Shi C, Yan ZY, Pan X *et al.* A posteriori approach for community detection. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 26(5): 792–805 Sept. 2011. DOI 10.1007/s11390-011-0178-z

A Posteriori Approach for Community Detection

Chuan Shi[1] (石　川), *Member*, *CCF*, *IEEE*, Zhen-Yu Yan[2] (闫震宇), *Member*, *IEEE*, Xin Pan[1] (潘　欣)
Ya-Nan Cai[1] (蔡亚男), and Bin Wu[1] (吴　斌), *Member*, *CCF*

[1] *School of Computer, Beijing University of Posts and Telecommunications, Beijing 100876, China*

[2] *Research Department, Fair Isaac Corporation (FICO), San Rafael, CA, 94903, U.S.A.*

E-mail: shichuan@bupt.edu.cn; yan_zhen_yu@hotmail.com; panyx2006@qq.com; diandacainan@gmail.com
wubin@bupt.edu.cn

Received October 5, 2010; revised June 28, 2011.

Abstract Conventional community detection approaches in complex network are based on the optimization of a priori decision, i.e., a single quality function designed beforehand. This paper proposes a posteriori decision approach for community detection. The approach includes two phases: in the search phase, a special multi-objective evolutionary algorithm is designed to search for a set of tradeoff partitions that reveal the community structure at different scales in one run; in the decision phase, three model selection criteria and the Possibility Matrix method are proposed to aid decision makers to select the preferable solutions through differentiating the set of optimal solutions according to their qualities. The experiments in five synthetic and real social networks illustrate that, in one run, our method is able to obtain many candidate solutions, which effectively avoids the resolution limit existing in priori decision approaches. In addition, our method can discover more authentic and comprehensive community structures than those priori decision approaches.

Keywords complex network, community detection, multi-objective evolutionary algorithm, modularity

1 Introduction

Analysis of large complex networks, such as social network, World Wide Web, telecommunication network and biological network, have drawn great interest in various research communities. One of the key problems in the field is "how to describe/explain its community structure"[1]. This topic is important because these communities often play special roles in the network systems. Detecting communities (or modules) can be a way to identify substructures corresponding to important functions.

A loose definition of the community is the group of nodes that are densely interconnected but only sparely connected with the rest of the network[2-3]. Many methods and algorithms have been developed for community detection (CD)[4-5]. Most contemporary community detection algorithms choose a cost function that measures the quality of community partitions first, and then optimize this function through searching the solution space. For example, community detection with the modularity, a popular quality function proposed by Newman[1], is equivalent to a modularity optimization. Some other quality functions have also been proposed, such as the "cut" function in spectral method[6] and the "description length" in information theoretic-based method[7]. From the perspective of decision making, these algorithms can be regarded as a priori preference articulation, that is, a single objective function is designed beforehand and the algorithm returns a single solution as results.

Although these priori approaches achieve great successes in artificial and real networks, they have some fundamental drawbacks. These algorithms attempt to optimize just one quality function and this confines the solution to a particular community structure property. And thus, it often causes a fundamental discrepancy that different algorithms may produce distinct solutions for the same network. Moreover, these priori approaches have the resolution limit problem. Fortunato and Barthelemy[8] showed mathematically that the modularity optimization has a resolution limit, that is, modularity optimization fails to find small communities in large networks, which raises important concerns about the reliability of the modules detected using these techniques, or more broadly using any other single quality functions. In order to avoid the resolution limit existing in the modularity optimization, some other quantitative measures have also been proposed, for example, the Hamiltonian-based method introduced

Regular Paper

Supported by the National Natural Science Foundation of China under Grant Nos. 60905025, 61074128, 61035003, the National High Technology Research and Development 863 Program of China under Grant No. 2009AA04Z136.

by Reichardt and Bornholdt (RB)[9] and a multiple resolution procedure proposed by Arenas, Fernandez and Gomez (AFG)[10]. However, these methods still have two disadvantages: high time complexities due to many runs through tuning the parameters, and the similar resolution limit due to a single objective used[11]. In addition, many contemporary algorithms require priori information: the number of communities, which is usually unknown for real networks. Last but not the least, a single fixed community partition returned by most contemporary algorithms may not be suitable for the networks with multiple potential structures. For example, a fixed community partition cannot reveal the hierarchical or overlapping structures. However, the overlapping and hierarchical structures[12] are pervasive in real networks[12].

Generally, there are three decision making diagrams when multiple objectives present: priori, posteriori, and progressive[13]. The conventional community detection approaches fall in the priori decision making category, that is, the Decision Maker (DMer) firstly designs an objective function that captures the notion of community, and then optimizes the objective. A single solution is usually returned. The progressive methods usually make a decision during the optimization process. In the posteriori decision approach, the DMer is usually presented with a set of optimal candidate solutions obtained through searching the solution space and then the DMer makes a decision to choose the proper solutions from that set. With the posteriori approach, a community detection algorithm returns a set of solutions that contain community partitions with different sizes, which may avoid the disadvantages existing in the priori approaches. Moreover, the real social networks usually are complex and uncertain. They not only have the complex hierarchical or overlapping structures, but also may evolve with time or other factors. As a consequence, a single community partition returned by the priori approaches hardly discovers the real dynamic community structure and it is hard to control in progressive approaches. In contrast, the posteriori approaches are easy to control and they return a set of optimal solutions with the structural and functional information in different angles, which provides DMers more choices to select proper models and analyze the internal implicit structures.

After a review of the related work, this paper proposes a posteriori decision approach for CD. As a powerful posteriori decision approach, the multi-objective evolutionary algorithm is applied to detect the community structure in this paper. The approach includes two phases. In the first phase, a special multi-objective evolutionary algorithm (MOEA) is designed to generate a set of optimal solutions. Then, in the second phase, we propose three model selection methods and the Possibility Matrix to differentiate the set of optimal solutions, which assists the DMers to select the preferable ones from them. The method will be validated not only in the synthetic hierarchical, overlapping and random networks but also two social networks (i.e., Karate network and coauthorship network).

2 Related Work

Many different algorithms have been designed to analyze the community structure in complex networks. The algorithms use methods and principles of physics, artificial intelligence, graph theory and even electrical circuits[5]. One of the most known algorithms proposed so far is the Girvan-Newman (GN) algorithm that introduces a divisive method by iteratively cutting the edge with the greatest betweenness value[1]. Some improved algorithms have been proposed[14-15]. These algorithms are based on a foundational measure criterion of community, modularity, proposed by Newman[1]. The larger the value is, the more accurate the community partition would be. As a consequence, the community detection becomes a modularity optimization problem. Because the search for the optimal (largest) modularity value is an NP-complete problem[16], many heuristic search algorithms have been applied to solve the optimization problem, such as extremal optimization[17], simulated annealing[3] and genetic algorithm[18].

Some other criteria are also proposed as the optimization objective. The Hamiltonian-based method introduced by Reichardt and Bornholdt (RB)[9] is based on considering the community indices of nodes as spins in a q-state Potts model. Recently, Arenas, Fernandez and Gomez (AFG)[10] proposed a multiple resolution procedure that allows the modularity optimization to go deep into the structure. These methods vary the thresholds by using a tuning parameter in their criteria and investigate the community structure at various resolutions. The modularity can be regarded as a special case of these two criteria. In addition, Fosvall and Bergstrom[7] proposed an information-theoretic formulation for the concept of modularity, in which community structures are an optimal compression of its topology. Although these criteria could effectively assess the quality of communities, the recent research shows that the optimization based on a single criterion (i.e., the priori approaches) has a fundamental disadvantage. Fortunato and Barthelemy[8] found that the modularity optimization may fail to identify modules smaller than a scale which depends on the total size of the network and on the degree of interconnectedness of the modules, even in the cases where modules are unambiguously defined. Kumpula *et al.*[11] further discussed a similar limited resolution when a global energy-like quantity is optimized, for example, the former two criteria (RB[9] and AFG[10]).

