A Multi-Objective Approach for Community Detection in Complex Network

Chuan Shi, Cha Zhong, Zhenyu Yan, Yanan Cai, Bin Wu

Abstract— **Detecting community structure is crucial for uncovering the links between structures and functions in complex networks. Most contemporary community detection algorithms employ single optimization criteria (e.g., modularity), which may have fundamental disadvantages. This paper considers the community detection process as a Multi-Objective optimization Problem (MOP). Correspondingly, a special Multi-Objective Evolutionary Algorithm (MOEA) is designed to solve the MOP and two model selection methods are proposed. The experiments in artificial and real networks show that the multi-objective community detection algorithm is able to discover more accurate community structures.**

I. INTRODUCTION

Community Detection (CD) in complex networks has attracted a lot of attention in recent years. The main reason is that communities are supposed to play a special role in the often stochastic dynamics of the systems under consideration; and detecting communities (or modules) can be a way to identify substructures which could correspond to important functions. Loosely speaking, these communities are groups of nodes that are densely interconnected but only sparely connected with the rest of the network [1,2].

There have been many successful algorithms to analyze the community structure in complex network. The contemporary community detection algorithms can be roughly classified into two categories: optimization based methods and heuristic methods. The optimization based methods (e.g., GN fast [3], spectral method [4]) convert the CD into an optimization problem and the heuristic methods convert the CD into the design of heuristic rules (e.g., the edge betweenness in GN [5]). In fact, the heuristic methods usually also need a measure criterion to stop the iteration process. For example, the maximum modularity Q is used as the stopping criterion in GN [5]. And thus the community detection problem can be regarded as a single-objective optimization problem. Without loss of generality, we assume it is a minimum problem. Most contemporary algorithms for CD are based on the single-objective optimization. Different algorithms vary in the optimization function, for example, the modularity Q in GN [5], the "cut" function in spectral method [4] and the "description length" in the information-theoretic based method [6].

Chuan Shi, Cha Zhong, Yanan Cai, Bin Wu is with Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, China 100876. (email: shichuan@bupt.edu.cn)

Zhenyu Yan is with the Research Department, Fair Isaac Corporation, San Rafeal, CA, USA, 94903.(email: yan_zhen_yu@hotmail.com)

The single objective based community detection algorithms have achieved great successes in both theory and applications. However, they also have some crucial disadvantages. These single-objective algorithms attempt to optimize just one of such criteria and this confines the solution to a particular community structure property. And thus it often causes a fundamental discrepancies that the different algorithms produce distinct solutions on the same networks. Moreover, the single-objective optimization algorithms may fail when the optimization criteria are inappropriate. An example is the resolution limit existing in the modularity Q: the modularity optimization may fail to identify modules smaller than a scale even in cases where modules are unambiguously defined [7]. Similar resolution limits also exist in some other single-objective algorithms [8]. In addition, many single-objective algorithms require prior information: the number of communities, which is usually unknown for real networks.

In order to alleviate disadvantages in single-objective community detection algorithms, a natural approach may be to consider community detection as a multi-objective optimization problem. That is to say, we simultaneously optimize multiple objective functions to obtain more accurate and comprehensive community structure. Compared to the contemporary single-objective algorithms, the multi-objective approaches for CD have obvious advantages in concept. Firstly, the community detection with multiple criteria is more consistent with human's intuition. The CD problem can be regarded as a graph clustering. The concept of a cluster is a generalization of what human perceive, as densely connected "patches" within data space, whereas human's intuition is inherently difficult to capture by means of single objective [13]. Secondly, besides the optimal solution found by a single-objective algorithm, a multi-objective algorithm is able to find the optimal solutions corresponding to the tradeoffs between the different objectives. Finally, some researchers have began to be aware that enumerating the modules in a network is a tradeoff among multi-objectives. Fortunato et al. believed that finding the maximum modularity is to look for the ideal tradeoff between the number of modules and the value of each term [7,11]. Rosvall and Bergstrom also thought that enumerating

978-1-4244-8126-2/10/$26.00 ©2010 IEEE

the modules in a network has an inevitable tradeoff between the amount of the structure information of a network and its description length [6]. Although these researchers have been aware of the intrinsic tradeoff, most algorithms still optimize the single objective which simply combines the conflict components (e.g., the linear combination in modularity Q).

Since the community detection process can be regarded as a Multi-objective Optimization Problem (MOP), this paper designs a special Multi-Objective Evolutionary Algorithm (MOEA), named Multi-Objective Community Detection algorithm (MOCD), to generate the Pareto optimal solution set of the MOP. Furthermore, this paper proposes two model selection methods to assist the decision makers (DMers) in selecting the proper solutions from the Pareto optimal solution set. Two experiments on artificial and real networks illustrate that the community structures discovered by the MOCD are more accurate than those generated by three well-established single-objective algorithms.

