Z-Score-Based Modularity for Community Detection in Networks

Atsushi Miyauchi*, Yasushi Kawase

Graduate School of Decision Science and Technology, Tokyo Institute of Technology, Ookayama 2-12-1, Meguro-ku, Tokyo 152-8552, Japan

* miyauchi.a.aa@m.titech.ac.jp

Abstract

Identifying community structure in networks is an issue of particular interest in network science. The modularity introduced by Newman and Girvan is the most popular quality function for community detection in networks. In this study, we identify a problem in the concept of modularity and suggest a solution to overcome this problem. Specifically, we obtain a new quality function for community detection. We refer to the function as *Z-modularity* because it measures the Z-score of a given partition with respect to the fraction of the number of edges within communities. Our theoretical analysis shows that Z-modularity mitigates the resolution limit of the original modularity in certain cases. Computational experiments using both artificial networks and well-known real-world networks demonstrate the validity and reliability of the proposed quality function.

Introduction

Many complex systems can be represented as networks. Analyzing the structure and dynamics of these networks provides meaningful information about the underlying systems. In fact, complex networks have attracted significant attention from diverse fields such as physics, chemistry, biology, and sociology [1, 2].

An issue of particular interest in network science is the identification of community structure [3]. Roughly speaking, a *community* (also referred to as a *module*) is a subset of vertices more densely connected with each other than with nodes in the rest of the network. Note that no absolute definition of a community exists because any such definition typically depends on the specific system at hand. Detecting communities is a powerful way to discover components that have some special roles or possess important functions. For example, consider the network representing the World Wide Web, where vertices correspond to web pages and edges represent the hyperlinks between pages. Communities in this network are likely to be the sets of web pages dealing with the same or similar topics.

There are various methods to detect community structure in networks, which can be roughly divided into two types. First, there are methods based on some conditions that should be satisfied by a community. The most fundamental concept is a clique. A *clique* is a subset of vertices wherein every pair of vertices is connected by an edge. As even a singleton or an edge is

a clique, we are usually interested in finding a *maximum clique* or a *maximal clique*, i.e., cliques with maximum size and cliques not contained in any other clique, respectively. Although the definition of a clique is very intuitive, it is too strong and restrictive to use practically. In 2004, Radicchi et al. [4] introduced more practical definitions: a community in a strong sense and a community in a weak sense. A subset S of vertices is called a *community in a strong sense* if for every vertex in S, the number of neighbors in S is strictly greater than the number of neighbors outside S. On the other hand, a subset S of vertices is called a *community in a weak sense* if the sum, over all vertices in S, of the number of neighbors in S is strictly greater than the number of cut edges of S. Thus, if a subset of vertices is a community in a strong sense, then it is also a community in a weak sense. Recently, Cafieri et al. [5] proposed an enumerative algorithm to list all partitions of the set of vertices into communities in a strong sense with moderate sizes.

Second, but perhaps more importantly, there are methods that maximize a globally defined quality function. The best known and most commonly used quality function is *modularity*, which was introduced by Newman and Girvan [6]. Here let $G = (V,E)$ be an undirected network consisting of $n = |V|$ vertices and $m = |E|$ edges. The modularity, a quality function for partition $\mathcal{C} = \{C_1, \ldots, C_k\}$ of V (i.e., $\bigcup_{i=1}^{k} C_i = V$ and $C_i \cap C_j = \emptyset$ for $i \neq j$), can be written as

$$Q(\mathcal{C}) = \sum_{C \in \mathcal{C}} \left(\frac{m_C}{m} - \left(\frac{D_C}{2m} \right)^2 \right),$$

where m_C is the number of edges in community C, and D_C is the sum of the degrees of the vertices in community C. The modularity represents the sum, over all communities, of the fraction of the number of edges in the communities minus the expected fraction of such edges assuming that they are placed at random with the same distribution of vertex degree.

Many studies have examined modularity maximization. In 2008, Brandes et al. [7] proved that modularity maximization is NP-hard. This implies that unless P = NP, no modularity maximization method that simultaneously satisfies the following exists: (i) finds a partition that maximizes modularity exactly (ii) in time polynomial in n and m (iii) for any network. To date, a major focus in modularity maximization has been designing accurate and scalable heuristics. In fact, there are a wide variety of algorithms based on greedy techniques [6, 8, 9], simulated annealing [10–12], extremal optimization [13], spectral optimization [14, 15], mathematical programming [16–19], and other techniques. Note that to reduce computation time, a few pre-processing techniques have been proposed [20]. Moreover, to improve the quality of partitions obtained by such heuristics, some post-processing algorithms have also been developed [21].

Although modularity maximization is the most popular and widely used method in practice, it is also known to have some serious drawbacks; i.e., the *resolution limit* [22] and *degeneracies* [23]. The former means that modularity maximization fails to detect communities smaller than a certain scale depending on the total number of edges in a network even if the communities are cliques connected by single edges. The latter means that there exist numerous nearly optimal partitions in terms of modularity maximization, which makes finding communities with maximum modularity extremely difficult. The resolution limit particularly narrows the application range of modularity maximization because most real-world networks consist of communities with very different sizes. To avoid this issue, some multiresolution variants of the modularity have been adopted in practical applications [24–26]. In these variants, the resolution level can be tuned freely by adjusting certain parameters. However, once the resolution level is determined, communities larger than the determined resolution level tend to be divided and smaller communities tend to be merged. Therefore, such multiresolution variants also fail to detect real community structure [27].

In this study, we identify a problem in the concept of modularity and suggest a solution to overcome this problem. Specifically, we obtain a new quality function for community detection. We refer to this function as *Z-modularity* because it measures the Z-score of a given partition with respect to the fraction of the number of edges within communities. Our theoretical analysis shows that Z-modularity mitigates the resolution limit of the original modularity in certain cases. In fact, Z-modularity never merges adjacent cliques in the well-known ring of cliques network with any number and size of cliques. Computational experiments using both artificial networks and well-known real-world networks demonstrate the validity and reliability of the proposed quality function.

Note that there are many other quality functions based on modularity or other concepts [28–33]. Most of them are collected in reference [3].

Methods

Definition of Z-modularity

Modularity simply computes the fraction of the number of edges within communities minus its expected value. The definition is quite intuitive; thus, it is the most popular and widely used quality function in practice.

However, we identify a problem with the concept of modularity. Here consider two partitions C_1 and C_2. Assume that the fraction of the number of edges within communities of C_1 and C_2 are 0.2 and 0.6, respectively. In addition, assume that their expected values are 0.1 and 0.5, respectively. Then, we see that these two partitions share the same modularity value (i.e., $Q(C_1) = Q(C_2) = 0.1$). The key question is as follows: should these two partitions receive the same quality value? Our answer is that it must depend on the variance of the probability distribution of the fraction of the number of edges within communities of C_1 and C_2. Fig 1 illustrates an example. In this case, we wish to assign a higher quality value to C_1 because it is statistically much rarer than C_2. This simple but critical observation forms the basis of our quality function.

Given an undirected network $G = (V, E)$ consisting of $n = |V|$ vertices and $m = |E|$ edges, and a partition C of V, we aim to quantify the statistical rarity of partition C in terms of the fraction
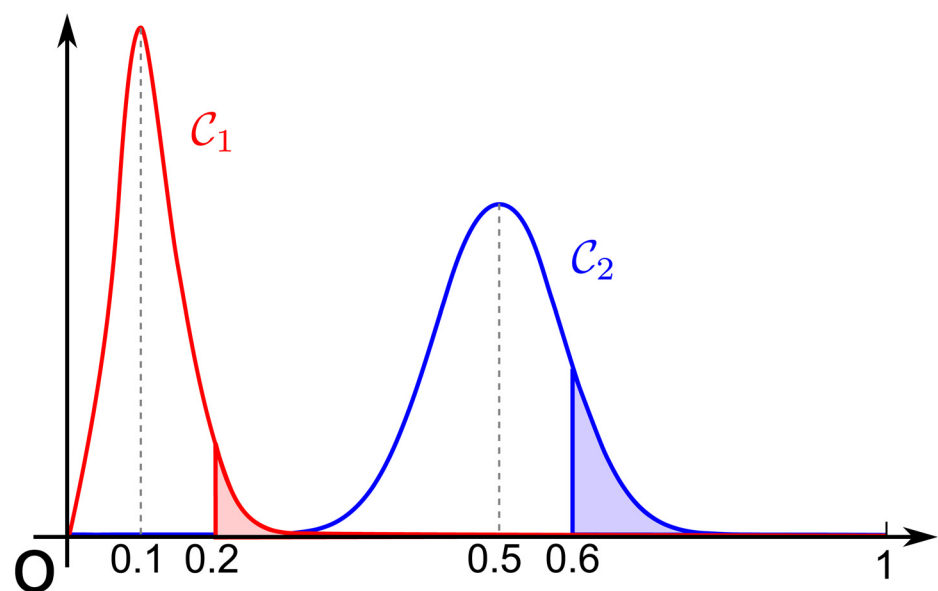