Recently, several researches pointed out that real networks have hierarchical structures with ubiquitous overlapping communities. Shen *et al.*[19] proposed novel covariance and correlation matrix to detect multi-scale community structures based on dimension reduction. Lancichinetti[20] *et al.* proposed a local optimization method of a fitness function, in which the hierarchical and overlapping community structures are revealed by peaks in the fitness function and parameter tuning. The maximal clique and an extension of modularity into the overlapping scenario are also applied to detect the overlapping and hierarchical communities[21-22]. These approaches usually optimize a priori objective function, and thus they can be considered as priori approaches.

The genetic algorithm (GA), as an effective optimization technique, has also been used for community detection. In order to optimize the modularity, the GAs in [18, 23] use the cluster centers and the locus-based adjacency as the encoding scheme, respectively. Pizzuti proposed another GA to optimize the "community score" criterion[24-25]. These algorithms have the advantage that the number of communities can be automatically determined during the evolutionary process. However, they also have the resolution limit, since a single objective is optimized. More recently, some researchers regard the CD as a multi-objective optimization problem (MOP) and solve the MOP with MOEAs[26-27]. Pizzuti[27] proposed the MOGA-Net to optimize the community score and community fitness. Shi *et al.*[26] proposed an MOEA to optimize two components of modularity Q. These two multi-objective methods show their advantages in detecting more accurately community structures. However, they do not further explore the benefits and properties of the multiple Pareto optimal solutions returned by the multi-objective methods.

3 Posterior Decision for Community Detection

In this section, we propose a posteriori approach: multi-objective evolutionary algorithm for community detection (MOCD). The approach consists of two phases. The first community detection phase applies a special MOEA to discover communities, and returns a set of optimal solutions. The second model selection phase proposes three community selection criteria and a Possibility Matrix method to assist DMer's decision making.

The rationality of applying MOEA to CD is as follows. Firstly, CD can be regarded as a special case of clustering problems, because it has the similar definition with clustering. The concept of a cluster is a generalization of what humans perceive, as densely connected "patches" within data space, whereas a human intuition is inherently difficult to capture by means of single objective[28]. As a consequence, the single-objective solution may not holistically reveal the intrinsic structure. In addition, some researchers have also been aware that enumerating the modules in a network is a tradeoff among multi-objectives[7-8,16]. For example, Fortunato *et al.*[8] believed that finding the maximum modularity is a tradeoff between the number of modules and the value of each term, and Rosvall and Bergstrom[7] also thought that enumerating the modules in a network has an inevitable tradeoff between the amount of the structure information of a network and its description length. Thus defining CD as an MOP reflects the inherent characteristics of CD. On the other hand, Evolutionary Algorithm (EA) becomes an increasingly popular approach for solving MOPs and many MOEAs have been suggested[29]. This is because the MOPs usually have no single optimal solution, which makes EA returning a set of promising solutions preferable to an algorithm returning only one solution based on some weighting of the objectives. The MOEAs use Pareto dominance to guide the search, and return a set of non-dominated solutions (i.e., Pareto optimal solutions) as results. As a successful posteriori decision approach, MOEA has the potential to avoid the disadvantages existing in those priori approaches, and thus this paper employs MOEA as the basic algorithm framework.

3.1 Community Detection Phase

This subsection designs a special MOEA suitable for CD. Although there are many successful MOEAs, when they are applied to CD, many components need to be redesigned according to CD's characteristics. These components include objective functions, genetic representation, operators, etc.

3.1.1 Algorithm Framework

This paper selects a well-known MOEA, NSGA-II[30], to form the basis of MOCD's community discovery phase. NSGA-II transforms the M objectives to a single fitness assignment by the creation of a number of fronts, sorted according to non-domination. During the fitness assignment, the individuals are divided into different fronts according to their dominating relationships. After each front has been created, its members are assigned crowded distance to be used later for niching. In each generation, N new individuals are generated, where N is the population size. The N best individuals are selected for the next generation from the combination of the new-generated individuals and the individuals in the current generation. In this way, a huge elite set can be kept from generation to generation. The main framework of the algorithm is illustrated in Algorithm 1. The detailed implementation can be seen in [30].

Algorithm 1. Main Framework of MOCD

```
procedure
  generate $P_0$ at random
  set $P_0 = (F_1, F_2, \ldots) = \textit{non-dominated-sort}(P_0)$
  for all $F_i \in P_0$ do
    $\textit{crowding-distance-assignment}(F_i)$
  end for
  set $t = 0$
  while (not done) do
    generate child population $Q_t$ from $P_t$
    set $R_t = P_t \cup Q_t$
    set $P_0 = (F_1, F_2, \ldots) = \textit{non-dominated-sort}(P_0)$
    set $P_{t+1} = \varPhi$
    set $i = 1$
    while $(|P_{t+1}| + |F_i| < N)$ do
      $\textit{crowding-distance-assignment}(F_i)$
      set $P_{t+1} = P_{t+1} \cup F_i$
      set $i = i + 1$
    end while
    sort $F_i$ on crowding distances
    $P_{t+1} = P_{t+1} \cup F_i[1 : (N - |P_t + 1|)]$
    set $t = t + 1$
  end while
  return $F_1$
end procedure
```

Algorithm 1 is the main framework. To apply NSGA-II to the community detection problem, there is much work to do. Two or more objective functions should be determined according to the characteristics of CD. Moreover, a community partition should be encoded through a genetic representation, and the corresponding genetic variation operators need to be designed. These choices are nontrivial and are crucial to the performance and particularly to the scalability of the algorithm. The design of an effective EA for CD requires a close harmonization of the encoding, the operators, and the objective functions, and so that the effective search space can be reduced and the search can be effectively guided. Our choices for these components, determined after extensive experimentation, are described next.

3.1.2 Objective Functions

Objective function (i.e., fitness function), which guides the search process, is one of the most important components in MOEAs. The objective functions quantify the optimality of a solution, so we should select optimization objectives that reflect the fundamentally different aspects of a good community partition. Modularity is a foundational quality index for CD. Given a simple graph $G = (V, E)$, we have the following definition[1]:

$$Q(C) = \sum_{c \in C} \Big[\frac{E(c)}{m} - \Big(\frac{\sum_{v \in c} \deg(v)}{2m} \Big)^2 \Big], \qquad (1)$$

where the sum is over the modules of the partition, $|E(c)|$ is the number of links inside module c, m is the total number of links in the network, C is a partition result, and $deg(v)$ is the degree of the node v in module c (it is same in the following sections). According to the definition, in order to maximize the modularity Q, we should maximize the first term, while minimizing the second term. Maximizing the first term increases the number of edges contained within clusters (i.e., "densely interconnected"). Minimizing the second term tends to split the graph into many clusters with small total degrees each (i.e., "sparely connected with the rest"). These two complementary terms reflect two fundamental aspects of a good partition. The modularity is an intrinsic trade-off between these two objectives.

