is a NP-complete problem [7], many heuristic search algorithms have been applied to solve the optimization problem, such as extremal optimization, and simulated annealing [4].

To capture the intuition of communities, many other objective functions have been proposed. In the graph theory community, the conductance and expansion [20] criteria are proposed to capture a notion of "surface area-to-volume". In the physical community, Reichardt and Bornholdt [13] proposed a criteria that considers the community indices of nodes as spins in a q-state Potts model. Based on information theory, Fosvall and Bergstrom [15] proposed "description length" to capture an optimal compression of networks' topology. Leskovec et al. [6] summarized some of popular objective functions. Different from the mainstream criteria measuring the quality of whole partition (i.e., all communities with any size), they consider the criteria to capture the best community for given size [6].

Many methods have been proposed to optimize a chosen objective. These methods can be roughly classified into two categories: heuristic methods and optimization based methods. Some heuristic rules are designed, such as edge betweenness in GN [5] and link clustering coefficient [27]. Many optimization technologies (e.g., spectral method, genetic algorithm, simulated annealing and extremal optimization) are applied to optimize a objective, such as Q [5]. The Genetic Algorithm (GA), as an effective optimization technique, has also been used for community detection. It has a special advantage that the number of communities can be automatically determined during the evolutionary process. Tasgin and Bingol [12] first applied GA for CD, in which the objective function is the modularity Q and the encoding scheme is the cluster centers. The GA-Net [10] proposed by Pizzuti optimizes the "community score" criteria and uses the the locus-based adjacency scheme. Recently, Shi et al. proposed a new GA, named GACD. GACD employs the modularity Q as the fitness function and applies a special locus-based adjacency encoding schema to represent the community partitions. Extensive experiments have validated GACD's effectiveness [11].

III. Objective Functions

Many objective functions have been proposed to capture the intuition of communities. We summarize some of the objective functions that are already widely used in community detection literatures or can be potentially used for community detection.

Let $G(V, E)$ be an undirected graph with $n = |V|$ nodes and $m = |E|$ edges. Let C be a partition with l communities and S be the set of nodes in one community, where $C = \{S_1, S_2, \cdots, S_l\}$. n_S is the number of nodes in S, $n_S = |S|$; m_S the number of edges in S, $m_S = |(u, v) : u \in S, v \in S|$; and c_S, the number of edges on the boundary of S, $c_S = |(u, v) : u \in S, v \notin S|$; and $d(u)$ is the degree of node u. Here we consider the quality of a partition.

- **Conductance**: $P_1(C) = \sum_{S\in C} \frac{c_S}{2m_S+c_S}$ measures the fraction of total edge volume that points outside the cluster [9,20].
- **Expansion**: $P_2(C) = \sum_{S\in C} \frac{c_S}{n_S}$ measures the number of edges per node that point outside the cluster [19].
- **Cut Ratio**: $P_3(C) = \sum_{S\in C} \frac{c_S}{n_S(n-n_S)}$ is the fraction of all possible edges leaving the cluster [21].
- **Normalized Cut**: $P_4(C) = \sum_{S\in C}(\frac{c_S}{2m_S+c_S} + \frac{c_S}{2(m-m_S)+c_S})$ [9].
- **Maxmum-ODF(Out Degree Fraction)**: $P_5(C) = \sum_{S\in C} max_{u\in S} \frac{|\{(u,v):v\notin S\}|}{d(u)}$ is the maximum fraction of edges of a node pointing outside the cluster [22].
- **Average-ODF**: $P_6(C) = \sum_{S\in C} \frac{1}{n_S} \sum_{u\in S} \frac{|\{(u,v):v\notin S\}|}{d(u)}$ is the average fraction nodes' edges pointing outside the cluster [22].
- **Flake-ODF**: $P_7(C) = \sum_{S\in C} \frac{|\{u:u\in S, |\{(u,v):v\in S\}|<d(u)/2\}|}{n_S}$ is the fraction of

nodes in S that have fewer edges pointing inside than to the outside of the cluster [22].

