Network Science (2024), 1–17
doi:10.1017/nws.2024.1

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

A generalized hypothesis test for community structure in networks

Eric Yanchenko and Srijan Sengupta

Department of Statistics, North Carolina State University, Raleigh, NC, USA
Corresponding author: Eric Yanchenko; Email: ekyanche@ncsu.edu

Abstract

Researchers theorize that many real-world networks exhibit community structure where within-community edges are more likely than between-community edges. While numerous methods exist to cluster nodes into different communities, less work has addressed this question: given some network, does it exhibit *statistically meaningful* community structure? We answer this question in a principled manner by framing it as a statistical hypothesis test in terms of a general and model-agnostic community structure parameter. Leveraging this parameter, we propose a simple and interpretable test statistic used to formulate two separate hypothesis testing frameworks. The first is an asymptotic test against a baseline value of the parameter while the second tests against a baseline model using bootstrap-based thresholds. We prove theoretical properties of these tests and demonstrate how the proposed method yields rich insights into real-world datasets.

Keywords: Assortative mixing; bootstrap; community detection; random graphs

1. Introduction

Networks are everywhere in the modern world. From social media (Kane et al., 2014; Guo et al., 2020) to infrastructure (Mason and Verwoerd, 2007) to epidemiology (Leitch et al., 2019), many fields gather and analyze network data. The growth of the discipline of network science in the past two decades has brought along with it an interesting phenomenon: despite being observed in vastly different fields, many networks share similar structural properties (Fotouhi et al., 2019). One of the most common structural properties is *community structure*, or the occurrence of tightly knit groups of nodes. Because communities can yield insights into node characteristics shared within these groups, various *community detection* methods have been proposed to assign nodes to communities, including spectral methods (Ng et al., 2002; Rohe et al., 2011; Jin, 2015; Sengupta and Chen, 2015) and greedy algorithms (Clauset et al., 2004; Blondel et al., 2008). While the community detection problem has received substantial attention, less work has considered whether a given network demonstrates a statistically meaningful community structure. In other words, is the near-ubiquity of community structure in observed networks informative about the underlying data-generating mechanisms, or is it simply spurious false positive results due to random noise or some unrelated feature such as degree heterogeneity?

As a motivating example, consider the classical influence maximization (IM) task where the goal is to select a small number of seed nodes such that the spread of influence is maximized on the entire network (Kempe et al., 2003; Yanchenko et al., 2023). In networks lacking community structure, Osawa and Murata (2015) showed that seeding nodes based on simple, centrality-based heuristics yields seed sets with substantial spread. On the other hand, when networks exhibit a community structure, selecting nodes from the same community results in ineffective seed sets

due to nodes sharing many similar neighbors. In this case, more sophisticated seeding algorithms are needed. Thus, understanding the significance of the community structure in the network is paramount for adequate seed selection in the IM problem.

This work aims to develop a formal statistical hypothesis test for community structure. The framework of statistical hypothesis testing consists of four fundamental components: (1) the model parameter of interest, (2) a test statistic that is typically based on an estimator of the model parameter, (3) a null model that reflects the absence of the property of interest, and (4) a rejection region for the test statistic. While previous literature exists on this problem, there has been a minimal emphasis placed on ingredients (1) and (3). In particular, none of the existing work identifies an underlying model parameter and the null model has not been studied thoroughly. Two popular choices for nulls are the Erdős–Rényi (ER) (Bickel and Sarkar, 2016; Yuan et al., 2022) and configuration model (Lancichinetti et al., 2010; Palowitch et al., 2018; Li and Qi, 2020). While these testing methods carry rigorous statistical guarantees, choosing these null models means that they effectively test against the null hypothesis that the network is generated from a specific null model, rather than testing against the null hypothesis that there is no community structure. This leads to problems in empirical studies because the ER model, for example, is so unrealistic that almost all real-world networks would diverge from it, leading to many false positives. On the other hand, it may be possible for a configuration model to have a small amount of community structure itself. Thus, a careful treatment of the null model is essential for a statistical test to be relevant for applied network scientists.

The main contributions of this paper follow the four components of the hypothesis testing framework. From first principles, we describe a model-agnostic parameter based on expected differences in edge densities which forms the basis of our statistical inference framework. Second, we propose an intuitive and interpretable test statistic which is directly connected to the model parameter. We leverage the model parameter and test statistic to formulate two types of hypothesis tests. The first is based on a user-specified threshold value of the parameter, which induces a model-agnostic test. For the second type, instead of specifying a baseline value of the parameter, the user specifies a baseline model or network property to test against. We derive theoretical results for the asymptotic cutoff in the first test and bootstrap cutoff in the second. Finally, we apply our method to well-studied real-world network datasets in the community structure literature. The results are insightful, as our method yields rich, new insights about the underlying network structure. Source code for this work is available on GitHub: https://github.com/eyanchenko/NetHypTest. The roadmap for the rest of this paper is as follows: in Section 2, we propose the model parameter and corresponding estimator, as well as present the first (asymptotic) hypothesis test. Section 3 discusses the baseline model test with a bootstrap threshold. We apply the method to synthetic data in Section 4 and real-world datasets in Section 5. We close by discussing the method in Section 6.