This paper is arranged as follows. Section 2 introduces the related work. Section 3 detailedly describes the multi-objective community detection algorithm. The experiments on artificial and real networks are done to validate the effectiveness and efficiency of the algorithm in Section 4. Section 5 concludes the paper.

II. Related Work

A. Community Detection in Complex Network

Complex systems in various domains may be modeled as complex networks, such as the internet, WWW, social networks and citation networks. Most of these networks are generally sparse in global yet dense in local, which can be described as that the nodes within the groups have higher density of edges while nodes among groups have lower density of edges. Those "groups" are called the communities, which are often the key elements to reveal many hidden features of a given network [1,2]. Hence, community identification is a fundamental step to understand the overall structural and functional properties of large networks.

There have been many algorithms to analyze the community structure in complex network. The algorithms use methods and principles of physics, artificial intelligence, graph theory and even electrical circuits [18]. One of the most known algorithms proposed so far is Girvan-Newman (GN) algorithm that introduces a divisive method by iteratively cutting the edge with the greatest betweenness value [5]. Some improved algorithms have been proposed [3,17]. These algorithms are based on a foundational measure criterion of community, modularity Q, proposed by Newman [5]. The larger the value of Q is, the more accurate a partition into communities is. As a consequence, the community detection becomes a modularity optimization problem (i.e., a single-objective optimization problem). Since the search for the optimal (largest) modularity value is a NP-complete problem [11], many heuristic search algorithms have been applied to solve the optimization problem, such as extremal optimization [19], simulated annealing [2].

Some other criteria are also used as the optimization objective. The Hamiltonian-based method introduced by Reichardt and Bornholdt (RB) is based on considering the community indices of nodes as spins in a q-state Potts model [9]. Recently, Arenas, Fernandez and Gomez (AFG) proposed a multiple resolution procedure that allows the optimization of modularity to go deep into the structure [10]. These methods vary the thresholds by using a tuning parameter in their criteria and investigate the community structure at variable resolutions. The modularity Q is the special case of these two criteria. In addition, Fosvall and Bergstrom proposed an information-theoretic foundation for the concept of modularity in networks [6], in which the network is composed of modules by finding an optimal compression of its topology. Although these criteria could effectively assess the quality of the community, the recent research show that the optimization based on single criterion has a fundamental disadvantage. Fortunato and Barthelemy 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 cases where modules are unambiguously defined [7]. Kumpula et al. further discussed the similar limited resolution of community detection methods where a global energy-like quantity is optimized, for example, the former two criteria (RB and AFG) [8].

The Genetic Algorithm (GA), as an effective optimization technique, has also been used for community detection. Tasgin and Bingol first applied GA (GATB) for CD, in which the objective function is the modularity Q and the encoding scheme is the cluster centers. Different from GATB, the GA proposed by Shi et al. [15] uses the locus-based adjacency as the encoding scheme, respectively. GA-Net proposed by Pizzuti optimizes the "community score" criteria and applies the the locus-based adjacency scheme [21,22]. Pizzuti further extended her algorithm to solve the overlapping community problem [23]. These algorithms have the advantage that the number of communities can be automatically determined during the evolutionary process. However, these algorithms also have the resolution limit, since the single objective is applied.

B. Evolutionary Algorithm for Multi-objective Optimization

Multi-objective optimization problems (MOPs) are those problems that involve simultaneous optimization of two or more than two objectives (often competing) and usually there is no single optimal solution [12]. A MOP is formally defined as follows:

Definition 1 [12]: General MOP: an MOP minimizes $F(\vec{x}) = (f_1(\vec{x}), \cdots, f_m(\vec{x}))$ subject to $g_i(\vec{x}) \leq 0$, $i = 1, \cdots, k$, $\vec{x} \in \Omega$ (Ω is the decision variable space). An MOP solution minimizes the components of the m-dimensional objective vector $F(\vec{x})$, where $\vec{x} = (x_1, \cdots, x_n)$ is an n-dimensional decision variable vector from some universe Ω.

It is usually difficult or even impossible to assign priorities as in single objective optimization problems (SOPs). This makes an algorithm returning a set of promising solutions preferable to an algorithm returning only one solution based on some weighting of the objectives. For this reason, there has been an increasing interest in applying Evolutionary Algorithms (EA) to MOPs in the past twenty years. An important notion is embraced, which can be defined as follows.

Definition 2 [12]: Pareto dominance: If a vector $U = (u_1, \cdots, u_m)$ Pareto dominates $V = (v_1, \cdots, v_m)$, denotes as $U \preceq V$, that is $U \preceq V$ if and only if $\vec{u}$ is partially less than $\vec{v}$, i.e.,$\forall i \in \{1, \ldots, m\}$: $u_i \leq v_i \wedge \exists i \in \{1, \ldots, m\} : u_i < v_i$.

Most contemporary research on MOP is based on Pareto dominance. A decision vector $\vec{x}_u$ is said to be Pareto optimal if and only if there is no $\vec{x}_v$ for which $F(\vec{x}_v) \preceq F(\vec{x}_u)$. The set of all Pareto optimal decision vectors is called the Pareto optimal set. The corresponding set of the objective vector is called the nondominated set, or Pareto front.

