Exploring the Limitations of Quality Metrics in Detecting and Evaluating Community Structures

Mohsen Arab
Department of Computer Science
Yazd University, Yazd, Iran
mohsen.arab63@gmail.com

Mahdieh Hasheminezhad*
Department of Computer Science
Yazd University, Yazd, Iran
hasheminezhad@yazd.ac.ir

Received: 2017/06/20 Revised: 2018/02/08 Accepted: 2018/03/14

***Abstract*— The discovery and analysis of community structures in networks has attracted increasing attention in recent years. However there are some well-known quality metrics for detecting and evaluating communities, each of them has its own limitations. In this paper, we first deeply discuss these limitations for community detection and evaluation based on the definitions and formulations of these quality metrics. Then, we perform some experiments on the artificial and real-world networks to demonstrate these limitations. Analyzed quality metrics in this paper include modularity, performance, coverage, normalized mutual information (NMI), conductance, internal density, triangle participation ratio and cut ratio. Comparing with previous works, we go through the limitations of modularity with much more accurate details. Moreover, for the first time, we present some limitations of NMI. In addition, however it is known that performance has tendency to get high values in large graphs, we explore this limitation by its formulation and discuss several specific cases in which performance even on small graphs gets high scores.**

Keywords—Limitations of quality metrics; community detection; quality function; social networks; data mining

1. INTRODUCTION

Detection and analysis of communities in networks has attracted a lot of attention in recent years. Although, there is no well-accepted definition for a community, it is well-known that most real-world networks display community structures. Loosely speaking, a community (also called cluster , module or group) is a subset of nodes which are more densely connected to each other than with the rest of the graph. Identification of communities enables us to find groups of nodes with similar features or function. Therefore, if one knows information provided by a small part of a community, it can be simply extrapolated to other nodes of the community.

Community structures are found in wide variety of complex systems such as social networks, the Internet, food webs, biological networks, computer science, engineering, economics, politics and so on. Finding groups of related people in social networks, doing recommendations based on relations in a group in e-commerce, classifying gene expression data and studying the spread of a disease in a population in bioinformatics are some applications of community detection.

There have been devoted a lot of efforts to solve the problem of finding communities in a network. Different algorithms have been devised. An overview can be found in [1]. An important challenge is finding a way for evaluation of the quality of a partition found by an algorithm. Quality metrics (or quality functions) such as modularity [2], performance [3], conductance [4], Normalized Mutual Information (NMI) [5] etc., have been presented. Thus, the quality of a partition can be estimated in terms of the value related to that metric.

In 2006, authors of [6] showed that modularity as the most popular quality metric expresses some limitations. It has been revealed that in modularity optimization strategy, small communities may not be found. This limitation is called resolution limit. To resolve resolution limit, several multi-resolution methods have been proposed [7, 8]. These methods present modified versions of modularity with tunable resolution parameters. But, these methods have their own intrinsic limitations as well [9, 10]. In [11], Benjamin et al. showed that modularity maximization finds so many different partitions whose modularity values are very close to each other. In [12], the authors expressed that in large graphs, performance values with great probability get high scores since real-world graphs are often sparse.

In this paper, our most important contributions are as follows: we show the limitations of modularity and performance with much more accurate details using their formulations. To the best of our knowledge, limitations of NMI have not been discussed in the field so far. We will show that NMI like modularity is not scalable. Therefore, in large graphs its values approach one. Finally, we define and propose two characteristics of a good quality metric called "sensitivity to link density" and "scalability". Then we will show that while only performance has characteristic of "sensitivity to link density", none of NMI, modularity and performance are scalable.

This paper is organized as follows. Section 2 reviews previous works about the limitations of quality metrics and also our contribution in the field with more details. In section 3, some elementary definitions and conventions in the field are listed. In section 4, we review some existing measures designed to evaluate how good a particular partition of a network is. Then, in section 5, we will discuss some limitations of current measures for evaluation of community detection algorithms. In section 6, we define and present two characteristics of a good quality metric for evaluation of communities. Section 7 reports some experimental results. Finally, in section 8, conclusion is stated.

2. Related Works

In 2006, Fortunato and Barthelemy showed that modularity expresses a limitation called resolution limit[6]. That is, modularity optimization strategies may result in merging communities smaller than a scale. The authors of [6] in order to show this limitation considered two weak communities M_1 and M_2 in a network and discussed in what situations, merging these two communities cause modularity value to increase. However, they considered just two simple extreme cases. In first case, two communities were disconnected from the rest of the network and also the number of links connecting them was double of the number of internal links of each of them. In second case, there is a single link connecting them to each other and also there is a single link connecting each of them to the rest of the graph. In fact, these two simple extreme cases cannot reflect the limitation of modularity for merging small communities very well. Instead, in this paper, for merging two corresponding weak communities, we consider a general case. Then using formulations and illustrations we will show that how increasing the number of links of the rest of the graph can cause modularity optimization to fail in detecting small communities. This is done by finding a lower bound for the number of links between two communities under which modularity maximization merge them.

To resolve resolution limit, several multi-resolution methods have been proposed [7, 8]. These methods present modified versions of modularity with tunable resolution parameters. In fact, using this parameter, one can set the size of communities to arbitrary values, from very large to very small. But, these methods have their own intrinsic limitations as well [9, 10]. In fact, the authors of [9] discussed that multi-resolution modularity is not capable of detecting right partition in practical application.

In [11], Benjamin et al. showed that modularity maximization results in detecting so many partitions whose modularity values are very close to absolute maximum value of the measure, but they may be topologically quite different from each other. They discussed that if a network is sparse and the number of communities approaches infinity, by considering the resolution limit of modularity, one can say that modularity value approaches one. In this paper, in accordance with this finding, we obtain similar but more accurate results. That is, we first show that if every community is smaller than a scale of the graph size, then we can find a lower bound for maximum value of modularity. Then, we will show that by merging small communities of a partition, there will be either of these two cases for modularity value of the resulted partition: 1) it increases. 2) It decreases very little.

In [12], the authors expressed that in large graphs, performance values with great probability get high scores since real-world graphs are often sparse. In this paper, we deeply go through this limitation using the formulation of performance. In fact, using both its definition and also running some tests on both artificial and real-world graphs we will show that even in small graph this characteristic exists.

3. E lementary Definitions

Throughout of this paper, let $G = (V, E)$ represent a connected, undirected, and un-weighed graph where V is the set of nodes and E is the set of all edges of G. Let $n = |V|$, $m = |E|$. Let also $P = \{C_1, C_2, \dots, C_p\}$ be a partition of graph into p communities. Also note that in this paper we repeatedly use words graph and network instead of each other.

Suppose d_v is the degree of node v. Let d_v^{in} and d_v^{out} be the number of neighbors of v within and outside of its community, respectively. Suppose $E(C_i)$ is the number of edges inside the community C_i and $Ext(C_i)$ is the number of edges connecting community C_i to the rest of the graph. $E(C_i)$ and $Ext(C_i)$ are also called the number of intra-community and inter-community edges of community C_i respectively. Then:

$$E(C_i) = \frac{\sum_{v \in C_i} d_v^{in}}{2}, \quad Ext(C_i) = \sum_{v \in C_i} d_v^{out} . \tag{1}$$

For a set C of nodes, $E(C)$ and $Ext(C)$ are defined similarly.