- **Q**:$P_8(C) = \sum_{S \in C}(\frac{m_S}{m} - (\frac{m_S+c_S}{2m})^2)$ measures the number of within-community edges, relative to a null model of a random graph with the same degree distribution [5].
- **Description Length**: $P_9(C) = n \log l + \frac{1}{2}l(l+1)\log m + \log(\prod_{i=1}^{l}\binom{n_i(n_i-1)/2}{m_i}\prod_{i>j}\binom{n_i n_j}{c_{ij}})$ where n_i and m_i is the number of nodes and edges in cluster i, respectively; c_{ij} is the number of edges between the cluster i and j. The criteria regard the community as a optimal compression of network's topology [15].
- **Community Score**: $P_{10}(C) = \sum_{S \in C}(2m_S/n_S)^2$ measures the density of a sub-matrices based on volume and row/column means [10].
- **Internal Density**: $P_{11}(C) = \sum_{S \in C}(1 - \frac{m_S}{n_S(n_S-1)/2})$ is the internal edge density of the cluster [19].

This paper roughly classifies the objective functions into three categories. The first category contains four criteria (i.e., *Conductance*, *Expansion*, *Cut Ratio*, *Normalized Cut*) from graph theory community. They all consider the "cut" in the graph, since they all contain c_S in the numerator. So we call them the cut-based criteria. The three criteria ended with "ODF", i.e., *Maximum-ODF*, *Average-ODF*, *Flake-ODF*, all consider the degree of nodes in a community, and thus we call them degree-based criteria. Finally, the remaining criteria are classified into one category. All these objectives attempt to capture a group of nodes with better internal connectivity than external connectivity, and thus they all can be potentially used in community detection [6]. Moreover, some objective functions are not considered, for example, the Hamiltonian-based method [13] and a multiple resolution procedure [14]. The objective functions in both of the two methods require tuning parameters. Because the parameters are hard to choose in applications, we did not include them in this paper. Note that some criteria need to be maximized (e.g., Q and *Community Score*). In order to handle all the objectives in a uniform format, we convert these objectives into a minimum problem for convenience. Note that the conversion does not affect the results.

IV. Comparison of Different Objectives

For the minimum problem $\min_{C \in \Omega} P(C)$, many technologies have been applied to optimize them, such as simulated annealing, genetic algorithm, extremal optimization, and spectral methods. For a fair comparison, we use the same optimizer for testing all the objective functions. Particularly, we choose a genetic algorithm proposed by Shi et al., named GACD [11], as the optimizer.

A. Resolution limit in Objective Functions

Fortunato and Barthelemy [16] pointed out that the modularity optimization has the disadvantage of resolution limit, that is, "the modularity optimization may fail to identify modules smaller than a scale which is determined by the size of the network and the degree of interconnectedness of the modules, even in cases where modules are unambiguously defined". Here we examine that whether or not other criteria also have the resolution limit problem. We consider a widely used benchmark network (called $Km - n$ network) made of m identical complete graphs (cliques) with size n, in which these cliques are disjoint from each other with one link [16]. It is obvious that the best partition have m communities (i.e., a clique as a community). However, many algorithms, such as the methods based on Q optimization, tend to incorrectly combine some connected cliques as one community. We optimize all the criteria (objective functions) using GACD on a *K30-5* network. The parameters are settled as follows: the running generation and population size both are 100; the ratio of crossover and mutation are 0.6 and 0.4, respectively (these two parameters are same for all experiments). As shown in Figure 1, we can find that all objective functions have resolution limit except *Community Score* and *Internal Density*. The cut-based criteria and *Description Length* have more serious resolution limit than the others. The *Community Score* always find the correct partition. And the *Internal Density* divides a clique into smaller communities, which shows it tends to over-divides the networks in opposite to the resolution limit problem.

We consider that these findings are reasonable. The definition of *Internal Density* tends to extract the small cliques. For the *Km-n* networks, if more cliques are combined, the denominator on cut-based criteria become larger, and the numerator remains constant. And thus these criteria become smaller. The definitions of cut-based criteria make them to tend to combine small communities into large ones.

B. Real Networks and Measures

To study the properties of communities identified by different objective functions, we optimize the objective functions on the 12 real networks shown in Table 1. These networks with medium and large sizes are from the popular common data sources [23,24]. For all the networks, the population size and running generation in GACD both are settled as 200.

