PLOS ONE

RESEARCH ARTICLE

Statistical test for detecting community structure in real-valued edge-weighted graphs

Tomoki Tokuda*

Okinawa Institute of Science and Technology Graduate University, 1919-1, Tancha, Onna-son, Okinawa, Japan

* tomoki.tokuda@oist.jp

Abstract

We propose a novel method to test the existence of community structure in undirected, real-valued, edge-weighted graphs. The method is based on the asymptotic behavior of extreme eigenvalues of a real symmetric edge-weight matrix. We provide a theoretical foundation for this method and report on its performance using synthetic and real data, suggesting that this new method outperforms other state-of-the-art methods.

Citation: Tokuda T (2018) Statistical test for detecting community structure in real-valued edge-weighted graphs. PLoS ONE 13(3): e0194079. https://doi.org/10.1371/journal.pone.0194079

Editor: Sergio Gómez, Universitat Rovira i Virgili, SPAIN

Received: October 11, 2017

Accepted: February 23, 2018

Published: March 20, 2018

Data Availability Statement: All real data files are available from UC Irvine Machine Learning Repository database (http://archive.ics.uci.edu/ml/index.php).

Funding: The author received no specific funding for this work.

Competing interests: The author has declared that no competing interests exist.

Introduction

Clustering objects based on their similarities is a basic data mining approach in statistical analysis. In particular, graphical data (or network data) that reflect relationships between nodes, are often acquired in various scientific domains such as protein-protein interaction, neural networks, and social networks [1], which potentially provide useful information on the underlying structure of the system in question.

Specifically, our interest is to detect possible 'community', or cluster structure of undirected graphs, which is defined as block structure of a graph (Fig 1A), where the corresponding edge-weight matrix consists of several cluster blocks (four cluster blocks in Fig 1B). To detect such structure, a number of clustering methods have been proposed in the statistical physics and information theory literature [2–4]. Mainly, there are four approaches: graph partitioning, hierarchical clustering, partitional clustering, and spectral clustering [1, 4, 5].

However, the conventional framework for analysis of community structure is typically an unsigned graph in which an edge weight is constrained to be non-negative. Recently, increased attention has been paid to analyzing signed graphs that allow negative weights [6]. Indeed, in real data, it is often essential to account for negative, as well as positive relationships, for a better understanding of the underlying community structure in a graph such as a social network. Most methods in the literature, however, address this problem in a rather limited framework in which edge weights within a cluster are positive while those between clusters are negative (i.e., weakly balanced structure) [6]. On the other hand, how to cluster nodes in a more general framework, such as negative edge weights within a cluster, remains an open question [7].

In the present paper, we address the question of community detection in a real-valued graph. Let us consider a general framework for community structure as follows. We assume that edge weights are independently generated from a generative model that is specific to a

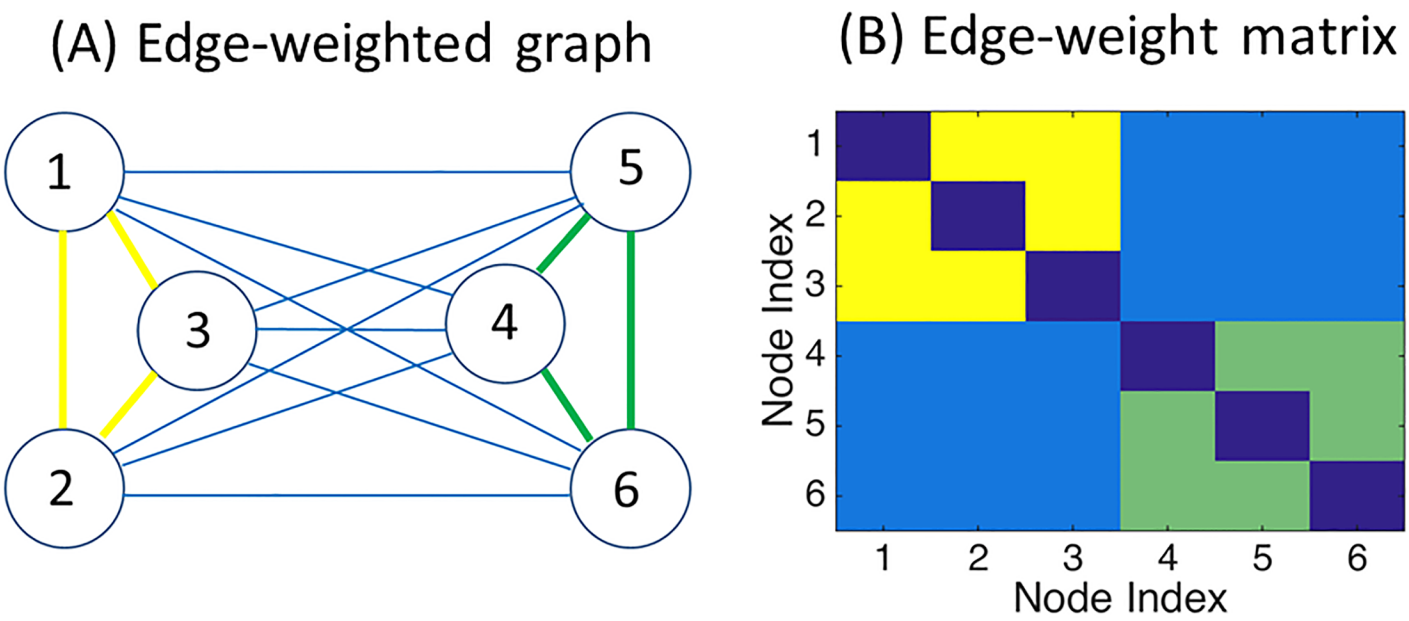

Fig 1. Illustration of two-way community structure in a graph. Panel (A): Graphical representation (edge-weighted graph). Panel (B): Matrix representation (edge-weight matrix), where strengths of relationships between nodes are denoted in color.

https://doi.org/10.1371/journal.pone.0194079.g001

particular cluster block, which characterizes a distribution of edges in each cluster block. Further, we assume that these distributions are distinguishable in terms of their mean and variance. For this framework, as a first step toward addressing a clustering problem, we aim to develop a statistical method for testing the existence of underlying community structure.

From the theoretical point of view, there is the issue of detectability of community structure. In the case of unweighted graphs, this issue has been intensively studied because of both mathematical and physical interest [8–10]. In the situation in which an edge connection is generated by a probability $P_{ab} = c_{a,b}/n$ where $c_{a,b}$ is constant and n is the number of nodes, it has been shown that it is impossible for any algorithm to detect underlying community structure (as $n \to \infty$) under certain circumstances. Further, it is shown that instead of a conventional adjacency matrix, a non-backtracking matrix, which represents non-backtracking walks in a network, provides a better platform for detection algorithms [11]. In the present paper, however, we focus on the case in which a generative model for edges has fixed parameters, irrespective of the number of nodes n. In the context of unweighted graphs, this suggests that $P_{a,b} = c_{a,b}$. In this situation, it was shown that it is possible to detect community structures (in case of bisection) as n goes to ∞ [12, 13]. In the present paper, we consider such a case.

Regarding statistical tests on community structure, several methods have been proposed in the context of unsigned (weighted or unweighted) graphs [1]. A common approach to this problem is to evaluate the stability of cluster solutions when the data in question are noisy [14, 15]. If similar cluster solutions are obtained for graphs with some perturbation of edge-weights, this suggests the stability of the cluster solution for the original graph, providing the evidence of the community structure. The bootstrap method employs [16] a similar approach. A second approach is based on comparisons of cluster solutions for the original graphs with solutions of randomly permuted graphs. As a statistic for testing significance, the entropy of graph configurations [17], or 'C-score' focusing on the lowest internal degrees [18] have been proposed. The common feature of these state-of-the-art methods is that a cluster solution to a given graph is required for testing. In other words, the test result depends on the clustering

method employed. In this sense, these methods test the significance of a resulting cluster solution, rather than the existence of community structure itself. For the general framework of our interest, such an approach is not applicable because appropriate clustering methods are not readily available. In [19–21], the spectral homophily of a multi-type random network has been proposed to capture connectivity between communities. This method uses the second largest eigenvalue of a symmetric matrix with expected fractions of the links where the partition in communities is exogenous. However, it is not straightforward to apply their results to our setting of real-valued edge-weights. Moreover, we consider a situation in which the partition in communities is not exogenous.

We propose a general method for testing community structure of edge-weighted graphs with real-valued weights, which does not require a cluster solution. Our method is based on the asymptotic behavior of eigenvalues of the normalized weight matrix of graph, which is described by Wigner semicircular law when there is no community structure. As in our approach, in the case of binary-valued graphs, a statistical test for community structure has recently been proposed [22], based on the exact asymptotic behavior of (maximum) eigenvalues. However, that method is not directly applicable to real-valued graphs that account for both mean and variance, because the Bernoulli distribution assumed in their method cannot properly capture these quantities. Our method provides a nontrivial extension of community structure detection to real-valued graphs, and broad applications to network data. In the following sections, first, a theoretical foundation for our method is provided. Second, it is shown that our method outperforms other methods with synthetic data. Third, we apply our method to real data.

Method

Our statistical test on community structure is based on the probability distribution of eigenvalues of the normalized edge-weighted matrix (we define 'normalization' later). We make the best use of asymptotic results on such a distribution when there is no community structure, which has been intensively studied in the field of Random Matrix Theory of Theoretical Physics [23]. In this section, we provide a theoretical foundation for our statistical test.