For the sake of convenience, for any partition P of a graph, the number of intra- and inter-community links is denoted by m_{in} and m_{out} respectively. In other words

$$m_{in} = \sum_{C_i \in P} E(C_i) = \frac{\sum_{v \in G} d_v^{in}}{2} \tag{2}$$

And

$$m_{out} = \frac{\sum_{C_i \in P} Ext(C_i)}{2} = \frac{\sum_{v \in G} d_v^{out}}{2} \tag{3}$$

Let $m_r = m_{in}$, if the corresponding partition is real partition of the network. In this paper, C_i means the i'th community, but *c(i)* is referred as the community containing node i. The terms *n(c(i))* and $n(C_i)$ denote the number of nodes inside the communities $c(i)$ and C_i respectively. Finally, for simplicity in writing, instead of using $n(C_i)$, $E(C_i)$ and $Ext(C_i)$, the parameters n_i, E_i and Ext_i are used.

Definition of community

In fact the first problem in network partitioning is how to define exactly what a community is. However there are so many definitions for the concept of a community, there exists no universally accepted one. Some authors classified these in three classes of definitions: local, global and based on node similarity [1,13]. Exploring local definition in more depth, leads us to two subclasses: self-referring and comparative definitions [13,14].

The first subclass requires only a set of nodes and the relations between them to decide whether or not call the set a community. That is, a community is defined only in reference to itself. The simplest one is a clique, i.e. a subset in which there is an edge between any two nodes. As clique is very hard to satisfy in reality, some softer definitions have been presented such as: *n*-clique, *n*-clan, *n*-club and *k*-plex and *k*-core [15,16]. An *n*-clique is a maximal sub-graph in which the largest geodesic distance between any two nodes is no greater than n . An *n*-clan is an *n*-clique in which the largest geodesic distance between any

two nodes is no greater than n considering only the paths within the sub-graph. *n*-club is defined as maximal sub-graph with diameter n. *k*-plex and *k*-core is defined based on nodal degree.

On the other hand, comparative definition comes from the intuitive notion that a community will be denser in terms of edges than its surroundings. In fact, comparative definitions lend them-selves much more easily to search for communities in large complex networks. A group U is called a strong community, if for each member v of U: $v_v^{in} > v_v^{out}$. U is said to be a weak community if $\sum_{v \in U} d_v^{in} > \sum_{v \in U} d_v^{out}$ [16].

4. Principle Presented Quality Metrics

In the following some of the most important quality metrics for measuring the quality of a network partition is presented.

4-1 Modularity

Modularity[2] as the most well-known quality metric is introduced by Newman and is formulated as follows

$$Q = \sum_{C_i \in P} e_{ii} - {a_i}^2 \quad (4)$$

where e_{ii} is the real fraction of edges inside community C_i , whereas a_i^2 is the expected fraction of edges within the community C_i if one redistributes edges randomly in communities. Therefore, based on above definition, in random graphs and also in the networks in which found communities are different from its original structure, modularity has low value.

4-2. Performance

Performance (Perf)[3] of a partition P is defined as

$$Perf(P) = \frac{|\{\{u,v\} \in E,\ c(u)=c(v)\}| + |\{\{u,v\} \notin E,\ c(u) \neq c(v)\}|}{n(n-1)/2} \quad (5)$$

or more simply as

$$Perf(P) = \frac{m_{in} + f(m_{out})}{n(n-1)/2} \quad (6)$$

where $f(m_{out})$ is defined as follows

$$\frac{\sum_{1 \le i,j \le p,\ i \neq j} n_i n_j}{2} - m_{out} \quad (7)$$

$Perf(P)$ counts the number of correctly interpreted pairs of nodes in the partition *P*, i.e. two adjacent nodes belonging to the same community, or two non-adjacent nodes belonging to different communities [1, 3, 12, 17]. By definition: $0 \le Perf(P) \le 1$.

One drawback of this measure is that in the large sparse networks, there is a great possibly that the number of nonadjacent nodes which belong to different communities, becomes so high. Therefore, $Perf(P)$ might be biased to high scores in such networks.

4-3. Coverage

The coverage(*P*) [1,12,17] of a partition P of a graph is the ratio of the number of intra-community edges to the total number of edges:

$$coverage(P) = \frac{m_{in}}{m} \quad (8)$$

By definition, the coverage of a partition whose all communities are disconnected from each other is 1, since all edges fall inside communities. Intuitively, the greater the value of coverage, the better the quality of partition. It is worthy to note that, whereas min-cut is not essentially a good partition, it has the maximum coverage value. As a result, additional constrains such as the number of communities, etc., seems to be essential to be considered in order to obtain a good partition.

4-4. Normalized Mutual Information

Danon et al. used a measure borrowed from information theory called Normalized Mutual Information (or *NMI*) [5] to evaluate quality of community structures:

$$\frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B} N_{ij} \log(\frac{N_{ij}N}{N_i N_j})}{\sum_{i=1}^{c_A} N_i \log(\frac{N_i}{N}) + \sum_{j=1}^{c_B} N_j \log(\frac{N_j}{N})} \quad (9)$$

This measure is based on definition of confusion matrix *N*, where rows corresponds to real communities and columns with found communities. Each element of this matrix, N_{ij} is the number of nodes in real community i that has been assigned to the community j by a community detection algorithm.

As it is obvious from this definition, real community of each node should already be known. In the other words, despite the other previously presented quality criteria, this measure cannot be applied on real-works networld, since real community structures are un-known.

4-5. Conductance

Conductance [12,17,18] of a cut $(C, V \backslash C)$ in a graph is defined as follows

$$\emptyset(C) = \frac{Ext(C)}{\min(\sum_{v \in C} d_v, \sum_{v \in V \backslash C} d_v)} \quad (10)$$

It compares the size of a cut (i. e., the number of edges of cut) in either of the two induced sub-graphs.

The conductance $\emptyset(G)$ of a graph G, is the minimum conductance value over all cuts of G.

$$\emptyset(G) = min_{C \subseteq V}\, \emptyset(C) \quad (11)$$

The intra-community conductance $\alpha(P)$ is the minimum conductance value over all induced sub-graphs *G(Ci)* :

$$\alpha(P) = min_{C_i \in P}\, \emptyset(G(C_i)) \quad (12)$$

Low value of intra-community conductance indicates the existence of at least one community which is too coarse. Inter-community conductance $\delta(P)$ is the complement of the maximum conductance value over all induced cuts *(Ci ,V\Ci)* . More formally:

$$\delta(P) = 1 - max_{C_i \in P}\, \emptyset(C_i) \quad (13)$$

Lower values of inter-community conductance might be as a sign of existence of strong connection between at least one community and the rest of the graph. Therefore, a partition is good if it has both high values of intra- and inter-community conductance at the same time [12] .

4-6. Internal Density

Internal density [19] for a set C of nodes is the ratio between the number of internal edges of C and the number of all possible internal edges:

$$f(C) = \frac{m(C)}{n(C)(n(C)-1)/2} \quad (14)$$

4-7. Triangle Participation Ratio

Triangle Participation Ratio(TPR) [19] is simply the fraction of nodes of C that participate at least in one triad whose nodes are completely in C.

$$f(C) = \frac{|\{u:u\in C,(v,w):v,w\in C,(u,v)\in E,(u,w)\in E,(v,w)\in E,E\neq\emptyset\}|}{n(C)} \quad (15)$$

4-8. Cut Ratio

Cut ratio [19] is the fraction of existing edges out of all possible edges connecting the set to the rest of the network:

$$f(C) = \frac{Ext(C)}{n(C)(n-n(C))} \quad (16)$$

5. Limitations of Quality Metrics

In this section the limitations of the most well-known quality metrics for community detection and evaluation are discussed.