In this paper, we select these two terms as the objective functions. In order to formulate the problem as a minimal optimization problem, we revise the first term. The first objective function minimizes 1 minus the inter-link strength of a partition, and it is called *inter* objective.

$$\textit{inter}(C) = 1 - \sum_{c \in C} \frac{|E(c)|}{m}. \qquad (2)$$

The second objective function minimizes the intra-link strength of a partition, and it is called *intra* objective.

$$\textit{intra}(C) = \sum_{c \in C} \Big(\frac{\sum_{v \in C} deg(v)}{2m} \Big)^2. \qquad (3)$$

According to the two definitions, we found that

$$Q(C) = 1 - \textit{inter}(C) - \textit{intra}(C). \qquad (4)$$

The important reason in the choice of these objective functions is their potential to balance each other's tendency to increase or decrease the number of communities, enabling the use of a representation that does not fix the number of communities (k). While the value associated with the *inter* objective necessarily improves with an increasing number of communities, the opposite is the case for *intra* objective. The conflict of the two objective functions tradeoffs the two objectives during the optimization, keeps the number of communities dynamically and avoids the convergence of trivial solutions (the detailed analysis can be seen in [29]). More objective functions can be used. However, our experiments indicate that the additional objective functions do not necessarily lead to better solutions, but may result in some practical difficulties, such as the larger search space and more candidate solutions. We will explore more potential objective functions in the further work.

3.1.3 Genetic Representation

The biological and social complex networks are usually represented as graphs consisting of nodes and links, and then the communities to be detected are groups of nodes. When the EA is applied to CD, a community partition needs to be encoded in a character string (i.e., genotype) with the genetic representation, and inversely a genotype (i.e., a solution of the problem or an individual in the population) can also be decoded into a community partition. This paper employs the locus-based adjacency representation[30] illustrated in Fig.1. In this graph-based representation, each genotype g consists of n genes $g_1, g_2, \ldots, g_n$ and each 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 nodes 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 that belong to the same connected component are then assigned to one community. Using a simple backtracking scheme, this decoding step can be performed in linear time[23].

According to the genetic structure, we found that the encoding approach cannot represent all partitions of nodes. For example, the genetic representation cannot combine two disconnected subgraphs into one community, and thus one may argue that the solutions with a good community structure may not be in the solution space constructed by the genetic representation. Recently, Brandes *et al.*[16] have analyzed the basic structural properties of the clustering with maximum modularity and proposed that "a clustering of maximum modularity does not include disconnected clusters". Although the modularity optimization has the resolution limit[8], the community partition with a large modularity usually is a good solution. Because the genetic representation contains all varieties of connected subgraphs, these properties promise that the community with a good structure can be represented with the genetic representation.

The locus-based adjacency encoding scheme is suitable for CD due to the following advantages. Most importantly, there is no need to fix the number of communities in advance, as it is automatically determined in the decoding step. Many methods need some priori knowledge such as the number of communities or threshold settled, whereas MOCD does not require any priori information. Another important advantage of this scheme is that the search space constructed by the representation is reduced significantly. In the former GA-based method[18], Tasgin and Bingol used a number ranging from 1 to n to represent the community a node belongs to, which results in a search space with the complexity $O(n^n)$. Brandes *et al.*[16] cast the problem of maximizing modularity into an integer linear program (ILP) with complexity $O(n^{2^n})$ in the search space. The complexity of the search space constructed by the locus-based adjacency representation is $O(d^n)$ (d is the degree of nodes). The complexity is much smaller than that of the other representations, because d is much smaller than n for most real problems. With the reduced search space, MOCD can obtain more accurate

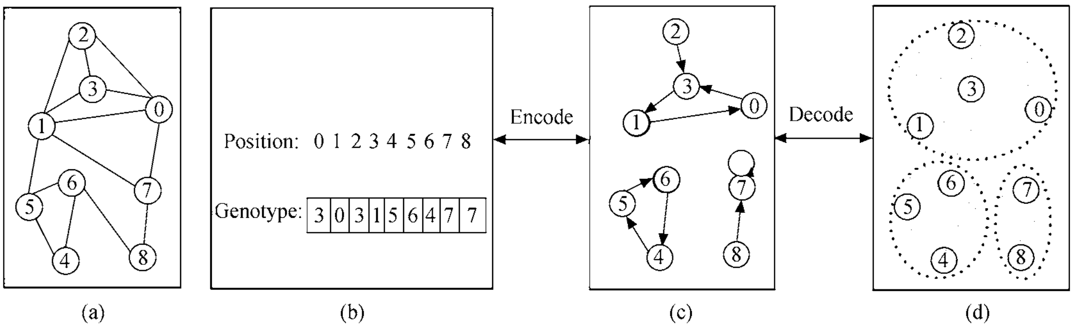

Fig.1. Illustration of the locus-based adjacency representation. (a) The topologic of the graph representing a complex network. (b) One possible genotype. (c) How (b) is translated into the graph structure, for example node 0 links to node 3, since gene g_0 is 3. (d) The partition result.

solutions in less time. Furthermore, this representation is suitable for the standard crossover operators such as uniform one-point or two-point crossover. A genetic algorithm with the locus-based adjacency encoding scheme in [23, 31] (called GACD) have validated the effectiveness of the encoding scheme. For most benchmark problems, GACD finds the maximal Q values, whereas its running time is much shorter than that of GN and Tasgin and Bingol's GA[23].

3.1.4 Operators and Initialization

Based on the locus-based adjacency representation, the crossover operation in MOCD is done by intersecting two chromosomes randomly selected from the population. For simplicity, the two chromosomes are called source and destination, respectively. Firstly, a gene is selected randomly from the source chromosome, and then we iteratively search for the gene values that the gene links to, and transfer these values in source chromosome to the corresponding genes in the destination chromosome. The exchange of gene segments is bidirectional. The crossover operator is prone to replicate the good structures generated by evolution to the new individual. Moreover, it is able to effectively generate the individual with different structures. The operator's computational complexity is $O(l)$ (where l is the length of the gene segment, namely the size of the community selected) l is usually smaller than n.

In the mutation operation, we randomly select some genes and replace them with other randomly selected adjacent nodes.

In the initialization, we randomly generate individuals with the predefined number *size* (see Line 2 of Algorithm 1). For each individual, each gene g_i randomly takes one of its adjacent nodes.

In Line 9 of Algorithm 1, two parameters λ_{cro} and λ_{mut} are used to control the ratio of crossover and mutation operations, respectively, and $\lambda_{cro} + \lambda_{cro} = 1$. In our implementation, the halting criterion (see Line 8 of Algorithm 1) is that the running generation is equal to a predefined value *gen*.

3.2 Model Selection Phase

MOCD does not return a single solution, but a set of Pareto optimal solutions. These community partitions correspond to different tradeoffs between the two objectives and also consist of communities of different sizes. Domain expertise can be leveraged to make the final decision through analyzing the alternative solutions. This is crucial to a problem with unknown structure, like CD. In addition, the DMer may desire that the set of candidate solutions can be further narrowed or some representative ones can be recommended. In this subsection, we therefore propose some methods to evaluate the quality of the Pareto optimal partition solutions. These methods are able to further identify some promising partitions from the optimal solutions.

Formally, let *CSet* be the set of community partitions (i.e., the optimal solution set returned by MOCD), C be a partition in *CSet*, and there are k communities in the partition C: $C = c_1 \cup c_2 \cup \cdots \cup c_k$. A partition result is also called a clustering model M.