2. Model parameter and baseline value testing framework

2.1 Notation

For this work, we will only consider simple, unweighted and undirected networks with no self loops. Consider a network with n nodes and let A denote the $n \times n$ adjacency matrix where $A_{ij} = 1$ if node i and node j have an edge, and 0 otherwise. We write $A \sim P$ as shorthand for $A_{ij}|P_{ij} \sim \text{Bernoulli}(P_{ij})$ for $1 \leq i < j \leq n$ and define a community assignment to be a vector $\boldsymbol{c} \in \{1, \ldots, K\}^n$ such that $c_i = k$ means node i is assigned to community $k \in \{1, \ldots, K\}$. We also introduce the following notation: for scalar sequences a_n and b_n, $a_n = O(b_n)$ means that $\lim_{n\to\infty} a_n/b_n \leq M$ for some constant M (which could be 0) and $a_n = o(b_n)$ means that $\lim_{n\to\infty} a_n/b_n = 0$. For a sequence of random variables, $X_n = O_p(Y_n)$ means that $X_n/Y_n \to_P M$ (which could be 0) and $X_n = o_p(Y_n)$ means that $X_n/Y_n \to_P 0$.

https://doi.org/10.1017/nws.2024.1 Published online by Cambridge University Press

2.2 Expected Edge Density Difference (E2D2) parameter and estimator

The first step of the hypothesis test is to identify the model parameter. Since there is no universal metric to quantify community structure, we construct one from the first principles that applies to a large class of network models. For an *observed* network, a natural global measure of the strength of the community structure is the difference between the intra- and inter-community edge densities (see page 83–84 of (Fortunato, 2010)). The larger this difference, the more prominent the community structure is in the network. Now, this definition makes sense at the *sample* level for a realized network, but we seek the model parameter at the *population* level, or the parameter that generates this network. For this, we propose the *Expected Edge Density Difference* (E2D2) parameter. Consider a network model P and let $\boldsymbol{c}$ be a community assignment. Define

$$\bar{p}_{\text{in}}(\boldsymbol{c}) = \frac{1}{\sum_{k=1}^{K}\binom{n_k}{2}}\sum_{i<j} P_{ij}\,\mathbb{1}(c_i = c_j) \quad \text{and} \quad \bar{p}_{\text{out}}(\boldsymbol{c}) \quad = \frac{1}{\sum_{k>l} n_k n_l}\sum_{i<j} P_{ij}\,\mathbb{1}(c_i \neq c_j), \tag{1}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Here, $\bar{p}_{\text{in}}(\boldsymbol{c})$ and $\bar{p}_{\text{out}}(\boldsymbol{c})$ are the average *expected* intra- and inter-community edge densities, respectively, hence the name *Expected* Edge Density Difference. This definition is sensible only when $1 < K < n$ for if $K = 1$, then $\bar{p}_{\text{out}}(\boldsymbol{c})$ is ill defined and the same is true for $\bar{p}_{\text{in}}(\boldsymbol{c})$ if $K = n$. Intuitively, a data-generating mechanism with large $\bar{p}_{\text{in}}(\boldsymbol{c}) - \bar{p}_{\text{out}}(\boldsymbol{c})$ is likely to produce a network with a large difference in observed intra- and inter- community edge density and, therefore, prominent community structure. This difference, however, should be adjusted with respect to the overall sparsity of the network and the number of groups each node can be assigned. So, we propose the E2D2 parameter $\gamma(\boldsymbol{c}, P)$ as:

$$\gamma(\boldsymbol{c}, P) := \frac{1}{K}\frac{\bar{p}_{\text{in}}(\boldsymbol{c}) - \bar{p}_{\text{out}}(\boldsymbol{c})}{\bar{p}}, \tag{2}$$

where $\bar{p} = \sum_{i>j} P_{ij}/\binom{n}{2}$ is the overall probability of an edge between nodes in the network. We also define

$$\tilde{\gamma}(P) = \max_{\boldsymbol{c}}\{\gamma(\boldsymbol{c}, P)\} \tag{3}$$

as the maximum value of the E2D2 parameter where the maximization is taken over the candidate community assignments $\boldsymbol{c}$. In the Supplemental Materials, we sketch a proof showing that $\bar{p}_{\text{in}}(\boldsymbol{c}) - \bar{p}_{\text{out}}(\boldsymbol{c}) \leq \bar{p}K$. Thus, the K^{-1} term ensures that the E2D2 parameter is always less than one. Indeed, $\gamma(\boldsymbol{c}, P) = 1$ if and only if $\bar{p}_{\text{out}} = 0$ and there are K equally sized communities. By construction, this parameter has a natural connection to the intuitive notion of community structure since larger values correspond to more prominent levels of community structure in the data-generating process.