Fig 1. Probability distributions.

doi:10.1371/journal.pone.0147805.g001

of the number of edges within communities. To this end, we consider the following edge generation process over V. Place N edges over V at random with the same distribution of vertex degree. Then, when we place an edge, the probability that the edge is placed within communities is given by

$$p = \sum_{C \in \mathcal{C}} \left(\frac{D_C}{2m} \right)^2.$$

Note that this edge generation process is the same as the null-model (also known as the *configuration model* [34]) used in the definition of modularity, with the exception of the sample size. We simply wish to estimate the probability distribution of the fraction of the number of edges within communities. Thus, unlike the null-model, the sample size N is not necessarily equal to the number of edges m.

Let X be a random variable denoting the number of edges generated by the process within communities. Then, X follows the binomial distribution $B(N,p)$. By the central limit theorem, when the sample size N is sufficiently large, the distribution of X/N can be approximated by the normal distribution $\mathcal{N}(p, p(1-p)/N)$. Thus, we can quantify the statistical rarity of partition $\mathcal{C}$ in terms of the fraction of the number of edges within communities using the Z-score as follows:

$$Z(\mathcal{C}) = \frac{\sum_{C \in \mathcal{C}} \frac{m_C}{m} - \sum_{C \in \mathcal{C}} \left(\frac{D_C}{2m} \right)^2}{\sqrt{\sum_{C \in \mathcal{C}} \left(\frac{D_C}{2m} \right)^2 \left(1 - \sum_{C \in \mathcal{C}} \left(\frac{D_C}{2m} \right)^2 \right)}}.$$

The sample size N never depends on a given partition; thus, it is omitted in the denominator. For partition $\mathcal{C}$ with $p = 0$ or 1, we define $Z(\mathcal{C}) = 0$. We refer to this quality function as *Z-modularity*.

Remarks on Z-modularity

The numerator of Z-modularity is none other than modularity. Although it does not make sense to compare the Z-modularity value and the modularity value for a given partition, they have the same sign. In fact, since the denominator of Z-modularity is always positive, Z-modularity takes a positive value if and only if so does modularity. Thus, the positive (resp. negative) value of Z-modularity implies that the fraction of the number of edges within communities is greater (resp. less) than the expected fraction of such edges in the above edge generation process.

Here we provide upper and lower bounds on Z-modularity. As shown in Brandes et al. [7], for any partition $\mathcal{C}$, the modularity value $Q(\mathcal{C})$ falls into the interval $[-1/2, 1]$. On the other hand, Z-modularity has an upper bound of $\sqrt{n}$ and a lower bound of $-n$, as shown below. Recall that $p = \sum_{C \in \mathcal{C}} \left(\frac{D_C}{2m} \right)^2$. As for the upper bound, since $p \geq 1/n$ (unless $p = 0$), we have

$$Z(\mathcal{C}) = \frac{\sum_{C \in \mathcal{C}} \frac{m_C}{m} - p}{\sqrt{p(1-p)}} \leq \frac{1-p}{\sqrt{p(1-p)}} = \sqrt{\frac{1}{p} - 1} \leq \sqrt{n-1} < \sqrt{n}.$$

As for the lower bound, since $p \leq \left(\frac{2m-1}{2m} \right)^2 + \left(\frac{1}{2m} \right)^2 = 1 - \frac{1}{m} + \frac{1}{2m^2} \leq 1 - \frac{1}{2m} < 1 - \frac{1}{n^2}$ (unless $p = 1$), we have

$$Z(\mathcal{C}) = \frac{\sum_{C \in \mathcal{C}} \frac{m_C}{m} - p}{\sqrt{p(1-p)}} \geq -\sqrt{\frac{p}{1-p}} > -n.$$

Finally, we mention the computation time concerning Z-modularity. Although Z-modularity is of a more complicated form than modularity, the computation time of the function value is almost the same. On the other hand, if we replace modularity with Z-modularity in some modularity maximization algorithms, their running time may change drastically because, for instance, the number of iterations to converge to a local optimum may increase or decrease substantially. However, as an example, our computational experiments confirmed that the simulated annealing algorithm proposed by Guimerà and Amaral [10] and its Z-modularity maximization version run in almost the same time.

Theoretical Results

Fortunato and Barthélemy [22] pointed out the resolution limit of modularity. This resolution limit means that modularity maximization fails to detect communities that are smaller than a certain scale depending on the total number of edges in a network. This phenomenon occurs even if the communities are cliques connected by single edges. Here we theoretically analyze Z-modularity from a resolution limit perspective. As a result, we demonstrate that Z-modularity mitigates the resolution limit of the original modularity in certain cases.

Ring of cliques network

First, we consider a *ring of cliques network* that consists of a number of cliques connected by single edges (Fig 2). Assume that each clique consists of p (≥ 3) vertices and the number of cliques is q (≥ 2). Then, the network has $n = p \cdot q$ vertices and $m = q \cdot (1 + p(p-1)/2)$ edges. Fortunato and Barthélemy [22] showed that modularity maximization would merge adjacent cliques if q is larger than a certain value depending on p. However, adjacent cliques are never merged in a partition with maximal Z-modularity value, as shown below.

Let $\mathcal{C}^*$ be the partition of V into the cliques. In addition, let $\mathcal{C} = \{C_1, \ldots, C_l\}$ ($1 < l < q$) be a partition of V such that each C_i consists of a series of s_i (≥ 1) cliques and $q = \sum_{i=1}^{l} s_i$. Then, Z-modularity for $\mathcal{C}^*$ and $\mathcal{C}$ are calculated by

$$Z(\mathcal{C}^*) = \frac{1 - q/m - 1/q}{\sqrt{(1 - 1/q)/q}} \quad \text{and} \quad Z(\mathcal{C}) = \frac{1 - l/m - t}{\sqrt{t(1 - t)}},$$

respectively, where $t = \sum_{i=1}^{l} (s_i/q)^2$. By the Cauchy–Schwarz inequality, we have $1 > t = \sum_{i=1}^{l} (s_i/q)^2 \geq (\sum_{i=1}^{l} (s_i/q))^2/l = 1/l$. To analyze the behavior of the function $Z(\mathcal{C})$ in details, we rewrite $Z(\mathcal{C})$ using two variables x and y as follows:

$$f(x, y) = \frac{1 - y/m - x}{\sqrt{x(1 - x)}}.$$

Then, the derivative of $f(x,y)$ with respect to x is

$$\frac{\partial}{\partial x} f(x, y) = \frac{-x \cdot y/m - (1 - y/m)(1 - x)}{2 \cdot (x(1 - x))^{3/2}} < 0$$