5-1. limitations of Modularity

One of the most popular criteria for computing the quality of a network partition is modularity (Q) which is introduced by Newman [2]. Modularity is defined as

$$Q = \sum_{C_i \in P} e_{ii} - {a_i}^2 \quad (17)$$

where

$$e_{ii} = \frac{E_i}{m} \quad (18)$$

and also

$$a_i = \frac{\sum_{v\in C_i} d_v}{\sum_{v\in G} d_v} = \frac{2E_i+Ext_i}{2m} \quad (19)$$

In [18], equivalently, modularity is formulated as

$$\frac{m_{in}}{m} - \frac{1}{4m^2}\sum_{C_i\in P}(\sum_{v\in C_i} d_v)^2 \quad (20)$$

In [20], it is shown that after merging two communities C_i and C_j, the change in modularity value equals to

$$\Delta Q = \frac{E_{ij}}{m} - 2a_i a_j \quad (21)$$

where E_{ij} is the number of links connecting two communities C_i and C_j.

In the following one important drawbacks of modularity, i.e. merging small communities in big networks is expressed.

If one accept modularity as a measure for evaluating quality of a partition found by a community detection algorithm, thus whatever the modularity is more close to one, the better the quality of found partition will be. One of the limitation of modularity is in large networks where modularity maximization strategy tends to merge small communities despite of lacking enough relationship in terms of number of links connecting them.

In what follows limitation of modularity maximization strategy in large networks will be discussed, by using modularity formula and also the condition under which merging two communities will increase modularity. Suppose that there are two communities C_1 and C_2 connected together with some links. Let E_1 and E_2 be the number of links within communities C_1 and C_2 respectively. Let V_1 and V_2 define the set of nodes inside the two communities. Also let E_{12} be the number of links connecting them.

Let G_r be the induced sub-graph $G[V\backslash\{V_1 \cup V_2\}]$. Let also the number of links inside G_r be denoted by E_r. Also suppose that the numbers of links connecting communities C_1 and C_2 with G_r are referred as $E_{1,r}$ and $E_{2,r}$ respectively (See the Fig. 1).

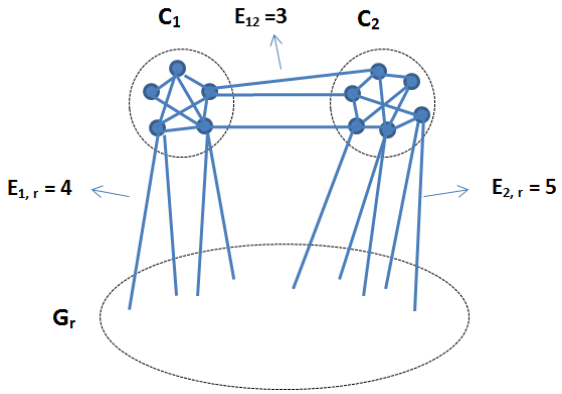

Fig. 1- An sample scheme for merging two communities

Let us define four parameters x , y , a and b as follows. $x = E_2/E_1, y = E_r/E_1, a = E_{1,r}/E_1, b = E_{2,r}/E_1$. Equivalently:

$$E_2 = xE_1,\ E_r = yE_1,\ E_{1,r} = aE_1,\ E_{2,r} = bE_1 \quad (22)$$

without loss of generality, one can assume that $E_2 \geq E_1$, i.e., $x \geq 1$. Suppose that ΔQ denotes the change in modularity value, if two communities C_1 and C_2 are merged.

By using (19) and (21) , ΔQ is positive, if and only if

$$\frac{E_{12}}{m} - 2\times\frac{2E_1+Ext_1}{2m}\times\frac{2E_2+Ext_2}{2m} > 0 \quad . \quad (23)$$

With setting $Ext_1 = aE_1 + E_{12}$, $Ext_2 = bE_1 + E_{12}$, $m = (1+x+y+a+b)E_1 + E_{12}$ and also with some simple computations, the condition under which ΔQ is positive can be written as

$$(a+b+2y)E_1E_{12} + {E_{12}}^2 > (2+a)(2x+b){E_1}^2 + (a+b)E_1E_{12} \quad (24)$$

With adding term $(\frac{a+b}{2}+y)^2{E_1}^2$ to two sides of above inequality, it can be simplified as

$$(E_{12} + (\frac{a+b}{2}+y)E_1)^2 > ((2+a)(2x+b) + (\frac{a+b}{2}+y)^2){E_1}^2 \ . \quad (25)$$

Finally, from that, the following relation for have $\Delta Q > 0$ is clearly obtained

$$E_{12} > (\sqrt{(2+a)(2x+b) + \left(\frac{a+b}{2}+y\right)^2} - \left(\frac{a+b}{2}+y\right))E_1 \ . \quad (26)$$

Suppose that $y \geq x$, i.e. G_r has more links than each of these two communities. Let consider the two communities to be weak communities. Therefore, $Ext_1 < 2E_1$ and $Ext_2 < 2E_2$ from which two upper bounds for a and b, i.e. $a \leq 2$ and $b \leq 2x$ is obtained. By supposing x and y as fixed numbers, let call right part of inequality (26) as $f(a,b)$. Therefore, the above inequality can be simplified as $E_{12} > f(a,b)$. If one can find an upper bound for $f(a,b)$ denoted by T, i.e., $T \geq f(a,b)$ for all values of a and b in their ranges, then the relation $E_{12} > T$ is one sufficient condition for satisfying the inequality (26) and in turn having $\Delta Q > 0$. It can simply be proved that one candidate for T is $f(a=2, b=2x)$. Therefore, the following relation $E_{12} > f(a=2, b=2x)$ is one sufficient condition for having $\Delta Q > 0$ after merging the two communities:

$$E_{12} > (\sqrt{16x + (y+x+1)^2} - (y+x+1))E_1 \qquad (27)$$

Let simplify above relation as $E_{12} > zE_1$. Therefore, $E_{12} > zE_1$ is a sufficient condition for having $\Delta Q > 0$. Let consider $E_1 = E_2$, i.e. $x = 1$, for simplicity. In this situation, one sufficient condition for which ΔQ is positive can be more simplified as

$$E_{12} > (\sqrt{16 + (y+2)^2} - (y+2))E_1 \qquad (28)$$

Now, in this situation if $y = 10$, $y = 20$, $y = 50$ and $y = 100$, then $E_{12} > 0.65E_1$, $E_{12} > 0.37E_1$, $E_{12} > 0.16E_1$ and $E_{12} > 0.08E_1$ are sufficient conditions respectively in order to have ΔQ positive (see Fig. 2).

If a=0 and b=0, that is if these two communities are disconnected from the rest of graph, the sufficient condition (27) for having $\Delta Q > 0$ can be more simplified as

$$E_{12} > (\sqrt{4 + y^2} - y)E_1 \qquad (29)$$

Now, one sufficient condition to witness positive ΔQ can be more easily satisfied. That is, if $y = 10$, $y = 20$, $y = 50$ and $y = 100$, then $E_{12} > 0.2E_1$, $E_{12} > 0.1E_1$, $E_{12} > 0.04E_1$ and $E_{12} > 0.02E_1$ are sufficient conditions respectively in order to have $\Delta Q > 0$.

Fig. 3 shows the limitation of modularity maximization for large graphs for $1 \leq x \leq 5$ and $1 \leq y \leq 100$. By fixing x, when y increases, the sufficient condition for merging two corresponding communities ($E_{12} > zE_1$) is more easily satisfied. As it is seen from (27), (28) and (29) and Fig. 3, modularity maximization in large networks will merge small communities together despite existing enough relationships between them.