Maximum Q Criterion. The criterion selects the model with maximum modularity Q. Because of the relationship of Q and two objective functions (see (4)), it is easy to select the model with maximum Q, and the corresponding model is called M_Q.

$$M_Q = \underset{C \in CSset}{\mathrm{argmax}}\{1 - \mathit{inter}(C) - \mathit{intra}(C)\}. \tag{5}$$

Strong Community Criterion. According to the strong community definition given by Radicchi *et al.*[14], each node i in each community c is validated whether to satisfy the strong definition. If the ratio of communities satisfying the strong definition is larger than the predefining threshold λ_{str}, the corresponding partition result is called strong partition, and the set comprising all the strong partitions is called *StrMSet*. ($k_i^{in}(c)$ is the number of edges connecting node to other nodes belonging to c. $k_i^{out}(c)$ is the number of connections toward nodes in the rest of the network.)

$$\begin{aligned} &\mathit{StrCSet} = \{c \,|\, k_i^{in}(c) > k_i^{out}(c), \quad \forall i \in c\}, \\ &\mathit{StrRatio}(C) = \frac{|\mathit{StrCSet}|}{|C|}, \\ &\mathit{StrMSet}(C) = \{C | \mathit{StrRadio}(C) > \lambda_{str}\}. \end{aligned} \tag{6}$$

According to the definition, we can find that *StrMSet* contains those good partitions in which the ratio of strong communities is larger than λ_{str}.

Weak Community Criterion. Similarly, according to the weak community definition[14], for each partition result, each community could be verified whether to satisfy the weak definition. If the ratio of communities satisfying the weak definition is larger than the predefining threshold λ_{weak}, the corresponding partition result is called weak partition, and the set comprising all the weak partitions is called *WeakMSet.*

$$\begin{aligned} &\mathit{WeakCSet} = \Big\{c \,\Big|\, \sum_{i \in c} k_i^{in}(c) > \sum_{i \in c} k_i^{out}(c)\Big\}, \\ &\mathit{WeakRadio}(C) = \frac{|\mathit{WeakCSet}|}{|C|}, \\ &\mathit{WeakMSet} = \{C | \mathit{WeakRadio}(C) > \lambda_{weak}\}. \end{aligned} \tag{7}$$

The definition shows that *WeakMSet* includes those partitions whose weak community ratio is larger than λ_{weak}.

In the definitions, two parameters λ_{str} and λ_{weak} need to be settled in the range from 0 to 1 beforehand, and they control the size of *StrMSet* and *WeakMSet*. If the networks have obvious community structures, these two parameters are settled with large values, otherwise with small values. These three criteria reflect the quality of solutions from different perspectives. The maximum Q criterion recommends an optimal solution M_Q in terms of Q to DMers. When $\lambda_{str} = \lambda_{weak}$, $StrMSet \subseteq WeakMSet$. And solutions in *StrMSet* have more obvious community structure than those in *WeakMSet*, because the definition of strong community is more restrictive than that of weak community.

In order to illustrate the statistical characteristics of multi-solutions, a Possibility Matrix is proposed to describe the probability that a pair of nodes belong to the same community.

Possibility Matrix. The rows and columns of the matrix correspond to the indices of nodes. For a partition solution, if two nodes are in the same community, the corresponding matrix value is 1, or else it is 0. For multi-solutions, the single Possibility Matrix of each solution is added as an accumulated Possibility Matrix. The matrix can be converted to a gray graph in which higher value corresponds to darker gray.

4 Experiments

We validated the effectiveness of MOCD through three synthetic networks and two real social networks. The experiments were carried out on a 3 GHz and 1 GB RAM computer running Windows XP.

4.1 Synthetic Network

4.1.1 *Hierarchical Network*

The hierarchical network is a K40-4 network consisting of a ring of cliques, connected through single link. The network has 40 cliques, and each clique is a complete graph with 4 nodes and 6 links. In the network, it is clear that there are 40 unit communities, and the connected cliques can also be considered as a community. The network has been used by Fortunato and Barthelemy to research the resolution limit in optimization of modularity[8].

We ran MOCD with the following parameters: the population size was 100, the running generation was 100, the crossover ratio was 0.6, and the mutation ratio was 0.4, λ_{str} and λ_{weak} both were 1. We also ran two popular priori approaches on the network: the betweenness-based heuristic algorithm GN[1] and the GA-based modularity optimization algorithm GACD[23]. Note that GACD has the same parameters with MOCD. In this experiment, the running times of MOCD, GN, and GACD are 26, 41, and 21 seconds, respectively.

GN obtained a solution with 16 communities, and GACD reached the maximal Q value 0.881 with 15 communities. In both solutions, some connected cliques were combined. According to the construction process, these two solutions can both be regarded as the correct partitions. However, they both failed to reveal the hierarchical characteristic of the network. MOCD obtained 100 non-dominated solutions which are illustrated in Fig.2(a). Please note that the *inter* and *intra* values are normalized (it is the same in the following section). There are 78 correct partitions in *StrCSet* with the number of communities from 26 to 40. There are two special models in these solutions. As illustrated in Fig.2(b), the model M_Q (labeled I in Fig.2(a)) reveals the 26 communities with the highest granularity. Another special strong community solution (labeled II), shown in Fig.2(c), reveals all 40 cliques with the lowest granularity. Most solutions lie between these two solutions. The accumulated Possibility Matrix of all solutions are illustrated in Fig.2(d). We can clearly find the hierarchical structure in which some large communities

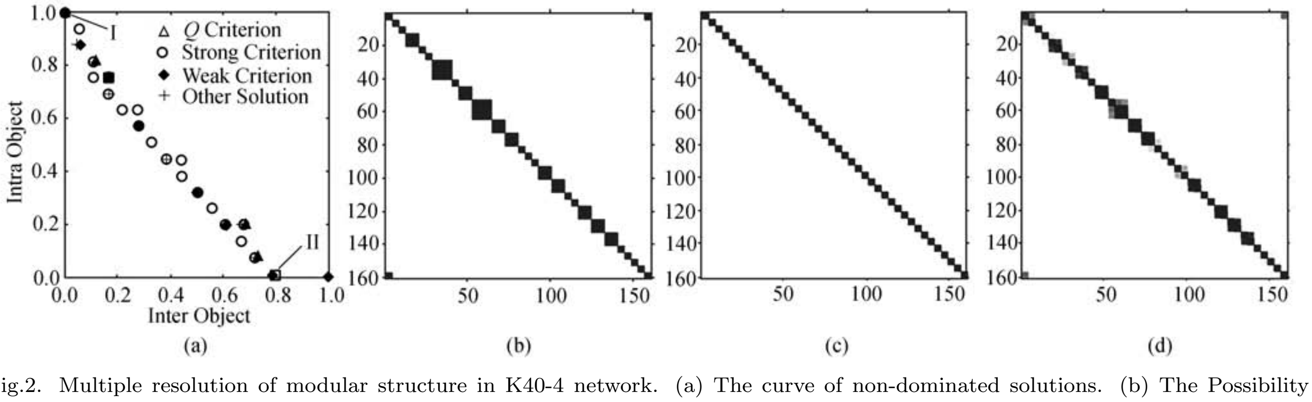

Fig.2. Multiple resolution of modular structure in K40-4 network. (a) The curve of non-dominated solutions. (b) The Possibility Matrix of the solution labeled with I. (c) The Possibility Matrix of the solution labeled with II. (d) The accumulated Possibility Matrix of all solutions.

may contain some connected cliques. Compared to one solution with the higher granularity returned by GN and GACD, MOCD can find the communities with different scales in one run, which reveals more structural information.