Setting

We consider a clustering problem of nodes for undirected edge-weighted graphs $G = (V, E)$ where V consists of n vertices $\{v_1, \ldots, v_n\}$, and E is represented by the edge-weight matrix $\boldsymbol{W}_n$, which is a $n \times n$ symmetric (real Hermitian) matrix with elements $w_{i,j} = w_{j,i} \in \mathbb{R}$ and $w_{i,i} = 0$ ($\mathbb{R}$ denotes a set of real numbers). Let us assume that there are K clusters of nodes, denoting them as $c_1, \ldots, c_K$. We define a cluster block (k, k') as a set of weights $w_{i,j}$ such that nodes i and j belong to the cluster c_k and $c_{k'}$, respectively: $v_i \in c_k$ and $v_j \in c_{k'}$ $(1 \leq k, k' \leq K)$. Here, we assume that each off-diagonal weight $w_{i,j}$ is independently drawn from a certain distribution. With this assumption, we define a K-way community structure as characterized by different distributions in $K \times K$ cluster blocks. To elaborate this definition, we assume the following distribution for each cluster block:

$$\begin{aligned} w_{i,j} &\sim g_{k,k'} \quad (i \neq j) \\ g_{k,k'} &= \mu_{k,k'} + g \times \sigma_{k,k'}, \end{aligned} \tag{1}$$

where $v_i \in c_k$, $v_j \in c_{k'}$, and g is a certain probability distribution. This definition suggests that a pair of parameters $(\mu_{k,k'}, \sigma_{k,k'})$ characterizes each cluster block, hence, community structure

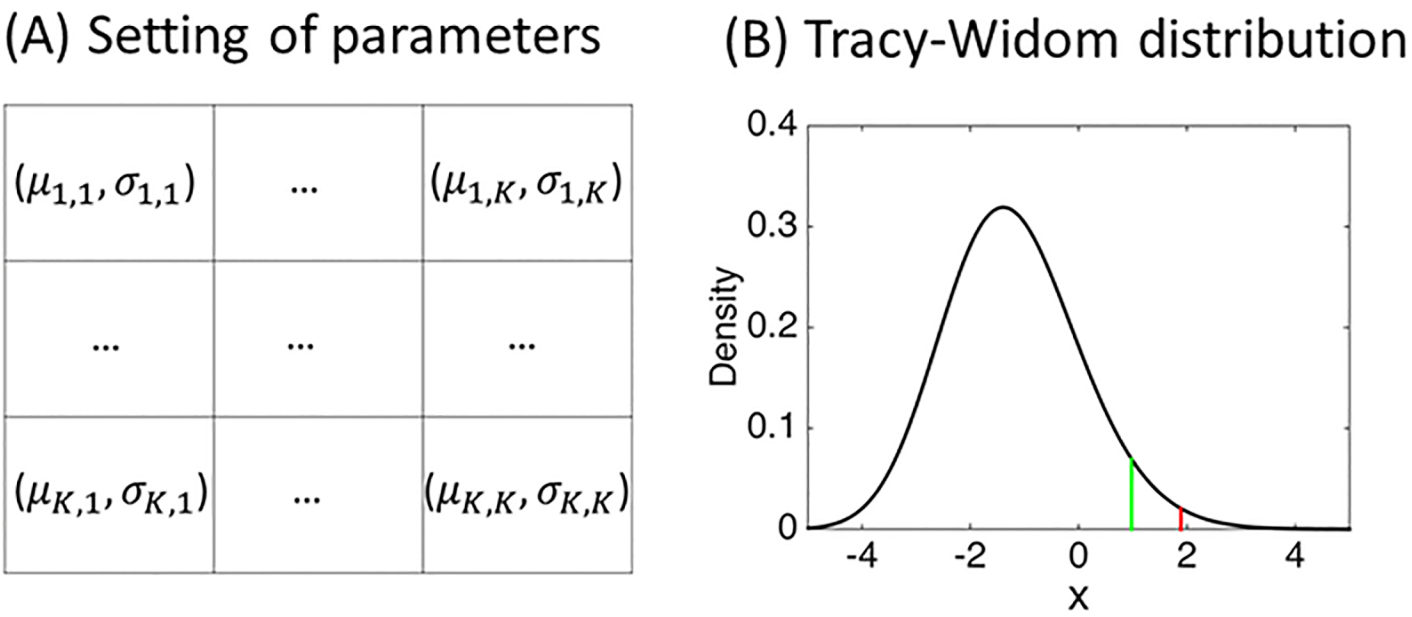

Fig 2. Setting of community structure and Tracy-Widom distribution. Panel (A): Illustration of the setting of community structure in a matrix representation where nodes are arranged in the order of cluster labels. Each cluster block is characterized by mean μ and standard deviation σ with cluster block index (k, k'). Panel (B): The density function of the Tracy-Widom distribution for Gaussian orthogonal ensembles with $\beta = 1$ (the first derivative of $F_1(x)$ in Eq (7)), generated by the function *dtw* in R-package {RMTstat}. The critical values at significance level $\alpha = 0.05$ and $\alpha/4 = 0.0125$ are 0.979 (green line) and 1.889 (red line), respectively.

https://doi.org/10.1371/journal.pone.0194079.g002

(Fig 2A). Note that in this definition we exclude the degenerate case in which $\mu_{k,k'} = constant$ and $\sigma_{k,k'} = 0$ such that variances become zero for the whole set of $\{w_{i,j}\}$.

Since the community structure of interest is based on differences of weight distributions, it is translation- and scale-invariant for all weights. Hence, to simplify the problem, as a preprocess, we standardize off-diagonal elements of $\boldsymbol{W}_n$ using all off-diagonal weights $w_{i,j}(i \neq j)$ so that the mean is zero and the variance one. We denote as S the mapping that standardizes the edge-weight matrix in this way, transforming each element of the matrix as

$$\begin{aligned} S: \quad & w_{i,j} \to (w_{i,j} - \mu)/\sigma \quad \text{for } i \neq j \\ & w_{i,i} \to 0, \end{aligned} \tag{2}$$

where μ and σ are the mean and the standard deviation of the whole off-diagonal elements $\{w_{i,j}\}$. Practically, these mean and standard deviation may be replaced by the empirical counterparts μ_{emp} and σ_{emp}. For the standardized edge-weight matrix $S(\boldsymbol{W}_n)$, we assume that the mean and the standard deviation of g in Eq (1) are zero and one, respectively. In this setting, the mean and the standard deviation in cluster block (k, k') are $\mu_{k,k'}$ and $\sigma_{k,k'}$, respectively. The differences of these parameters distinguish between clusters in terms of the first and second moments, while controlling higher moments than two. Using this setting of community structure, we define no community case as a single community with $K = 1$ where $\mu_{k,k'} = 0$ and $\sigma_{k,k'} = 1$ for $S(\boldsymbol{W}_n)$. Note that since g is arbitrary, including a mixture distribution of a certain distribution family, our definition of no community structure includes the case in which each weight is generated from a specific distribution in a list of distributions in random order. Importantly, when we shuffle the off-diagonal elements $\boldsymbol{W}_n$ at random (in element-wise manner), the community structure always disappears. Indeed, in such a case, each element $w'_{i,j}$ of the shuffled matrix $\boldsymbol{W}'_n$ independently and identically follows the mixture distribution

consisting of different components, i.e., $\sum_{k,k'} \pi_{k,k'} g_{k,k'}$ where $\pi_{k,k'}$ is the proportion of elements of cluster block (k, k') for the original matrix $\boldsymbol{W}_n$. We use this property for our statistical test as an alternative way to estimate confidence intervals.

Statistical test

In this section, we develop a statistical test for the existence of community structure defined in the previous section (i.e., $K = 1$ vs. $K > 1$). We base our test on the asymptotic behavior of the eigenvalues of $S(\boldsymbol{W}_n)$ (n goes to ∞) when there is no community structure. A useful result of Random Matrix Theory in this context is that if the elements of an infinite dimensional symmetric matrix $\boldsymbol{X}$ independently follow a certain distribution with mean zero and variance one, then the empirical (random) distribution of the eigenvalue λ of $\boldsymbol{X}_n/\sqrt{n}$, where $\boldsymbol{X}_n$ is the principal submatrix of $\boldsymbol{X}$ for the first n rows and columns, converges almost surely to a Wigner semicircular distribution as n goes to ∞ (semicircular law).

$$f_{sc}(\lambda) \equiv \frac{1}{2\pi}\sqrt{4-\lambda^2}.$$

Note that this law holds for any generative distribution of the elements in matrix $\boldsymbol{X}$ (as long as independently drawn), which is referred to as the universality property of the law. Also, this law holds even if we replace diagonal elements with zero's, as in our case. Further, strong Bai-Yin theorem suggests that with the additional condition of the distribution of each element (namely, the existence of a fourth moment), the largest magnitude of eigenvalues is almost certainly bounded by 2. These two theorems imply that the largest magnitude of eigenvalues almost surely converges to two ([24], p.136).

In order to apply this property to our context, we consider a normalization mapping of edge-weight matrix $\boldsymbol{W}_n$, transforming each element of the matrix as

$$T: \qquad w_{i,j} \rightarrow S(w_{i,j})/\sqrt{n}, \tag{3}$$

where S is the standardization mapping in Eq (2). Now, let us assume that the elements in an edge-weight matrix $\boldsymbol{W}_n$ are generated as in Eq (1). In this setting, if the largest magnitude of eigenvalues of $T(\boldsymbol{W}_n)$ does not converge to two, then, there should be some K-way community structure in the graph ($K > 1$) because of our assumption in Eq (1) (Note that without the assumption in Eq (1), this property does not hold. For instance, one can make a scale-free graph where the eigenvalues do not follow the semicircular law [25]). However, the converse argument does not necessarily hold. That is to say, the fact that the convergence of the largest magnitude of eigenvalues to two does not imply that there is no community structure (i.e., $K = 1$). A simple counter example is given as follows (proof in S1 Appendix).