One of the key advantages of the E2D2 parameter is that it is general and model-agnostic. By *model-agnostic*, we mean that the definition applies to any network-generating model where edges are conditionally independent given the model parameters P_{ij}. This encompasses a wide-range of models including: ER (Erdős and Rényi, 1959), Chung–Lu (CL) (Chung and Lu, 2002), stochastic block models (SBMs) (Holland et al., 1983), DCBMs (Karrer and Newman, 2011), latent space models (Hoff et al., 2002), random dot product graphs (Athreya et al., 2017) and more. Some other models such as the Barabási-Albert (Barabási and Albert, 1999), configuration (Fosdick et al., 2018), and exponential random graph models (Robins et al., 2007), however, do not fit into this framework. While by no means applicable to *all* network models, the general definition of the E2D2 parameter, requiring only the conditional independence of edges, gives the parameter great flexibility.

https://doi.org/10.1017/nws.2024.1 Published online by Cambridge University Press

The second ingredient in our hypothesis testing recipe is an estimator of the E2D2 parameter. Using the same notation as above, define

$$\hat{p}_{\text{in}}(\boldsymbol{c}) = \frac{1}{\sum_{k=1}^{K} \binom{n_k}{2}} \sum_{i<j} A_{ij}\, \mathbb{1}(c_i = c_j) \quad \text{and} \quad \hat{p}_{\text{out}}(\boldsymbol{c}) \quad = \frac{1}{\sum_{k>l} n_k n_l} \sum_{i<j} A_{ij}\, \mathbb{1}(c_i \neq c_j), \tag{4}$$

Then we estimate $\gamma(\boldsymbol{c}, P)$ from (2) as:

$$T(\boldsymbol{c}, A) := \frac{1}{K} \frac{\hat{p}_{\text{in}}(\boldsymbol{c}) - \hat{p}_{\text{out}}(\boldsymbol{c})}{\hat{p}}, \tag{5}$$

where $\hat{p} = \sum_{i>j} A_{ij} / \binom{n}{2}$ and $\hat{p}_{\text{in}}(\boldsymbol{c})$, $\hat{p}_{\text{out}}(\boldsymbol{c})$ and $\hat{p}$ are the sample versions of $\bar{p}_{\text{in}}(\boldsymbol{c})$, $\bar{p}_{\text{out}}(\boldsymbol{c})$, and $\bar{p}$, respectively. In other words, $T(\boldsymbol{c}, A)$ is the *observed* edge density difference, the sample version of the E2D2 parameter. Below we find the maximum of this test statistic over all possible community labels $\boldsymbol{c}$ so we also introduce the notation:

$$\tilde{T}(A) = \max_{\boldsymbol{c}} \{T(\boldsymbol{c}, A)\}. \tag{6}$$

Since $\hat{p}$ depends on A but not on $\boldsymbol{c}$, $\tilde{T}(A)$ maximizes the intra-community edge probability over the candidate values of $\boldsymbol{c}$ with a penalty for larger inter-community edge probability, akin to the objective function in Mancoridis et al. (1998).

This metric has a natural connection to the well-known Newman–Girvan *modularity* quantity (Newman, 2006). The modularity, $Q(\boldsymbol{c}, A)$, of a network partition $\boldsymbol{c}$ is defined as:

$$Q(\boldsymbol{c}, A) = \frac{1}{m} \sum_{i<j} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \mathbb{1}(c_i = c_j) \tag{7}$$

where d_i is the degree of node i and m is the total number of edges in the network. Rearranging (7), we see that

$$Q(\boldsymbol{c}, A) = \frac{1}{m} \sum_{i<j} A_{ij} \mathbb{1}(c_i = c_j) - \frac{1}{2m^2} \sum_{i<j} d_i d_j \mathbb{1}(c_i = c_j). \tag{8}$$

The connection is now immediate: the first term is the (scaled) number of intra-community edges and has a one-to-one relationship with $\hat{p}_{\text{in}}(\boldsymbol{c})$. The second term can be thought of as the penalty term for the expected number of edges for a random network with given degree sequence. In light of these similarities, the proposed estimator has a key advantage compared to modularity. The penalty term for modularity assumes the configuration model as the null model, that is, comparing the strength of the community structure against a random network with identical degree sequence. The penalty term for the proposed method $\hat{p}_{\text{out}}(\boldsymbol{c})$, however, is not model-dependent. Hence, any model could be chosen as the null model, giving the proposed estimator far greater flexibility than modularity. The numerator of $T(\boldsymbol{c}, A)$ can also be written in the general modularity formulation of Bickel and Chen (2009).

