On Selection of Objective Functions in Multi-Objective Community Detection

Chuan Shi
Beijing University of Posts and Telecommunications
Beijing China 100876
shichuan@bupt.edu.cn

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

Yanan Cai
Beijing University of Posts and Telecommunications
Beijing China 100876
diandacainan@gmail.com

Zhenyu Yan
Fair Isaac Corporation(FICO)
CA USA 94903
yan_zhen_yu@hotmail.com

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

ABSTRACT

There is a surge of community detection of complex networks in recent years. Different from conventional single-objective community detection, this paper formulates community detection as a multi-objective optimization problem and proposes a general algorithm NSGA-Net based on evolutionary multi-objective optimization. Interested in the effect of optimization objectives on the performance of the multi-objective community detection, we further study the correlations (i.e., positively correlated, independent, or negatively correlated) of 11 objective functions that have been used or can potentially be used for community detection. Our experiments show that NSGA-Net optimizing over a pair of negatively correlated objectives usually performs better than the single-objective algorithm optimizing over either of the original objectives, and even better than other well-established community detection approaches.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-Data Mining

General Terms

Algorithm

Keywords

Complex network, community detection, multi-objective optimization, evolutionary algorithm

1. INTRODUCTION

Recently a large amount of research has been devoted to the task of defining and identifying communities in social and information networks. Loosely speaking, communities are groups of nodes that are densely interconnected but only sparely connected with the rest of the network [3]. 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. As a consequence, the community detection problem (Ω, O) can be formally defined as a Single-objective Optimization Problem (SOP): determine the partition C^* for which

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

where Ω is the set of feasible partitions, C is a community structure of a given network G and $O : \Omega \to R$ is an objective function. Without loss of generality, we assume O is to be minimized. Most conventional community detection algorithms are based on the SOP. Different algorithms vary in the objective function O and optimization techniques.

These single-objective community detection algorithms have been widely applied to both artificial and real problems. However, they also face some fundamental difficulties. These single-objective algorithms attempt to optimize just one objective function and this confines the solution to a particular community structure property. Moreover, these algorithms may fail when the optimized objectives are inappropriate. In addition, one single fixed community partition returned by the single-objective algorithms may not be suitable for the networks with multiple potential structures (e.g., hierarchical and overlapping structures).

It might be more natural and reasonable to consider the community structure from different angles (i.e. multiple optimized objectives) at the same time. That is, in the multi-objective community detection problem $(\Omega, O_1, O_2, \cdots, O_t)$, we aim to discover the community structure C^* for which

$$O(C^*) = \min_{C\in\Omega}(O_1(C), O_2(C), \cdots, O_t(C)) \tag{2}$$

where t is the number of objectives and O_i represents the i-th objective. With the introduction of multi-objective, there is usually no single best solution for this optimization task, but

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

instead, the notion of Pareto optimality should be embraced. For two partitions $C_1, C_2 \in \Omega$, the partition C_1 is said to dominate the partition C_2 (denoted as $C_1 \preceq C_2$) if and only if

$$\begin{aligned} &\forall i \in \{1, \cdots, t\}\ O_i(C_1) \leq O_i(C_2) \\ &\wedge \exists i \in \{1, \cdots, t\}\ O_i(C_1) < O_i(C_2) \end{aligned} \tag{3}$$

A partition $C \in \Omega$ is said to be Pareto optimal if and only if there is no other partition dominating C. The set of all Pareto optimal partitions is the Pareto optimal set and the corresponding set in the objective space is called the non-dominated set, or Pareto front.

Compared to single-objective approaches, the multi-objective community detection has many advantages. (1) The optimal solutions of the single-objective community detection problems defined by $(\Omega, O_1), \cdots, (\Omega, O_t)$ are always comprised by the Pareto optimal set of the multi-objective problem defined by $(\Omega, O_1, \cdots, O_t)$. (2) The multiple objectives can measure characteristics of community structure from different angles, and thus it helps to avoid the risk that one single objective may only be suitable to a certain kind of networks. (3) The multi-objective community detection usually returns a set of community partitions according to the multiple optimized objectives. These community partitions reveal community structure from different angles, which help to discover complex and comprehensive community structures.