It is difficult to evaluate the quality of communities, since real community structures are unknown. Many contemporary measures use an indicator to evaluate the quality of a partition. Different from these measures, we attempt to measure the quality of communi-

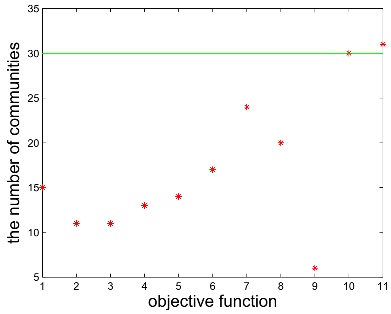


Figure 1. Illustration of resolution limit problem of objective functions on *K30-5* network. The green line in figures are the correct number of communities with the finest granularity. The figure shows the number of communities identified by different objectives. 1-11 represents the objective function $P_1 - P_{11}$, respectively.

ties as a function of their size, which provides a much finer resolution to examine the results of different objective functions. For a community partition C, we consider the average measures of the communities with size k (note that the community size k may be not sequential).

$$\begin{aligned} f(k) &= \frac{\sum_{S_i \in C_k} Measure(S_i)}{|C_k|} \\ C_k &= \{S_i \mid |S_i| = k, S_i \in C\} \\ C &= \{C_k \mid k \in \{1, \cdots, n\}\} \end{aligned} \tag{2}$$

We select two popular measures, Conductance and Shortest Path in graph theory, to evaluate the quality of a community. Conductance can be thought of as the ratio between the number of edges inside the cluster and the number of edges leaving the cluster [9,20]. A smaller Conductance means the ratio of edges pointing outside is smaller, which indicates that nodes in the community sparsely connect with outside. The Shortest Path is the average shortest path of nodes in the community (note that the code schema in GACD promises that all nodes in a community are connected). A smaller Shortest Path shows that the nodes in the community are closely (densely) connect with each other. Together the two measures reflect two aspects of a good community (i.e., densely interconnected and sparsely connected with outside).

$$\begin{aligned} & Conductance(S) = c_S / min(Vol(S), Vol(V \backslash S)) \\ & ShortestPath(S) = \frac{2\sum_{i,j \in S} dis(i,j)}{n(n+1)} \\ & P(k) = \frac{|C_k|}{|C|} \end{aligned} \tag{3}$$

(where $Vol(S) = \sum_{u \in S} d(u)$, $dis(i,j)$ is the shortest path of node i and j, $C = \{S_1, S_2, \cdots S_l\}$). Here, $P(k)$ measures the ratio of communities with size k, which is used in Figure 5(a).

Our measures (especially Conductance based measure) are similar to the Network Community Profile (NCP) proposed by Leskovec [6]. They both are size-resolved measure and conductance-based. However, they have a lot of differences. First, they have different aim. NCP asks for approximation to the best cluster (i.e., a cluster with the smallest conductance) for every possible size. Conductance-based measure calculates the average conductance of communities with the same size. Second, different from the lower-envelope curves (i.e., smallest conductance) of communities with size k illustrated by NCP, Conductance-based measure shows the average conductance of communities with the same size. And thus, we think our measure more comprehensively reflects the quality of communities. Moreover, our measures are more suitable for the evaluation of current algorithms, since they both are based on clusters of any size, rather than the cluster for every possible size.

Because the measure values fluctuate greatly in some situation, we smooth these data with the mid-value method to illustrate them more clearly. The process decreases the fluctuation without changing the trend. In addition, the experimental results do not include those of the degree-based criteria, since we find that these three criteria all are very prone to divide the whole network as one community. It is not very surprising, because their objective values reach the smallest value (i.e., approximate to 0) in this situation. It indicates that the degree-based criteria might not be suitable for community detection.

C. *Experimental Results on Real Networks*

The experimental results on Conductance-based measure are illustrated in Figure 2. Considering the value of conductance, we can find *Community Score* has the largest conductance for almost all networks, which indicates that its communities have more edges connecting with outside. Q has a small conductance for most networks, especially when the community size is large. The cut-based criteria have the similar results and their conductances are smaller compared with the other criteria. We now look at the shapes of conductance curves. For most of networks, the curves generally increase but with some dropping nodes. In other words, the curves have "V" shapes. The bottom nodes in "V" shapes are among 10-100. We also can find that most objectives have the "V" shape at the approximately same community size. The similar "V" shapes have also been found by Leskovec et al. [6] in NCP curves. It indicates that a cluster is more