Using the experimental data, we analyzed the relationship of the objective values and the number of communities as shown in Fig.3(a). It is obvious that with the increase of the number of communities, the *inter* values increase, whereas the *intra* values decrease. It validates that the two objective functions are conflicting and complementary and the modularity Q is a trade-off between these two objectives. As for the Q value, it seems to decrease with the increase of the number of communities. In order to observe their relationship more clearly, the relationship of the number of communities and Q values of solutions in *StrMSet* is shown in Fig.3(b). It is clear that with the increase of the number of communities the Q value trends to become small. As the experiments illustrated, the priori approaches (e.g., GN and GACD) could only reveal the communities with large sizes. In fact, all the community partitions with small sizes discovered by MOCD are also correct. The experiment further confirms the resolution limit in the priori approaches with single objective[8]: methods based on optimizing the modularity measure or other single criterion may fail to identify modules smaller than some thresholds. Compared with those priori approaches, MOCD can discover the hierarchical network with different scales (i.e., both small and large sizes).

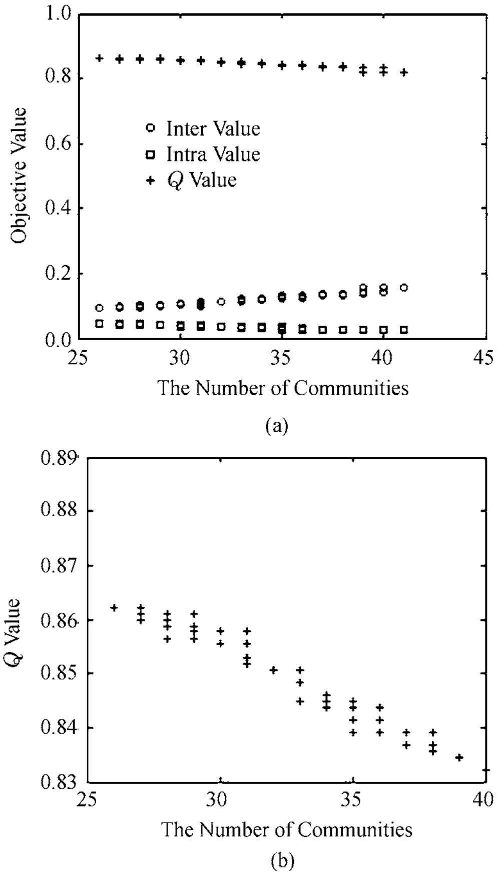

Fig.3. The relationship of the number of communities and the objective values. (a) The relationship on the optimal solution set. (b) The relationship on the strong community set.

4.1.2 Overlapping Network

The second experiment was on an overlapping network. The network consists of two large communities A and B, each containing 128 nodes, which have on average 12 internal links per node. Within A and B, a subgroup of 32 nodes exist, which we denoted by a and b, respectively. Every node within this subgroup had six of its 12 intra community links with the 31 other members of this subgroup. The two subgroups a and b had on average three links per node with each other. Additionally, every node had one link with randomly chosen nodes from the network. It is clear that the network has two large communities (i.e., A and B) and one overlapping community $a\&b$ between A and B. The similar network has been used by Reichardt and Bornholdt to discover the overlapping network[9].

MOCD was settled with the following parameters: the population size was 200, the running generation was 500, the crossover ratio was 0.6, and the mutation ratio was 0.4, λ_{str} was 0.3 and λ_{weak} was 0.5. GN and GACD were also run on this network, and GACD was equipped with the same parameters in MOCD. The running time of MOCD, GN and GACD were 214, 312, and 198 seconds, respectively.

GN and GACD both revealed the large communities A and B accurately. However, they were not able to discover the overlapping structure. MOCD obtained 200 non-dominated solutions which are illustrated in Fig.4(a). All the solutions were divided into four types: 1 solution for M_Q, 9 solutions in *StrCSet*, 22 solutions in *WeakCSet*, and 165 other solutions. We also found two special clustering models in the figure. The model M_Q (labeled with I in Fig.4(a)) reveals the same community structure as that of GN and GACD (i.e., two large communities A and B as illustrated in Fig.4(b)). Another special partition was a strong community solution (labeled with II) as shown in Fig.4(c). The partition consisted of three communities: two large communities, an overlapping community that was constituted by the nodes in a and b. The result shows that MOCD not only finds the obvious large community structure, but also reveals the implicit overlapping community in one run.

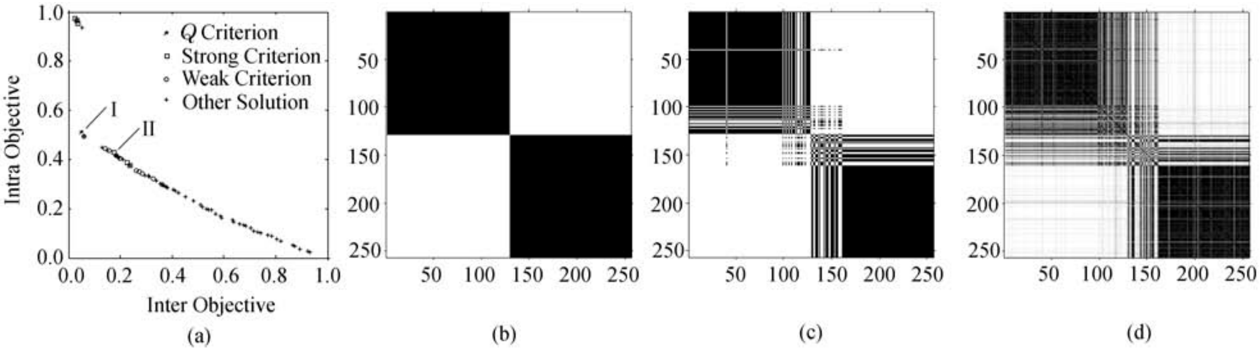

Fig.4. Multiple resolutions of modular structure in the overlapping network. (a) The curve of non-dominated solutions. (b) The Possibility Matrix of the solution labeled with I with a gray graph. (c) The Possibility Matrix of the solution labeled with II. (d) The accumulated Possibility Matrix of all the solutions.

The overlapping community can be easily identified through an aggregation of all the solutions obtained from MOCD. Fig.4(d) shows the accumulated Possibility Matrix of all the solutions of MOCD. We can see that an overlapping community which spans from node 98 to node 160 lies between the two large communities. A single solution obtained by any priori approach, such as GN or GACD, can hardly discover the overlapping structures. Whereas, the accumulated Possibility Matrix can easily reveal it with aggregation of all the optimal solutions obtained by MOCD. In [9], Reichardt and Bornholdt also found the two partitions in Figs. 4(b) and 4(c) at $\gamma = 0.5$ and $\gamma = 1$, respectively. However, in order to discover the correct partition, many runs should be done to find the proper γ. Compared with their method, MOCD obtains many partitions including the correct partitions in one run and the accumulated Possibility Matrix is able to statistically reveal the hidden but informative structure.

4.1.3 Random Network

In order to further validate the performance of MOCD, we compared MOCD with four popular algorithms on a set of random networks with known structures. Because conventional priori methods only return a single solution, here we only use one single solution with the maximum Q selected from the solutions set returned by MOCD. We name the solution MOCD-Q. The baseline methods include: 1) GN[1]: the betweenness-based heuristic algorithm; 2) GN Fast[15]: the improved version of GN; 3) GACD[23]: the GA-based modularity optimization algorithm; and 4) INFO[7]: the information-theoretic framework based algorithm. The random networks are the newly proposed benchmark graphs[32] that account for the heterogeneity in the distributions of node degrees and community sizes. As suggested in [32], the benchmark graphs are set as follows: the number of nodes is $N = 1500$; the average degree is $k = 25$ and the maximum degree is not more than 80; the degree and the community size distributions are power laws, with exponents $\gamma = 2$ and $\beta = 2$, respectively. μ is the mixing parameter that controls the fraction of the links of a node with the other nodes outside the community of the node. As μ increases, it becomes harder and harder to identify the community structure. The Normalized Mutual Information[32] (NMI) is used to evaluate the performance.