Many multi-objective evolutionary algorithms have been proposed and these algorithms have successfully solve some real problems [12]. Handle and Knowles have applied MOEA for clustering (MOEAC) [24] and their experiments demonstrated that the performances of MOEAC are better than a number of well-established single-objective clustering algorithms and ensemble techniques. However, no MOEAs are applied for community detection until now.

III. A Multi-objective Evolutionary Algorithm for Community Detection

This paper applies the evolutionary algorithm to solve the multi-objective community detection problem. This algorithm consists of the two main phases. In the detection phase, MOCD optimizes two complementary objectives and returns a set of Pareto optimal solutions which corresponds to different tradeoffs between these two objectives. In the model selection phase, MOCD employs two methods to select the most preferable solution from the Pareto optimal set.

A. Detection and Candidate Solution Generation Phase

We select an existing MOEA, the Pareto Envelope-based Selection Algorithm version 2 (PESA-II) [13], as the framework of the MOCD. In fact, other successful MOEAs can also be used or a new MOEA can be designed. Because the successful application of a MOEA depends on the design of its components according to the problem's characteristics [12], many components in the MOCD should be redesigned.

1) Algorithm Framework: PESA-II follows the standard principles of an EA with the difference that two populations of solutions are maintained: an internal population (IP) of fixed size, and an external population (EP). The IP explores new solutions and achieves these by the standard EA process of reproduction and variation. The EP is to exploit good solutions by maintaining a large and diverse set of the non-dominated solutions discovered during search. Selection occurs at the interface between the two populations, primarily in the update of EP. The detailed implementation can be seen in ref. [13]. There are five basic parameters in the algorithm and their meanings are illustrated here:

ipsize and *epsize* are the size of *IP* and *EP*.

p_c and p_m are the ratio of crossover and mutation.

gen is the running generation.

To apply PESA-II to the community detection problem, much work should be done. Two or more objective functions should be determined according to the characteristics of CD. Moreover, a community structure should be encoded with a genetic representation, and the corresponding genetic variation operators need to be designed. These choices are crucial for the performance and particularly for the scalability of the algorithm. Our choices for these components are determined after extensive experimentation.

2) Objective Functions: For the evaluating objectives, we are interested in selecting those reflecting fundamentally different aspects of a good community partition. Modularity is a foundational quality index for CD. Given a simple graph G=(V,E), we follow [5] and define the following equation.

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

, 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. Observing the equation, to maximize the modularity Q, we should maximize the first term and minimize the second term. To maximize the first term, many edges should be contained in clusters (i.e., "densely interconnected"). To minimize the second term, the graph is split into many clusters each with small total degrees each (i.e., "sparely connected with the rest"). These two complementary terms reflect two fundamental aspects of a good partition, and the modularity Q 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 minimum optimization problem, we revise the first term. The first objective function minimizes 1 minus the intra-link strength of a partition, and it is called $intra$

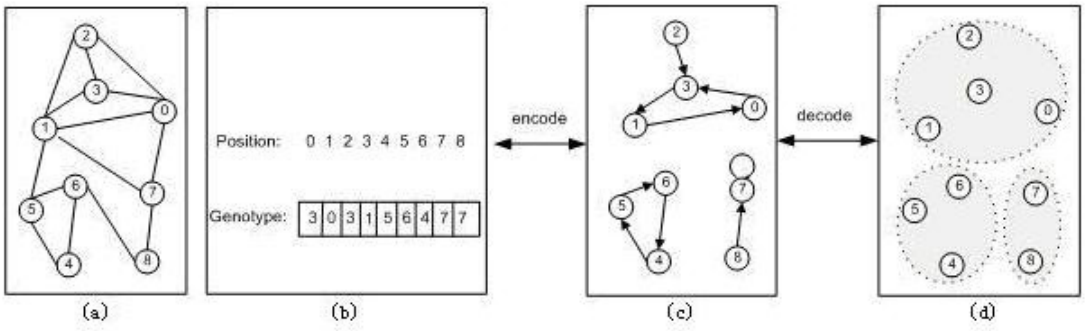

Fig. 1. Illustration of the locus-based adjacency representation. (a) shows the topology of a complex network. (b) shows one possible genotype. (c) shows how the genotype in (b) is translated into a graph structure, for example node 0 links to node 3, because the value of gene g_0 is 3. (d) shows the partition result.

objective.

$$intra(C) = 1 - \sum_{c \in C} \frac{|E(c)|}{m} \quad (2)$$

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

$$inter(C) = \sum_{c \in C} (\frac{\sum_{v \in c} deg(v)}{2m})^2 \quad (3)$$

According to the two definitions, we deduce that

$$Q(C) = 1 - intra(C) - inter(C) \quad (4)$$

The motivation of selecting two components of the modularity Q as the objective functions rather than other criteria are stated as follows. Firstly, these two functions have the 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. With an increasing number of communities, the fewer edges fall in communities (i.e., $E(c)$ becomes smaller), and thus the $intra$ objective value tends to increase. The opposite trend happens to the $inter$ objective. Secondly, compared to the single-objective algorithms based on the modularity optimization, the multi-objective algorithm based on these two components of modularity have better performances as the experiments will illustrate, which effectively confirms the advantages of the multi-objective algorithm. Finally, after many experimentations, we find these two functions are more empirically suitable.