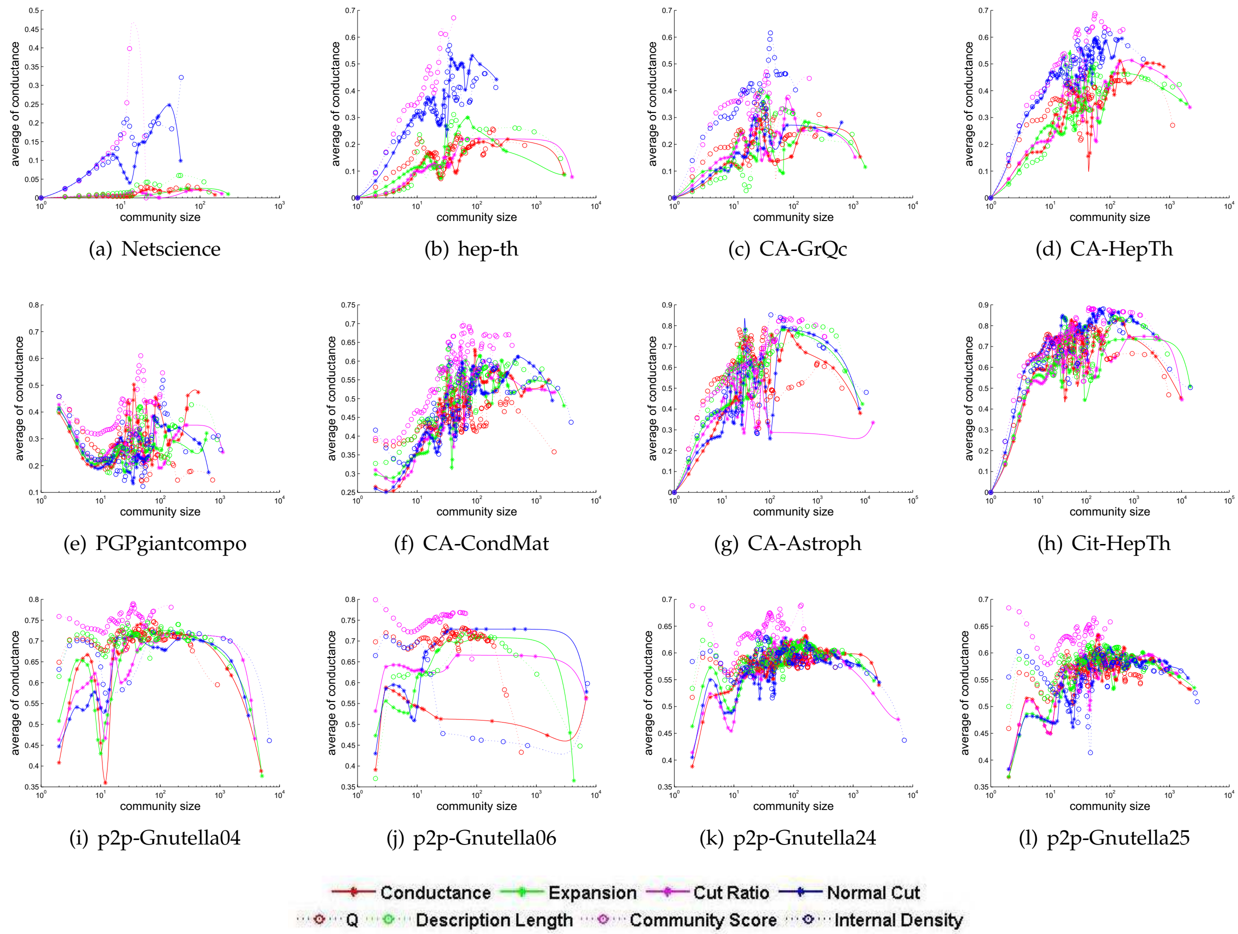


Figure 2. The average of conductance of communities identified by eight objective functions on real networks from P1-P12. The results of the degree-based criteria are not included, because most of their results are not meaningful.

community-like when its size ranges from 10 to 100. Figure 2 also shows that the cut-based criteria and *Internal Density* generate some huge communities with small conductance. It is not particularly surprising, because these huge communities have many edges connecting with outside as well as more edges inside.