2.3 Algorithm for computing the E2D2 estimator

The E2D2 estimator is also of independent interest as an objective function for community detection. Finding the maximum of $T(\boldsymbol{c}, A)$ is a combinatorial optimization problem with $O(K^n)$ solutions. Thus, an exhaustive search is clearly infeasible for even moderate n so we propose a greedy, label-switching algorithm to approximate $\tilde{T}(A)$. We briefly explain the ideas here and present the full algorithm in Algorithm 1. First, each node is initialized with a community label $c_i \in \{1, \ldots, K\}$. Then, for each node i, its community assignment is switched with all neighboring communities. The new label of node i is whichever switch yielded the largest value of the E2D2

https://doi.org/10.1017/nws.2024.1 Published online by Cambridge University Press

Algorithm 1. Greedy

Result: Community labels $\boldsymbol{c}$
Input: $n \times n$ adjacency matrix A, number of communities K
Initialize labels $\boldsymbol{c} \in \{1, \ldots, K\}^n$
$run = 1$
while $run > 0$ **do**
  $run = 0$
  Randomly order nodes
  for i *in* $1, \ldots, n$ **do**
    Find neighboring communities, K_i, of node i: $K_i = \{k_1, \ldots, k_{|K_i|}\}$
    Swap label of node i with all $k_j \in K_i$: $\boldsymbol{c}_j^* = \boldsymbol{c}$, $(\boldsymbol{c}_j^*)_i = k_j$
    $\boldsymbol{c}^* = \arg\max_j \{T(\boldsymbol{c}_j^*, A)\}$
    if $T(\boldsymbol{c}^*, A) > T(\boldsymbol{c}, A)$ **then**
      $\boldsymbol{c} \longleftarrow \boldsymbol{c}^*$
      $run = 1$
    end
  end
end

estimator (or it is kept in the original community if none of the swaps increased $T(\boldsymbol{c}, A)$). This process repeats for all n nodes. The algorithm stops when all nodes have been cycled through and no labels have changed. The current labels, $\boldsymbol{c}$, are then returned. The assumption that K is known is rather strong and unrealistic for most real-world networks. We view it as reasonable here, however, since the goal of this work is not primarily to propose a new community detection algorithm. In practice, any off-the-shelf method can be used to estimate K before using the proposed algorithm.

2.4 Baseline value test

We now leverage the E2D2 parameter and estimator to formulate our first hypothesis test. Because this parameter is interpretable and meaningful as a descriptor of the network-generating process, we consider the scenario where the researcher has a problem-specific benchmark value of the E2D2 parameter that she would like to test against. In other words, the baseline value has domain-relevant meaning as "no community structure." Then we must determine whether *any* assignment exceeds this threshold so the formal test is

$$H_0 : \tilde{\gamma}(P) \leq \gamma_0 \; \textit{vs.} \; H_1 : \tilde{\gamma}(P) > \gamma_0, \tag{9}$$

for some $\gamma_0 \in [0, 1)$. Naturally, we reject H_0 if

$$\tilde{T}(A) = \max_{\boldsymbol{c}} \{T(\boldsymbol{c}, A)\} > C \tag{10}$$

for some cutoff C that depends on the network size n and null value γ_0.

Now, to obtain a test with level α, we should set C as the $(1 - \alpha)$ quantile of the null distribution of $\tilde{T}(A)$. This depends on the data-generating matrix P under the null hypothesis, and so implicitly also depends on $\tilde{\gamma}(P) = \gamma_0$. But this is a difficult task since the test statistic is the maximum taken over $O(K^n)$ possible community assignments and these random variables are highly correlated.

https://doi.org/10.1017/nws.2024.1 Published online by Cambridge University Press

We propose to sidestep this difficult theoretical problem with an asymptotic cutoff. We first make the following three assumptions.

A1. For any candidate community assignment, at least two community sizes must grow linearly with n.

A2. The number of communities K_n is known, but is allowed to diverge.

A3. $\log^{1/2} K_n/(n^{1/2}\bar{p}K_n) \to 0$ *as* $n \to \infty$.