3) Genetic Representation and Operators: This paper employs the locus-based adjacency representation [14,15], in which each genotype g consists of n genes $g_1, g_2, \ldots, g_n$ and each g_i can take one of the adjacent nodes of node i. Thus, a value of j assigned to the ith gene, is then interpreted as a link between node i and j. In the resulting solution, they will be in the same community. Figure 1 illustrates an example of the genetic representation. The locus-based adjacency encoding scheme has been validated to be effective for community discovery [15].

We choose the uniform two-point crossover because it is unbiased with respect to the ordering of genes and can generate any combination of alleles from the two parents. In the mutation operation, we randomly select some genes and assign 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 its adjacent nodes of node i.

B. Model Selection Phase

As noted previously, MOCD does not return a single solution, but a set of Pareto optimal solutions. These solutions correspond to different tradeoffs between the two objectives and also consist of the communities with different sizes. The DMers may desire that the set of candidate solutions could be narrowed down to those of interest to him or her. This section proposes two automated methods for assessing the quality of solutions and identifying the promising solution.

Maximum Q criterion. The criterion selects the model with maximum modularity Q. Because of the relationship of Q and two objective functions (see Equation 4), we can directly select the model with maximum Q according to these two objective values. (SF is the candidate solution set, i.e., the Pareto front)

$$S_{Max-Q} = \underset{C \in SF}{\operatorname{argmax}} \{1 - intra(C) - inter(C)\} \quad (5)$$

Max-Min Distance criterion. Since the physical meaning of Q is the fraction of edges that falls within communities, minus the expected value of the same quantity if the edges fall at random without regard for the community structure, the Q evaluates the extent to which the community structure deviates from randomness [5]. The similar principle can also be used for the model selection. Firstly, MOCD can be run on the real

network and a random network with the same scale, and the real candidate solution set (i.e., real Pareto front) and the random control solution set (i.e., random Pareto front) can be obtained, respectively. And then we select the solution in the real Pareto front with the most deviation from the solutions in the random Pareto front as the best solution. Since there are multiple solutions in the real and random Pareto front, we need to quantitatively evaluate the deviations between any two solutions in the two sets. Here a heuristic rule is applied: the deviation of a solution in the real Pareto front is evaluated by the minimum Euclidean distance between the solution and a solution in the random Pareto front, and then the solution in the real Pareto front with the largest deviation is selected. The model selection process can be formulated as the following equation:

$$\begin{aligned} &dis(C, C') = \\ &\sqrt{(intra(C) - intra(C'))^2 + (inter(C) - inter(C'))^2} \\ &diff(C, CF) = min\{dis(C, C') | C' \in CF\} \\ &S_{Max-Min-Dis} = \underset{C \in SF}{\text{maxarg}}\{diff(C, CF)\} \end{aligned} \tag{6}$$

where CF and SF represent the random and real Pareto fronts, respectively. In fact, the purpose of the random control solution set is to obtain an estimate of the values of $intra$ and $inter$ that would be expected for unstructured network and the *Max-Min Distance* criterion evaluates the difference between the real objective values and the expected ones.

IV. Experiments

This section will validate the effectiveness and efficiency of the multi-objective community detection algorithm through the artificial and real networks. The experiments are carried out on a 3GHz and 1G RAM computer running Windows XP.

A. Artificial Networks

To validate the quality of the solutions selected by the model selection methods, we first use artificial networks with a known community structure. These networks have 128 vertices grouped in four communities of 32 vertices [5]. Each vertex has on average z_{in} edges to vertices in the same community and z_{out} edges to vertices in other communities, keeping an average degree $z_{in}+z_{out} = 16$. The network is called the symmetric network. The experiments further vary the network structures in the following ways. The first variation, called the vertex asymmetric network, merges three of the four groups in the benchmark test to form a series of test networks each with one large group of 96 vertices and one small group with 32 vertices. In the second variation, the benchmark networks, called the edge asymmetric network, compose two groups each with 64 vertices, but with different average degrees of edges (8 and 24) per vertex. As the average number of edges z_{out} increases, it becomes harder and harder to identify the group structure. To compare the quality of solutions, the experiments use the Fraction of Vertices Identified Correctly (FVIC), which has been used in many researches [5,18,19]. The larger the FVIC is the better partition is. The FVIC can be calculated as follows.

$$\begin{aligned} &olSet(c, c') = \{v | v \in c \wedge v \in c'\} \\ &maxOlSet(c, C_K) = \max_{c' \in C_K} \{|olSet(c, c')|\} \\ &FVIC = \textstyle\sum_{c \in C_F} \dfrac{maxOlSet(c, C_K)}{N} \end{aligned} \tag{7}$$

, where C_F and C_K represent the found and known community partition, respectively; c and c' are a community in C_F and C_K, respectively. N is the number of vertices in the graph.