Figure 3 illustrates the results of the Shortest Path based measure. Although all objectives have the same shape, they still have some differences. For most networks, the Q has the obviously smaller shortest path, and the *Description Length* have larger shortest path for large communities. The cut-based criteria have similar performance. Observing the shape of curves, we can find most curves have the similar trend that the shortest path increases first and then decreases as the community size increases. The largest average shortest path is between 5 and 12 when the community size ranges from 100 to 1000. When the community size is larger than 1000, the shortest path usually become smaller. The shapes show that the communities also exists the small-world phenomenon.

We further illustrate the distribution of community size in Figure 4 (see the definition of $P(k)$ in Equation 4). It is obvious that their distributions are broad with a tail that can be fairly well approximated by the power law. Most communities are very small and the size of most communities are smaller than 100. In order to observe their distribution more clearly, we statistically estimate their distribution in Figure 5. From Figure 5(a), we can find that *Community Score* and Q find the maximum number of communities and the cut-based criteria reveal the minimum number of communities for most networks. We further show the average size of the smallest 50% and the largest 10% communities identified by eight objectives on the 12 networks in Figure 5(b) and (c), respectively. It shows that most communities are very small. There are no obvious difference on the size of small communities identified

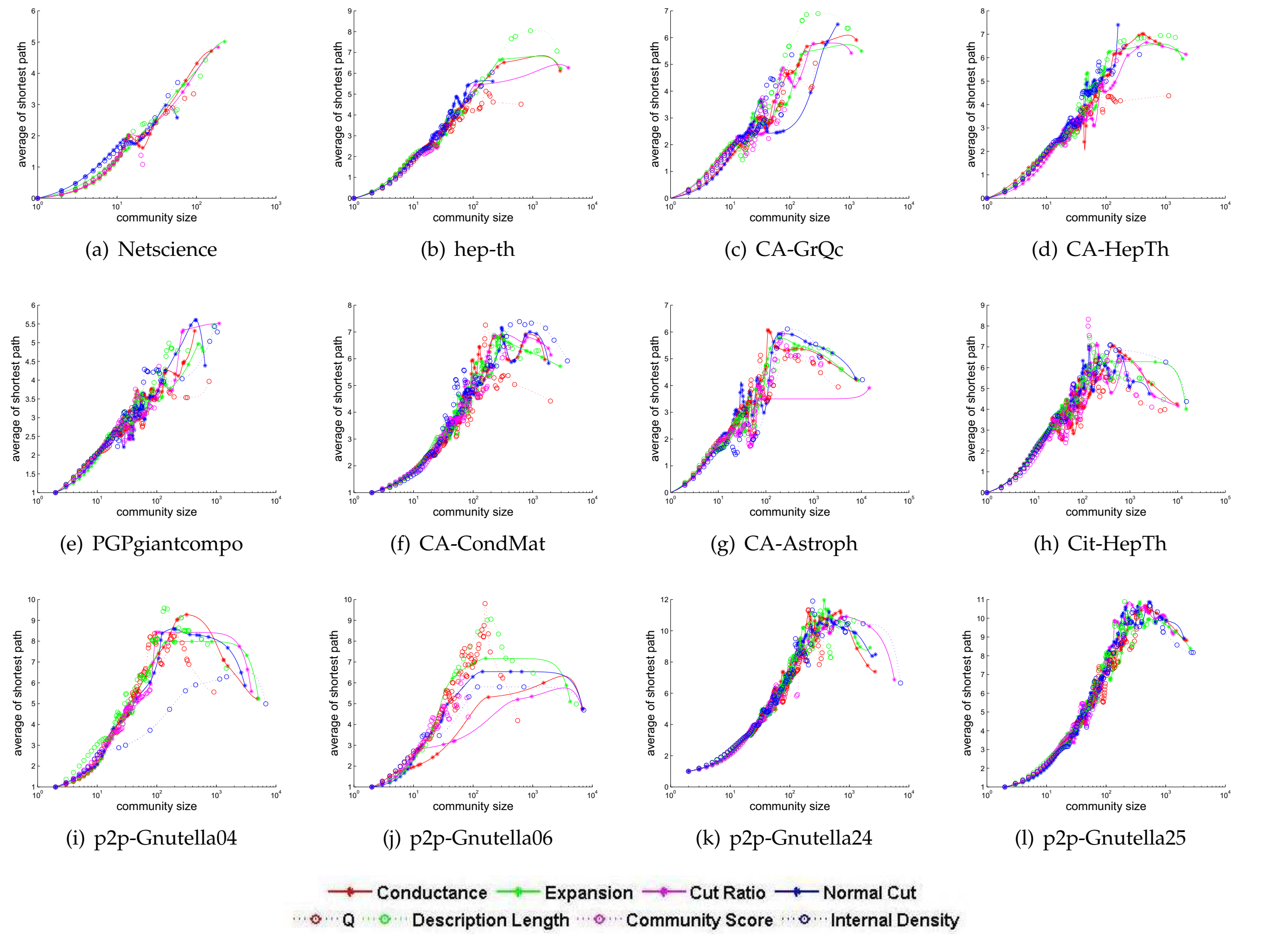


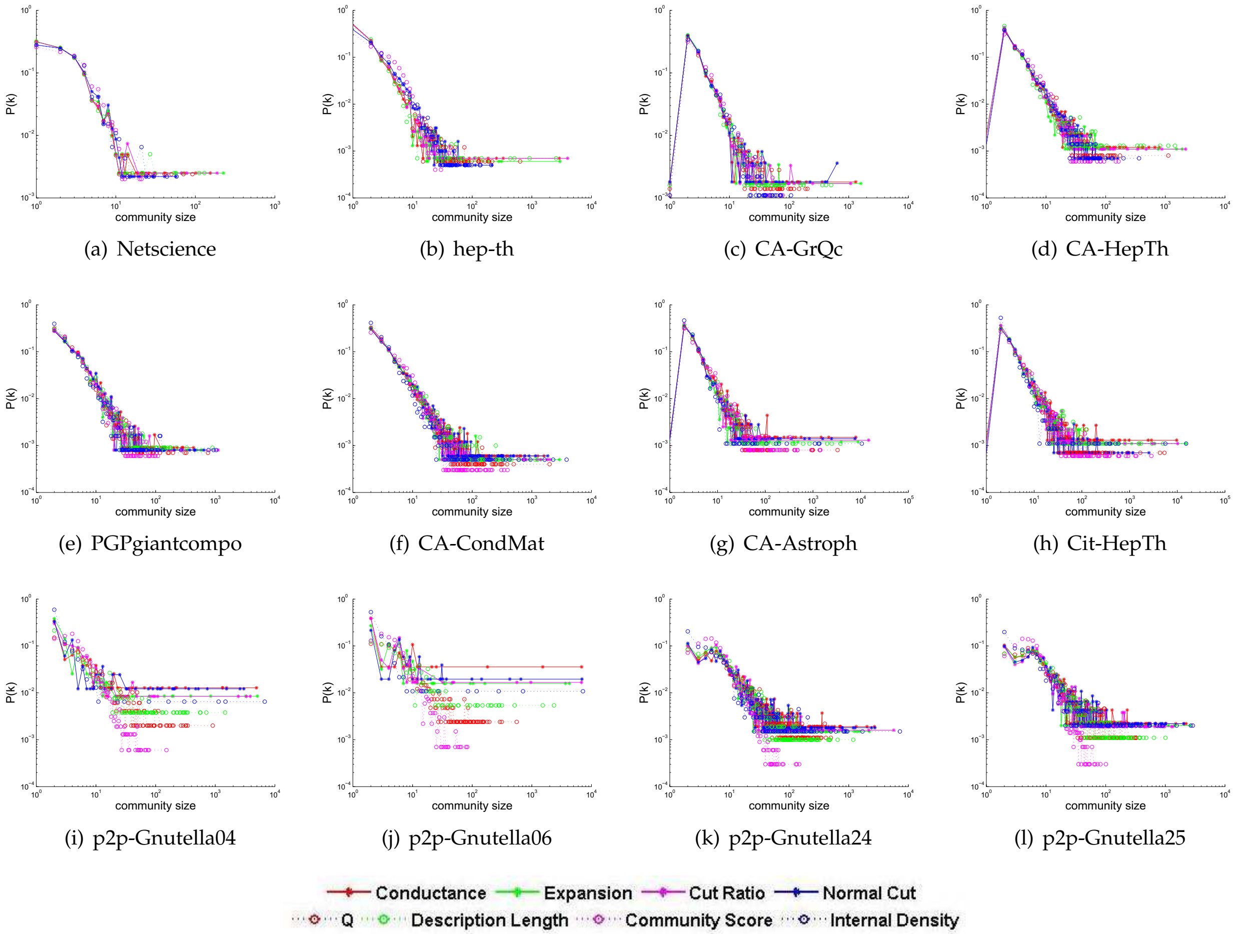
Figure 3. The average of shortest path of communities identified by eight objective functions on real networks from P1-P12. The results of the degree-based criteria are not included, because most of their results are not meaningful.