A1 lays down a basic requirement for any legitimate candidate community assignment, since otherwise, one community will dominate the entire network, while A2 shows that our theory holds even if the number of communities K_n goes to infinity. A3 is a sparsity requirement for $\bar{p}$ that depends on the asymptotics of K_n. If $K_n \equiv K$ is fixed (e.g., (Bickel and Chen, 2009; Sengupta and Chen, 2018)), then we require $n^{1/2}\bar{p} \to \infty$. A typical sparsity assumption is that $\bar{p} = O(n^{-1})$ such that the expected number of edges in the network grows linearly with n. Our result requires a stronger condition which means we are in the *semi-dense* regime. While this is not ideal, proofs in the *dense* regime (fixed p) are common in the literature (e.g., (Bickel and Sarkar, 2016)). Our work, in fact, holds under less stringent conditions, that is, $n^{1/2}\bar{p} \to \infty$ but allows $\bar{p} \to 0$. Indeed, this assumption implies that the test has larger power when the network is denser and/or has more communities since this leads to a smaller cutoff, as we see in the formal result that now follows.

Theorem 2.1. *Let $A \sim P$ and consider testing $H_0 : \tilde{\gamma}(P) \leq \gamma_0$ as in (9). Let A1 and A2 be true and consider the cutoff:*

$$C = \left(\gamma_0 + \frac{k_n}{K_n\hat{p}}\right)(1+\epsilon) \tag{11}$$

where $k_n = \{(\log K_n)/n\}^{1/2}$ and arbitrarily small $\epsilon > 0$ chosen by the user. Then when the null hypothesis is true ($\tilde{\gamma}(P) \leq \gamma_0$), the type I error goes to 0, that is, for any $\eta > 0$,

$$\lim_{n\to\infty} \mathsf{P}\{\tilde{T}(A) > C \mid H_0\} \leq \eta.$$

If the alternative hypothesis is true ($\tilde{\gamma}(P) > \gamma_0$), then the power goes to 1, that is,

$$\lim_{n\to\infty} \mathsf{P}\{\tilde{T}(A) > C \mid H_1\} > 1 - \eta.$$

A proof of the theorem, as well as proofs of all subsequent theoretical results, are left to the Supplemental Materials. The proof approximates the cutoff under the null hypothesis using a union bound and then leverages Hoeffding's inequality to show that the probability of failing to reject under the alternative hypothesis goes to 0. The cutoff depends on K and, for theoretical purposes, we assume that K is known. In practice, we run a community detection algorithm on the network (e.g., Fast Greedy algorithm of (Clauset et al., 2004)) and then use the number of communities returned by this algorithm, $\hat{K}$, to find $\tilde{T}(A)$ and construct the threshold. Additionally, A3 ensures that the $k_n/(K\hat{p})$ term converges to 0 such that $C \to \gamma_0$ as $n \to \infty$. In practice, we confirm the necessity of this assumption as this test has larger power for denser networks (large $\hat{p}$) and more communities.

This test is fundamentally different from existing ones in the literature because the practitioner chooses the value of γ_0 for her particular problem, inducing a model-agnostic test. In other words, the null hypothesis is not a baseline model, but instead a baseline quantity of community structure in the network-generating process. Since the E2D2 parameter has a natural connection to community structure, this test is also more easily interpretable with respect to this feature. Indeed, rejecting the null hypothesis means that the data-generating matrix for the observed network has greater community structure (as measured by the E2D2 parameter) than some baseline value (γ_0).

https://doi.org/10.1017/nws.2024.1 Published online by Cambridge University Press

A natural question that arises is how to choose γ_0. We stress that this choice depends on the domain and question of interest. There are some special cases, however, that yield insights into selecting a meaningful value. For example, assume that the practitioner believes that her network has two roughly equal-sized communities. Then setting $\gamma_0 = (\mu - 1)/(\mu + 1)$ is equivalent to testing whether the average intra-community edge density is more than $\mu > 1$ times larger than the average inter-community edge density, that is, $\bar{p}_{\text{in}} \geq \mu \bar{p}_{\text{out}}$.

The special case of $\gamma_0 = 0$ is also worthy of further discussion. It is trivial to show that if P is from an ER model (Erdős and Rényi, 1959) where $P_{ij} = p$ for all i, j, then $\tilde{\gamma}(P) = 0$. We show in the Supplementary Materials, however, that the converse of this statement is also true, that is, $\tilde{\gamma}(P) = 0$ *only if* P is from an ER model. This means that setting $\gamma_0 = 0$ is equivalent to testing against the null hypothesis that the network is generated from an ER model. In other words, any other network model will reject this test when $\gamma_0 = 0$. But there are many models (e.g., CL (Chung and Lu, 2002), small world (Watts and Strogatz, 1998)) which may not be ER but also do not intuitively have community structure. This connection between the model-agnostic E2D2 parameter and its behavior under certain model assumptions motivates the test in the following section.