Example 1. *Let $\boldsymbol{W}_n$ be a $n \times n$ symmetric edge-weight matrix that has K-way community structure with the same cluster size (n/K) as defined in the previous section. Suppose that $\mu_{k,k'} = 0$ for $\forall k, k'$, $\sigma^2_{k,k'} = 0$ for $k \neq k'$, and $\sigma^2_{k,k} = 1$. Then, the largest magnitude of eigenvalues of $T(\boldsymbol{W}_n)$ almost surely converges to two as n goes to ∞.*

Nonetheless, in our setting, we can show that an additional condition on the eigenvalue distribution for an exponentially mapped edge-weight matrix ensures that the converse argument also holds. For this purpose, we introduce the exponential mapping *Exp* that transforms each element of $\boldsymbol{W}_n$ as

$$\begin{aligned} Exp: \qquad & w_{i,j} \rightarrow \exp(t \times w_{i,j}) \quad \text{for } i \neq j \\ & w_{i,i} \rightarrow 0, \end{aligned} \tag{4}$$

where $t \in \mathbb{R}$ is a tuning parameter (we do not explicitly denote the dependence of Exp on t because of cluttering). Subsequently, we define the normalization mapping T_e for the exponentially transformed matrix as

$$T_e : \quad w_{i,j} \to S(Exp(w_{i,j}))/\sqrt{n}. \tag{5}$$

Now, the following theorem provides a necessary and sufficient condition for the existence of community structure (proof in S2 Appendix).

Theorem 1. *Let $\mathbf{W}_n$ be a $n \times n$ weight matrix defined in the previous section with the fixed proportion of cluster sizes $(r_1, \ldots, r_K)$ and the pairs of fixed parameters $\{(\mu_{k,k'}, \sigma_{k,k'})\}(k, k' = 1, \ldots, K)$. Suppose that there exists the moment-generating function $M(t)$ in an open interval containing zero for g (g is defined in Eq (1)). Then, the following statements (C1) and (C2) are equivalent*:

(C1) There is no community structure (i.e., $K = 1$)

(C2) Each of the largest magnitudes of eigenvalues of $T(\mathbf{W}_n)$ and $T_e(\mathbf{W}_n)$ for any non-zero real value $t_0 \neq 0$ almost surely converges to two, as n goes to ∞.

Theorem 1 motivates us to use the largest magnitude of eigenvalues of edge-weight matrix to establish a statistical test on the null hypothesis H_0:

$$H_0 : \text{There is no community structure.} \tag{6}$$

Practically, to test the null hypothesis H_0, we focus on positive and negative extreme values of eigenvalues. The largest eigenvalue may deviate positively from two, while the smallest eigenvalue may deviated negatively from -2.

Comments.

1. Strictly speaking, the independent assumption on weights is broken if we transform them by T or T_e using an empirical mean and standard deviation μ_{emp} and σ_{emp}. For simplicity, however, we ignore such an effect in the present paper.
2. Our method is not applicable to a directed graph, because in that case an edge-weight matrix becomes non-symmetric; hence, Theorem 1, which is based on properties of eigenvalues of symmetric matrices, does not hold.
3. The spectral method in [19–21] takes a slightly different approach from our method. In the context of unweighted graph, they consider a symmetric matrix representing an expected fraction of edges between two communities (hence, the summation of entries in a row is one). Further, the size of the matrix in their approach is $K \times K$ while that of our method is $n \times n$. Moreover, in their approach, the second largest eigenvalue is considered because the largest eigenvalue is constant (always one). As other community detection methods, it is not trivial to generalize their method to real-valued graphs.
4. Theorem 1 implies that if there is no community structure (i.e., edges are i.i.d. generated), the ratio of the second largest eigenvalue to the largest eigenvalue converges to one. This is a general property to a symmetric matrix. This result is contrasted with Brody's conjecture that for a positive non-symmetric matrix in which all edges are i.i.d. generated, the ratio of the second largest eigenvalue to the largest eigenvalue goes to zero [26–28].
5. The exponentially transformed matrix in Theorem 1 does not replace non-backtracking matrix in [11]. Rather, the exponentially transformed matrix serves to capture differences of variances in generative models.

6. In the current setting, means and variances in generative models are fixed. One may wonder how much this condition can be relaxed. From the argument about unweighted graphs by [12, 29], we speculate that one may be able to relax the condition such that the difference of means may be larger than $\Omega(\log n/n)$. Indeed, assuming that this condition holds after standardizing the edge-weight matrix, Eq (3) in S2 Appendix becomes $\lambda_1(M_n) \geq A(\log n)^2$. Hence, in this relaxed condition as well as in non-relaxed condition, it holds that if there is community structure with equal variances the largest magnitude of eigenvalues does not converge to two (the first part of the proof of Theorem 1). Moreover, using the relaxed condition, it can be shown that for the unstandardized matrix, mean differences should be also larger than $\Omega(\log n/n)$. Note that from the prime number theorem [30], the reciprocal of $\log n/n$ denotes the number of prime numbers less than n. This observation provides us the following interpretation of the results. If we assume that the community size is the same across different communities, the number of prime numbers denotes the number of irreducible topologies of communities (for example, four-community structure may be reduced to two-community structure by paring two communities while three-community structure is not reducible). We speculate that such an irreducible community structure is easier to detect than the remainder of structures and that if such structures are more available, the detection of community structure becomes easier. This interpretation suggests that the number of prime numbers may be inversely related to detectable differences of means, which is consistent with the lower bound $\Omega(\log n/n)$.

7. One may wish to tune the value of t_0 as follows. We consider two-community structure, assuming that edges within communities are generated by $N(\mu, \sigma_1^2)$ while edges between communities by $N(\mu, \sigma_2^2)$. The first moment (mean) of the exponentially transformed variable generated from $N(\mu, \sigma)$ is given by $\exp(\mu t + \sigma^2 t^2/2)$ while the second moment is $\exp(2\mu t + 2\sigma^2 t^2)$. Using these results, we can analytically evaluate the difference in means of the exponentially transformed variables between $N(\mu, \sigma_1^2)$ and $N(\mu, \sigma_2^2)$ normalized by the square root of the average of variances. It can be shown that in this case μ is irrelevant for the normalized difference. So, one may choose t_0 that maximizes the normalized difference for a given σ_1 and σ_2, or, simply the ratio $r = \sigma_1^2/\sigma_2^2$. From Fig 3, one may choose t_0 between 0.5 and 1.

The behavior of the largest eigenvalue has been well studied in the literature when elements of the edge-weight matrix $\boldsymbol{W}_n$ are independently generated by certain symmetric distributions g (typically Gaussian, otherwise, its density function may be even with less heavier tails than Gaussian distributions) with mean zero and variance one for non-diagonal elements and with mean zero and variance two for diagonal elements. In this setting, the largest eigenvalue λ_{max} asymptotically follows the Tracy-Widom distribution for Gaussian orthogonal ensembles with parameter $\beta = 1$:

$$\lim_{n\to\infty} P(\lambda_{max} \leq 2 + x/n^{2/3}) = F_1(x), \tag{7}$$

where $F_1(x) \equiv \exp\{-(1/2)\int_x^\infty q(y)dy\}(F_2(x))^{1/2}$ with $F_2(x) \equiv \exp\{-\int_x^\infty (y-x)q^2(y)dy\}$ where $q(x)$ is the solution of Painlevé II equation $d^2q/dx^2 = xq + 2q^3$ with the boundary condition $q(x) \sim \mathrm{Ai}(x)$ as $x \to \infty$ [31, 32]. Note that the Tracy-Widom distribution is for the maximum eigenvalue of a specific type of symmetric matrix (e.g., Gaussian ensembles) while the semicircular law holds for the distribution of eigenvalues in a general type of symmetric matrix (Wigner ensembles). Moreover, in our framework, the diagonal elements are all zero, which is a slightly different situation than the conventional assumption for the Tracy-Widom

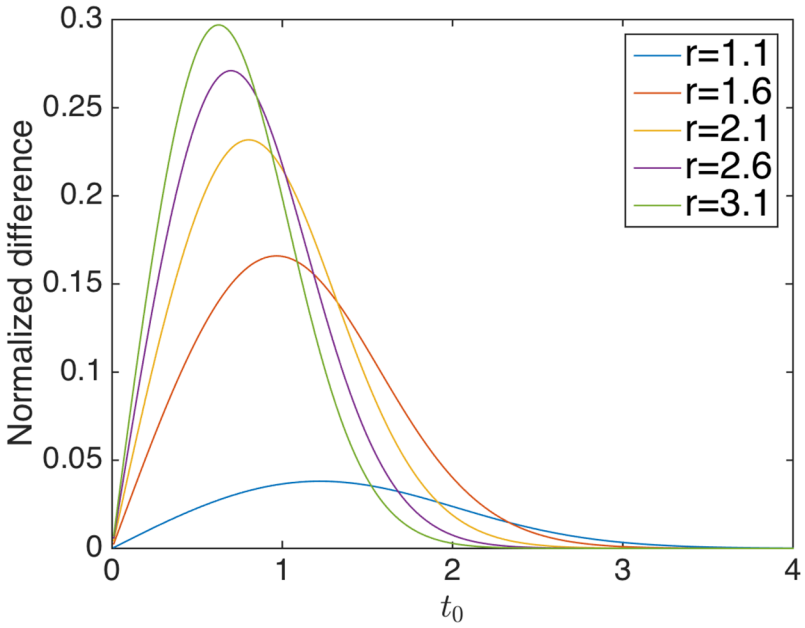

Fig 3. Normalized differences of exponentially transformed variables between normal distributions $N(\mu, \sigma_1)$ and $N(\mu, \sigma_2)$. The X-axis denotes t_0 in Theorem 1 and the Y-axis normalized difference. We set $r = \sigma_1^2/\sigma_2^2$ to 1.1, 1.6, 2.1, 2.6, and 3.1.

https://doi.org/10.1371/journal.pone.0194079.g003

distribution. Nevertheless, because of the universality property of the Tracy-Widom distribution ([33], Theorem 21.4.3), we can safely apply Eq (7) to our context (obviously, our context satisfies the condition of universality that the diagonal part should be symmetric with a sub-Gaussian tail).