Five algorithms are included in the experiments. The first algorithm is based on the information-theoretic framework proposed by Rosvall and Bergstrom [6] (called INFO). The second one is the betweenness-based heuristic algorithm proposed by Newman and Girvan [5] (called GN). The third one is genetic algorithm based modularity optimization algorithm proposed by Shi et al. [15] (called GACD). The other two algorithms are MOCD with $Max\ Q$ model selection (called MOCD-Q) and $Max\text{-}Min\ Distance$ model selection (called MOCD-D). The parameters setting of MOCD are as following: $ipsize$ is 100, $epsize$ is 100, gen are 200, and p_c and p_m are 0.6 and 0.4, respectively. GACD uses the same parameters with those of MOCD. Note that, in the comparison experiments, MOCDs and GACD evaluate the same number of individuals. The results are the average values of 100 network realizations.

Figure 2 presents the FVIC results of the five algorithms. When z_{out} is small, all algorithms find the correct community partition. As z_{out} increases, these five algorithms have different performances, and their differences become more distinct. We can observe that the MOCD based algorithms have the highest FVIC in most conditions and INFO is better than GN and GACD for the asymmetric networks. Comparing the results in the symmetric networks with those in the asymmetric networks, we can find that it is more difficult for all algorithms to discover the community structure in the asymmetric networks, especially for GN. However, the asymmetric networks have less impact on the MOCD based algorithms. The NCs found by these five algorithms are illustrated in Figure 3. Similar to the results of FVIC, the NCs obtained with different algorithms are correct for all problems when z_{out} is small, whereas they have more deviations from the correct values as z_{out} increases. The NCs found by MOCD-D are the closest to the correct values in

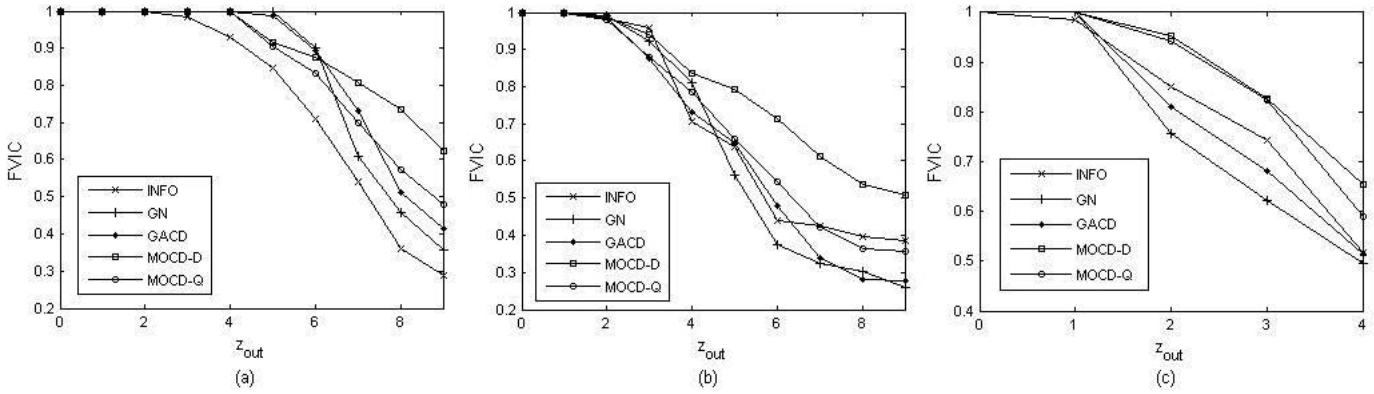

Fig. 2. Benchmark performance for symmetric and asymmetric group detection measured as Fraction of Vertices Identified Correctly (FVIC). (a), (b) and (c) are the results of the symmetric networks, the vertex asymmetric networks, and the edge asymmetric networks, respectively.