The parameters in MOCD-Q and GACD are the same: the population size is 200, the running generation is 200, the crossover ratio is 0.6, and the mutation ratio is 0.4. The experimental results are shown in Fig.5. It is clear that MOCD-Q reveals the most accurate community structures for most networks compared with other methods, and the superior of MOCD-Q even becomes more obvious when the community structure becomes fuzzier.

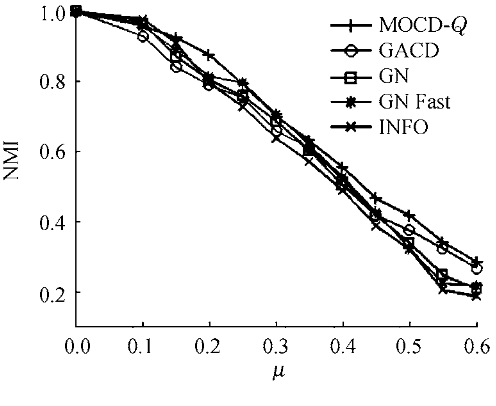

Fig.5. Comparison of MOCD with other methods on the random networks with known structure.

4.2 Social Network

We now turn to two real world examples to see whether these structural properties can indeed be found in real networks.

4.2.1 Karate Network

The famous Karate club network analyzed by Zachary is widely used as a benchmark to test the community detection methods[1,17,23]. The network consists of 34 members of a Karate club with nodes and 78 edges representing friendship between members of the club which was observed over a period of two years. Due to a disagreement between the club's administrator and its instructor, the club split into two groups. The question we concern is that if we can detect the real groups.

The following parameters were used in MOCD: the population size was 50, the running generation was 100, the crossover ratio was 0.6, and the mutation ratio was 0.4, λ_{str} was 0.5 and λ_{weak} was 0.7. GN and GACD were also ran on this network. GN found five communities which are distinguished with the color of the interior of nodes in Fig.6(b). GACD divided the network into 4 groups with the maximal Q value 0.419, which are distinguished with the shape of nodes. They both fail to find the correct partition. Fig.6(a) illustrates the 50 non-dominated solutions returned by MOCD. The number of communities of those solutions ranges from 1 to 6. Note that the 50 solutions returned by MOCD actually include the partition results returned by GN and GACD. We labeled these two solutions with III and I, respectively. Moreover, MOCD successfully revealed the real partition which was denoted by label II in Fig.6(a). In all, MOCD not only found the community structures discovered by GN and GACD, but also revealed the true structure.

4.2.2 Coauthorship Network

The real coauthorship network with 2 122 nodes and 5 678 edges reflects the coauthorships of the Beijing University of Posts and Telecommunications (BUPT)[33]. The nodes and edges represent authors and coauthorship relations, respectively. This network was used because the real community partition is known to authors, and thus it is easy to validate the partition results. MOCD was settled with the following parameters: the population size was 200, the running generation was 800, the crossover ratio was 0.6, the mutation ratio was 0.4, λ_{str} was 0.2 and λ_{weak} was 0.5. The running time was 550 seconds. GN was also run on this network and it obtained a partition result in 2 109 seconds. As shown in Fig.7(a), GN obtained one solution that partitioned each node into a certain community. MOCD obtained 200 solutions and all the solutions were accumulated to form the Possibility Matrix shown in Fig.8(a). Note that, in order to demonstrate the community structure clearly, the order of nodes in the Possibility Matrix is sorted by their order in the partition result of the model M_Q.

From the graph, we can observe that MOCD reveals the obvious community structures with the lower granularity and some nodes are not definitely clustered as a community. These communities can be roughly categorized into three categories: 1) coherent structure; 2) hierarchical structure; and 3) overlapping structure. In order to analyze their practical meanings, we select three representative communities from these three categories, respectively and they are shown from Fig.8(b) to Fig.8(d). To compare the qualities of the results of MOCD and GN, we also select three corresponding communities from Fig.7(a). In other words, the three pairs of communities in Fig.8(a) and Fig.7(a) have the same core members. Fig.8(b) displays a coherent community, which can hardly be further divided. The actual situation is consistent with the experimental result.

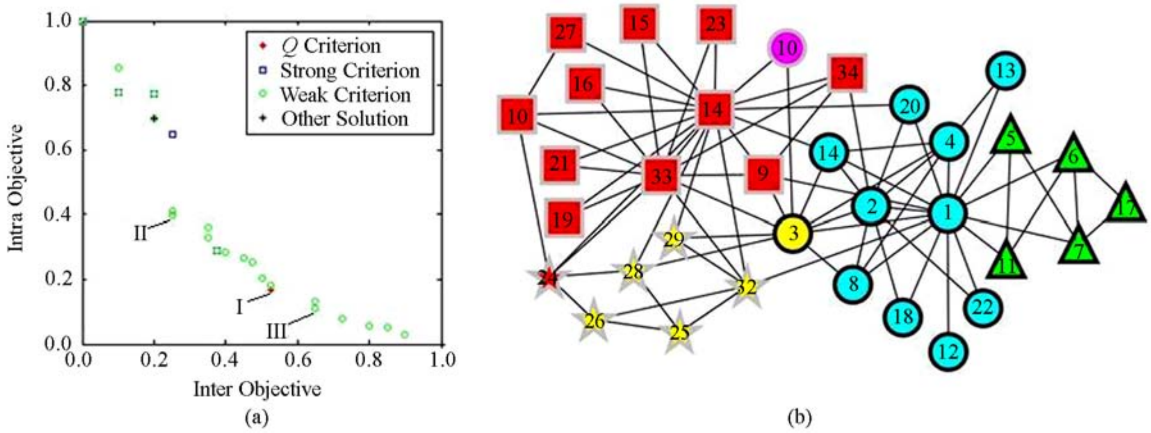

Fig.6. Multiple solutions of modular structure in Karate network. (a) The curve of non-dominated solutions. (b) Three different partitions. The differences of partitions are made by the color of the boundary of nodes, the color of the interior of nodes, and the shape of nodes, which correspond to the results of real partition & the solution labeled with II, GN & the solution labeled with III, and GACD & the solution labeled with I, respectively.

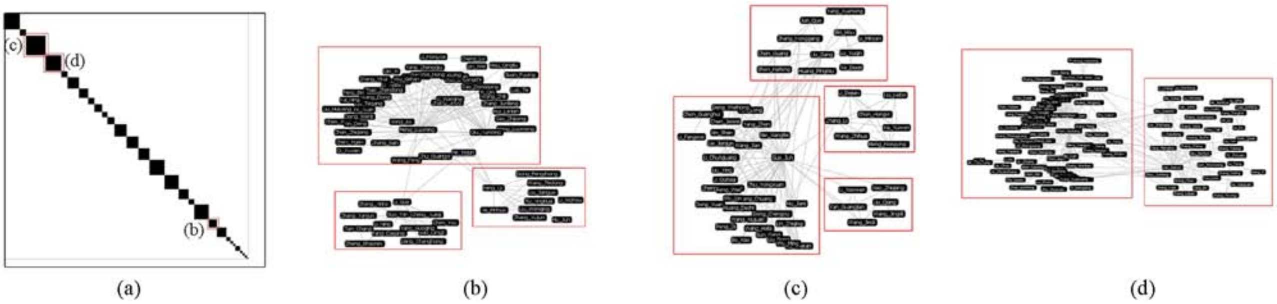

Fig.7. Partition result in the real coauthor network with GN. (a) The Possibility Matrix of GN's partition. (b)∼(d) The practical meaning of the selected communities in (a).