for $0 < x < 1$ and $1 \leq y \leq m$. Thus, we obtain

$$f(1/l, l) \geq f(t, l).$$

Moreover, the derivative of $f(1/y,y)$ with respect to y is

$$\frac{\partial}{\partial y} f(1/y, y) = \frac{(m - 3y)(y - 1) + y}{2m \cdot (y - 1)^{3/2}} > 0$$
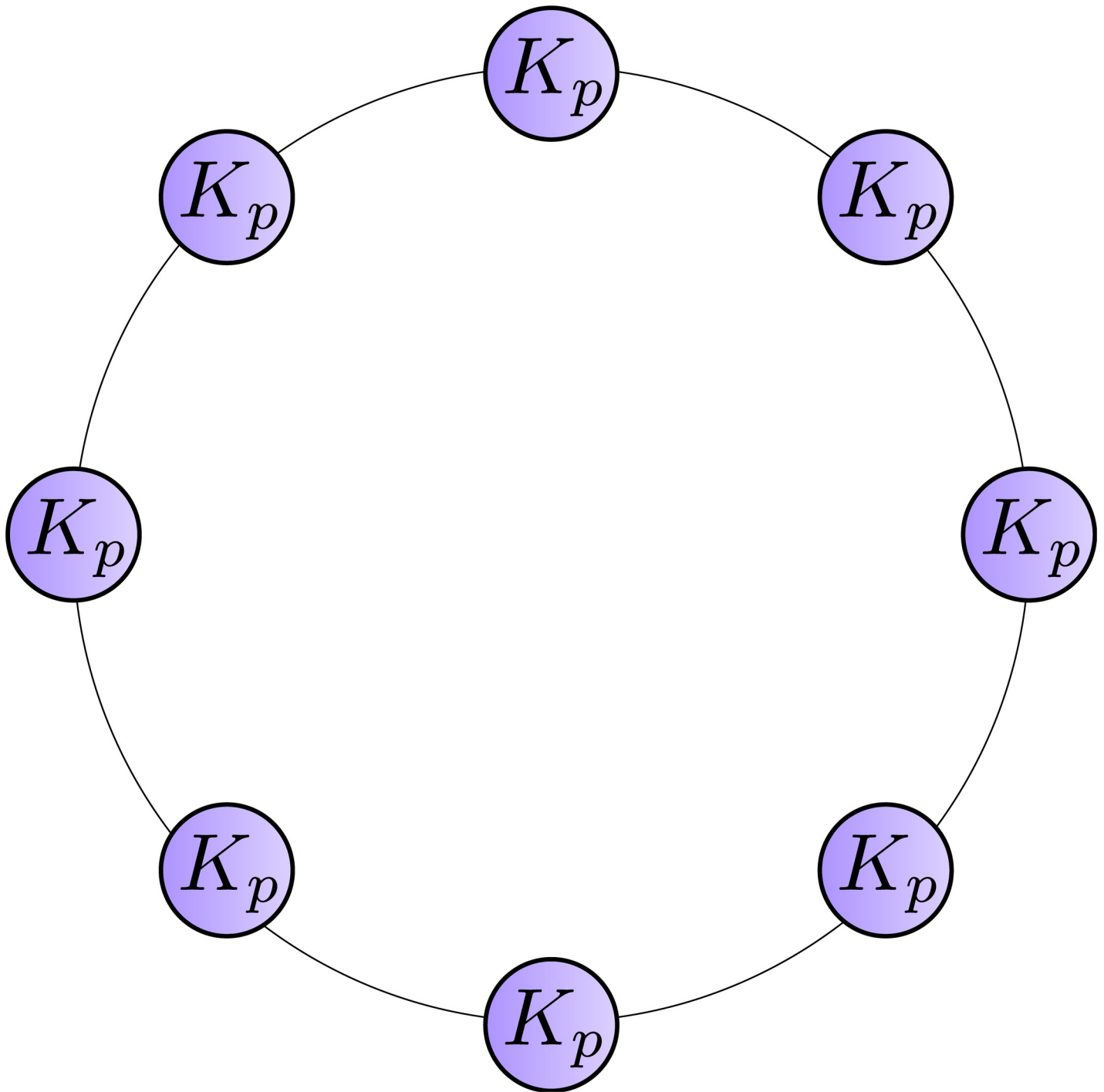
Fig 2. Ring of cliques network. K_p represents a clique with p vertices.

Table 1. Numerical examples of modularity and Z-modularity for some ring of cliques networks.

n	m	p	q	$Q(\mathcal{C}^*)\%$	$Q(\mathcal{C})\%$	$Z(\mathcal{C}^*)\%$	$Z(\mathcal{C})\%$
100	220	5	20	**0.8591**	0.8548	**3.942**	2.848
200	440	5	40	0.8841	**0.9045**	**5.663**	4.150
400	880	5	80	0.8966	**0.9295**	**8.070**	5.954
5000	11000	5	1000	0.9081	**0.9525**	**28.73**	21.32

doi:10.1371/journal.pone.0147805.t001

for $1 < y < m/3$. Thus, we have

$$f(1/q, q) > f(1/l, l),$$

since $1 < l < q \leq m/4$ by $m = q \cdot (1 + p(p-1)/2) \geq 4q$. Therefore, we have

$$Z(\mathcal{C}^*) = f(1/q, q) > f(1/l, l) \geq f(t, l) = Z(\mathcal{C}),$$

which means that maximizing Z-modularity never merges adjacent cliques.

[Table 1](#) lists the values of modularity and Z-modularity of partitions $\mathcal{C}^*$ and $\mathcal{C}$ ($s_i = 2$ for $i = 1, \ldots, l$) for some ring of cliques networks. As can be seen, the modularity of $\mathcal{C}$ is greater than that of $\mathcal{C}^*$ when the number of cliques is large, which is consistent with Fortunato and Barthélemy [22]. On the other hand, as we proved above, Z-modularity of $\mathcal{C}^*$ is certainly higher than that of $\mathcal{C}$ for every number of cliques.

Network with two pairwise identical cliques

Here we consider a *network with two pairwise identical cliques* that consists of a pair of cliques C_1 and C_2 with q vertices each and a pair of cliques C_3 and C_4 with p ($< q$) vertices each. These four cliques are connected by single edges, as described in [Fig 3](#). This network has $n = 2(p + q)$ vertices and $m = p(p-1) + q(q-1) + 4$ edges.

Consider two partitions $\mathcal{C}_A = \{C_1, C_2, C_3, C_4\}$ and $\mathcal{C}_B = \{C_1, C_2, C_3 \cup C_4\}$. Note that partition $\mathcal{C}_A$ is a more natural community structure that we would like to identify. Unfortunately, maximizing Z-modularity may choose $\mathcal{C}_B$, i.e., $Z(\mathcal{C}_A) < Z(\mathcal{C}_B)$ holds for some pair of p and q. However, if modularity maximization adopts $\mathcal{C}_A$, then so does Z-modularity, i.e., for any pair of p and q, if $Q(\mathcal{C}_A) > Q(\mathcal{C}_B)$ holds, then $Z(\mathcal{C}_A) > Z(\mathcal{C}_B)$ also holds. This fact follows directly from the definitions of Z-modularity and the original modularity.
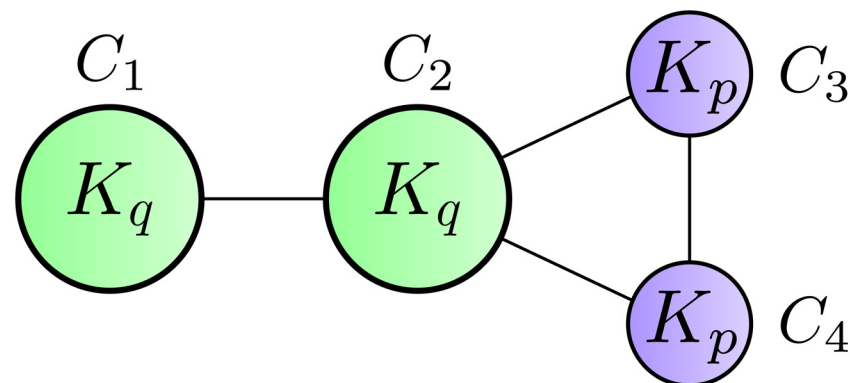

Fig 3. Network with two pairwise identical cliques. K_p and K_q represent cliques with p and q vertices, respectively.

doi:10.1371/journal.pone.0147805.g003

Table 2. Numerical examples of modularity and Z-modularity for some networks with two pairwise identical cliques.

n	m	p	q	$Q(\mathcal{C}_A)$	$Q(\mathcal{C}_B)$	$Z(\mathcal{C}_A)$	$Z(\mathcal{C}_B)$
26	80	5	8	**0.6618**	0.3385	**1.443**	1.345
42	264	5	16	0.5650	**0.5653**	**1.144**	1.143
74	1016	5	32	0.5182	**0.5190**	1.037	**1.039**
138	4056	5	64	0.5047	**0.5049**	1.009	**1.010**

doi:10.1371/journal.pone.0147805.t002

Table 2 lists the values of modularity and Z-modularity of partitions $\mathcal{C}_A$ and $\mathcal{C}_B$ for some networks with two pairwise identical cliques. We can confirm that both modularity and Z-modularity tend to merge C_3 and C_4 as the sizes of C_1 and C_2 become large. However, there is the case where only Z-modularity could divide C_3 and C_4. Therefore, we see that Z-modularity again mitigates the resolution limit of modularity in this case.