In order to effectively solve the Multi-objective Optimization Problem (MOP), we propose a general solution, NSGA-Net, which simultaneously optimizes multiple objective functions with an evolutionary algorithm. As a general multi-objective community detection solution, NSGA-Net can optimize over any multiple objective functions. For this new community detection paradigm, one important issue is that what type of objective functions should be optimized to improve the accuracy of community partition. To solve this issue, we first study correlation relations among 11 popular objective functions and divide the relations between any two objective functions into three categories: positively correlated, independent, and negatively correlated. Then we compare NSGA-Net optimizing over six pairs of objective functions from these three types of correlations (two pairs for each type) to a SOP based approach optimizing over the original single objective. Experiments demonstrate that NSGA-Net only with negatively correlated objectives usually leads to a better performance than that can be achieved by any of the original objectives. We also show that with a pair of negatively correlated objectives, the NSGA-Net performs better than most conventional community detection algorithms.

2. NSGA-NET

In order to effectively solve the multi-objective community detection problem, we propose the NSGA-Net solution based on Evolutionary Algorithm (EA). EA has been proven to be an effective method for MOP. Evolutionary Multi-objective Optimization (EMO) has become one of the main research fields in the EA community, which has also been applied in data mining [11]. Conventional EMO algorithms are designed for numerical optimization problems. When we solve a real problem with EMO, many components of EA need to be redesigned.

Multi-objective optimization mechanism. In this paper, we select NSGA-II [2] as the multi-objective optimization mechanism in NSGA-Net. Four parameters govern the run of NSGA-Net: the population size *popSize*, the running generation *gen*, the ratio of crossover *croRat* and the ratio of mutation *mutRat*.

Genetic representation. We apply the *locus-based adjacency* [4] to represent a partition. In this graph-based representation, each genotype g consists of n genes $g_1, g_2, \cdots, g_n$ and each g_i can take one of the adjacent nodes of node i. Thus, a value of j assigned to the i-th gene, is then interpreted as a link between node i and j. In the resulting solution, they will be in the same community. The decoding of this representation requires the identification of all connected components. All nodes belong to the same connected component are then assigned to one community.

Genetic operation and initialization. The uniform two-point crossover is employed. In the mutation operation, NSGA-Net randomly selects some genes and assigns them with other randomly selected adjacent nodes. In the initialization process, we randomly generate some individuals. For each individual, each gene g_i randomly takes one of the adjacent nodes of node i.

Model selection. NSGA-Net returns a set of solutions, which provides Decision Makers (DMers) with more choices. Sometimes DMers may desire that the set of candidate solutions could be narrowed down to those of most interest. In this paper, we therefore propose the novel *Max-Min Distance* model section method to select one single recommendation solution from the Pareto front. Inspired by Gap statistic [4], the *Max-Min Distance* method selects the solution model that mostly deviates from the null models generated by NSGA-Net by running on random networks with the same distribution. Concretely, NSGA-Net firstly runs on the real network and randomly generated networks with the same scale. Thus the optimal solution set on the real network (called *CandSet*) and the corresponding random network (called *RandSet*) can be obtained, respectively. For each solution in *CandSet*, we calculate the minimum-distance with solutions in *RandSet*, and then we select the solution in *CandSet* with the maximum minimum-distance as the recommendation solution. Here, Euclidean distance is employed. Intuitively, the recommendation solution is the most different one from the solutions in *RandSet*.

Objective function. Still NSGA-Net has an important component unsolved: objective functions. As a general multi-objective community detection solution, NSGA-Net can apply any multiple objective functions. With NSGA-Net, we will examine the general performance of the multi-objective community detection. Furthermore, we explore what type of objectives is suitable for the multi-objective paradigm.