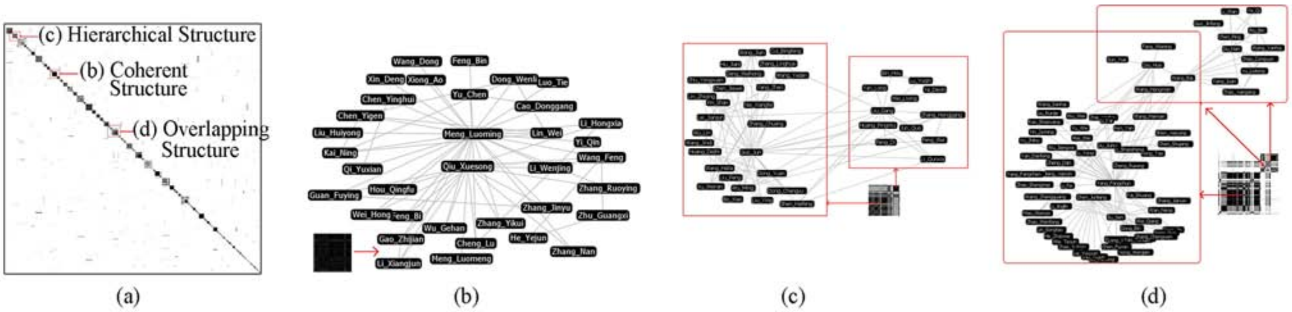

Fig.8. Multiple resolutions of modular structure in the real coauthor network with MOCD. (a) The Possibility Matrix of all solutions. (b) The practical meaning of the coherent structure. (c) The practical meaning of the hierarchical structure. (d) The practical meaning of the overlapping structure.

The members in this community, led by both Luo-Ming Meng and Xue-Song Qiu, are coherent. Their research interests mainly focus on network management and almost all the persons are in the same department. As for the GN's partition shown in Fig.7(b), besides Meng and Qiu's group, the community also includes two other relatively independent groups. In fact, the lower right group is led by Feng Qi, an expertise in telecommunications. Although he was a member of the network management group led by Meng and Qiu before, he is now a leader of a new group. The same thing also happens to the lower right group. Compared with GN, MOCD identifies the community more correctly.

Fig.8(c) shows an example of the community with a hierarchical structure. In the real situation, all the members in this community are in the same lab where Jun Guo is the director. Gang Liu, another professor, in the lab leads a different team. Guo's team and Liu's team share the similar research interest, which causes the formation of two sub-communities under this community. Fig.7(c) displays Jun Guo's research community obtained by GN. Apparently, this community includes more sub-communities, such as Jing-Di Wang's and Li Zhang's groups, which are not closely related to Guo's group in the real situation. Thus, GN only gives a large community. It cannot correctly reveal those sub-communities within it. MOCD is able to uncover more detailed community information or show its hierarchical structure.

Fig.8(d) shows an overlapping community. Most members of the community are the faculty members in the School of Computer. These members belong to two laboratories with an overlapping person Bai Wang. These two laboratories are the intelligent network group in the State Key Laboratory of Networking and Switching Technology (SKLNST) led by Fang-Chun Yang and Jun-Liang Chen, and the Telecommunication Software Engineering Group (TSEG) with professor Liu-Tong Xu and Bin Wu. The overlapping part is caused by Bai Wang who moved from SKLNST to TSEG in 2002. As shown in Fig.7(d), the corresponding partition in GN still displays a general structure that reveals two groups in SKLNST which are led by Junliang Chen and Jianxin Liao, respectively. It failed to reveal the implicitly overlapping part. In all, MOCD more correctly detects the complex community structures than GN.

4.3 Discussion

In the experiments, five networks including the synthetic and social networks were used to validate the effectiveness of MOCD. The optimal solutions returned by MOCD successfully discovered the underlying hierarchical and overlapping structures that is hard to be discovered by any single partition returned by the priori approaches. The experiments also show that MOCD can avoid the resolution limit existing in the priori approaches (e.g., GN and GACD), because MOCD is able to find small and independent communities. We think the advantages of MOCD can be explained as follows. The real social networks usually are complex and uncertain, because the data of the network are not clean and full of noise. And thus it is nearly impractical to describe the community structure with a fixed partition. MOCD solves the problem by providing many solutions in one run. These tradeoff solutions describe the community structure from different angles. A single solution of them may ignore some real structures, but their aggregation (i.e., the Possibility Matrix) can statistically offset the noise and uncertainty and reveal the true and comprehensive information.

MOCD requires some parameters settled before running. There are two types of parameters: four parameters for MOCD (i.e., population size, running generation, and the ratio of crossover and mutation), and two parameters for model selection (i.e., λ_{str} and λ_{weak}). Selection of GA-related parameters (i.e., first four parameters) can follow the general rules in that of MOEA[29]. Problems with large scale may require larger population size and more running generations to get good performance. A large ratio of crossover is helpful to convergence, but it may result in premature. The ratio of mutation has the opposite effect, that is, it helps to maintain the population diversity but slow down the convergence speed. A large crossover ratio and a small mutation ratio are usually used in MOEAs. Many experiments have confirmed that MOEAs with rational parameters could generate steady solutions[29]. Due to the limited space, we do not validate it with experiments in this paper. In the experiments, we chose the appropriate parameters based on the problem scales and did not specially tune them. The parameters for model selection are used to control the size of *StrMSet* and *WeakMSet*, which does not affect the quality of solutions. We set the proper parameters for better demonstration purpose in the experiments. Generally speaking, the networks will have more obvious community structure with larger λ_{str} and λ_{weak}.

The fitness evaluation function (i.e., calculating the values of the objectives) is the most time-consuming process in the algorithm. Calculating the objective functions has the complexity $O(m)$, and the decoding process has the complexity $O(n)$. As a consequence, the fitness evaluation based on an individual has the complexity $O(m + n)$. The whole complexity of MOCD is $gs^2(m+n)$ which is linear with the scale of the network. (g is the running generation, and s is the population size.) Note that the framework of MOCD (i.e., NSGA-II) has the complexity $O(gs^2)$[30]. More running generations and larger population size are usually desirable for large scale problems and lead to longer running time. However, increasing the population size or running generation does not yield better results at some point. As the constant parameters, these values (i.e., g, s) do not increase the time-complexity of the algorithm. As we know, most community detection algorithms have a large time-complexity[5]. Compared with these algorithms, the complexity of MOCD is small. Some multi-resolution methods (e.g., RB[9] and AFG[10]) apply the optimization technology (e.g., genetic algorithm, simulated annealing algorithm) to obtain a solution. To obtain multi-resolutions, these algorithms should be run many times by tuning parameters. However, MOCD obtains many solutions with only one run.

Similar to the contemporary GAs for CD[18,23-25], MOCD is also a heuristic search algorithm based on GA. A difference lies in the two objective functions of MOCD. Due to the difference, MOCD, as a posteriori approach, can avoid the resolution limit of those traditional GA-based approaches. With the great success of MOEAs, Handle and Knowles[28] have applied MOEA for clustering (MOEAC). MOCD and MOEAC both have the same MOEA framework. Because of the different problem characteristics, MOCD and MOEAC have many differences in objective functions, operators and model selection methods. Similar to MOCD, RB[9] and AFG[10] also provide multiple solutions through tuning a parameter. In fact, two components in RB are very similar to the *intra* and *inter* objectives, so RB and MOCD can be considered as two implementation of single- and multi-objective for the CD optimization problem, respectively. In order to obtain multi-resolutions, MOCD only requires one run, whereas RB and AFG should be run for many times by tuning the parameter. Recently, some researchers have proposed to detect communities with MOEA. Our method is different from MOGA-Net proposed Pizzuti[27] in many aspects, including objective functions, operators, and model selection methods. Similar to Shi *et al.*'s method[26] in the community detection phase, our approach has different model selection methods. Our approach firstly proposes effective model selection methods to explore the potential benefits of multi-solutions returned by multi-objective based methods in detecting the complex community structures (i.e., hierarchical and overlapping structures).