Experimental Results and Discussion

The purpose of our computational experiments is to evaluate the validity and reliability of the quality function Z-modularity. To this end, throughout the experiments, we maximize Z-modularity using a simulated annealing algorithm. Note that our algorithm is obtained immediately by changing the objective function from modularity to Z-modularity in the algorithm proposed by Guimerà and Amaral [10]. The implementation of their algorithm can be found on Lanci-chinetti's web page [35], and we use it with default parameters with the exception of the above change of objective function. Our experiments are conducted on various artificial networks and on well-known real-world networks.

Artificial networks

First, we report the results of computational experiments with artificial networks. We compare partitions obtained by maximizing Z-modularity with partitions obtained by modularity maximization on a wide variety of networks. The modularity is also maximized by the simulated annealing algorithm proposed by Guimerà and Amaral [10]. We deal with three types of artificial networks: the planted l-partition model, the Lancichinetti–Fortunato–Radicchi (LFR) benchmark, and the Hanoi graph. For the planted l-partition model and the LFR benchmark, once their parameters are set, the ground-truth community structure is fixed. Thus, we can evaluate the quality of the obtained community structure by comparison with the ground-truth using some measure.

To this end, we adopt the *normalized mutual information* for two partitions, which was introduced by Fred and Jain [36]. The normalized mutual information for two partitions $\mathcal{C}_1$ and $\mathcal{C}_2$ of n vertices is defined as follows:

$$I_{\mathrm{norm}}(\mathcal{C}_1, \mathcal{C}_2) = \frac{2I(\mathcal{C}_1, \mathcal{C}_2)}{H(\mathcal{C}_1) + H(\mathcal{C}_2)},$$

where

$$I(\mathcal{C}_1, \mathcal{C}_2) = \sum_{C_1 \in \mathcal{C}_1} \sum_{C_2 \in \mathcal{C}_2} \frac{|C_1 \cap C_2|}{n} \log_2 \left(\frac{n \cdot |C_1 \cap C_2|}{|C_1| \cdot |C_2|} \right)$$

and

$$H(\mathcal{C}) = -\sum_{C \in \mathcal{C}} \frac{|C|}{n} \log_2 \frac{|C|}{n}.$$

The normalized mutual information ranges from 0 to 1. For two partitions $\mathcal{C}_1$ and $\mathcal{C}_2$, the higher the normalized mutual information is, the more similar they are (and vice versa). In fact, $I_{\mathrm{norm}}(\mathcal{C}_1, \mathcal{C}_2) = 1$ if $\mathcal{C}_1$ and $\mathcal{C}_2$ are identical, and $I_{\mathrm{norm}}(\mathcal{C}_1, \mathcal{C}_2) = 0$ if they are independent. This measure has often been used to evaluate community detection methods. For example, see the computational experiments in references [37, 38].

Planted l-partition model. The planted l-partition model was introduced by Condon and Karp [39]. In this model, n vertices are divided into l equally sized groups (with size $c = n/l$). Two vertices in the same group are connected by probability p_{in}, whereas two vertices in different groups are connected by probability $p_{\mathrm{out}} (< p_{\mathrm{in}})$. Throughout the experiments, we set $p_{\mathrm{in}} = 0.5$. We construct four networks corresponding to combinations of two different network sizes ($n = 1000$ or 5000) and two different community sizes ($c = 20$ or 50). The parameter p_{out} starts with 0.01 and then increases in stages.

The results are shown in Fig 4. As can be seen, Z-modularity outperforms the original modularity in all four cases. In particular, Z-modularity provides much more superior results compared to modularity for networks consisting of relatively small communities.

LFR benchmark. In the planted l-partition model, each group in a generated network forms the Erdő–Rényi random graph [40]. Thus, all vertices have approximately the same degree. Moreover, all groups have exactly the same size. These phenomena are rarely observed in networks in real-world systems. As a more realistic model, the LFR benchmark was proposed by Lancichinetti, Fortunato, and Radicchi [41] for the case of unweighted and undirected
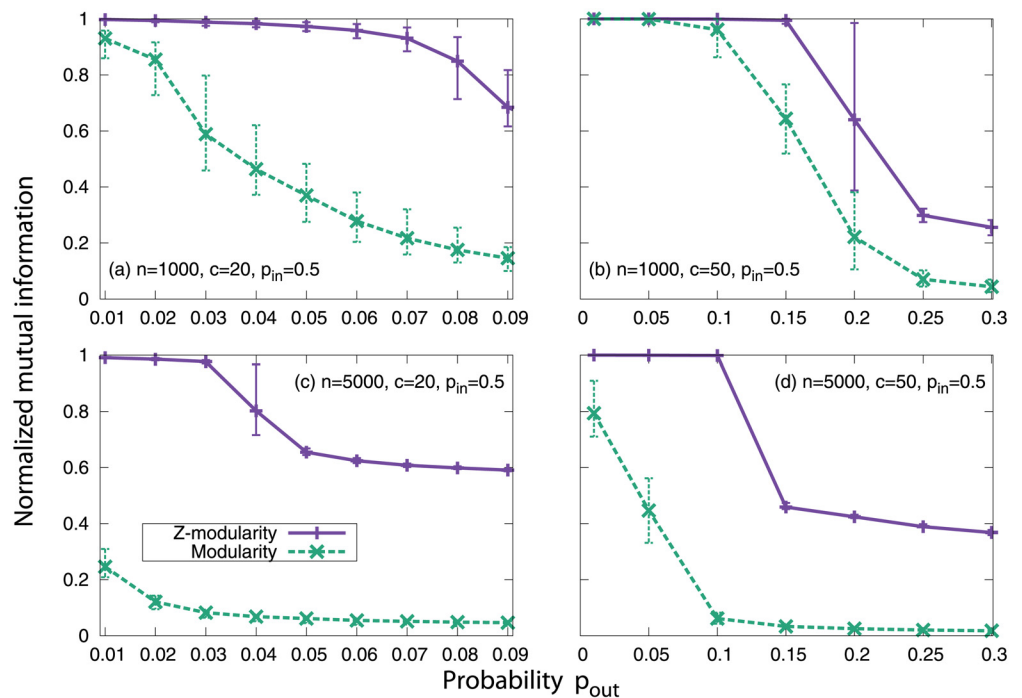

Fig 4. Results for the planted l-partition model. Each point is the result of averaging over 100 network realizations. The top and bottom bars represent the maximum and minimum values, respectively.

networks. The LFR benchmark was then extended to the case of directed and weighted networks with overlapping communities [42]. We now use the original unweighted and undirected case without overlapping communities.

In the model, degree distribution and community size distribution follow power laws with exponents γ and β, respectively. Furthermore, we can specify the number of vertices n, average degree $\langle k \rangle$, maximum degree k_{max}, minimum community size c_{min}, maximum community size c_{max}, and mixing parameter μ. In particular, mixing parameter μ indicates the mixing ratio of communities, i.e., the higher μ is, the more densely connected the communities are. The model constructs a network consistent with the specified parameters. For more details, see reference [41]. In our experiments, we set the parameters as in references [37, 38] as follows: $\gamma = -2$, $\beta = -1$, $\langle k \rangle = 20$, and $k_{max} = 50$. We construct eight networks corresponding to combinations of two different network sizes ($n = 1000$ or 5000) and four different ranges of community size ((c_{min}, c_{max}) = (10,50), (20,100), (30,150), or (40,200)).

The results are illustrated in Fig 5 ($n = 1000$) and Fig 6 ($n = 5000$). For the smaller networks ($n = 1000$), the mutual information values obtained by maximizing Z-modularity are lower than those obtained by modularity maximization when $\mu \leq 0.5$ for all community size settings. This trend is significant when the network consists of relatively large communities (e.g., (c_{min}, c_{max}) = (30,150) and (40,200)). On the other hand, for larger networks ($n = 5000$), Z-modularity outperforms the original modularity for all community size settings. From the above, we see that Z-modularity is particularly suitable for identifying community structure when a network consists of relatively small communities.