Using the Tracy-Widom distribution in Eq (7), we set confidence intervals for our statistical test as follows. For the normalized edge-weight matrix $T(\boldsymbol{W}_n)$, we set the confidence interval CI_{max} of the largest eigenvalue λ_{max} at level α. Since the violation of the semicircular law occurs as the positive deviation from the expected value, we consider the one-sided confidence interval as $(-\infty, q)$ where q is a critical value at significant level α, i.e., $P(\lambda_{max} \geq q|H_0) = \alpha$, which is estimated by $F_1(x)$ in Eq (7) (refer to the shape of its first derivative in Fig 2B). If the generative distribution g is not symmetric or is heavy-tailed, one may evaluate the distribution of the largest eigenvalues by means of a permutation test for $T(\boldsymbol{W}_n)$. Though the permutation test may provide an accurate confidence interval, it is not computationally efficient because we need to compute eigenvalues a large number of times. Therefore, when the number of nodes is large, one may opt for the Tracy-Widom distribution to efficiently obtain confidence intervals. In addition to the largest eigenvalue, we also test the smallest eigenvalue λ_{min}, which may violate the semicircular law (what matters is indeed the largest magnitude of eigenvalue). In this case, the confidence interval CI_{min} is given by $(-q, \infty)$. In similar fashion, we test the largest and the smallest eigenvalue of the exponentially normalized weight matrix. We first standardize the data and then apply the mapping T_e where we set t_0 to 1/2 as default. This results in the transformed matrix $T_e(S(\boldsymbol{W}_n))$ (we denote the confidence intervals as CI'_{max} and CI'_{min}, respectively). Since this procedure involves a series of four statistical tests, we set the level of significance to $\alpha/4$ for each test, taking into account the Bonferroni correction (Algorithm 1; $\mathbb{I}(a)$ is an indicator function: 1 for correct a; 0 otherwise).

Algorithm 1. Testing the existence of community structure

Input: Edge-weight matrix $\boldsymbol{W}$, confidence intervals CI_{max}, CI_{min}, CI'_{max} and CI'_{min} at level $\alpha/4$.

$s \leftarrow 0$

$s \leftarrow s + \mathbb{I}$ (max. eigenvalue of $T(\boldsymbol{W}) \in CI_{max}$)

$s \leftarrow s + \mathbb{I}$ (min. eigenvalues of $T(\boldsymbol{W}) \in CI_{min}$)

$s \leftarrow s + \mathbb{I}$ (max. eigenvalue of $T_e(S(\boldsymbol{W})) \in CI'_{max}$)

$s \leftarrow s + \mathbb{I}$ (min. eigenvalue of $T_e(S(\boldsymbol{W})) \in CI'_{min}$)

if $s = 4$ **then**

  Accept H_0

else

  Reject H_0

end if

Simulation study on synthetic data

In this section, we report on a simulation study to evaluate the performance of our method. First, we investigate the validity of using $F_1(x)$ in Eq (7) to approximate the distribution of the maximum eigenvalue λ_{max} when n is finite. Second, we investigate the power of our method when the null hypothesis H_0 is not true.

Third, we compare the performance of our method outlined in Algorithm 1 with other methods. Basically, existing methods consist of two steps. In the first step, a clustering solution for a given graph is produced by a (arbitrary) clustering method. The resulting solution is subsequently compared with clustering solutions for randomized graphs, and is further evaluated with a specific statistic. In this study, we adapt one of the state-of-the-art methods based on clustering entropy ('CE', originally designed for a unweighted graph) [14]: $S = -\frac{1}{L}\sum_{(i,j)}\{p_{i,j}\log_2 p_{i,j} + (1-p_{i,j})\log_2(1-p_{i,j})\}$ where L is the total number of edges in the graph, and $p_{i,j}$ is 'in-cluster probability' that measures the proportion of concordance of cluster memberships of nodes i and j between the given graph and the randomized graph over a number of different noisy contaminations (we set the number of such contaminations to 100). Regarding clustering, to the best of our knowledge, there is no clustering method that is specifically designed to detect community structure based on differences of distribution patterns. As a bail-out procedure, we consider one of the state-of-the-art methods for signed networks: Signed spectral clustering based on a normalized, signed Laplacian method ('SignedSpec'), which is designed to detect weakly balanced structure of graphs, i.e., positive weights within clusters and negative weights between clusters [6]. We also consider conventional spectral clustering (normalized Laplacian method, 'ConvSpec'), which is applicable to graphs with positive weights. To apply the method 'ConvSpec' in our context, we transform an edge-weight matrix into a positively-weighted matrix by subtracting $\min_{i,j} w_{i,j}$ from each weight. Note that the method 'ConvSpec' is equivalent to the method 'SignedSpec' when edge weights are all positive.

Data generation

For the data structure in this simulation study, we adopted that in [34], setting the number of clusters to five and cluster size to ($10s$, $20s$, $30s$, $40s$, $50s$), where we manipulated integer s. In this setting, we have $5 \times 5 = 25$ cluster blocks. In each cluster block, weights were independently drawn from a Gaussian distribution $N(\mu_{k,k'}, \sigma_{k,k'})$ where $\mu_{k,k'}$ and $\sigma^2_{k,k'}$ are the mean and the variance for a cluster block (k, k'). We generated 100 datasets for each setting.

Results

When the number of nodes ranges from 150 to 1500, the distribution function $F_1(x)$ in Eq (7) provides a good approximation of the critical value at a significance level of $\alpha = 0.05$ under the null hypothesis H_0 (Fig 4A). Since the function $F_1(x)$ provides the asymptotic probability distribution, this result suggests that the function $F_1(x)$ also provides a good approximation of the critical value when the number of nodes exceeds this range. In regard to statistical power, it is implied that our method can readily detect the existence of community structure when means $\mu_{k,k'}$ in each block differ by at most 0.3 ($3 \times 0.05 + 3 \times 0.05$) when $\sigma_{k,k'} = 1$ with the number of nodes being 750 (Fig 4B). On the other hand, the power may not be sufficient when differences