If for each community C_i, there exist an integer f such that $\sum_{v \in C_i} d_v \leq \frac{2m}{f}$, thus using (20) it is not very hard to obtain the following lower bound for modularity

$$Q \geq \frac{m_{in}}{m} - \frac{1}{f} \qquad (30)$$

If all communities are weak and for each community C_i: $E_i \leq m/(2f)$, then the relation $\sum_{v \in C_i} d_v \leq 2m/f$ and in turn the above lower bound holds. For example if all found communities are weak and also each community consists

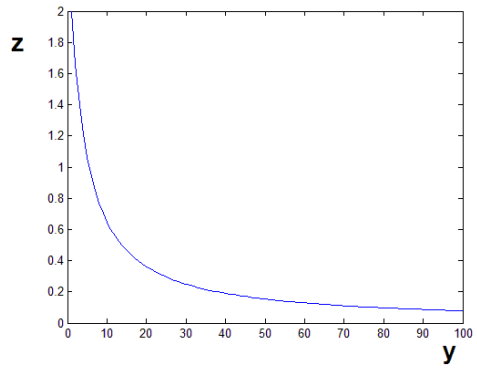

Fig. 2- Limitation of modularity, when $E_1 = E_2$ (x=1).

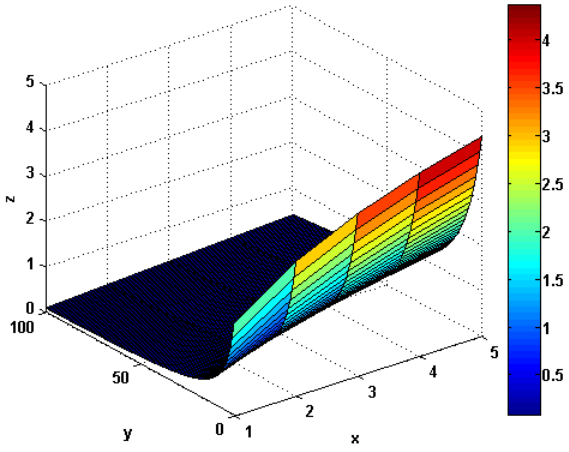

Fig. 3- Limit of modularity maximization.

of at most one percent of all edges of the graph (i.e. *f=50*), then $Q \geq \frac{m_{in}}{m} - 0.02$.

Suppose that Q_r and Q_f indicate the modularity values of real and found partitions of a graph respectively. It is clear that by merging some weak communities, the resulted community will be weak as well. Therefore, if real communities are weak and one can find an integer f such that for each community $C_i: E_i \leq m/(2f)$, then by merging real small communities, as long as the relation $E_i \leq m/(2f)$ holds for each found community, one of the two following cases holds: $Q_f > Q_r$ or $Q_r - \frac{1}{f} \leq Q_f \leq Q_r$. This is because in this case: $Q_r \geq \frac{m_r}{m} - \frac{1}{f}$, $Q_f \geq \frac{m_{in}}{m} - \frac{1}{f}$ and $m_{in} \geq m_r$. Therefore, in this case, by merging real weak communities, the relations $Q_r - \frac{1}{f} \leq Q_f \leq 1$ and $\frac{m_r}{m} - \frac{1}{f} \leq Q_f \leq 1$ hold.

In large graphs one can find larger values of such *f*. Whatever real communities have higher internal relationships and lower external ones, then lower bound of Q_f, i.e. $\frac{m_r}{m} - \frac{1}{f}$, will be more close to one. This shows that by modularity maximization, one may find so many partitions with high modularity values but with very different structures at the same time.

5-2. Limitation of Performance

First let investigate in what condition performance value will increase after merging two communities C_1 and C_2. The performance values before and after merging are denoted by $Perf(P_1)$ and $Perf(P_2)$ respectively. The following are straightforward based on the definition of performance:

$$Perf(P_1) = \frac{E_1 + E_2 + E_r + n_1n_2 + n_2n_r + n_1n_r - (E_{12} + E_{1,r} + E_{2,r})}{n(n-1)/2}$$

$$Perf(P_2) = \frac{E_1+E_2+E_r+E_{12}+(n_1+n_2)n_r-(E_{1,r}+E_{2,r})}{n(n-1)/2} \quad (31)$$

The change in performance value after merging, equals to

$$\Delta Perf = Perf(P_2) - Perf(P_1) = \frac{2E_{12}-n_1n_2}{n(n-1)/2} \quad (32)$$

Now, the condition under which merging two communities results in higher performance value can be stated as

$$E_{12} > \frac{n_1n_2}{2} \quad (33)$$

That means that there should exist more than half of maximum possible links between two communities, in order to $\Delta Perf > 0$ when combining them together which is very hard to be satisfied in reality.

Therefore, unlike modularity , when using performance as quality metric, sub-graph G_r have not any role in deciding whether or not combining these two communities cause this measure to increase. Thus, different from modularity maximization, using performance maximization in large graphs one cannot expect to witness some drawbacks such as merging small communities without enough relation .

The next thing to consider is the range of values which performance can get . The following relation is straightforward

$$f(m_{out}) = \frac{n(n-1)}{2} - m_{out} - \sum_{C_i \in P} \frac{n_i(n_i-1)}{2} \quad (34)$$

With replacing above formula in (6) and also with using (2) and (3) the following is obtained.

$$Perf(P) = 1 - \frac{\left(\sum_{C_i \in P} {n_i}^2\right)-n}{n(n-1)} + \frac{\sum_{v\in G} d_v^{in} - \sum_{v\in G} d_v^{out}}{n(n-1)} \quad (35)$$

Suppose that for each community C_i , the relation $n_i \le \frac{n}{f}$ holds for some integer number f. In this situation,

$$\sum_{C_i \in P} {n_i}^2 \le \frac{n^2}{f} \quad (36)$$

Therefore, following relation holds:

$$Perf(P) > 1 - \frac{\frac{n^2}{f}-n}{n(n-1)} + \frac{\sum_{v\in G} d_v^{in} - \sum_{v\in G} d_v^{out}}{n(n-1)} \quad (37)$$

Since $\frac{\left(\frac{n^2}{f}-n\right)}{n(n-1)} < \frac{1}{f}$ for each positive integer f, then the following simple result clearly hold

$$Perf(P) > 1 - \frac{1}{f} + \frac{\sum_{v\in G} d_v^{in} - \sum_{v\in G} d_v^{out}}{n(n-1)} \quad (38)$$

In case one, suppose that every community C_i is a weak community. In this case, based on definition of comweak munities the following holds:

$$Perf(P) > 1 - \frac{1}{f} . \quad (39)$$

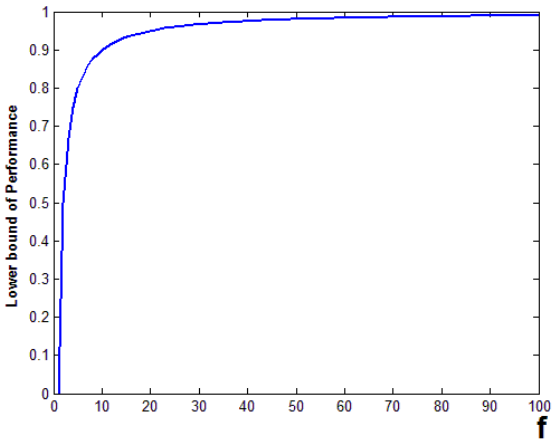

Fig. 4- Limitation of performance.

It is obvious that by merging several weak communities together, the resulted community is still a weak community. This indicates that as long as merging small real weak communities results in forming a community C_i with $\frac{n(C_i)}{n} \le \frac{1}{f}$, (39) will hold.