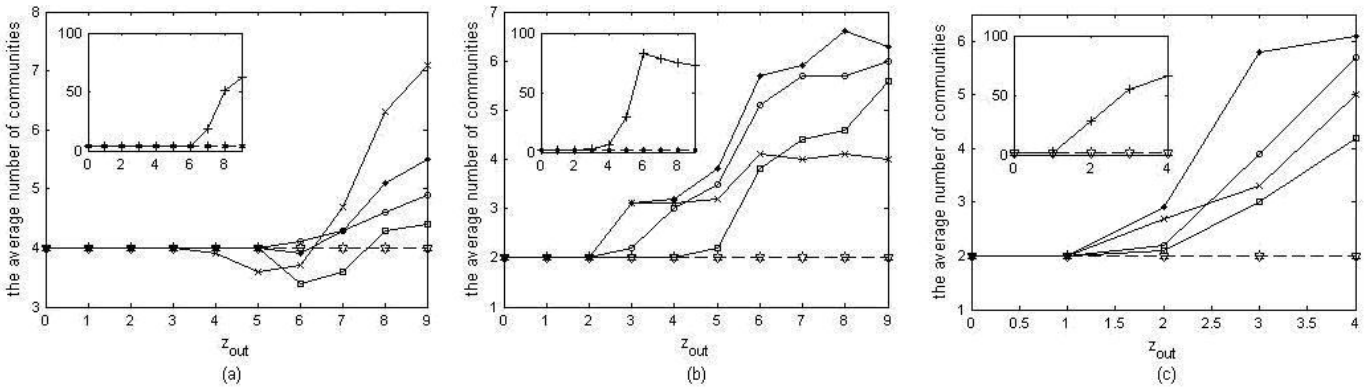

Fig. 3. The number of communities found for the symmetric and asymmetric networks. (a), (b) and (c) are the results of the symmetric networks, the vertex asymmetric networks, and the edge asymmetric networks, respectively. The broken line with label ☆, as baseline, is the correct number of communities.

most conditions. When z_{out} is large, GN divides the graph into so many communities that its FVIC declines rapidly in Figure 2. In summary, GN and GACD are more effective in symmetric networks, whereas INFO is more effective in asymmetric networks. The results are consistent with those in ref. [6]. Generally speaking, the MOCD based algorithms are more effective than the other algorithms both in symmetric and asymmetric networks.

Through the experiments, we find that the MOCD based algorithms (especially MOCD-D) have the highest FVICs and the best NCs (i.e., the closest to the correct NCs) for most problems. To study the reason behind the superior performance of the MOCDs, we compare GACD with the MOCDs (especially MOCD-Q). A number of components of these two algorithms are same (e.g., GA framework, fitness function, genetic representation, and the same number of individuals evaluated), except that GACD uses Q as the single criterion function, whereas MOCDs treat the two components of Q as two criterion functions. Therefore, the superior performance of MOCDs should be driven by the multi-criterion functions. We consider two reasons may account for the effectiveness of the multi-criterion functions. Firstly, the multiple objectives can measure the community structure comprehensively and avoid the risk that one single objective may only be suitable to a kind of networks (e.g., GN is only suitable for symmetric networks). Secondly, the multi-objective optimization process tradeoffs the multiple conflicting objectives, which can effectively avoid being trapped to local optima. Between MOCD-Q and MOCD-D, it is clear that MOCD-D has better performances. This shows that *Max-Min Distance* may be a better model selection criterion than *Max Q*. In fact, $Max\text{-}Min\ Distance$ criterion selects the model with the largest deviation from the random network with the same scale, which indicates the selected model has the most remarkable community structure. As for INFO and MOCD, they are based on the information-theoretic framework and the multi-objective framework, respectively. The experiments illustrate that the multi-objective framework may be more effective than the information-theoretic framework.

TABLE I

TEST PROBLEMS AND PARAMETERS SETTINGS IN GACD AND MOCD.

		Karate (P1)	Lesmis (P2)	Polbooks (P3)	Adjnoun (P4)	Football (P5)	Celegans nearal(P6)	Celegansmetabolic(P7)	Netscience(P8)	Power (P9)	Hep-th (P10)
Number of nodes		34	77	105	112	115	297	453	1589	4941	8361
Number of edges		78	254	441	425	613	2345	2025	2742	6594	15751
GACD	*pop*	50	50	50	50	50	100	100	200	300	400
	gen	50	50	100	100	100	100	100	200	300	400
MOCD	*ep*	50	50	50	50	50	100	100	100	100	100
	ip	50	50	50	50	50	100	100	200	300	400
	gen	50	50	100	100	100	100	100	200	300	400

TABLE II

THE EXPERIMENTAL RESULTS FOR REAL SOCIAL NETWORKS. CNUM IS THE NUMBER OF COMMUNITIES, SRAT IS THE RATIO OF STRONG COMMUNITIES, WRAT IS THE RATIO OF WEAK COMMUNITIES, AND TIME IS THE RUNNING TIME (THE UNIT IS SECOND).