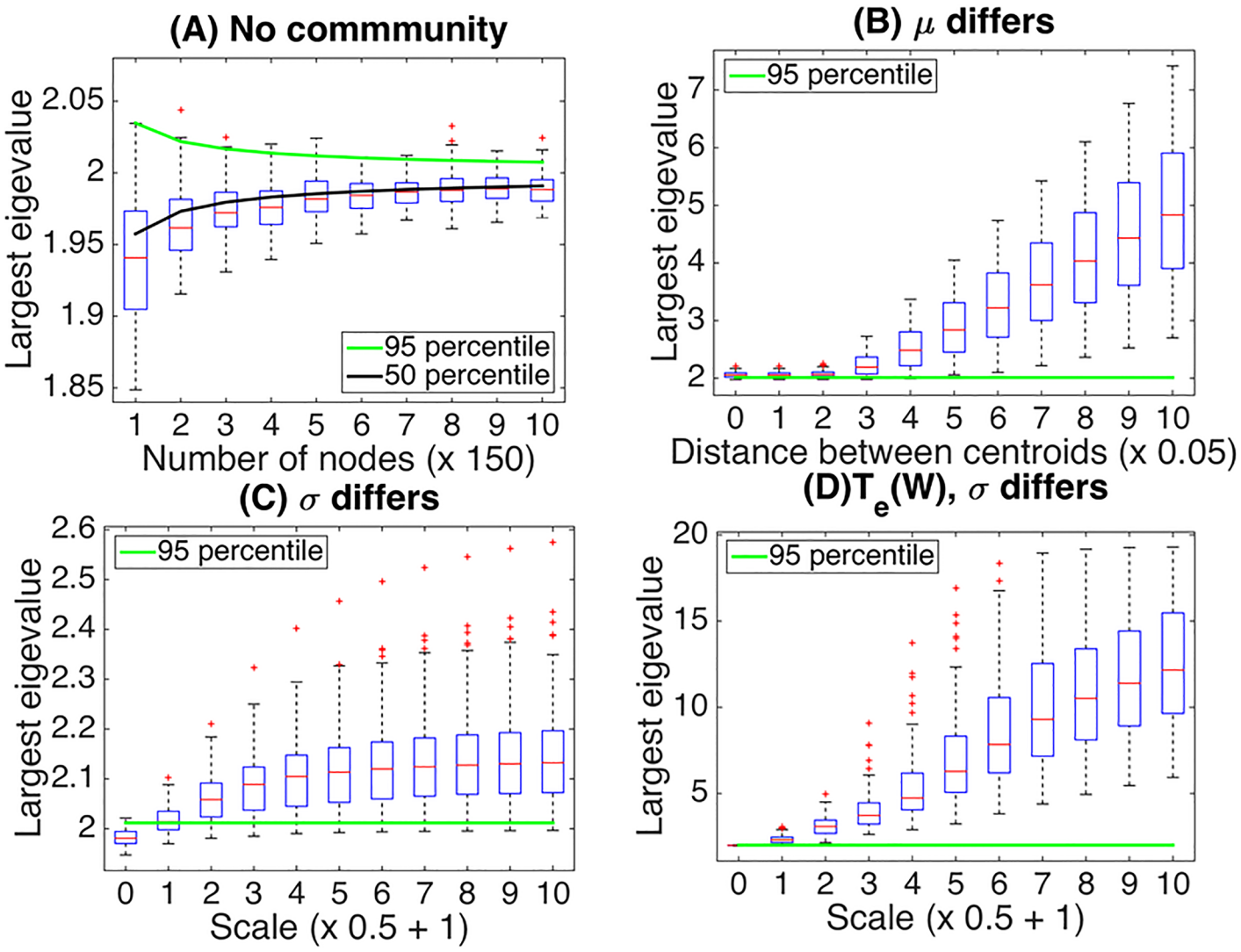

Fig 4. Boxplots represent distributions of the largest eigenvalues for various settings. Panel (A): No-community case ($K = 1$) of Gaussian ensembles for different number of nodes from 150 to 1500 in x-axis. Panel (B): Five-way community case with the number of nodes 750 and cluster size (50, 100, 150, 200, 250). Each cluster block is characterized by means of a Gaussian distribution (while fixing variance = 1), which is randomly chosen from $\{-\mu, \mu\}$ with equal probabilities. The value of μ is manipulated from 0 to 0.5 of width 0.1 in x-axis. Panel (C): A five-way community case characterized by variance (while fixing mean = 0), which is randomly chosen from $\{1, \sigma^2\}$ with equal probabilities. The value of σ is manipulated from 1 to 6 of width 1 in x-axis. Panel (D): A five-way community case in the same setting as in (C), but each edge-weight matrix is transformed by the exponential mapping *Exp* in Eq (4) with $t_0 = 1/2$. In all panels, the green line denotes the 95 percentile of the largest eigenvalue under the null hypothesis H_0 in (6).

https://doi.org/10.1371/journal.pone.0194079.g004

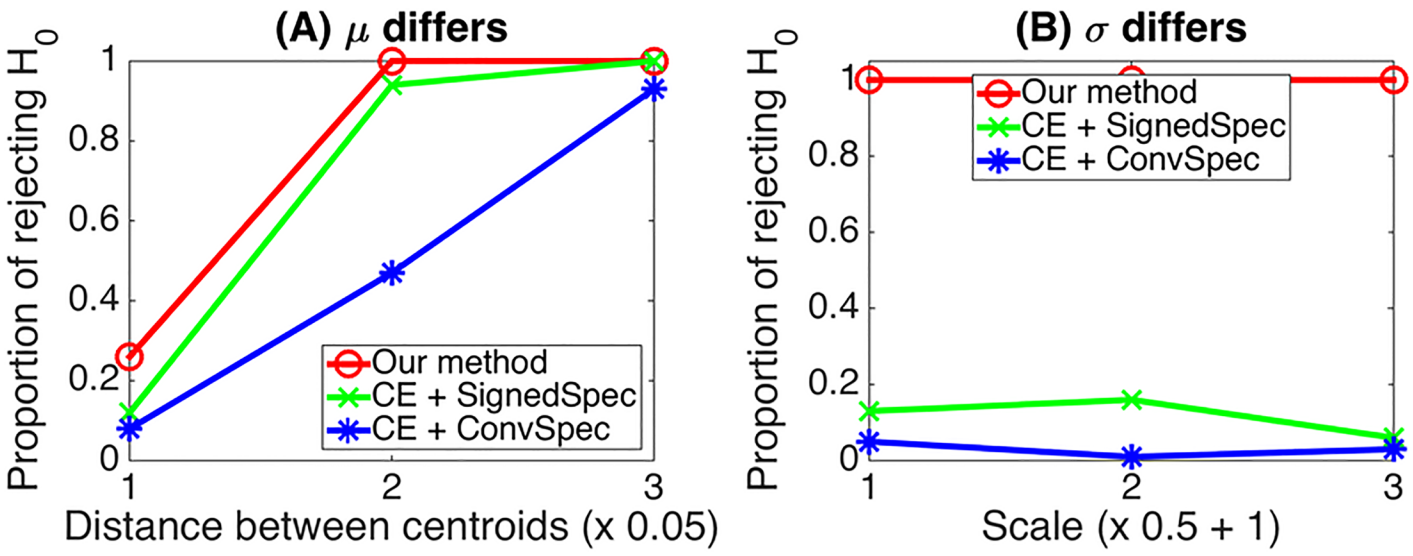

Fig 5. Comparison of the power of the test for three different methods. Our method, the clustering entropy method for the resulting cluster solution using the signed spectral clustering method (CE + SignedSpec), and the clustering entropy method using conventional spectral clustering (CE + ConvSpec). The true community structure is set as follows: cluster size (50, 100, 150, 200, 250); means and variances are manipulated in x-axis of Panel (A) and (B) as in Fig 4B and 4C, respectively.

https://doi.org/10.1371/journal.pone.0194079.g005

among cluster blocks are characterized by variances $\sigma^2_{k,k'}$ (Fig 4C). However, the application of our method to the exponentially transformed matrix by *Exp* considerably improves the power (Fig 4D). All these results suggest good performance of our method in testing for the existence of community structure in a graph.

Lastly, we compare the performance of our method with the remaining methods. We applied our method as outlined in Algorithm 1 to synthetic data, setting α to 0.05 (hence, $\alpha/4 = 0.0125$). When the community structure is characterized by mean differences, the performance of our method is comparable with the clustering entropy method with signed spectral clustering (CE + SignedSpec), while it outperforms the clustering entropy method with conventional spectral clustering (CE + ConvSpec) (Fig 5A). On the other hand, when the community structure is characterized by scale differences, our method considerably outperforms other methods (Fig 5B).

Application to real data 1

In this section, we test our method on real data. The objective is to evaluate the performance of our method when it is applied to various types of real graph data.

Data

First, we applied our method to the following benchmark graph datasets: Karate club, *Karate* [35]; co-authorships in network science, *Co-authours* [36]; Tribal relationships in highland New Guinea, *Gahuku-Gama* [37]. The datasets of *Karate* and *Co-authours* are binary (i.e., {0, 1}), while the edges in the dataset of *Gahuku-Gama* take discrete signed values, {−1, 0, 1}. The number of nodes for these datasets are 34, 1589, and 16, respectively. These datasets have been well studied in terms of detecting community structure [7].

Second, we applied our method to a real-valued edge-weighted graph: resting state functional MRI (*fMRI*) data [38]. The original dataset consists of the level of BOLD (Blood-Oxygen-Level Dependent) signals at short intervals, which reflects neural activity at tiny regions of

the brain, called 'voxel' (4949 voxels in this dataset). We pre-processed this dataset by evaluating temporal correlations among these voxels and carrying out Fisher's z-transformation for them, which results in a 4949 edge-weight matrix $\boldsymbol{W}$. The objective in this dataset is to test our method on a real-valued, edge-weight matrix and to draw useful inferences from the analysis.

Results

For the first group of real datasets, our method finds some community structure (i.e., $K > 1$), whether we estimate critical values using the Tracy-Widom distribution or a permutation test (Fig 6). Note that in the binary case, we always obtain the same results for the original matrix and for the exponentially transposed matrix, because $T(\boldsymbol{W}) = T_e(S(\boldsymbol{W}))$. So, we tested only $T(\boldsymbol{W})$ in *Karate* and *Co-authors* datasets, setting the significance level to $\alpha/2$.

It is observed in Fig 6 that confidence intervals largely match between the Tracy-Widom distribution and the permutation test for the *Karate* and *Co-authors* datasets. On the other hand, there is some discrepancy between these for *Gahuku-Gama* data. A possible explanation for this is due to the small number of nodes in the dataset: the Tracy-Widom distribution describes the asymptotic behavior of the eigenvalue when n goes to ∞.