Therefore, one might find so many partitions whose com munities consist of several small real communities with just one mentioned condition in size and still performance value is greater than $1 - \frac{1}{f}$. This lower bound of performance value based on values of f is displayed in Fig. 4 .

In case two, suppose that there is just one restricting condition $\frac{n(C_i)}{n} \le \frac{1}{f}$ for each community. In this case, communities are not necessarily weak. The following relation clearly holds:

$$\sum_{v\in G} d_v^{in} - \sum_{v\in G} d_v^{out} > -2m \quad (40)$$

Also as $m = n\frac{D}{2}$ (D is average degree of network) , so:

$$Perf(P) > 1 - \frac{1}{f} - \frac{D}{n-1} \quad (41)$$

Therefore, for a sparse network with $n = 1000$ and $D = 20$, and with the condition $\frac{n(C_i)}{n} \le \frac{1}{20}$ for each community, Perf(P) > 0.93 which is extremely high. Suppose that P_s indicates the partition in which each community consists of one single node. In this case, $m_{in} = 0$ and $m_{out} = m$. Therefore, $f(m_{out}) = \frac{n(n-1)}{2} - m$. Performance of this partition equals to $Perf(P_s) = 1 - \frac{2m}{n(n-1)}$. Since $m = n\frac{D}{2}$, the following is straightforward:

$$Perf(P_s) = 1 - \frac{D}{n-1} \quad (42)$$

For real-world networks which are sparse ($D \ll n$), $Perf(P_s)$ is very close to one. For example, if n =1000 and D =20, $Perf(P_s)$ is 0.98 which is very high. Increasing n cause performance value to get more close to one. Therefore, high values of performance does not necessarily indicate the goodness of a partition.

5-3. Limitations of NMI

Consider the real partition $P = \{C_1, C_2, \dots, C_p\}$ of a network into p communities. Let consider two cases:

Case 1: splitting communities. First suppose that a community detection algorithm Alg_1 has divided each real community C_i into s equal-size sub-communities $C_{i,1}$, $C_{i,2}$,..., $C_{i,s}$. That is $(C_{i,1}) = n(C_{i,2}) = \cdots = n(C_{i,s})$. Let define I as the set consisting of indices of real communities, i.e. $I = \{1,2, \dots, p\}$. Corresponding NMI value for this case equals to

$$\frac{2N\log(N) - 2\sum_{i\in I} n_i \log(n_i)}{2N\log(N) - 2\sum_{i\in I} n_i \log(n_i) + N\log(s)} \quad (43)$$

In the above equation, it is not very hard to show that by fixing s,when p (number of communities) increases, NMI value increases as well. This can be more clear to understand this fact if the sizes of real communities are considered to be equal which means $n(C_1) = n(C_2) = \cdots = n(C_p)$. By this supposition, each N_{ij} elements equals to $n/(ps)$ and also each N_i and N_j elements equal to n/p and $n/(ps)$ respectively. NMI value for this sub-case is

$$\frac{2\log(p)}{2\log(p) + \log(s)} \quad (44)$$

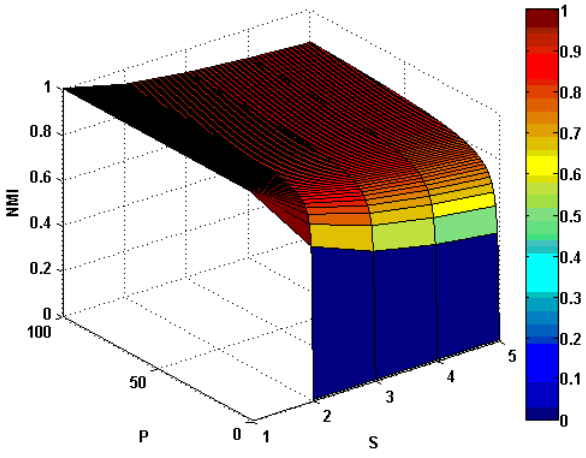

Fig. 5- Limitation of NMI for splitting case.

In above example it can be clearly understood that by having fixed s, if one increases p (again, number of communities of the network), NMI value increases (see Fig. 5). Fig. 6 demonstrates equation (44) when $S = 2$. This is actually a drawback of NMI to see this behavior. To see why this is actually a drawback, let consider partition P_1 consisting of one arbitrary community.

Let make R copies of this community to form a partition P_2 containing R separated similar communities. Most famous community detection methods such as modularity maximization and label propagation algorithm [16] and so on, follow the same behavior on each community of P_2 as on the one single community of P_1. For example, if they divide the single community of partition P_1 into w sub-communities, then they will follow the same approach on each community of partition P_2.

In fact intuitive idea tells us to do that as well. But *NMI* for partition P_2 sets a higher value as a sign of better accuracy in community detection because of just increasing number of communities of partition P_2, not based on better quality of detected communities.

Case 2: merging communities. Let $n(C_1) = n(C_2) = \cdots = n(C_p)$ hold for each community of real partition. Suppose that each detected community of the found partition of algorithm Alg_2 consists of k communities of the real partition. *NMI* value of this case is as follows

$$\frac{2\log(p) - 2\log(k)}{2\log(p) - \log(k)} \quad .(45)$$

Similar to previous case, in this case, by fixing k and increasing number of the real communities one can get higher *NMI* values (see Fig. 7 for $k = 2$).

5-4. Limitations of Coverage

This is very obvious that this measure cannot be a good quality metric neither for finding communities nor for evaluating the detected communities. Assigning all nodes of a network as one big communities leads to having maximum Coverage value. Therefore, some additional information such as the number of communities and so on is needed for detecting communities. As in reality such information does not exist about the real communities, this measure cannot be helpful for community detection.

5-5. Limitations of Internal Density

The most important drawback of this quality metric is that it just considers internal relations between nodes of a community without any attention to external ones. Consider a partition whose communities consist of pairs of nodes connected with an edge. This quality metric evaluates this partition with value "one" as the best possible partition, without any regard to their external relation to the rest of the network. So, it is necessary to use also those quality metrics which take into account external relations of communities. That is, in addition to high internal density, a community should have low external ones. But even in this case an import issue will be how to participate both of these two measures in final formula.

5-6. Limitations of Cut Ratio

Despite other previously defined quality metrics, low values of Cut Ratio indicate better community detection. Cut Ratio has two main drawbacks. The first one is that it just considers external connections of each community. The second one is that because of large value of its denominator it usually gets low values which cannot reflex correctly the external strength of communities.

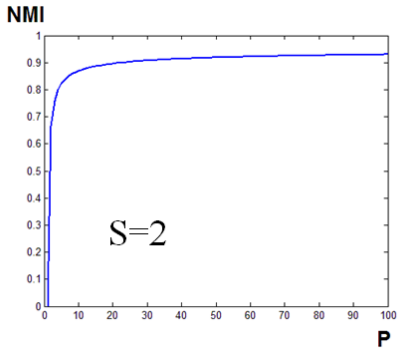

Fig. 6- Limitation of NMI for splitting case when s=2.

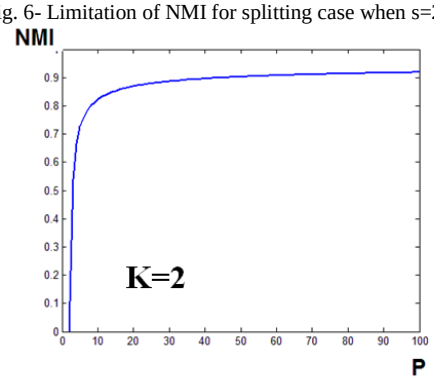

Fig. 7- Limitation of NMI for merging case when k=2.