Table I
REAL NETWORKS AND THEIR SIZE

	Net-science (P1)	Hep-th (P2)	CA-GrQc (P3)	CA-HepTh (P4)	PGPgiantcompo (P5)	CA-CondMat (P6)	CA-AstroPh (P7)	Cit-HepTh (P8)	P2p-Gnutella04 (P9)	P2p-Gnutella06 (P10)	P2p-Gnutella24 (P11)	P2p-Gnutella25 (P12)
Nodes	1589	8361	5242	9877	10680	23133	18772	27770	10876	8717	26518	22687
Edges	2742	15751	28980	51971	24316	186936	396160	352807	39994	31525	65369	54705

by different objectives. However, it is not case for large communities. For most networks, the cut-based criteria always find larger communities, and *Community Score* and Q have the opposite trend. Similar to the cut-based criteria, *Description Length* also tend to find a small number of networks with large size. It is interesting for *Internal Density* that simultaneously find many small communities and some large communities.

In all, the experiments show that *Community Score* and Q divide the networks with finer granularity (i.e., more communities with smaller size). The cut-based criteria and *Description Length* reveal the community structure with coarser granularity (i.e., fewer communities with large size). The behavior of *Internal Density* has the characteristics of both of them. On one hand, *Internal Density* divides the networks into many small communities with finer granularity than those of *Community Score* and Q; on the other hand, it generates some large communities like those of the cut-based criteria. Although the large communities identifies by cut-based criteria have small conductance and shortest path, these large communities are not meaningful, which indicates that these criteria may not be very suitable for networks with small communities.

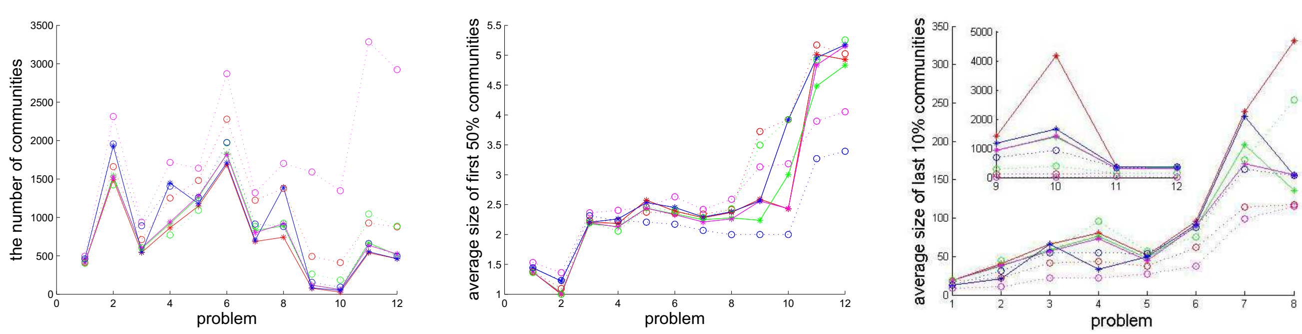


(a) Netscience (b) hep-th (c) CA-GrQc (d) CA-HepTh

(e) PGPgiantcompo (f) CA-CondMat (g) CA-Astroph (h) Cit-HepTh

(i) p2p-Gnutella04 (j) p2p-Gnutella06 (k) p2p-Gnutella24 (l) p2p-Gnutella25

Figure 4. The distribution of the size of communities identified by eight objective functions on real networks from P1-P12. The results of the degree-based criteria are not included, because most of their results are not meaningful.



(a) the number of communities (b) average size of smallest 50% communities (c) average size of largest 10% communities

Figure 5. The distribution of the size of communities identified by different objective functions. The lines have the same meaning with those in Figure 4.

Moreover, the cut-based criteria have very similar distribution in all three different aspects, which shows that they are intrinsically similar. In addition, in all these experiments and results, the modularity Q has the stable and good performance. It might explain why Q is the most effective and popular criterion.