For *fMRI* dataset, our test rejected the null hypothesis H_0, yielding the maximum and minimum eigenvalues as 31.0 and -7.2 for $T(\boldsymbol{W})$, and 31.8 and -10.9 for $T_e(S(\boldsymbol{W}))$, which provides strong evidence that community structure exists in this graph. Furthermore, we carried out our test for subsets of voxels in brain regions that are anatomically predefined, where the number of voxels ranges from 13 to 498. Our test results suggest that community structure may exist in each region (except for brain region 16) (Fig 7). This result supports the conjecture on

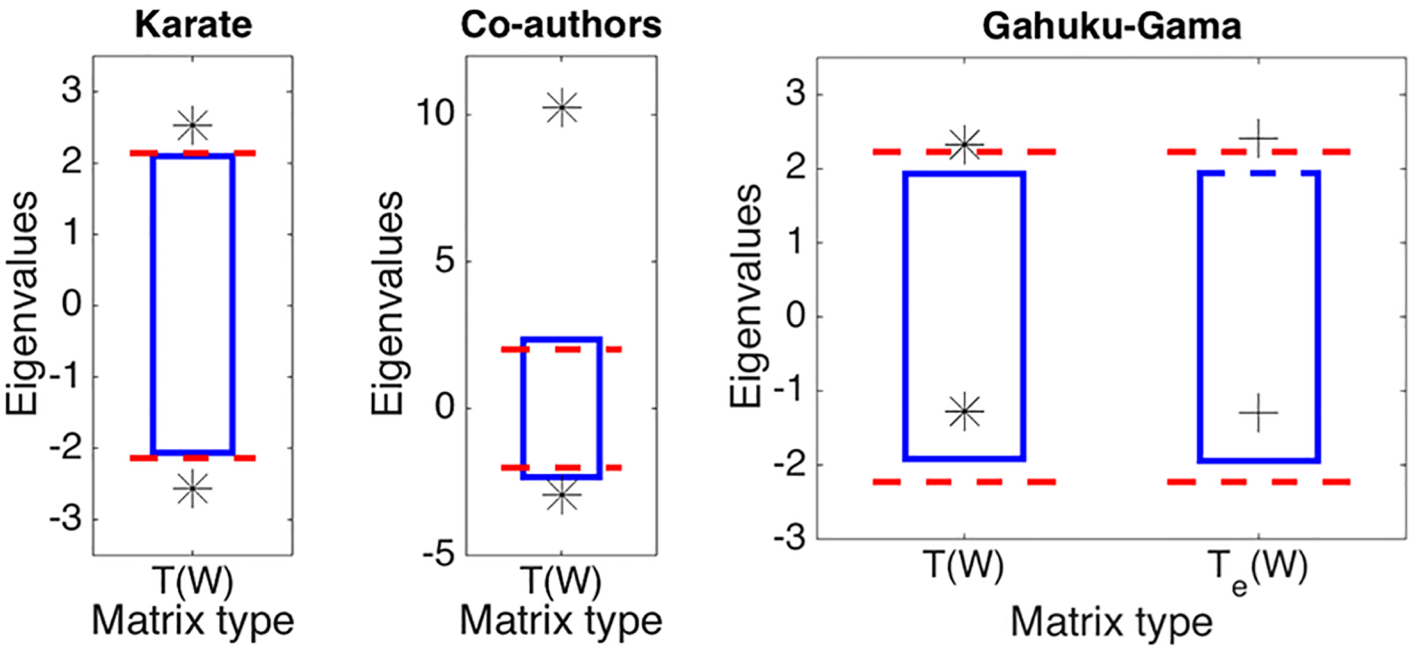

Fig 6. Results of application of our method to real datasets. *Karate*, *Co-authors*, and *Gahuku-Gama* from left to right panels. A star denotes the maximum or minimum eigenvalues of the normalized matrix $T(\boldsymbol{W})$, while a cross denotes those of the exponentially normalized matrix $T_e(S(\boldsymbol{W}))$. The top or bottom edges of boxes denote critical values of these eigenvalues at significance level $\alpha/2$ with $\alpha = 0.05$ for *Karate* and *Co-authors* datasets, and $\alpha/4$ for *Gahuku-Gama* dataset. These critical values resulted from a permutation test with 1000 randomized realizations. In contrast, red dashed lines denote critical values derived from the Tracy-Widom distribution $F_1(x)$.

https://doi.org/10.1371/journal.pone.0194079.g006

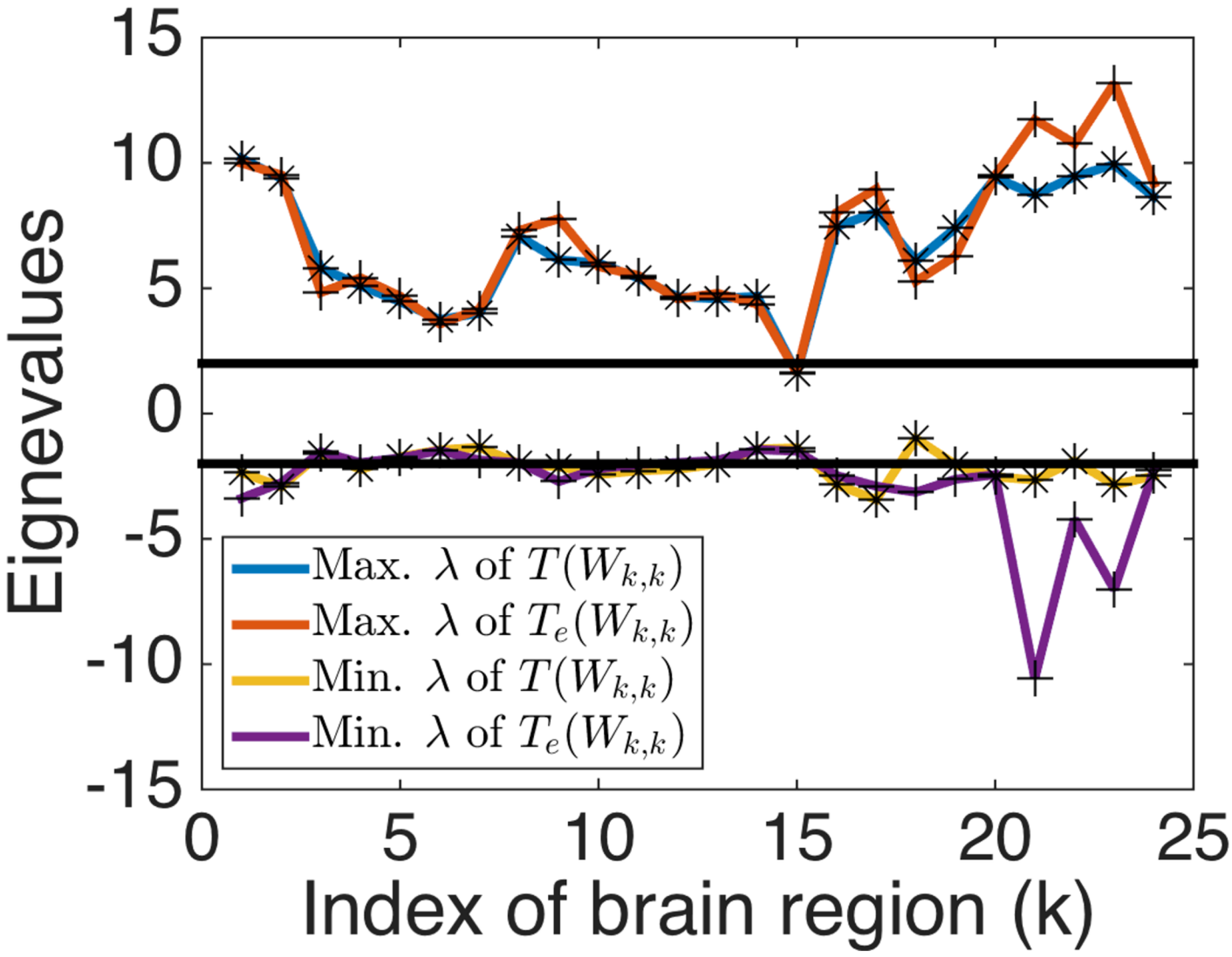

Fig 7. Results of application of our method to the fMRI dataset. Stars denote the maximum or minimum eigenvalues λ for normalized weight matrices by mapping T in various brains regions with an edge-weight matrix $\boldsymbol{W}_{k,k}$, indexed by the brain region k in the x-axis. Crosses denote counterparts for exponentially normalized weight matrices by the mapping T_e. Horizontal lines denote lines $y = -2$ and $y = 2$, which correspond to values at which the minimum and maximum eigenvalues asymptotically converge.

https://doi.org/10.1371/journal.pone.0194079.g007

heterogeneity of brain activities in anatomically defined brain regions, discussed in the neuroscience literature [39].

Application to real data 2

We consider further application of our method to real data, focussing in unweighted graphs. The object is to compare its performance and computation time with other relevant methods, which specialize in unweighted graphs. Though our method has been developed for weighted graphs, it works for unweighted graphs as well, because a unweighted graph is a special case of weighted graph. In addition to performance, we also compare computation time in this context.

Relevant methods

One of the most popular approaches to detection of communities in an unweighted graph is based on 'modularity' [40, 41], which is defined as

$$Q = \sum_k (e_{k,k} - a_k^2), \tag{8}$$

where $e_{k,k'}$ is one-half of the fractions of edges between cluster k and k', and $a_k = \Sigma_{k'} e_{k,k'}$. The modularity Q denotes deviation of the number of edges from possible random configurations, hence, serving as an objective function for finding a community structure. The algorithm of optimizing Q is to start with node-community (a community consisting of a single node) and to aggregate communities to increase Q in a similar fashion to a hierarchical clustering algorithm [42]. We use an algorithm of this kind proposed by [41], which is referred to as 'Newman'. On the other hand, Louvain methods [43–45] are a variant of the modularity-based methods, which optimizes Q (or, different type of Q) by means of iterating the following two steps. The first step is to optimize Q by aggregating communities in the aforementioned manner. The second step is to re-parameterize each community as a single node. These steps are alternatively carried out until no further increment in Q is possible. Here, we use one of the most popular methods by [43], referred to as 'Louvain'. For a threshold of detecting community structure, we use an analytical approximation of modularity for an Erdős-Rényi random graph with n nodes and probability p of connecting two nodes, which is given as $(1 - 2/\sqrt{n})(2/(pn))^{2/3}$ by [46]. Note that in a sparse graph, even without any community structure, modularity Q can take a large value. The analytical approximation captures this point, providing a useful criterion of community detection, though the confidence interval is not readily available. For another relevant method to modularity, we consider an approach based on eigenvalues of a modularity matrix by [47]. In a similar line to the graph Laplacian [48], this method partitions nodes based on the eigenvector of the modularity matrix corresponding to the largest positive eigenvalue. By repeatedly evaluating such an eigenvector, we continue to partition nodes until no further positive eigenvalue is obtained. Here, we use this method (referred to as 'Split') for the first partition of nodes, evaluating the largest eigenvalue of the modularity matrix. Furthermore, we consider a versatile approach: a Bayesian clustering method for communities in a graph by [49]. This method explicitly models community memberships as probabilistic parameters, which are optimized in a Bayesian manner (referred to as 'Bayesian'). Lastly, we include a bootstrap method by [16], which is combined with the community detecting method 'Newman', setting the proportion of disturbance to 5% (referred to as 'Bootstrap'). In this method, for simplicity, we evaluate concordances of community structure between the original graph and bootstrapped graphs by means of Adjusted Rand Index [50] in the same spirit as [51].