5-7. Limitations of Triangle Participation Ratio

The first limitation of this measure is its non-decreasing behavior. This means that when merging real communities together, as even nodes from different communities may share some common friends, this measure can increase. Second limitation of this measure can be the fact that in sparse networks with low density of edges, this quality metric may not be very helpful. Because this measure is mainly based on this intuitive idea that nodes inside communities, because of high internal densities of edges, share some common friends together.

5-8. Limitations of Conductance

Similar to other quality metrics, conductance has several limitations as well. First limitation of conductance is that despite all other previously defined quality metrics, computing intra-community conductance, i.e. $\alpha(P)$, is NP-Hard. Because, for each community C_i, all cuts of induced sub-graphs $G(C_i)$ should be considered for computing conductance of $G(C_i)$. This cannot be done in polynomial time. This property makes this measure impractical to be used in reality for evaluation of community structures.

The second limitation of conductance is that both of intra-community conductance and inter-community conductance do not give us much information about a partition if their corresponding values are low. For example, if in a partition there is just one community consisting of two or more sub-communities with weak connections between them, intra-community conductance $\alpha(P)$ will be low. In this case, no matter how many communities have such similar internal connections, $\alpha(P)$ is low (see Fig. 8). On the other hand, low value of inter-community conductance $\sigma(P)$ do not give us any information about this fact that how many communities have strong relationships with the rest of network. In this case it just tells us that there is at least one such community (see Fig. 9). .

6. PROPERTIES OF A GOOD QULITY METRIC

In this section, we define and present two properties of a good quality metric called σ.

1. Sensitivity to link density: Comparative definition of a community tells that a community is a set of nodes with higher internal relationships than external ones. Therefore, one can expect that either by increasing (decreasing) the internal link number or decreasing (increasing) the external link number of a community, the quality of that community and in turn the quality of corresponding partition boosts (drops). Let call this behavior as sensitivity to link density. From all previously mentioned quality metrics, only performance has this characteristic. For modularity, one can find examples where this characteristic does not hold. As an example, consider a partition $P_1 = \{C_1, C_2, C_3\}$ including three disconnected communities where each community has one internal link. Suppose that one community is chosen and one link in added to that community. Let call this new partition as P_2. The modularity values corresponding to partition P_1 and P_1 equals to $Q(P_1) = 0.666$ and $Q(P_2) = 0.625$ respectively. Therefore, while one expects that the inequality $Q(P_2) > Q(P_1)$ should hold, this is not the case. Since NMI does not consider links, it has not this characteristic as well.

As an special case of this characteristic, σ should get its optimum value on the best possible partition, i.e. the partition P_{d-c} consisting of disconnected cliques. But, as it can be inferred clearly from (20), modularity value would not equal to one in any possible situation. But, performance will get its maximum value one on the partition P_{d-c}.

2. **Scalability:** This trait indicates that with increasing the size of a graph, the accuracy of a quality metric for the evaluation of network partitions should not change. For example, in the simplest case, let consider a partition

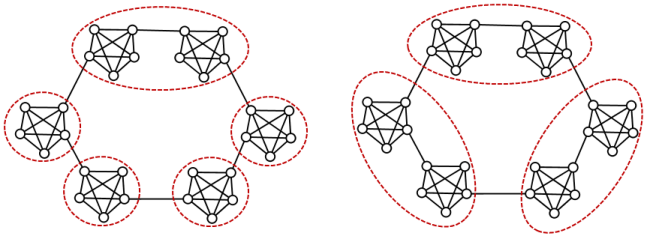

Fig. 8- Two different partitions of a network with equal intra-community conductance.

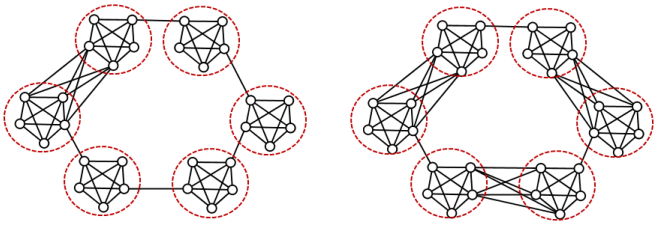

Fig. 9- Two different partitions of two networks with equal inter-community conductance.

$P_1 = \{C_1, C_2, \dots, C_p\}$ including p disconnected communities such that $E_1 = E_2 = \dots = E_p$. Increasing the number of community to $2p$ results in a partition called P_2. It is obvious that the quality of the communities in two partitions is the same, however using
either self-referring or comparative definitions of a community leads us to this conclusion as well. Modularity and Performance are not scalable on big graphs, since they evaluate partition P_2 higher than P_1. Modularity values on P_1 and P_2 are $1-\frac{1}{p}$ and $1-\frac{1}{2p}$ respectively. In addition to these two quality metrics, as it has been discussed earlier, NMI is not scalable as well, since its values approaches one with increasing just the number of communities.

7. Experiments

In this section, we are going to run some experiments to analyze how different quality metrics evaluate communities.

7-1. Artificail networks

For this purpose GN benchmark [21] is chosen. GN benchmark is a well-known artificial benchmark for community detection which is called Girvan-Newman benchmark (or GN benchmark). This network consists of 128 nodes with the equal degree $d=16$ for each node. There exists four communities which each of them has 32 nodes. Each node makes d_{in} edge connections to other nodes in its community randomly. The remaining d_{out} ($d_{out} = d - d_{in}$) edges will also be selected randomly with the other communities. The ratio d_{out}/d is called mixing parameter and is denoted by μ. As μ increases, the communities will be more difficult to detect.

At first step, starting from single nodes as singleton communities, from each community, a random set of K nodes called sub-community C is selected in order to evaluate quality metrics on them. k starts from 1 to 32, i.e. number of nodes inside each community of GN. The random selection of this set is carried out 50 times for each K on GN benchmark in order to increase accuracy. In this step, let define $S = n(C)/32$. After that, merging real communities is carried out in step two. In step two, at first each pair of communities are merged together. These pairs are chosen randomly as well. This is shown by setting $S = 2$. After that, all four communities are merged together and the quality metrics on the whole network is evaluated. In this case, $S = 4$.

In Fig. 10, evaluations of different quality metrics on GN benchmark can be seen. Performance (P), modularity (**Q**), internal density (ID), triangle participation ratio (TPR), cut ratio (CR) and inter-community conductance (C). As in cut ratio, low values indicates better quality of detected communities, for having a consistent evaluation, let define and use star version of this measure, i.e. cut ratio(*)=1-cut ratio. Therefore, in Fig. 10, CR (*) indicate cut ratio (*).

A good quality metric q should be such that with increasing S, it should increase until $S=1$. Because in this stage, by increasing S, sub-communities are getting larger and thus quality of detected communities becomes better. After this point, whatever real communities are merged together, q should follow a decreasing trend. Therefore, a good quality metric should have increasing trend for $S \leq 1$ and decreasing one for $S \geq 1$. As it is obvious from Fig. 10, modularity follows this pattern very well for GN benchmark. Performance and internal density has approximately fixed value for $S \leq 1$ and a decreasing trend for $S \geq 1$. Other quality metric have not good results. Note that, however in Fig. 10, $\mu=0.1$ but for $\mu=0.2$, $\mu=0.3$ and $\mu=0.4$, the results are similar (see Fig. 11 for μ=0.4).