3. Baseline model test

In the previous section, we derived a hypothesis test based on a user-defined benchmark value that does not refer to a null model. There may be situations, however, where the practitioner does not have a meaningful way to set the null parameter γ_0. In this case, we set the null hypothesis in reference to a particular null model and/or model property. If $P(\boldsymbol{\phi})$ is the true data-generating model for A defined by the parameters $\boldsymbol{\phi}$, then instead of testing $\tilde{\gamma}(P(\boldsymbol{\phi})) \leq \gamma_0$, the null hypothesis is now that $\tilde{\gamma}(P(\boldsymbol{\phi}))$ is less than or equal to the largest value of the E2D2 parameter under the null model. Since the specific set of parameters for the null model is typically unknown, they are estimated from the observed network as $\hat{\boldsymbol{\phi}}$. To fix ideas, we provide the following two examples.

ER model: The simplest null model to consider is the ER model where $P_{ij} = p$ for all i, j. Thus, the value of p completely defines an ER model. We estimate p by taking the average edge probability of the network, that is, $\hat{p} = \sum_{i<j} A_{ij}/\{n(n-1)/2\}$. Then this test determines whether the observed value of the E2D2 estimator is greater than what could arise if the network was generated from an ER model. Recall that this is equivalent to the test in (9) setting $\gamma_0 = 0$.

CL model: Another sensible null model to consider is the CL model (Chung and Lu, 2002) where $P_{ij} = \theta_i \theta_j$ for some weight vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ which uniquely defines the model. This model is similar to the configuration model except instead of preserving the exact degree sequence, it preserves the expected degree sequence. We estimate $\boldsymbol{\theta}$ using the rank 1 adjacency spectral embedding (ASE) (Sussman et al., 2012):

$$\hat{\boldsymbol{\theta}} = |\hat{\lambda}|^{1/2} \hat{\boldsymbol{u}} \tag{12}$$

where $\hat{\lambda}$ is the largest-magnitude eigenvalue of A and $\hat{\boldsymbol{u}}$ is the corresponding eigenvector. Now the test is whether the value of the E2D2 estimator is greater than what likely would have been observed if the CL model generated the observed network.

3.1 Bootstrap test

To carry out this test, we propose a bootstrap procedure. We describe the method for the CL null but full details for the ER and CL null can be found in Algorithm 2. Additionally, the bootstrap test can be trivially modified to test against many other null models.

This approach directly estimates the $1 - \alpha$ quantile of the null distribution of $\tilde{T}(A)$ with a parametric bootstrap and then uses this quantity as the testing threshold. In particular, we first

https://doi.org/10.1017/nws.2024.1 Published online by Cambridge University Press

Algorithm 2. Bootstrap hypothesis test

Result: p-value
Input: $n \times n$ adjacency matrix A, number of iterations B, null model $\mathcal{M}$
Compute $\tilde{T}(A) = \max_{\boldsymbol{c}}\{T(A, \boldsymbol{c})\}$ as in (6)
if $\mathcal{M}$=*ER* **then**
  Compute $\hat{p} = \sum_{i<j} A_{ij}/\{n(n-1)/2\}$
end
if $\mathcal{M}$=*CL* **then**
  Compute $\hat{\boldsymbol{\theta}} = \hat{\lambda}^{1/2}\hat{\mathbf{u}}$ as in (12)
end
for B *times* **do**
  if $\mathcal{M}$=*ER* **then**
    $A_b^* \leftarrow$ ER network with $\hat{p}$
  end
  if $\mathcal{M}$=*CL* **then**
    Draw $\hat{\boldsymbol{\theta}}_b^* = (\hat{\theta}_{b1}^*, \ldots, \hat{\theta}_{bn}^*)^T$ with replacement from $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)^T$ (12)
    $A_b^* \leftarrow$ CL network with $\hat{\boldsymbol{\theta}}_b^*$
  end
  Compute $\tilde{T}_b^* = \max_{\boldsymbol{c}}\{T(A_b^*, \boldsymbol{c})\}$
end
p-val$= \sum_b \mathbb{1}\{\tilde{T}_b^* \geq \tilde{T}(A)\}/B$

compute the test statistic $\tilde{T}(A)$ as in (6). Since the null distribution of $\tilde{T}(A)$ is unknown, we must simulate draws from this distribution in order to have a comparison with our observed test statistic. So we next estimate $\boldsymbol{\theta}$ with the ASE as in (12). Then, for $b = 1, \ldots, B$, we draw $\hat{\boldsymbol{\theta}}_b^* = (\hat{\theta}_{b1}^*, \ldots, \hat{\theta}_{bn}^*)^T$ with replacement from $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)^T$, generate a CL network A_b^* with $\hat{\boldsymbol{\theta}}_b^*$ and find $\tilde{T}_b^* = \max_{\boldsymbol{c}}\{T(A_b^*, \boldsymbol{c}\}$. The empirical distribution of $\{\tilde{T}_b^*\}_{b=1}^B$ serves as a proxy for the null distribution of $\tilde{T}(A)$ so the p-value is

$$p\text{-value} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{\tilde{T}_b^* \geq \tilde{T}(A)\} \tag{13}$$

and we reject H_0 if the p-value is less than a prespecified α.