For meaningful comparison of computation time among different methods, we ran these methods in the same programming language, Matlab, using publicly available codes for Newman in [52], Louvain in [53] and Bayesian in [54]. For our method, Split method, and Bootstrap method, we ourselves programmed corresponding Matlab codes.

Data

We consider the following real datasets: Social networks in Indian Villages [55, 56] with 203 nodes and 523 edges (referred to as 'IndianVillage'); Protein-protein interactions in budding yeast [57] with 2361 nodes and 6600 edges (referred to as 'Yeast'); Word associations based on empirical studies [58] with 10617 nodes and 63000 edges (referred to as 'FreeAssoc'; we transformed the original data into an undirected graph by adding edges if there is a connection

between nodes in either direction). In addition, we also consider inverted graphs in which the status of an edge is inverted in these datasets (i.e., if there is an edge, it is removed; otherwise, it is added). We expected that this would clarify differences of performance among the methods in question. Finally, we generate weighted versions of these datasets as follows. If there is an edge, a weight is randomly generated from $N(0, 1)$, otherwise from $N(0, 0.01)$. We apply our method and Bayesian method to these datasets (the remainder of methods are not applicable to a weighted graph).

Results

For the original datasets, the performance of our method is comparable to other relevant methods, because the existence of community structures is well detected (Table 1). On the other hand, our method suggests the existence of community structures for the inverted graphs as well. Bayesian method and Split method yielded similar results. However, the modularity-based methods Newman and Louvain suggest no community structures while the performance of Bootstrap method is in-between. These differences arise from different (implicit) assumptions in the methods. Our method and Bayesian method focus on differences of patterns in occurrence of edges in communities, while the modularity-based approaches focus only on high density of edges in communists. For practical usage, this implies that one should carefully choose a method, depending on what type of community structure one aims to detect. In case of weighted datasets, both our method and Bayesian method yield similar

Table 1. Results of application to unweighted graphs of real datasets: IndianVillage, Yeast and FreeAssoc. In the column of 'Type' in the table, 'Ori' denotes the original graph while 'Inv' the inverted graph. Further, 'Ori.w' denotes the weighted original graph while 'Inv.w' denotes the weighted inverted graph. For each cell in the table, computation time and a corresponding statistic to detect community structure are displayed. A star marker in digits denotes that the result supports the existence of community structure. These statistics and critical values are given as follows. For our method, the maximum magnitude of eigenvalues λ is used. The critical value is given by the Tracy-Widom distribution in Eq (7). For Newman and Louvain methods, modularity Q is used with the critical value 0.45, 0.48, and 0.29 for IndianVillage, Yeast and FreeAssoc, respectively, based on the analytical approximation of modularity for an Erdős-Rényi random graph. For Split method, a positive largest eigenvalue of modularity matrix λ' suggests community structure while a negative largest eigenvalue λ' non-community structure. For Bayesian method, the difference of marginal log-likelihood for $K = 1$ and $K = 2$ ('Dif'; subtraction of $K = 1$ case from $K = 2$ case) is used. A positive difference suggests community structure while a negative difference non-community structure. For Bootstrap method, we evaluate stability of community structure by means of Adjusted Rand Index (ARI) between the targeted graph and bootstrapped graphs (the number of replicates is set to 100). We compare the median of ARI (mARI) with the distribution of ARI when the targeted graph is randomized. If mARI falls within the 95% confidence interval, it suggests that there is no community structure. Seemingly, this method is not computationally efficient. We were not able to obtain the results for FreeAssoc within 72 hours.

Methods	Type	Datasets		
		IndianVillage	Yeast	FreeAssoc
Our method	Ori	0sec, $\lambda = 3.1^*$	0sec, $\lambda = 7.8^*$	3mn, $\lambda = 9.6^*$
	Inv	0sec, $\lambda = 3.1^*$	4sec, $\lambda = 7.8^*$	4mn, $\lambda = 9.6^*$
Newman	Ori	0sec, $Q = 0.53^*$	2mn, $Q = 0.56^*$	3hr, $Q = 0.39^*$
	Inv	0sec, $Q = 0.00$	2mn, $Q = 0.00$	3hr, $Q = 0.00$
Louvain	Ori	0sec, $Q = 0.53^*$	15sec, $Q = 0.56^*$	2mn, $Q = 0.42^*$
	Inv	1sec, $Q = 0.00$	2mn, $Q = 0.00$	1hr, $Q = 0.00$
Split	Ori	0sec, $\lambda' = 6.3^*$	3sec, $\lambda' = 17.0^*$	5mn, $\lambda' = 27.4^*$
	Inv	0sec, $\lambda' = 4.5^*$	3sec, $\lambda' = 10.1^*$	6mn, $\lambda' = 17.0^*$
Bayesian	Ori	2mn, Dif = $2.3e4^*$	23mn, Dif = $9.4e6^*$	4hr, Dif = $1.8e8^*$
	Inv	2mn, Dif = $4.9e3^*$	15mn, Dif = $8.2e6^*$	4hr, Dif = $1.6e8^*$
Bootstrap	Ori	27sec, mARI = 0.36^*	4hr, mARI = 0.50^*	> 72hr
	Inv	29sec, mARI = 0.15	6hr, mARI = 0.28^*	> 72hr
Our method	Ori.w	0sec, $\lambda = 2.7^*$	7sec, $\lambda = 2.7^*$	8mn, $\lambda = 2.5^*$
	Inv.w	0sec, $\lambda = 2.06^*$	7sec, $\lambda = 2.00$	8mn, $\lambda = 1.99$
Bayesian	Ori.w	4mn, Dif = $2.1e3^*$	43mn, Dif = $6.0e4^*$	19hr, Dif = $4.0e5^*$
	Inv.w	4mn, Dif = $-3.3e1$	48mn, Dif = $-4.9e1$	66mn, Dif = $-5.8e1$

https://doi.org/10.1371/journal.pone.0194079.t001

results on detection of community structures. Lastly, as regards computation time, our method outperforms the remainder of the methods. This is possibly due to that these methods go through a procedure to search for community memberships including the number of communities, while our method does not include such a procedure.

Discussion

We have proposed a novel method for a statistical test for the existence of community structure in an undirected graph that is characterized by the first and the second moments of a generative model for edge weights. This method can be considered a nontrivial extension of the recently proposed method [22] from a binary-valued to a real-valued graph. Unlike the existing methods for real-valued graphs, our method does not need a cluster solution. Hence, we can apply this method even to the nontrivial case of clustering in which edge weights take both positive and negative real values. Also, our approach avoids a nontrivial problem of determining the number of clusters. Further, our method is quite efficient in terms of computation time: We only need to evaluate the eigenvalues of an edge-weight matrix once if we use the Tracy-Widom distribution, which is due to the asymptotic results derived from Random Matrix Theory.

As the next step of analysis, one may wonder how to find community memberships when our test rejects the null hypothesis of $K = 1$. The present paper did not address this issue, but it would be quite useful to examine eigenvectors of the edge-weight matrix as in the case of spectral clustering. It is conjectured that some of the eigenvectors of $T(\boldsymbol{W})$ and $T_e(S(\boldsymbol{W}))$ may contain information on community memberships. In the future, it will be important to determine and to synthesize relevant eigenvectors for inferring underlying community structure.

Supporting information

S1 Appendix. A. Proof of Example 1.
(PDF)

S2 Appendix. B. Proof of Theorem 1.
(PDF)

Acknowledgments

I would like to thank Dr. Steven D. Aird at Okinawa Institute of Science and Technology Graduate University for his proof-reading of this article.

Author Contributions

Conceptualization: Tomoki Tokuda.

References

1. Fortunato S. Community detection in graphs. Physics Reports. 2010; 486(3):75–174. https://doi.org/10.1016/j.physrep.2009.11.002
2. Newman ME, Girvan M. Finding and evaluating community structure in networks. Physical Review E. 2004; 69(2):026113. https://doi.org/10.1103/PhysRevE.69.026113
3. Reichardt J, Bornholdt S. Statistical mechanics of community detection. Physical Review E. 2006; 74 (1):016110. https://doi.org/10.1103/PhysRevE.74.016110
4. Bolla M. Spectral Clustering and Biclustering: Learning Large Graphs and Contingency Tables. John Wiley & Sons; 2013.
5. Ng AY, Jordan MI, Weiss Y, et al. On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems. 2002; 2:849–856.