Here we investigate why the mutual information values obtained by maximizing Z-modularity are low when the community sizes are large. To this end, Fig 7 depicts the adjacency matrices of the LFR benchmark network with parameters $\gamma = -2$, $\beta = -1$, $n = 1000$, $\langle k \rangle = 20$, $k_{max} = 50$, $c_{min} = 40$, $c_{max} = 200$, and $\mu = 0.3$. The vertices are ordered according to both the
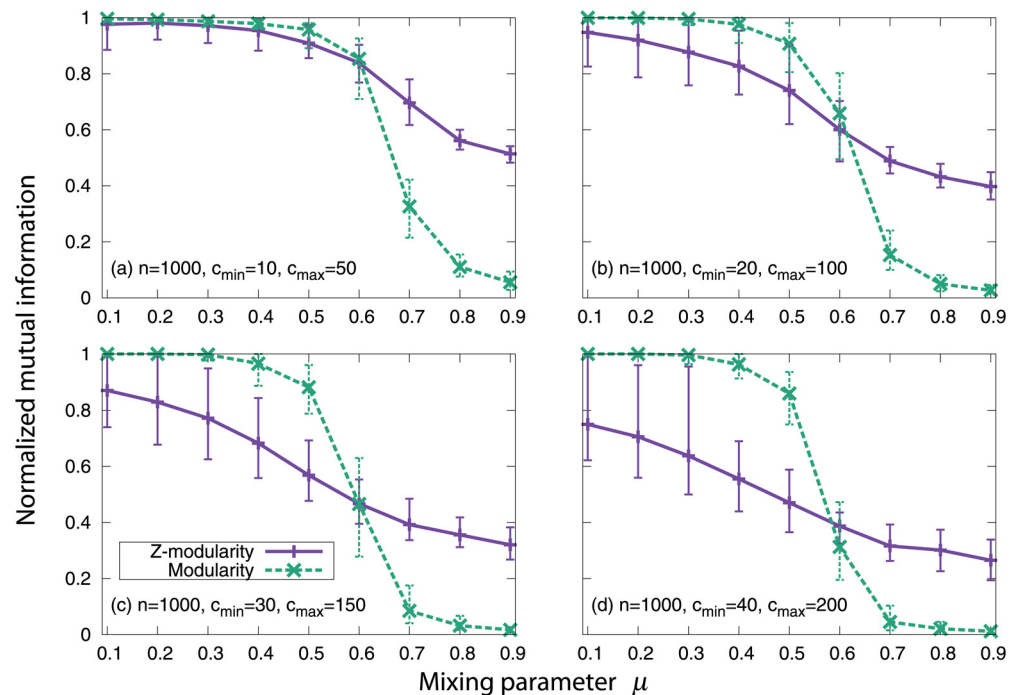

Fig 5. Results for the LFR benchmark (n = 1000). Each point is the result of averaging over 100 network realizations. The top and bottom bars represent the maximum and minimum values, respectively.
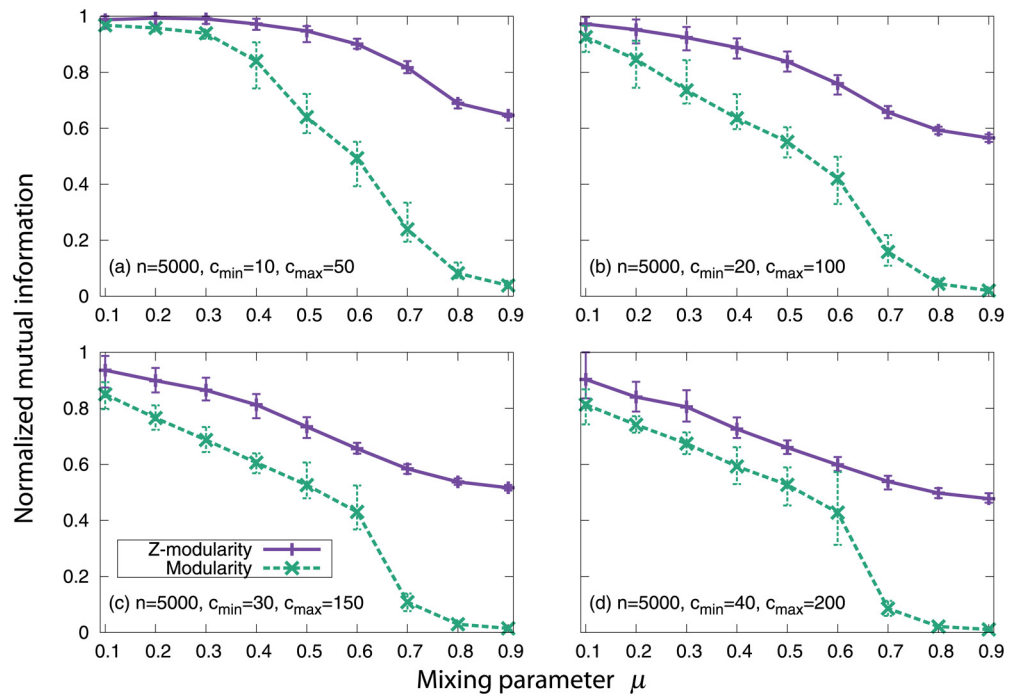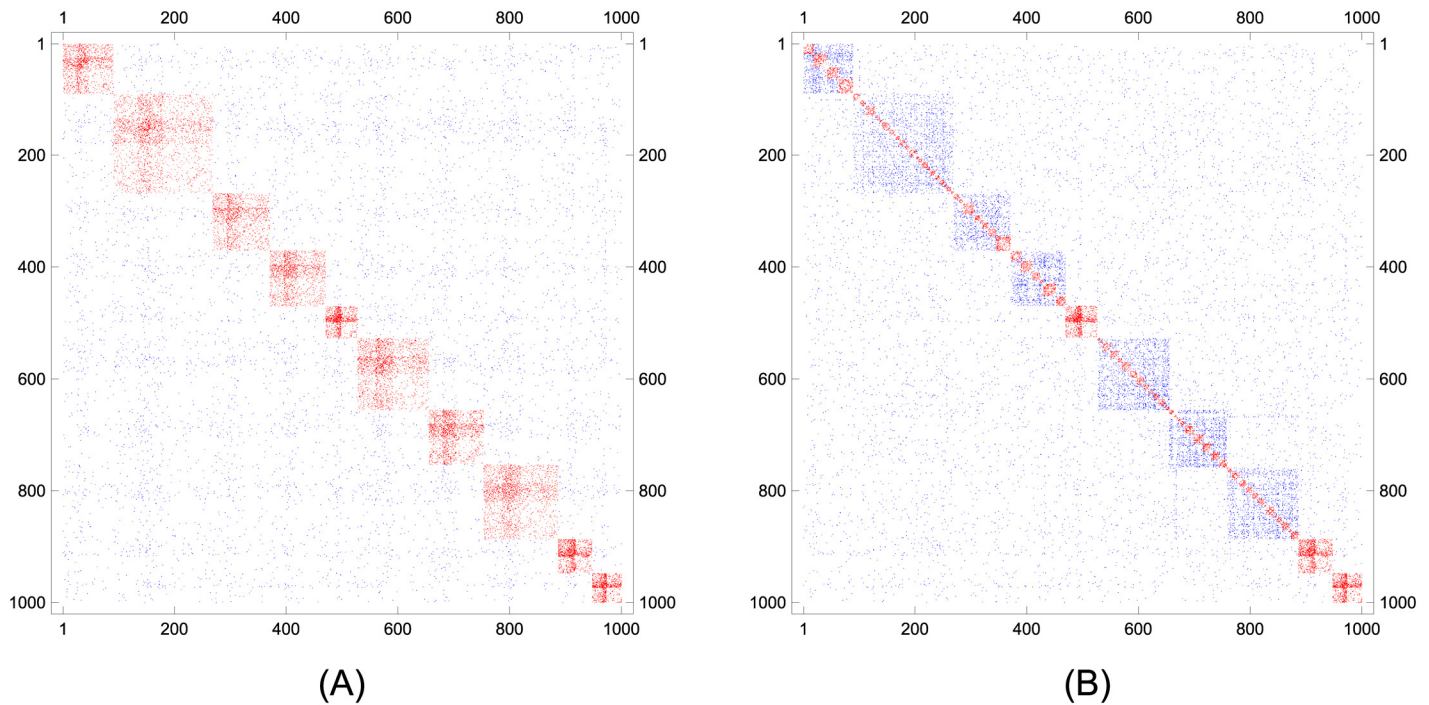
doi:10.1371/journal.pone.0147805.g005

Fig 6. Results for the LFR benchmark (n = 5000). Each point is the result of averaging over 100 network realizations. The top and bottom bars represent the maximum and minimum values, respectively.