		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
INFO	cNum	2.0	5.0	3.0	1.0[1]	9.0	7.0	25.4	396.0	12.0	–[2]
	sRat	0.5000	**0.4000**	0.3333	1.0000	**0.7777**	0.0000	0.0124	0.6767	**0.1666**	–
	wRat	1.0000	**1.0000**	1.0000	1.0000	**1.0000**	0.3333	0.3218	0.6767	**1.0000**	–
	time	0.1091	0.2856	3.7030	4.5547	1.5616	88.130	145.67	27.657	10779	–
GN	cNum	5.0	11.0	5.0	69.0	10.0	33.0	38.0	405.0	45.0	–
	sRat	0.5000	0.0909	0.0000	0.0000	0.5000	0.0000	0.0263	**0.9876**	0.1555	–
	wRat	0.6000	0.7272	1.0000	0.0144	1.0000	0.0909	0.3157	**1.0000**	1.0000	–
	time	0.0328	0.2029	0.4295	1.7235	1.4016	64.603	135.82	1.8170	11784	–
GACD	cNum	4.0	7.0	4.3	6.2	8.3	6.0	18.0	413.0	355.0	1601.0
	sRat	0.5000	0.2428	0.3200	0.0000	0.4945	0.0000	0.0536	0.6803	0.1267	0.4141
	wRat	1.0000	1.0000	0.9400	0.2547	0.9909	0.4696	0.3921	0.6900	0.9352	0.5203
	time	0.4264	0.8969	2.3860	2.4891	2.5548	12.855	19.989	828.94	2428.0	15256
MOCD-D	cNum	3.5	6.0	4.2	5.1	7.6	5.0	20.8	449.0	700.0	1837.0
	sRat	**0.5500**	0.3700	**0.3733**	**0.1000**	0.3741	**0.2000**	**0.0954**	0.6436	0.1485	**0.5679**
	wRat	**1.0000**	1.0000	**1.0000**	**0.3402**	1.0000	**0.5511**	**0.4442**	0.7037	0.9571	**0.6307**
	time	1.0764	2.0986	5.6749	5.4952	7.2204	54.461	69.412	1375.4	10505	51284
MOCD-Q	cNum	3.8	6.9	4.5	7.3	7.1	6.4	23.3	438.0	715.0	1832.0
	sRat	0.5000	0.3279	0.3200	0.1000	0.2886	0.1000	0.0844	0.6552	0.1454	0.4733
	wRat	1.0000	1.0000	1.0000	0.2914	1.0000	0.4723	0.3938	0.7077	0.9549	0.5343
	time	0.5454	1.1734	3.2438	3.0750	3.6687	28.776	39.114	882.09	5657.5	33291

[1] For P4, INFO always partitions the network into one community, and thus sRat and wRat both are 1. However, the partition is irrational.

[2] –represents the algorithm cannot solve the problem in twenty hours.

B. Real Networks

In order to further compare the performance of different algorithms, we use ten real social networks (these networks all are from ref. [25]). These test problems are widely used as benchmarks in community detection [2,4,11,12,15], and they have different scales with the number of vertices ranging from 34 to 8361. These test problems and the parameters setting in GACD and MOCDs are illustrated in Table 1 (p_c and p_m in GACD and MOCDs both are 0.6 and 0.4, respectively). Note that we do not make any effort in setting good parameters for GACD and MOCDs and the appropriate parameters are settled according to the scale of the problems. Moreover, the same numbers of individuals are evaluated in GACD and MOCDs. Since the community structure of most networks is unknown, we can only evaluate the quality of solutions from the structural characteristics. Here we use two popular criteria to measure the quality. According to the strong and weak community definition given by Radicchi et al. [17], each community c is validated based on whether satisfying the strong (or weak) community definition. The *ratio of strong (or weak) communities* is the fraction of communities in a partition C satisfying the strong (or weak) community definition.

$$
\begin{aligned}
&strRatio(C) = \frac{|\{c|k_i^{in}(c) > k_i^{out}(c) \forall i \in c \wedge \forall c \in C\}|}{|C|} \\
&weakRatio(C) = \\
&\frac{|\{c|\sum_{i\in c} k_i^{in}(c) > \sum_{i\in c} k_i^{out}(c) \forall c \in C\}|}{|C|}
\end{aligned}
\tag{8}
$$

, where c is a community in the partition C; $k_i^{in}(c)$ is the number of edges connecting node i to other nodes belonging to c, and $k_i^{out}(c)$ is the number of connections toward nodes in the rest of the network. These two criteria quantitatively evaluate how obvious the community structure is. The larger value means the better partition. According to the definitions, a strong community should be a weak community, and thus the *ratio of weak communities* is usually larger than the *ratio of strong communities*. Only one random Pareto front is generated for MOCD-D in the experiments. The results

are the average of ten runs.