5 Conclusion

From the decision's perspective, this paper proposes a posteriori decision approach for community detection. Observing the fact that the community detection is intrinsically a multi-objective optimization problem, in this paper, we propose a posteriori approach, multi-objective evolutionary algorithm for community detection (MOCD), to detect complex community structure. The approach includes two phases. In the first community detection phase, a special multi-objective evolutionary algorithm is designed to search the solution space and return a set of optimal solutions. To help the decision maker (DMer) select proper community partitions from the optimal solution set, the second model selection phase further proposes three model selection criteria and the Possibility Matrix that effectively differentiate these optimal solutions according to their qualities.

Five synthetic and social networks validate the effectiveness of MOCD. The hierarchical and overlapping networks experiments illustrate that MOCD is able to effectively reveal the implicit hierarchical and overlapping communities, simultaneously avoiding the resolution limit. The random networks with known structures also validate that MOCD can find more accurate community structures compared with the state-of-the-art methods. Two social networks further show the advantages of MOCD that it not only correctly finds the independent and compact communities, but also reveals the valuable underlying structure information (e.g., overlapping and hierarchical structure) which is consistent with the real situation.

This paper focuses on the concept of community discovery with a posteriori approach and its practical advantages. Many interesting issues need further research. One of them is how to effectively select models from the candidate solutions and another interesting work is to design more effective objective functions.

References

[1] Newman M E J, Girvan M. Finding and evaluating community structure in networks. *Physics Review E*, 2004, 69(2): 026113

[2] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D U. Complex networks: Structure and dynamics. *Physics Report*, 2006, 424(4/5): 175-308.

[3] Guimera R, Amaral L A N. Functional cartography of complex metabolic networks. *Nature*, 2005, 433: 895-900.

[4] Palla G, Dereyi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 2005, 435(7043): 814-818.

[5] Danon L, Diaaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiments*, 2005, (9): p09008.

[6] Pothen A, Sinmon H, Liou K-P. Partitioning sparse matrices with eigenvectors of graphs. *SIAM J.Matrix Anal App.*, 1990, 11(3): 430-452.

[7] Martin R, Carl T B. An information-theoretic framework for resolving community structure in complex networks. *PNAS*, 2007, 104(18): 7327-7331.

[8] Fortunato S, Barthelemy M. Resolution limit in community detection. In *Proc. the National Academy of Sciences*, 2007, 104(1): 36-41.

[9] Reichardt J, Bornholdt S. Statistical mechanics of community detection. *Physics Review E*, 2006, 74(1): 016110.

[10] Arenas A, Fernandez A, Gomez A. Multiple resolution of modular structure of complex networks. arXiv:physics/0703218v 1, 2007.

[11] Kumpula J M, Saramaki J, Kaski K, Kertesz J. Limit resolution and multiresolution models in complex network community detection. arXiv:0706. 2230v2, Jan. 25, 2008.

[12] Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multi-scale complexity in networks. *Nature*, 2010, 466: 761-764.

[13] Hwang C L, Masud A S M. Multiple Objective Decision Making-Methods and Applications. Berlin: Springer Verlag, Germany, 1979.

[14] Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. Defining and identifying communities in networks. *PNAS*, 2004, 101(9): 2658-2663.

[15] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks. *Physical Review E*, 2004, 70(6): 06611.

[16] Brandes U, Delling D, Gaetler M *et al.* On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2008, 20(2): 172-188.

[17] Duch J, Arenas A. Community detection in complex networks using extremal optimization. arXiv:condmat/0501368, 2005.

[18] Tasgin M, Bingol H. Community detection in complex networks using genetic algorithm,. arXiv:cond-mat/0604419, 2006.

[19] Shen H, Cheng X, Fang B. Covariance, correlation matrix, and the multiscale community structure of networks. *Phys. Rev. E*, 82(1): 016114.

[20] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 2009, 11(3): 033015.

[21] Shen H, Cheng X, Kai C, Hu M. Detect overlapping and hierarchical community structure in networks. *Physica A*, 2009, 388: 1706-1712.

[22] Shen H, Cheng X, Guo J. Quantifying and identifying the overlapping community structure in networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2009, P07042.

[23] Shi C, Wang Y, Wu B, Zhong C. A new genetic algorithm for community detection. *Complex Sciences*, 2009, 5(1): 1298-1309.

[24] Pizzuti C, GA-Net: A genetic algorithm for community detection in social networks. In *Proc. PPSN2008*, Dortmund, Germany, Sept. 13-17, 2008, pp.1081-1090.

[25] Pizzuti C. Community detection in social networks with genetic algorithms. In *Proc. GECCO2008*, Alanta, USA, Jul. 12-16, 2008, pp.1137-1138.

[26] Shi C, Zhong C, Yan Z, Cai Y, Wu B. A multi-objective optimization approach for community detection in complex network. In *Proc. CEC2010*, Barcelona, Spain, Jul. 18-23, 2010, pp.1-8.

[27] Pizzuti C. A multi-objective genetic algorithm for community detection in networks. In *Proc. ICTAI2009*, Newark, USA, Nov. 2-4, 2009, pp.379-386.

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

[29] Deb K. Multiobjective Optimization Using Evolutionary Algorithms. Wiley, UK, 2001.

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

[31] Shi C, Yan Z, Wang Y, Cai Y, Wu B. A genetic algorithm for detecting communities in large scale complex networks. *Advances in Complex Systems*, 2010, 13(1): 3-17.

[32] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(4): 046110.

[33] Du N, Wang B, Wu B. Community detection in complex networks. *Journal of Computer Science and Technology*, 2008, 23(4): 672-683.

Chuan Shi received the Ph.D. degree from the Institute of Computing Technology, Chinese Academic of Sciences, Beijing, in 2007. He is a member of CCF. He joined the School of Computer of Beijing University of Posts and Telecommunications as a lecturer in 2007, and is an associate professor at present. His research interests are in machine learning, data mining, and evolutionary computing. He has published more than 20 papers in refereed journals and conferences.

Zhen-Yu Yan received the Ph.D. degree in systems engineering in 2007 from the University of Virginia, USA. He is currently a lead scientist in the Research Department at Fair Isaac Corporation (FICO). His research interests include risk analysis, multi-objective optimization and decision making, and data mining. He has published more than 20 research papers on the related areas.

Xin Pan is an undergraduate student in Beijing University of Posts and Telecommunications. His research interests are in machine learning, data mining, and evolutionary computing.

Ya-Nan Cai is a Master's candidate in Beijing University of Posts and Telecommunications. Her research interests are in machine learning, data mining, and evolutionary computing.

Bin Wu received the Ph.D. degree from the Institute of Computing Technology, Chinese Academic of Sciences, Beijing, in 2002. He is a member of CCF. He joined the School of Computer of Beijing University of Posts and Telecommunications as a lecturer in 2002, and is an associate professor at present. His research interests are in data mining, complex network, and cloud computing. He has published more than 40 papers in refereed journals and conferences.