3. PERFORMANCES OF MULTI-OBJECTIVE COMMUNITY DETECTION

3.1 Objective Functions

Many objective functions have been proposed to capture the intuition of communities. We summarize 11 objective functions that are already widely used in community detection literatures or can be potentially used for community detection.

- **Conductance** (Q_1) measures the fraction of total edge volume that points outside the cluster [6].
- **Expansion** (Q_2) measures the number of edges per node that point outside the cluster [6].

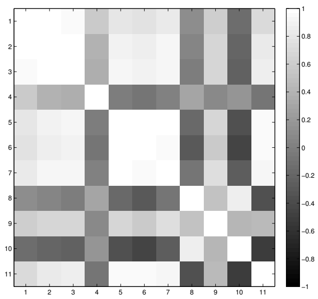


Figure 1: Pearson correlative coefficients of the 11 objectives. 1-11 represent the objective functions $O_1 - O_{11}$, respectively.

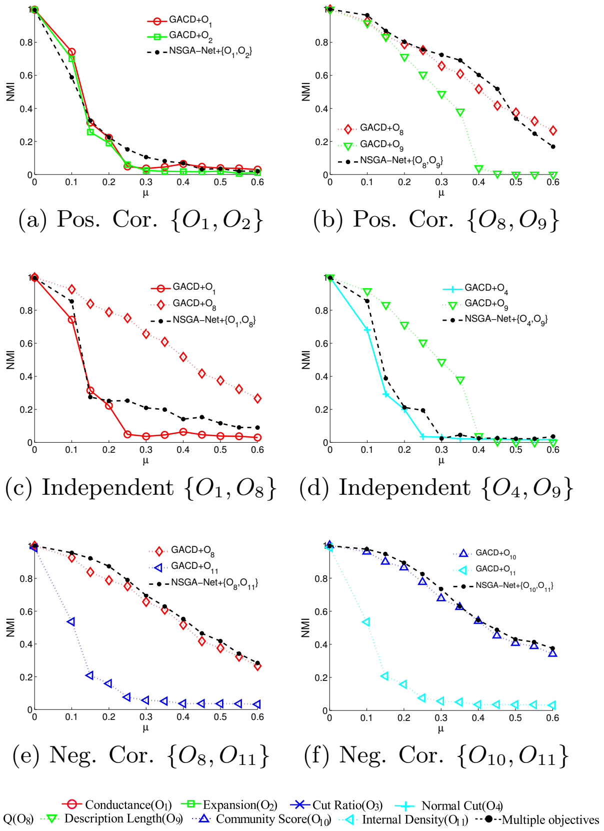
- **Cut Ratio** (Q_3) is the fraction of all possible edges leaving the cluster [6].
- **Normalized Cut** (Q_4) is the normalized fraction of edges leaving the cluster [6].
- **Maximum-ODF(Out Degree Fraction)** ((Q_5)) is the maximum fraction of edges of a node pointing outside the cluster [6].
- **Average-ODF** (Q_6) is the average fraction nodes' edges pointing outside the cluster [6].
- **Flake-ODF** (Q_7) is the fraction of nodes in S that have fewer edges pointing inside than to the outside of the cluster [6].
- **Q** (Q_8) measures the number of within-community edges, relative to a null model of a random graph with the same degree distribution [8].
- **Description Length** (Q_9) is the number of edges between the community i and j. The objective regards the community as a optimal compression of network's topology [7].
- **Community Score** (Q_{10}) measures the density of a sub-matrices based on volume and row/column means [9].
- **Internal Density** (Q_{11}) is the internal edge density of the cluster [6].

We roughly classify the objective functions into three categories. The first category contains the first four objectives from graph theory community, which are called the cut-based objectives. The three objectives ended with "ODF" are called the degree-based objectives. Finally, the remaining objectives are classified into one category. These objective functions come from different research fields, such as graph theory and physics. 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.