The experimental results are illustrated in Table.2. MOCD-D discovers the community structure with the highest accuracy for most real networks (e.g., P1, P3, P4, P6, P7, and P10), and INFO also has the highest accuracy for three real networks (e.g., P2, P5, and P9). The experiments validate the conclusions drawn in the above simulated networks again. Due to the multi-objective framework, MOCD-Q has the better performance than GACD for most problems (e.g., P2, P3, P4, P6, P7, P9, and P10). *Max-Min Distance* criterion is better than *Max Q* criterion, because the performance of MOCD-D is better that of MOCD-Q for almost all problems. Observing the running time in Table 2, we can find that although the running times of the GA based algorithms (i.e., GACD, MOCD-D, MOCD-Q) are longer than that of GN and INFO for the small-scale problems, such as P1-P5, it is not the case for the large-scale problems, especially for P10 where GN and INFO cannot obtain the results in the given time. An exception case is P8 which has a obvious community structure and is very suitable for GN. The experiments show that the running time of MOCDs are acceptable and it is especially suitable for large-scale networks. Although GACD and MOCD evaluate the same number of individuals, the multi-objective algorithms are more complicated than the single-objective algorithms, so the running times of MOCDs are longer than that of GACD. Since MOCD-D needs to run twice to obtain the real and random Pareto fronts, the running times of MOCD-D are near the twice of that of MOCD-Q. The *Max-Min Distance* criterion in MOCD-D has better performance at the cost of longer running times.

V. Conclusion

This paper considers the community detection problem as a Multi-Objective optimization Problem (MOP) and designs a special Multi-Objective Evolutionary Algorithm (MOEA) for the MOP (called MOCD). The method includes two phases. In the first phase, an existing MOEA, PESA-II, is adapted with two complementary objective functions and the locus-based adjacency genetic representation. To help the DMers select the proper community partitions from the optimal candidate solution set generated in the first phase, the second phase further proposes two model selection methods: *Max Q* and *Max-Min Distance*. The experiment on the artificial and real networks show that MOCD can discover more accurate community structure compared to the three representative single-objective algorithms: the heuristic algorithm GN [5], the optimization algorithm GACD [15], and the information-theoretic framework based algorithm INFO [6]. The future work can further explore to make use of the Pareto solutions and employ other objective functions.

VI. Acknowledgments

This work is supported by the National Science Foundation of China (No. 60905025, 90924029). It is also supported 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] S.Boccaletti, V.Latora, Y.Moreno, M.Chavez and D.U.Hwang, "Complex Networks: Structure and Dynamics," *Physics Report*, 424(4-5):175-308, 2006.

[2] R.Guimera and L.A.N.Amaral, "Functional Cartography of Complex Metabolic Networks," *Nature*, 433:895-900, 2005.

[3] A.Clauset, M.E.J.Newman and C.Moore, "Finding community structure in very large networks," *Physical Review E*, vol-70:06611.

[4] A.Pothen, H.Sinmon, K-P.Liou, "Partitioning Sparse Matrices with Eigenvectors of Graphs," *SIAM J. Matrix Anal App.*, Vol-11:430-452.

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

[6] R.Martin and T.B.Carl, "An information-theoretic framework for resolving community structure in complex networks," *PNAS* 2007, 104, 7327-7331.

[7] S.Fortunato and M.Barthelemy, "Resolution Limit in Community Detection," *Proceedings of the National Academy of Sciences*, vol.104, no.1, Jan. 2, 2007.

[8] J.M.Kumpula, J.Saramaki, K.Kaski and J.Kertesz, "Limit Resolution and Multiresolution Models in Complex Network Community Detection," arXiv:0706. 2230v2, 25 Jan 2008.

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

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

[11] U.Brandes, D.Delling, M.Gaetler, et al., "On Modularity Clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, issue 2, pages 172-188 2008.

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

[13] D.W.Corne, N.R.Jerram, J.D.Knowles, and M.J.Oates, "PESA-II: Region-based Selection in Evolutionary Multiobjective Optimization," in *Proc. Genetic Evol. Comput. Conf.* 2001, pp.283-290.

[14] Y.J.Park, M.S.Song, "A Genetic Algorithm for Clustering Problem," *Proc. 3rd Annu. Conf. Genetic Program*, 1998, pp. 568-575.

[15] C.Shi, Y.Wang, B.Wu, C.Zhong, "A New Genetic Algorithm for Community Detection," Complex 09, 5(1), 1298-1309.

[16] A.Arenas, A.Diaz-Guiler and C.J.Perez-Vicente, "Synchronization Reveals Topolotical Scales in Complex Networks," Physics Review Letter, 96 114102.

[17] F.Radicchi, C.Castellano, F.Cecconi, V.Loreto, D.Parisi, "Defining and Identifying Communities in Networks," *PNAS*, vol-101:2658.

[18] L.Danon, A.Diaaz-Guilera, J.Duch and A.Arenas, "Comparing Community Structure Identification," *Journal of Statistical Mechanics: Theory and Experiments*, 9, 2005.

[19] J.Duch, A.Arenas, "Community Detection in Complex Networks using Extremal Optimization," arXiv:cond-mat/0501368, 2005.

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

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

[22] C.Pizzuti, "Community Detection in Social Networks with Genetic Algorithms," in *GECCO'08*.

[23] C.Pizzuti, "Overlapped Community Detection in Complex Networks," in *GECCO'09* 859-866.

[24] J.Handle and J.Knowles, "An Evolutionary Approach to Multiobjective Clustering," *Transaction on Evolutionary Computation*, vol.11 no. 1, 2007.

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