7-2. Real-world networks

In this subsection, we are going to evaluate the accuracy of different quality metrics on real-world networks. In this section, we chose three following real-world networks:Zachary's karate club, dolphin and football. For each of these three networks, starting from small communities they will be merged in repetitive process until real communities are obtained. The primitive small communities are formed by assigning pairs of nodes with highest ratio of common friends into the same community. In merging process, in each repetition, at first all communities are unmarked. Then, in that repetition, for each community C_i, a community C_j with maximum number of links between them is selected. If both of them are unmarked, then they are marked with number C_i to be merged at the end of that repetition. For this section only

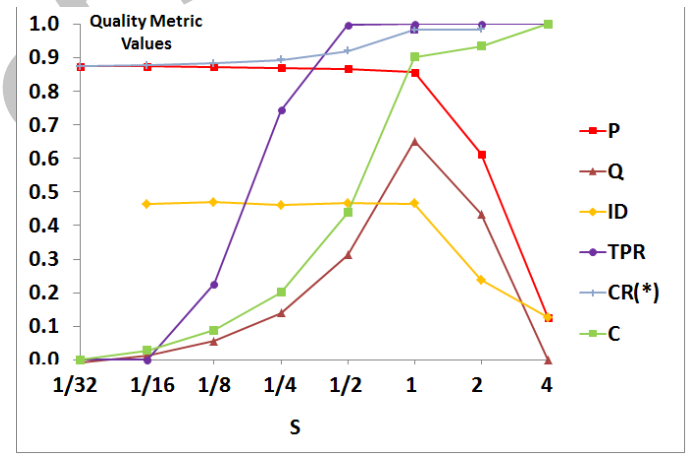

Fig. 10- Quality metrics values on GN benchmark for μ=0.1.

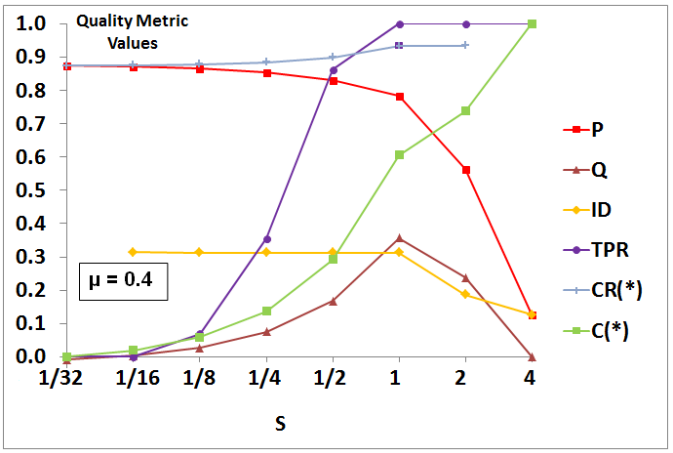

Fig. 11- Quality metrics values on GN benchmark for μ=0.4.

NMI, performance and modularity are tested, since the other quality metrics, as discussed earlier, have monotonic trend which does not reflect the quality of detected partitions.

Forming small primitive communities and also merging process are followed such that just sub-communities from the same real community can be merged. In this paper, in the illustrations and graphs, repetition one is associated with the result of detected primitive sub-communities. The next repetitions are the results of merging process.

Zachary's Karate Club: This network represents the friendship between 34 members of a karate club. After a conflict between club's administrator (node 34) and the club's instructor (node 1), the instructor leaved the club and started a new one with taking about the half of the original club's members with him. The resulting two groups can be considered as ground truth communities of this network for testing the accuracy of different community detection algorithms. These two original communities are specified with square and circle in Fig. 12.

In Fig. 13, the result of merging process of sub-communities of karate network is illustrated. For each repetition R, three values of NMI , performance (P) and modularity (Q) are displayed. Unlike performance, modularity values follow the increasing pattern of NMI which indicates truly the better quality of detected communities in next repetitions.

However modularity values have increasing behavior in Fig. 13, but maximizing modularity using simulated annealing (Sim. Ann.) finds four communities on this network, instead of two real communities. The detected communities using Sim. Ann. method are illustrated with different colors in Fig. 12

Dolphin network: This network displays the statistically significant frequent association between 62 bottlenose dolphins in Doubtful Sound, New Zealand. This network has two original communities which are specified with circle and square in Fig. 14.

The results of hierarchical pairwise merging of sub-communities are displayed in Fig. 15. As it can be seen, after 14 repetitions, two real communities are recovered by getting $NMI = 1$. The array Psize = [37, 26, 21, 17, 14, 12, 10, 8, 7, 6, 5, 4, 3, 2] indicates the sizes of resulted partitions in each repetition R. Thus, starting from 37 detected primitive sub-communities, in second and third repetition, 26 and 21 sub-communities are formed respectively and so on.

In fact, as it is clear from Fig. 15, by this hierarchical merging process, the maximum resulted modularity value is obtained in repetition 5 with 14 detected communities. This best found partition of this naive merging process based on obtained modularity value is very far from the real partition including two original communities.

Simulated annealing method which is used to maximize modularity has better result than the previously mentioned naive merging process (see Fig. 14) . This method

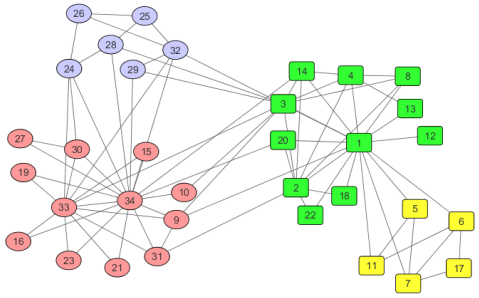

Fig. 12- The detected communities of modularity maximization using simulated annealing on karate network.

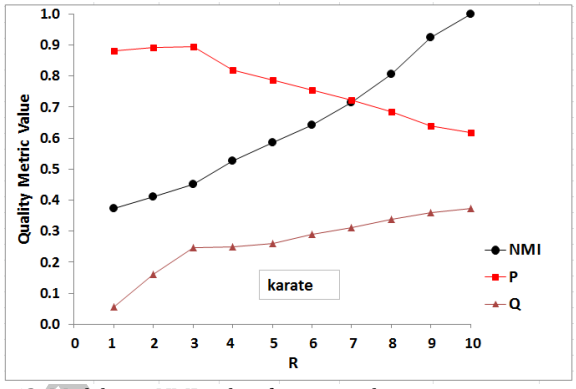

Fig. 13- Modularity, NMI and performance values in merging process on karate club network.

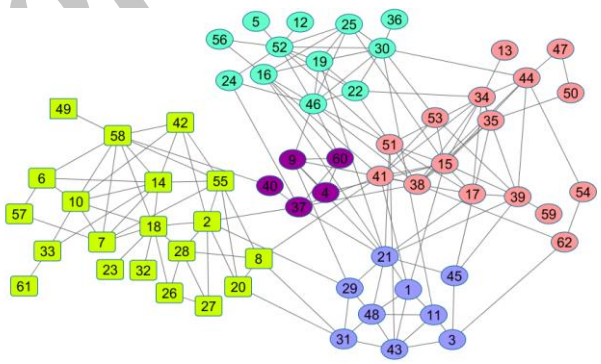

Fig. 14- The detected communities of modularity maximization using simulated annealing on dolphin network.

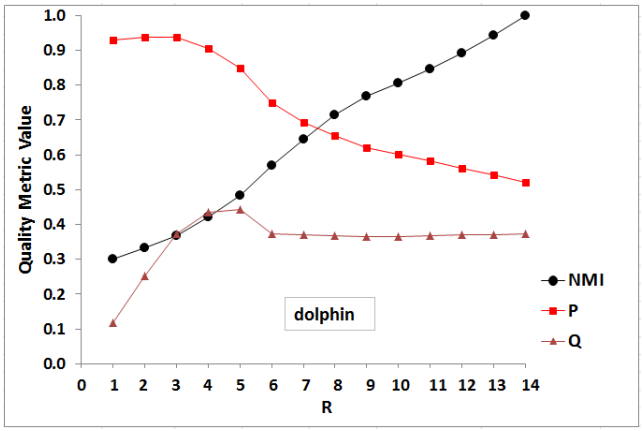

Fig. 15- Modularity, NMI and performance values in merging process on dolphin network.

detects five communities for this network. More specifically, it finds one original community as it is, but divides the second one into four smaller sub-communities.