We highlight a nuanced but important distinction between the interpretation of the baseline value and the baseline model test. For each bootstrap iteration, we re-estimate the number of groups K_b^* that is then used when computing $\tilde{T}_b^* = \max_{\boldsymbol{c}}\{T(A_b^*, \boldsymbol{c})\}$. Therefore, it is possible that for a given iteration, $K_b^* \neq K$, where K was estimated from the original network A. Thus, the results of the bootstrap test are *unconditional* on the value of K. Rejecting against the ER null, for example, means that the observed network has greater community structure than could be expected from a network generated from an ER model, and for any K used to compute $\tilde{T}(A)$. Conversely, the baseline value test results are *conditional* on K. Therefore, rejecting this test means that the observed network has stronger community structure than a network model with $\tilde{\gamma}(P) = \gamma_0$ could generate with K groups.

https://doi.org/10.1017/nws.2024.1 Published online by Cambridge University Press

Now, the bootstrap test naturally fits into the existing community detection testing literature as it considers a specific null model. Indeed, to the knowledge of the authors, this is the first[1] statistical hypothesis test for the significance of communities in a DCBM (Karrer and Newman, 2011) as the CL model serves as the null model of "no communities." Nevertheless, the bootstrap method's strength is its flexibility, owing in part to the general definition of the E2D2 metric, in that it can easily accommodate any null model.[2] Additionally, it allows for more general inference as multiple, realistic null distributions can be tested against, leading to a rich understanding of the network's community structure. The bootstrap test also yields an insightful visual tool where the observed value of the test statistic is plotted next to the bootstrap histogram. This tool helps in visualizing how the strength of community structure observed in the network compares to benchmark networks with the same density, or with the same degree distribution, and so on. We study these plots more in Section 5. Lastly, a fundamental challenge of bootstrapping networks is that, in general, we only observe a single network. If we knew the model parameters (e.g., p or $\boldsymbol{\theta}$) then it would be trivial to generate bootstrap replicates. Since the true model parameters are unknown, we must first estimate them and then generate networks using the estimated parameters. Thus, the quality of the bootstrap procedure depends on the quality of these estimates. In the following subsection, we formally prove certain properties of this procedure.

3.2 Bootstrap theory

We now turn our attention to theoretical properties of the bootstrap. We want to show that, if $A, H \sim P(\boldsymbol{\phi})$ and $\hat{A}^* \sim P(\hat{\boldsymbol{\phi}})$ where $\hat{\boldsymbol{\phi}}$ estimates $\boldsymbol{\phi}$ using A, then $\tilde{T}(\hat{A}^*)$ converges to $\tilde{T}(H)$. To have any hope of showing this result, $\hat{A}^*$ must be similar to A. We consider the *Wasserstein p-distance* and adopt the notation of (Levin and Levina, 2019). Let $p \geq 1$ and let A_1, A_2 be adjacency matrices on n nodes. Let $\Gamma(A_1, A_2)$ be the set of all couplings of A_1 and A_2. Then the Wasserstein p-distance between A_1 and A_2 is

$$W_p^p(A_1, A_2) = \inf_{\nu \in \Gamma(A_1, A_2)} \int d_{GM}^p(A_1, A_2) d\nu \tag{14}$$

where

$$d_{GM}(A_1, A_2) = \min_{Q \in \Pi_n} \binom{m}{r}^{-1} \frac{1}{2} \|A_1 - QA_2Q'\|_1 \tag{15}$$

where Π_n is the set of all $n \times n$ permutation matrices and $\|A\|_1 = \sum_{i,j} |A_{ij}|$. The following results show that $\hat{A}^*$ converges in distribution to A in the Wasserstein p-distance sense for both the ER and CL null. For all results in this section, assume that model parameters do not depend on n, that is, $p_n = p \not\to 0$.