3.2 Objective Correlations

We can observe that the definitions of some objectives are similar, such as the cut-based objectives. Namely, these objectives are correlated. Here we apply the Pearson correlation coefficients to describe their relations. Because it is difficult to analyze their correlations from the definitions directly, we perform experiments to estimate the Pearson correlation coefficients. The experiments are done with the following steps. (1) For a given network, we generate a set of random partitions. (2) For each partition, we calculate the values of the different objective functions. Thus each objective function has a vector of random samples. (3) We estimate the Pearson correlation coefficients among these objective vectors. (4) In order to reduce the estimation variance, we repeat step 1 to 3 many times and get the average values.

The results are illustrated in Figure 1. We can observe that the cut-based objectives are highly correlated (especially $O_1 - O_3$). It is the same case for the degree-based objectives. In addition, we notice that *InternalDensity* is negatively correlated with Q and *CommunityScore*. The relations of these objectives can be roughly classified into three categories in terms of their correlation coefficients: positively correlated (e.g., $\{O_1, O_2, O_3, O_4\}$, $\{O_5, O_6, O_7\}$, $\{O_8, O_9, O_{10}\}$), independent (e.g., $\{O_1, O_8\}$, $\{O_1, O_{10}\}$, $\{O_4, O_9\}$), and negatively correlated (e.g., $\{O_8, O_{11}\}$, $\{O_{10}, O_{11}\}$).



(a) Pos. Cor. $\{O_1, O_2\}$ (b) Pos. Cor. $\{O_8, O_9\}$

(c) Independent $\{O_1, O_8\}$ (d) Independent $\{O_4, O_9\}$

(e) Neg. Cor. $\{O_8, O_{11}\}$ (f) Neg. Cor. $\{O_{10}, O_{11}\}$

Figure 2: The NMI comparison of NSGA-Net optimizing over three types of objective functions (i.e., positively correlated, independent, negatively correlated) and GACD optimizing over original single objectives on artificial networks. The larger the NMI, the better the performance.

3.3 Performance Effect of Objective Selection

In this section, we will test the performances of the multi-objective community detection method (i.e., NSGA-Net) and find what kind of objectives are suitable for the method. Here we only consider two objectives, rather than more objectives, in order to focus on the effectiveness of the multi-objective method and reduce the complexity. From each of the three categories of objective correlations, we select two pairs as the optimized objectives in NSGA-Net. Particularly, for the positively correlated objectives we choose $\{O_1, O_2\}$, $\{O_8, O_9\}$; the independent objectives, $\{O_1, O_8\}$, $\{O_4, O_9\}$; and the negatively correlated objectives, $\{O_8, O_{11}\}$, $\{O_{10}, O_{11}\}$. For the SOP: $\min_{C \in \Omega} O(C)$, we choose GACD [12] as the single-objective community detection optimizer. NSGA-Net is equipped with the same parameters with GACD for a fair comparison. That is, $popSize = gen = 200$; $croRat = 0.6$; and $mutRat = 0.4$.

We use a popular artificial network with a known commu-

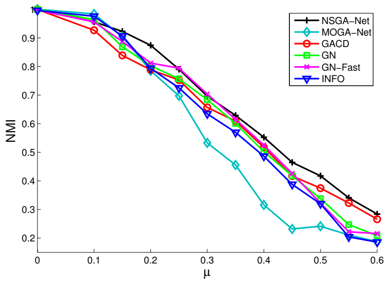


Figure 3: The comparison of NSGA-Net with a pair of negatively correlated objectives (i.e., O_8 and O_{11}) with other popular algorithms on artificial networks.

nity structure [5], which has the heterogeneity in the distributions of node degrees and community sizes. Same in ref. [5], each node in the benchmark graphs share a fraction $1-\mu$ of its links with the other nodes of its community and a fraction μ with the other nodes of the network. As μ increases, it becomes harder and harder to identify the community structure. To compare the built-in modular structure with the one delivered by different objectives, we adopt the Normalized Mutual Information (NMI), a measure of similarity of partitions borrowed from information theory [5].