Fig 7. Adjacency matrices for an LFR benchmark network. (A) Ground-truth partition: 10 communities. (B) Optimal partition for Z-modularity: 81 communities and I_{norm} = 0.6942.
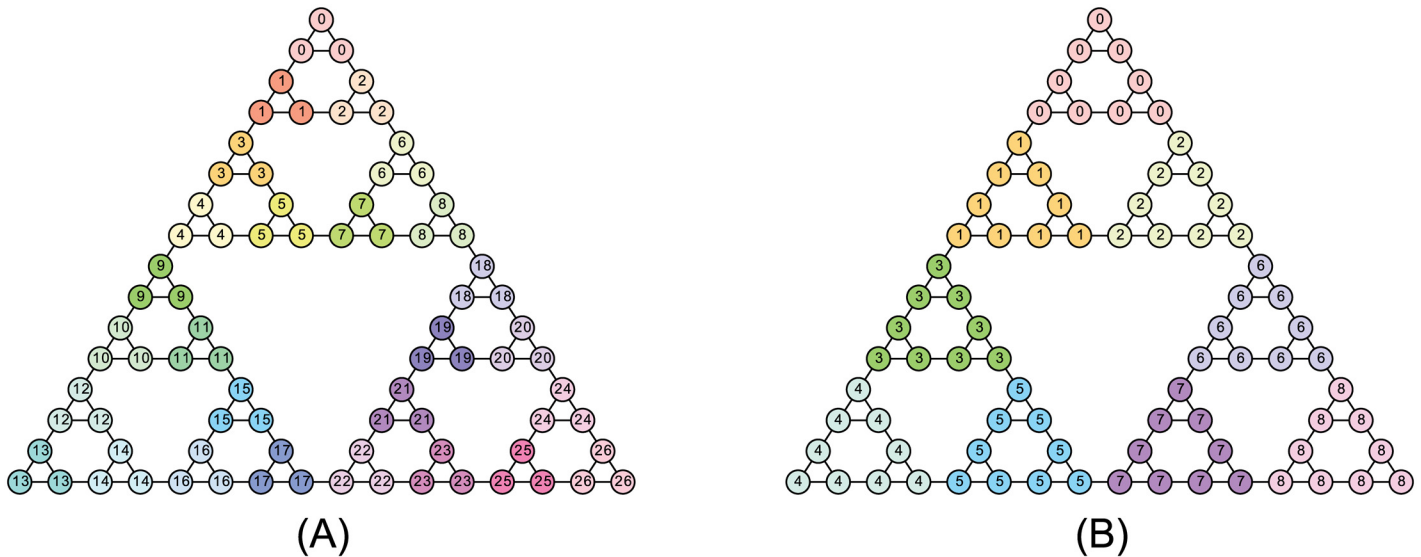
Fig 8. Community structure for Hanoi graph H_4. (A) Optimal partition for Z-modularity: 27 communities, $Z = 3.376$, and $Q = 0.6379$. (B) Optimal partition for modularity: 9 communities, $Z = 2.510$, and $Q = 0.7889$.

doi:10.1371/journal.pone.0147805.g008

ground-truth partition and the optimal partition for Z-modularity. The edges connecting vertices in the same community and in different communities are plotted with different colors, i.e., red and blue, respectively. As can be seen, maximizing Z-modularity divides the relatively large ground-truth communities because they contain much denser communities in the hierarchical structure by random behavior.

Hanoi graph. Here we demonstrate optimal partitions with respect to Z-modularity and the original modularity for the Hanoi graph, which is an example of networks with hierarchical organization. The Hanoi graph H_n corresponds to the allowed moves in the *tower of Hanoi* for n disks, which is a famous puzzle invented by Édouard Lucas in 1883. The Hanoi graph H_n has 3^n vertices and $3 \cdot (3^n - 1)/2$ edges. In the context of community detection in networks, Hanoi graph H_3 is used by Rosvall and Bergstrom [43]. Note that since the Hanoi graph does not have the ground-truth community structure, it is impossible to conclude whether the obtained partition is reasonable; we use this instance to observe the behavior of Z-modularity maximization.

The results for Hanoi graph H_4 are shown in Fig 8, where the label (and color) of each vertex represents the community to which the vertex belongs. As can be seen, maximizing Z-modularity leads to more detailed partition than modularity maximization. This result supports the trend observed in the experiments for the planted *l*-partition model and the LFR benchmark.

Real-world networks

Here we report the results of computational experiments with real-world networks; i.e., the Zachary's karate club network, the Les Misérables network, and the American college football network.

Zachary's karate club network. The first example is the famous karate club network analyzed by Zachary [44], which is often used as a benchmark to evaluate community detection methods. It consists of 34 vertices representing the members of a karate club in an American university, in addition to 78 edges representing friendship relations among individuals. Because of a conflict between the club administrator and the instructor, the club members split into two groups, one supporting the administrator and the other supporting the instructor. Therefore, these groups can be viewed as a ground-truth community structure.
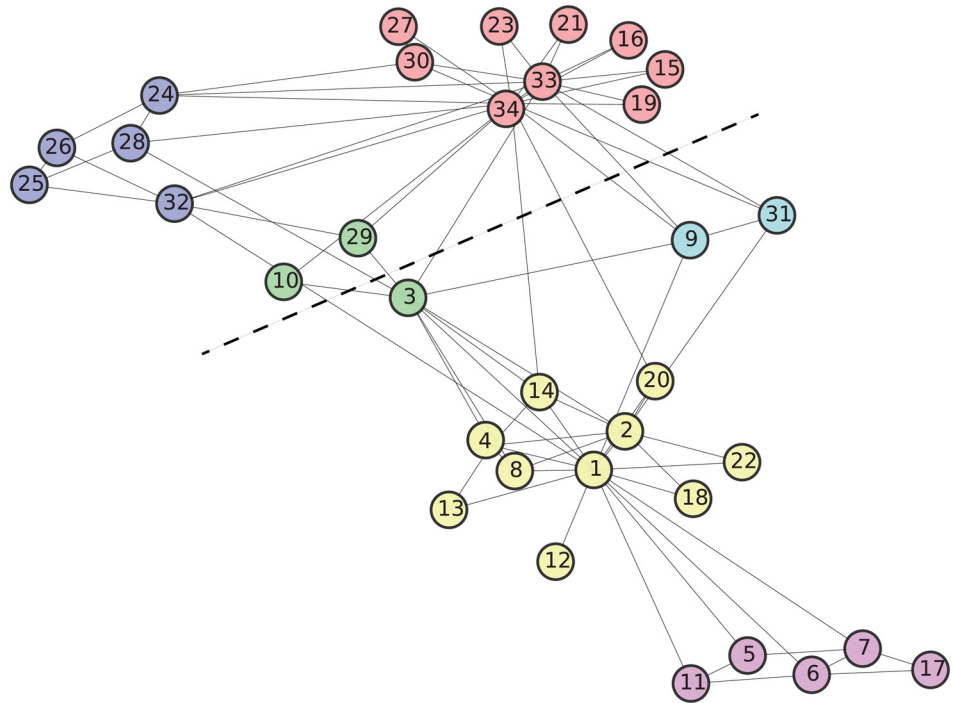
Fig 9. Community structure for Zachary's karate club network: 6 communities, Z = 0.9266, Q = 0.3882, and I_{norm} = 0.4796.