V. Conclusion

In this paper, we systematically examined a wide range of objective functions that are either widely used or potentially can be used for community detection. Through the experiments, we found that different types of objectives behave in different ways. The cut-based criteria, degree-based criteria and *Description Length* tend to divide the networks with coarser granularity. Namely the community size is large and the number is small. *Community Score* and Q have the opposite trend (i.e., large number of communities with small size). *Internal Density* is between those two types and can divide the networks into some large communities as well as many small communities. Experiments also show that most objectives have the resolution limit. That is, they all tend to combine some small communities into a large one except for *Community Score* and *Internal Density*. These findings will be beneficial for selecting the most appropriate objectives in the context of networks and applications.

Acknowledgments

This work is supported by in part by the National Science Foundation of China (No. 60905025, 90924029). It is also supported in part by the National High-tech R&D Program of China (No.2009AA04Z136) and the National Key Technology R&D Program of China (No.2006BAH03B05).

References

[1] G.W. Flake, S. Lawrence, C.L. Giles, F.M. Coetzee, Self-Organization and Identification of Web Communities, IEEE Computer 35(3) (2002) 66-71.

[2] E. Ravasz, A.L. Somera, A. Mongru, Z.N. Oltvai, A.L. Barabasi, Hierachical Organization of Modularity in Metabolic Networks, Science 297(5586) (2002) 1551 - 1555.

[3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D.U. Hwang, Complex Networks: Structure and Dynamics, Physics Report 424(4-5) (2006) 175-308.

[4] R. Guimera, L.A.N. Amaral, Functional Cartography of Complex Metabolic Networks, Nature 433 (2005) 895-900.

[5] M.E.J. Newman, M. Girvan, Finding and Evaluating Community Structure in Networks, Physics Review E, 69 (2004) 026113.

[6] J. Leskovec, K.J. Lang, M.W. Mahoney, Empirical Comparison of Algorithms for Network Community Detection, www 2010.

[7] U. Brandes, D. Delling, M. Gaetler, et al., On Modularity Clustering, IEEE Transactions on Knowledge and Data Engineering, 20(2) (2008) 172-188.

[8] C. Shi, C. Zhong, Z.Y. Yan, et al., A Multi-Objective Optimization Approach for Community Detection, WCCI2010.

[9] J. Shi and J. Malik. Normalized cuts and image segmentation. IEEE Transaction of Pattern Analysis and machine Intelligence. 22(8):888-905, 2000.

[10] C. Pizzuti, GA-Net: a genetic algorithm for community detection in social networks, in PPSN2008 1081-1090.

[11] C. Shi, Z.Y. Yan, Y. Wang, Y.N. Cai, and B. Wu, A Genetic Algorithm for Detecting Communities in Large-scale Complex Networks, Advance in Complex System, 13(1),(2010), 3-17.

[12] M. Tasgin H. Bingol, Community Detection in Complex Networks using Genetic Algorithm, arXiv:cond-mat/0604419, 2006.

[13] J. Reichardt, S. Bornholdt, Statistical Mechanics of Community Detection, Physics Review E 74(1)(2006) 016110.

[14] A. Arenas, A. Fernandez, S. Gomez, Analysis of the structure of complex networks at different resolution levels, arXiv: physics/ 0703218 v1, 2007.

[15] R. Martin, T.B. Carl, An information-theoretic framework for resolving community structure in complex networks, PNAS 104(2007) 7327-7331.

[16] S. Fortunato, M. Barthelemy, Resolution Limit in Community Detection, Proceedings of the National Academy of Sciences 104(1)(2007)36-41.

[17] K. Deb, Multiobjective Optimization using Evolutionary Algorithms, U.K: Wiley, 2001.

[18] A. Pothen, H. Sinmon, K-P. Liou, Partitioning Sparse Matrices with Eigenvectors of Graphs, SIAM J. Matrix Anal App. l(11)(1990)430-452.

[19] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Defining and Identifying Communities in Networks, PNAS 101(9)(2004) 2658-2663.

[20] R. Kannan, S. Vempala, and A. Vetta, On Clusterings: good, bad and Spectral. Journal of the ACM, 51(3):497-515, 2004.

[21] S. Fortunato. Community detection in graphs. arXiv:0906.0612, June 2009.

[22] G. Flake, S. Lawrence, and C. Giles. Efficient identification of web communities. In KDD'00, 150-160, 2000.

[23] http://snap.stanford.edu.

[24] http://www-personal.umich.edu/ mejn/netdata.