We first run NSGA-Net on the artificial networks. The comparison results of NSGA-Net optimizing over six pairs of objectives and GACD optimizing over original single objectives are shown in Figure 2. When the optimized objectives are positively correlated or independent, NSGA-Net's performances have no obvious differences from the performances of the optimization on each single objective with GACD. Most results of NSGA-Net are between those of the single objectives. However, it is obvious that NSGA-Net with a pair of negatively correlated objectives has better performance than the optimization on the original single objective, since their NMI are larger than those of the single objective in most conditions.

3.4 Comparison with Other Algorithms

We further validate the performance of NSGA-Net through comparing it with representative community detection algorithms. NSGA-Net is equipped with a pair of negatively correlated objectives O_8 and O_{11}. Other five algorithms are included in the experiments. It includes the betweenness-based heuristic algorithm [8] (named GN) and its improved version [1] (named GN Fast). The EA-based optimization algorithm [12] (named GACD) optimizes the O_8. The information-theoretic framework based algorithm (named INFO) [7] optimizes the O_9. Another multi-objective method MOGA-Net [10] is also included. In order to obtain one single recommendation solution, MOGA-Net also employs the *Max-Min distance* model selection method. NSGA-Net and MOGA-Net are set as the same parameters with GACD. The benchmark is the same artificial networks as before.

The experimental results are shown in Figure 3. In most conditions, NSGA-Net obviously performs better than other five algorithms including not only the single-objective algorithms but also the multi-objective method, MOGA-Net. An important difference between NSGA-Net and MOGA-Net lies in the objective functions. We think the absence of the sufficient negative correlation between objectives in MOGA-Net causes its bad performances.

4. CONCLUSION

In this paper, we study the multi-objective community detection problem and propose a novel solution NSGA-Net. Aiming to exploit the universal validity of the multi-objective solution for community detection and its requisition on objective functions, we first analyze the intrinsic correlations among 11 objectives. Then we compare the performances of NSGA-Net optimizing over different types of objectives to those of a single-objective based approach optimizing over the original single objective. The experiments show that NSGA-Net only with a pair of negatively correlated objectives remarkably improve the performance.

5. ACKNOWLEDGMENTS

It is supported by the National Natural Science Foundation of China (No. 60905025, 61074128, 61035003) and the 863 Program of China No.2009AA04Z136.

6. REFERENCES

[1] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(06611), 2004.

[2] K. Deb, A. Pratab, S. Agarwal, and T. MeyArivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transaction on Evolutionary Computation*, 6(2):182–197, 2002.

[3] M. Girvan1 and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[4] J. Handle and J. Knowles. An evolutionary approach to multiobjective clustering. *Transaction on Evolutionary Computation*, 11(1):56–76, 2007.

[5] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(046110), 2008.

[6] J. Leskovec, K. J. Lang, and M. W. Mahoney. Empirical comparison of algorithms for network community detection. In *WWW2010*, pages 631–640, 2010.

[7] R. Martin and T. B. Carl. An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, 104(18):7327–7331, 2007.

[8] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physics Review E*, 69(026113), 2004.

[9] C. Pizzuti. Ga-net: a genetic algorithm for community detection in social networks. In *PPSN2008*, pages 1081–1090, 2008.

[10] C. Pizzuti. A multi-objective genetic algorithm for community detection in networks. In *ICTAI09*, pages 379–386, 2009.

[11] C. Shi, X. Kong, P. S. Yu, and B. Wang. Multi-label ensemble learning. In *ECML/PKDD 2011*, 2011.

[12] 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):3–17, 2010.