College football network: This network is representation of 115 nodes and 613 links. Nodes represent football teams and the links between two nodes indicate that two corresponding teams have played a game together. Teams are divided into 11 conferences. Also, there are five independent teams which do not belong to any conferences. These five teams are specified with orange color in Fig. 16. Modularity values increase as number of repetition R increase (see Fig. 17). Modularity maximization using simulated annealing failed to find one community, but other communities were detected fairly well.

8. Conclusion

In this paper we analyzed deeply the lmitations of somefamous quality metrics for community detection and evaluation. We showed limitations of modularity maximization and performance with more accurate details than previous works. Moreover, for the first time, we showed that *NMI* has the scalability issue like modularity and performance. Moreover, we discussed the limitations of other quality metrics such as conductance, internal density and cut ratio, etc. In addition, we defined and proposed two characteristics of a good quality metric for community evaluation. Now, to remedy some of these limitations to some extent, we propose several possible approaches to be considered as future work. Firstly, in modularity optimization, communities bigger than a scale can be treated as new sub-graphs. Thus on these big sub-graphs, modularity optimization can be run again to find possible small communities. Secondly, using both intra-community and inter-community conductance can be helpful, if we consider two points: 1) intra-community conductance should be rewritten as the average conductance value of the graphs induced by each community. Similarly, inter-community conductance should be considered as the average conductance values of communities. 2) We use an appropriate hierarchical agglomerative algorithm for community detection such that starting from single nodes as communities, finally one big community is obtained as the whole graph. Then, the partition of the iteration with highest intra-community and lowest inter-community value can be taken as final output. Thirdly, presenting quality metrics which consider only local information of each community instead of global one, may lead to community detection and evaluation with higher accuracy.

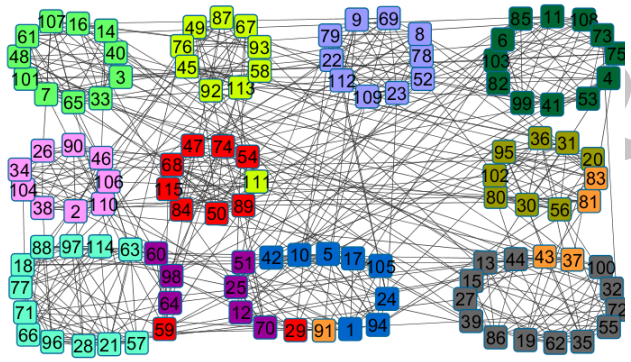

Fig. 16- The detected communities of modularity maximization using simulated annealing on football network.

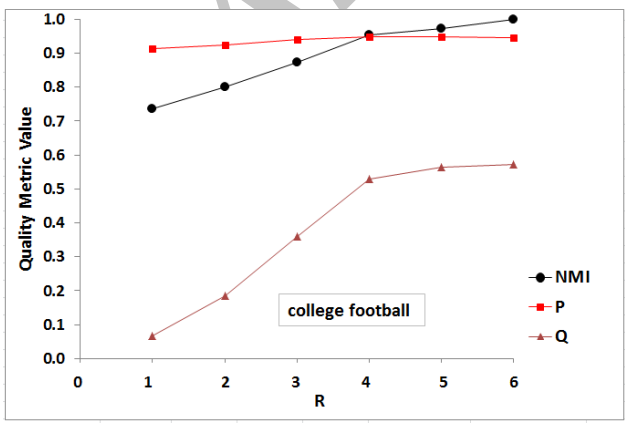

Fig. 17- Modularity, NMI and performance values in merging process on football network.

References

[1] Fortunato, Santo. "Community detection in graphs." Physics reports 486, no. 3 (2010): 75-174.

[2] Newman, Mark EJ, and Michelle Girvan. "Finding and evaluating community structure in networks." Physical review E 69, no. 2 (2004): 026113.

[3] Van Dongen, Stijn Marinus. "Graph clustering by flow simulation." PhD diss., 2001.

[4] Kannan, Ravi, Santosh Vempala, and Adrian Vetta. "On clusterings: Good, bad and spectral." Journal of the ACM (JACM) 51, no. 3 (2004): 497-515.

[5] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, Journal of Statistical Mechanics: Theory and Experiment 2005, P09008 (2005).

[6] Fortunato S, Barthelemy M. Resolution limit in community detection. Proceedings of the National Academy of Sciences USA 2007;104:36–41.

[7] J. Reichardt and S. Bornholdt, Phys. Rev. E 74, 016110 (2006).

[8] A. Arenas, A. Fernández, and S. Gómez, New J. Phys. 10, 053039 (2008).

[9] Lancichinetti A, Fortunato S. Limits of modularity maximization in community detection. Physical Review E 2011;84:066122.

[10] Xiang, J., & Hu, K. (2012). Limitation of multi-resolution methods in community detection. *Physica A: Statistical Mechanics and its Applications*, *391*(20), 4995-5003.

[11] Good, B.H., de Montjoye, Y.A. and Clauset, A., 2010. Performance of modularity maximization in practical contexts. *Physical Review E*, *81*(4), p.046106.

[12] Almeida, Hélio, Dorgival Guedes, Wagner Meira Jr, and Mohammed J. Zaki. "Is there a best quality metric for graph clusters?." In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 44-59. Springer Berlin Heidelberg, 2011.

[13] Fortunato, Santo, and Claudio Castellano. "Community structure in graphs." In Computational Complexity, pp. 490-512. Springer New York, 2012.

[14] Caldarelli, Guido. Large scale structure and dynamics of complex networks: from information technology to finance and natural science. Vol. 2. World Scientific, 2007.

[15] Wasserman, Stanley, and Katherine Faust. Social network analysis: Methods and applications. Vol. 8. Cambridge university press, 1994.

[16] Raghavan, Usha Nandini, Réka Albert, and Soundar Kumara. "Near linear time algorithm to detect community structures in large-scale networks." Physical review E 76.3 (2007): 036106.

[17] Brandes, Ulrik, Marco Gaertler, and Dorothea Wagner. "Experiments on graph clustering algorithms." In European Symposium on Algorithms, pp. 568-579. Springer Berlin Heidelberg, 2003.

[18] Moradi, Farnaz, Tomas Olovsson, and Philippas Tsigas. "An evaluation of community detection algorithms on large-scale email traffic." In International Symposium on Experimental Algorithms, pp. 283-294. Springer Berlin Heidelberg, 2012.

[19] Yang, Jaewon, and Jure Leskovec. "Defining and evaluating network communities based on ground-truth." Knowledge and Information Systems 42, no. 1 (2015): 181-213.

[20] Arab, Mohsen, and Mohsen Afsharchi. "Community detection in social networks using hybrid merging of sub-communities." Journal of Network and Computer Applications 40 (2014): 73-84.

[21] Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." Proceedings of the national academy of sciences 99, no. 12 (2002): 7821-7826.

Mohsen Arab is a PhD candidate of Computer Science in Yazd University. His work focuses mainly on the community detection algorithms in social networks.

Mahdieh Hasheminezhad is an assistant professor of Computer science Department in Yazd University. Her work focuses specifically on the graph algorithms. She did her PhD in Amirkabir University of Technology.