6. Kunegis J, Schmidt S, Lommatzsch A, Lerner J, De Luca EW, Albayrak S. Spectral analysis of signed graphs for clustering, prediction and visualization. In: SDM. vol. 10. SIAM; 2010. p. 559–570.
7. Yang B, Cheung WK, Liu J. Community mining from signed social networks. Knowledge and Data Engineering, IEEE Transactions. 2007; 19(10):1333–1348. https://doi.org/10.1109/TKDE.2007.1061
8. Decelle A, Krzakala F, Moore C, Zdeborová L. Inference and phase transitions in the detection of modules in sparse networks. Physical Review Letters. 2011; 107(6):065701. https://doi.org/10.1103/PhysRevLett.107.065701 PMID: 21902340
9. Decelle A, Krzakala F, Moore C, Zdeborová L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. Physical Review E. 2011; 84(6):066106. https://doi.org/10.1103/PhysRevE.84.066106
10. Mossel E, Neeman J, Sly A. Stochastic block models and reconstruction. arXiv preprint arXiv:12021499. 2012;.
11. Krzakala F, Moore C, Mossel E, Neeman J, Sly A, Zdeborová L, et al. Spectral redemption in clustering sparse networks. Proceedings of the National Academy of Sciences. 2013; 110(52):20935–20940. https://doi.org/10.1073/pnas.1312486110
12. Mossel E, Neeman J, Sly A. Reconstruction and estimation in the planted partition model. Probability Theory and Related Fields. 2015; 162(3-4):431–461. https://doi.org/10.1007/s00440-014-0576-6
13. Dyer ME, Frieze AM. The solution of some random NP-hard problems in polynomial expected time. Journal of Algorithms. 1989; 10(4):451–489. https://doi.org/10.1016/0196-6774(89)90001-1
14. Gfeller D, Chappelier JC, De Los Rios P. Finding instabilities in the community structure of complex networks. Physical Review E. 2005; 72(5):056135. https://doi.org/10.1103/PhysRevE.72.056135
15. Karrer B, Levina E, Newman ME. Robustness of community structure in networks. Physical Review E. 2008; 77(4):046119. https://doi.org/10.1103/PhysRevE.77.046119
16. Rosvall M, Bergstrom CT. Mapping change in large networks. PloS one. 2010; 5(1):e8694. https://doi.org/10.1371/journal.pone.0008694 PMID: 20111700
17. Bianconi G, Pin P, Marsili M. Assessing the relevance of node features for network structure. Proceedings of the National Academy of Sciences. 2009; 106(28):11433–11438. https://doi.org/10.1073/pnas.0811511106
18. Lancichinetti A, Radicchi F, Ramasco JJ. Statistical significance of communities in networks. Physical Review E. 2010; 81(4):046110. https://doi.org/10.1103/PhysRevE.81.046110
19. Golub B, Jackson MO. Does homophily predict consensus times? Testing a model of network structure via a dynamic process. Review of Network Economics. 2012; 11(3). https://doi.org/10.1515/1446-9022.1367
20. Golub B, Jackson MO. How homophily affects the speed of learning and best-response dynamics. The Quarterly Journal of Economics. 2012; 127(3):1287–1338. https://doi.org/10.1093/qje/qjs021
21. Golub B, Jackson MO. Network structure and the speed of learning measuring homophily based on its consequences. Annals of Economics and Statistics/ANNALES D'ÉCONOMIE ET DE STATISTIQUE. 2012; p. 33–48.
22. Bickel PJ, Sarkar P. Hypothesis testing for automated community detection in networks. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2016; 78(1):253–273. https://doi.org/10.1111/rssb.12117
23. Mehta ML. Random matrices. vol. 142. Academic Press; 2004.
24. Tao T. Topics in random matrix theory. vol. 132. American Mathematical Soc.; 2012.
25. Rodgers GJ, Nagao T. Complex networks. In: Akemann G, Baik J, Di Francesco P, editors. The Oxford Handbook of Random Matrix Theory. Oxford: Oxford University Press; 2011. p. 898–911.
26. Brody A. The second eigenvalue of the Leontief matrix. Economic Systems Research. 1997; 9(3):253–258. https://doi.org/10.1080/09535319700000018
27. Molnár G, Simonovits A. The subdominant eigenvalue of a large stochastic matrix. Economic Systems Research. 1998; 10(1):79–82. https://doi.org/10.1080/09535319800000007
28. Gurgul H, Wójtowicz T. On the economic interpretation of the Bródy conjecture. Economic Systems Research. 2015; 27(1):122–131. https://doi.org/10.1080/09535314.2014.979138
29. McSherry F. Spectral partitioning of random graphs. In: Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium. IEEE; 2001. p. 529–537.
30. Estermann T. Introduction to modern prime number theory. 41. Cambridge University Press; 2011.
31. Tracy CA, Widom H. On orthogonal and symplectic matrix ensembles. Communications in Mathematical Physics. 1996; 177(3):727–754. https://doi.org/10.1007/BF02099545

32. Tracy CA, Widom H. The distributions of random matrix theory and their applications. In: New Trends in Mathematical Physics. Springer; 2009. p. 753–765.
33. Arous GB, Guionnet A. Wigner matrices. In: Akemann G, Baik J, Di Francesco P, editors. The Oxford Handbook of Random Matrix Theory. Oxford University Press; 2011. p. 433–451.
34. Hsieh CJ, Chiang KY, Dhillon IS. Low rank modeling of signed networks. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge Discovery and Data Mining. ACM; 2012. p. 507–515.
35. Zachary WW. An information flow model for conflict and fission in small groups. Journal of Anthropological Research. 1977; p. 452–473. https://doi.org/10.1086/jar.33.4.3629752
36. Newman ME. Finding community structure in networks using the eigenvectors of matrices. Physical Review E. 2006; 74(3):036104. https://doi.org/10.1103/PhysRevE.74.036104
37. Read KE. Cultures of the central highlands, New Guinea. Southwestern Journal of Anthropology. 1954; p. 1–43. https://doi.org/10.1086/soutjanth.10.1.3629074
38. Mitchell T. type; 2005. Available from: http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/.
39. Birn RM, Saad ZS, Bandettini PA. Spatial heterogeneity of the nonlinear dynamics in the FMRI BOLD response. Neuroimage. 2001; 14(4):817–826. https://doi.org/10.1006/nimg.2001.0873 PMID: 11554800
40. Girvan M, Newman ME. Community structure in social and biological networks. Proceedings of the National Academy of Sciences. 2002; 99(12):7821–7826. https://doi.org/10.1073/pnas.122653799
41. Newman ME. Fast algorithm for detecting community structure in networks. Physical Review E. 2004; 69(6):066133. https://doi.org/10.1103/PhysRevE.69.066133
42. Johnson SC. Hierarchical clustering schemes. Psychometrika. 1967; 32(3):241–254. https://doi.org/10.1007/BF02289588 PMID: 5234703
43. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008; 2008(10):P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008
44. Traag VA, Krings G, Van Dooren P. Significant scales in community structure. Scientific Reports. 2013; 3. https://doi.org/10.1038/srep02930 PMID: 24121597
45. Traag VA, Aldecoa R, Delvenne JC. Detecting communities using asymptotical surprise. Physical Review E. 2015; 92(2):022816. https://doi.org/10.1103/PhysRevE.92.022816
46. Guimera R, Sales-Pardo M, Amaral LAN. Modularity from fluctuations in random graphs and complex networks. Physical Review E. 2004; 70(2):025101. https://doi.org/10.1103/PhysRevE.70.025101
47. Newman ME. Modularity and community structure in networks. Proceedings of the National Academy of Sciences. 2006; 103(23):8577–8582. https://doi.org/10.1073/pnas.0601602103
48. Chung FR. Spectral graph theory. 92. American Mathematical Soc.; 1997.
49. Aicher C, Jacobs AZ, Clauset A. Learning latent block structure in weighted networks. Journal of Complex Networks. 2014; 3(2):221–248. https://doi.org/10.1093/comnet/cnu026
50. Hubert L, Arabie P. Comparing partitions. Journal of Classification. 1985; 2(1):193–218. https://doi.org/10.1007/BF01908075
51. Dolnicar S, Leisch F. Evaluation of structure and reproducibility of cluster solutions using the bootstrap. Marketing Letters. 2010; 21(1):83–101. https://doi.org/10.1007/s11002-009-9083-4
52. Kehagias A. Community Detection Toolbox;. Available from: https://jp.mathworks.com/matlabcentral/fileexchange/45867-community-detection-toolbox?focused=3813773&tab=function.
53. Scherrer A. Matlab / C++ implementation of community detection algorithm;. Available from: https://github.com/jblocher/matlab-network-utilities/tree/master/Louvain.
54. Aicher C. The Weighted Stochastic Block Model;. Available from: http://tuvalu.santafe.edu/~aaronc/wsbm/.
55. Banerjee A, Chandrasekhar AG, Duflo E, Jackson MO. The diffusion of microfinance. Science. 2013; 341(6144):1236498. https://doi.org/10.1126/science.1236498 PMID: 23888042
56. Jackson MO, Rodriguez-Barraquer T, Tan X. Social capital and social quilts: Network patterns of favor exchange. The American Economic Review. 2012; 102(5):1857–1897. https://doi.org/10.1257/aer.102.5.1857
57. Bu D, Zhao Y, Cai L, Xue H, Zhu X, Lu H, et al. Topological structure analysis of the protein–protein interaction network in budding yeast. Nucleic Acids Research. 2003; 31(9):2443–2450. https://doi.org/10.1093/nar/gkg340 PMID: 12711690
58. Nelson DL, McEvoy CL, Schreiber TA. The University of South Florida free association, rhyme, and word fragment norms. Behavior Research Methods, Instruments, & Computers. 2004; 36(3):402–407. https://doi.org/10.3758/BF03195588