The partition obtained by maximizing Z-modularity is shown in Fig 9, where vertices with the same color represent a community. The label of each vertex represents an identification number of the member. For example, 1 and 34 represent the administrator and the instructor, respectively. The dashed line gives the partition of the network into the above two groups. The mutual information value of the obtained partition is not high. In fact, in comparison with the ground-truth community structure, the obtained partition consists of relatively small communities. This result is consistent with the trend observed in the experiments for artificial networks.

Les Misérables network. The second example is the network of the characters in the novel *Les Misérables* by Victor Hugo, compiled by Knuth [45]. It consists of 77 vertices representing the characters and 254 edges indicating the co-appearance of characters. Note here that since this network does not have the ground-truth community structure, it is impossible to evaluate the obtained partition using the mutual information value; we use this network to observe the behavior of Z-modularity maximization.

The partition obtained by maximizing Z-modularity is presented in Fig 10, where vertices with the same color represent a community. The label of each vertex represents the name of the character. Identified communities are likely to correspond to specific groups within the story. For example, the community consisting of 12 vertices (shaded with light brown) at the top left corner contains major characters belonging to the revolutionary student club *Friends of the ABC*.

American college football network. The third and final example is a network of college football teams in the United States, which was derived by Girvan and Newman [46]. There are 115 vertices representing the football teams, and 654 edges connecting teams that played each other in a regular season. The teams are divided into 12 groups referred to as *conferences* containing approximately 10 teams each. More games are played between teams in the same
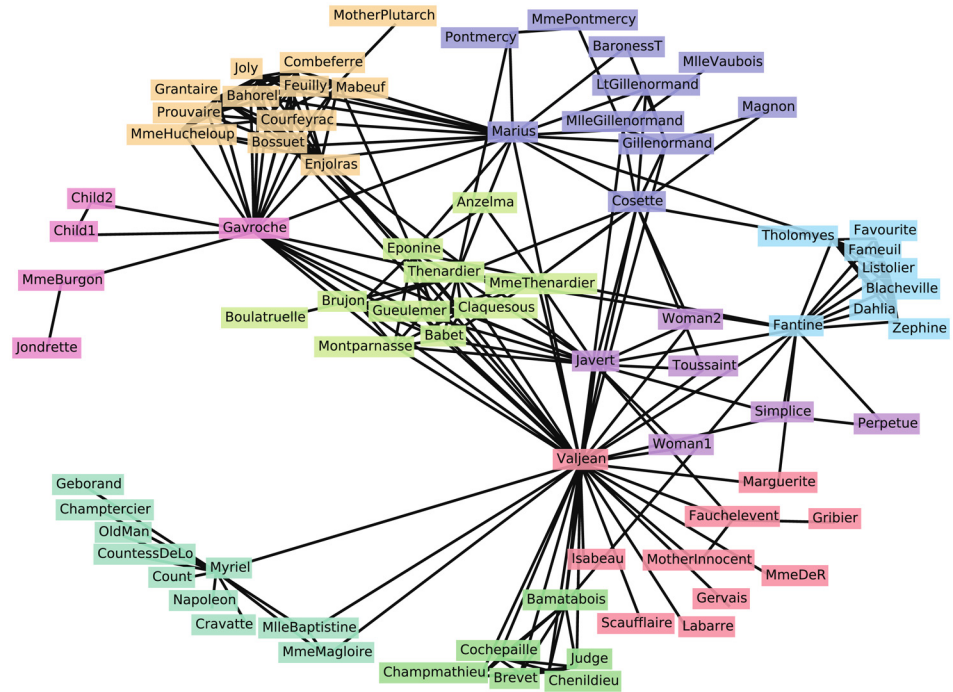
Fig 10. Community structure for Les Misérables network: 9 communities, Z = 1.490, and Q = 0.5245.

doi:10.1371/journal.pone.0147805.g010

conference than between teams in different conferences. Thus, the conferences can be viewed as a ground-truth community structure.

The partition obtained by maximizing Z-modularity is shown in Fig 11, where vertices with the same color represent a community. Note that the label of each vertex now represents the
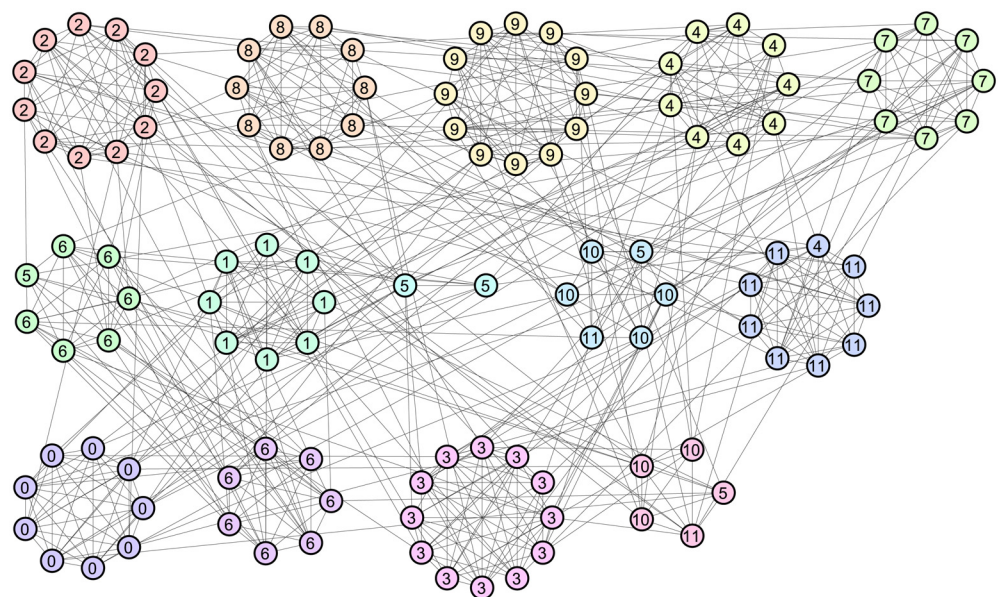

Fig 11. Community structure for American college football network: 14 communities, Z = 2.111, Q = 0.5738, and I_{norm} = 0.9205.

doi:10.1371/journal.pone.0147805.g011

conference to which the team belongs rather than an identification number of the team. Although some misclassifications are observed, Z-modularity correctly identifies 7 out of 12 conferences (i.e., conferences 0, 1, 2, 3, 7, 8, and 9). This result is outstanding in comparison with partitions obtained by modularity maximization. In fact, as reported in reference [16], only four conferences were correctly recovered by partition with a higher modularity value $Q = 0.6046$.

Conclusions

In this study, we have identified a problem in the concept of modularity and suggested a solution to overcome this problem. Specifically, we have obtained a new quality function Z-modularity that measures the Z-score of a given partition with respect to the fraction of the number of edges within communities. Theoretical analysis has shown that Z-modularity mitigates the resolution limit of the original modularity in certain cases. In fact, Z-modularity never merges adjacent cliques in the well-known ring of cliques network with any number and size of cliques. In computational experiments, we have evaluated the validity and reliability of Z-modularity. The results for artificial networks show that Z-modularity more accurately detects the ground-truth community structure than the original modularity in most cases. In particular, Z-modularity outperforms modularity for networks consisting of relatively small communities. Furthermore, the results for real-world networks demonstrate that Z-modularity leads to natural and reasonable community structure in practical use. Therefore, we conclude that Z-modularity could be another option for the quality function in community detection.

In the future, further experiments should be conducted to examine the performance of Z-modularity in more details. In fact, computational experiments in the present study have not used large-scale networks (with heterogeneous community sizes). This is due to the time complexity of the simulated annealing algorithm proposed by Guimerà and Amaral [10], on which our algorithm is based. To conduct computational experiments on large-scale networks, scalable Z-modularity maximization algorithms should be developed. For modularity maximization, there exist a wide variety of fast algorithms that perform well in practice. For example, the greedy algorithm proposed by Blondel et al. [8], which is known as the *Louvain method*, runs in time approximately linear in the size of the network. It should be noted that the corresponding Z-modularity maximization algorithm is derived directly by changing the objective function from modularity to Z-modularity. However, our preliminary experiments demonstrated that the Louvain method is not suitable for Z-modularity maximization. Indeed, the first aggregation step in the algorithm is not effective due to the term in the denominator of Z-modularity.

As another future direction, the physical interpretation of maximizing Z-modularity should be investigated. For example, it is known that modularity maximization can be interpreted as the problem of finding the ground state of a spin glass model [24].

Acknowledgments

Author Contributions

Conceived and designed the experiments: AM YK. Performed the experiments: YK. Analyzed the data: AM YK. Contributed reagents/materials/analysis tools: AM YK. Wrote the paper: AM.

References

1. Newman MEJ. The structure and function of complex networks. SIAM Rev. 2003; 45:167–256. doi: 10.1137/S003614450342480

2. Newman MEJ. Networks: An Introduction. Oxford University Press; 2010.

3. Fortunato S. Community detection in graphs. Phys Rep. 2010; 486:75–174. doi: 10.1016/j.physrep.2009.11.002

4. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. Defining and identifying communities in networks. Proc Natl Acad Sci USA. 2004; 101:2658–2663. doi: 10.1073/pnas.0400054101 PMID: 14981240

5. Cafieri S, Caporossi G, Hansen P, Perron S, Costa A. Finding communities in networks in the strong and almost-strong sense. Phys Rev E. 2012; 85:046113. doi: 10.1103/PhysRevE.85.046113

6. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E. 2004; 69:026113.

7. Brandes U, Delling D, Gaertler M, Görke R, Hoefer M, Nikoloski Z, et al. On modularity clustering. IEEE Trans Knowl Data Eng. 2008; 20:172–188. doi: 10.1109/TKDE.2007.190689

8. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech: Theory Exp. 2008; 2008:P10008. doi: 10.1088/1742-5468/2008/10/P10008

9. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Phys Rev E. 2004; 70:066111. doi: 10.1103/PhysRevE.70.066111

10. Guimerà R, Amaral LAN. Functional cartography of complex metabolic networks. Nature (London). 2005; 433:895–900. doi: 10.1038/nature03288

11. Massen CP, Doye JPK. Identifying communities within energy landscapes. Phys Rev E. 2005; 71:046101. doi: 10.1103/PhysRevE.71.046101

12. Medus A, Acuña G, Dorso C. Detection of community structures in networks via global optimization. Physica A. 2005; 358:593–604. doi: 10.1016/j.physa.2005.04.022

13. Duch J, Arenas A. Community detection in complex networks using extremal optimization. Phys Rev E. 2005; 72:027104. doi: 10.1103/PhysRevE.72.027104

14. Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci USA. 2006; 103:8577–8582. doi: 10.1073/pnas.0601602103 PMID: 16723398

15. Richardson T, Mucha PJ, Porter MA. Spectral tripartitioning of networks. Phys Rev E. 2009; 80:036111. doi: 10.1103/PhysRevE.80.036111

16. Agarwal G, Kempe D. Modularity-maximizing graph communities via mathematical programming. Eur Phys J B. 2008; 66:409–418. doi: 10.1140/epjb/e2008-00425-1

17. Cafieri S, Hansen P, Liberti L. Locally optimal heuristic for modularity maximization of networks. Phys Rev E. 2011; 83:056105. doi: 10.1103/PhysRevE.83.056105

18. Miyauchi A, Miyamoto Y. Computing an upper bound of modularity. Eur Phys J B. 2013; 86:302. doi: 10.1140/epjb/e2013-40006-7

19. Cafieri S, Costa A, Hansen P. Reformulation of a model for hierarchical divisive graph modularity maximization. Ann Oper Res. 2014; 222:213–226. doi: 10.1007/s10479-012-1286-z

20. Arenas A, Duch J, Fernández A, Gómez S. Size reduction of complex networks preserving modularity. New J Phys. 2007; 9:176. doi: 10.1088/1367-2630/9/6/176

21. Cafieri S, Hansen P, Liberti L. Improving heuristics for network modularity maximization using an exact algorithm. Discrete Appl Math. 2014; 163:65–72. doi: 10.1016/j.dam.2012.03.030

22. Fortunato S, Barthélemy M. Resolution limit in community detection. Proc Natl Acad Sci USA. 2007; 104:36–41. doi: 10.1073/pnas.0605965104 PMID: 17190818

23. Good BH, de Montjoye YA, Clauset A. Performance of modularity maximization in practical contexts. Phys Rev E. 2010; 81:046106. doi: 10.1103/PhysRevE.81.046106

24. Reichardt J, Bornholdt S. Statistical mechanics of community detection. Phys Rev E. 2006; 74:016110. doi: 10.1103/PhysRevE.74.016110

25. Arenas A, Fernández A, Gómez S. Analysis of the structure of complex networks at different resolution levels. New J Phys. 2008; 10:053039. doi: 10.1088/1367-2630/10/5/053039

26. Pons P, Latapy M. Post-processing hierarchical community structures: Quality improvements and multi-scale view. Theor Comput Sci. 2011; 412:892–900. doi: 10.1016/j.tcs.2010.11.041

27. Lancichinetti A, Fortunato S. Limits of modularity maximization in community detection. Phys Rev E. 2011; 84:066122. doi: 10.1103/PhysRevE.84.066122

28. Rosvall M, Bergstrom CT. An information-theoretic framework for resolving community structure in complex networks. Proc Natl Acad Sci USA. 2007; 104:7327–7331. doi: 10.1073/pnas.0611034104 PMID: 17452639

29. Li Z, Zhang S, Wang RS, Zhang XS, Chen L. Quantitative function for community detection. Phys Rev E. 2008; 77:036109. doi: 10.1103/PhysRevE.77.036109

30. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA. 2008; 105:1118–1123. doi: 10.1073/pnas.0706851105 PMID: 18216267

31. Chen J, Zaïane OR, Goebel R. Detecting Communities in Social Networks Using Max-Min Modularity. In: Apte C, Park H, Wang K, Zaki MJ, editors. Proceedings of the 2009 SIAM International Conference on Data Mining. vol. 3. SIAM; 2009. pp. 978–989.

32. Zhang S, Zhao H. Community identification in networks with unbalanced structure. Phys Rev E. 2012; 85:066114. doi: 10.1103/PhysRevE.85.066114

33. Zhang S, Zhao H. Normalized modularity optimization method for community identification with degree adjustment. Phys Rev E. 2013; 88:052802. doi: 10.1103/PhysRevE.88.052802

34. Molloy M, Reed B. A critical point for random graphs with a given degree sequence. Random Struct Algorithms. 1995; 6:161–180. doi: 10.1002/rsa.3240060204

35. Lancichinetti A. Codes. Available: https://sites.google.com/site/andrealancichinetti/software [Accessed 2 December 2015].

36. Fred ALN, Jain AK. Robust data clustering. In: Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 2. IEEE; 2003. pp. 128–133.

37. Lancichinetti A, Fortunato S. Community detection algorithms: A comparative analysis. Phys Rev E. 2009; 80:056117. doi: 10.1103/PhysRevE.80.056117

38. Lancichinetti A, Fortunato S. Erratum: Community detection algorithms: A comparative analysis [Phys. Rev. E 80, 056117 (2009)]. Phys Rev E. 2014; 89:049902(E). doi: 10.1103/PhysRevE.89.049902

39. Condon A, Karp RM. Algorithms for graph partitioning on the planted partition model. Random Struct Algorithms. 2001; 18:116–140. doi: 10.1002/1098-2418(200103)18:2%3C116::AID-RSA1001%3E3.0.CO;2-2

40. Erdős P, Rényi A. On random graphs I. Publ Math Debrecen. 1959; 6:290–297.

41. Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Phys Rev E. 2008; 78:046110. doi: 10.1103/PhysRevE.78.046110

42. Lancichinetti A, Fortunato S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys Rev E. 2009; 80:016118. doi: 10.1103/PhysRevE.80.016118

43. Rosvall M, Bergstrom CT. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. PLoS ONE. 2011; 6:e18209. doi: 10.1371/journal.pone.0018209 PMID: 21494658

44. Zachary WW. An information flow model for conflict and fission in small groups. J Anthropol Res. 1977; 33:452–473.

45. Knuth DE. The Stanford GraphBase: A Platform for Combinatorial Computing. Addison-Wesley, Reading, MA; 1993.

46. Girvan M, Newman MEJ. Community structure in social and biological networks. Proc Natl Acad Sci USA. 2002; 99:7821–7826. doi: 10.1073/pnas.122653799 PMID: 12060727