Lemma 3.1. *Let $A, H \sim ER(p)$ and $\hat{A}^* \sim ER(\hat{p})$ where $\hat{p} = \sum_{i,j} A_{ij}/\{n(n-1)\}$. Then,*

$$W_p^p(\hat{A}^*, H) = O(n^{-1}).$$

Lemma 3.2. *Let $A, H \sim CL(\boldsymbol{\theta})$ and $\hat{A}^* \sim CL(\hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}$ is found using (12). Then,*

$$W_p^p(\hat{A}^*, H) = O(n^{-1/2} \log n).$$

Both proofs are based on Theorem 5 in Levin and Levina (2019). These results show that for large n, the networks generated from the bootstrap model are similar to the networks generated from the original model. Since there is only one parameter to estimate in the ER model but n in the CL model, it is sensible that the rate of convergence is much faster for the former.

Next, we show that for a fixed $\boldsymbol{c}$, the distribution of the bootstrapped test statistic converges to the same distribution as the original test statistic. We only show this result for the ER null.

https://doi.org/10.1017/nws.2024.1 Published online by Cambridge University Press

First, we introduce some useful notation. Let $t(A, \boldsymbol{c})$ be the numerator of the E2D2 test estimator, that is,

$$t(A, \boldsymbol{c}) = \frac{1}{m_{\text{in}}} \sum_{i<c} A_{ij} \mathbb{1}(c_i = c_j) - \frac{1}{m_{\text{out}}} \sum_{i<c} A_{ij} \mathbb{1}(c_i \neq c_j) := \sum_{i<j} C_{ij} A_{ij}$$

where m_{in} and m_{out} are the total possible number of intra- and inter- community edges, respecitvely, and $C_{ij} = m_{\text{in}}^{-1}$ if $c_i = c_j$ and m_{out}^{-1} otherwise. Let $C_1 = \sum_{i<j} C_{ij}$ and $C_2 = \sum_{i<j} C_{ij}^2$. Additionally, let K be the number of groups that a node can be assigned, which is assumed to be fixed.[3] Then we have the following result.

Lemma 3.3. *Let $A, H \sim ER(p)$ and $\hat{A}^* \sim ER(\hat{p})$ where $\hat{p} = \sum_{i<j} A_{ij}/\{n(n-1)/2\}$ and consider a fixed $\boldsymbol{c}$. Furthermore, let $s_n^2 = p(1-p)C_2$. Then,*

$$\frac{1}{s_n}\{T(H, \boldsymbol{c}) - \gamma(H, \boldsymbol{c})\} \xrightarrow{d} \mathsf{N}(0, K^2 p^2)$$

and

$$\frac{1}{s_n}\{T(\hat{A}^*, \boldsymbol{c}) - \gamma(H, \boldsymbol{c})\} \xrightarrow{d} \mathsf{N}(0, K^2 p^2)$$

The proof is a simple application of the nonidentically distributed central limit theorem and iterated expectations. This lemma implies that the E2D2 estimator $T(A, \boldsymbol{c})$ consistently estimates the E2D2 model parameter $\gamma(A, \boldsymbol{c})$ for a particular $\boldsymbol{c}$. Additionally, the distribution of the test statistic converges to the same normal distribution, whether the network was generated from the original model or the bootstrap model. This result is more difficult to show for the CL null model. We cannot use the ideas from the proof of Lemma 3.3 because $\hat{\boldsymbol{\theta}}$ has a more complicated form and the bootstrap step is more involved than that of the ER null; nor can we use the results in Levin and Levina (2019) because the E2D2 estimator cannot be written as U-statistic.

Ideally, we would like to show that this result also holds when using the community assignment which maximizes the E2D2 estimator. Unfortunately, showing this convergence for arbitrary statistics is difficult (e.g., (Levin and Levina, 2019)). This is challenging in our particular case for several reasons. First, the E2D2 estimator $\tilde{T}(A)$ is the maximum of $O(e^n)$ statistics $T(A, \boldsymbol{c}_i)$, meaning the maximum is taken over a set of random variables which goes to infinity. Additionally, these variables are nontrivially correlated since they depend on the same adjacency matrix. Another angle to view the difficulty of this problem is that, for $\boldsymbol{c}^* = \arg\max_{\boldsymbol{c}}\{T(A, \boldsymbol{c}\}$, C_{ij}^* is now dependent on A_{ij}. Even computing the mean and variance of this estimator becomes difficult. Bootstrap theoretical results for the maximum of the test statistic is an important avenue for future work.

4. Hypothesis testing simulations

4.1 Settings

We now study the performance of the proposed method on synthetic data. Our primary metric of interest is the rejection rate of the test under different settings. We consider two settings for the baseline value test as well as settings for the baseline model test with both the ER and CL nulls. In each setting, we first fix $\tilde{\gamma}(P)$ and increase the number of nodes n. Then we fix n and increase $\tilde{\gamma}(P)$ and, in both cases, we expect an increasing rejection rate. We run 100 Monte Carlo simulations and compute the fraction of rejections. Our bootstrap method uses $B = 200$ bootstrap samples, and we fix the level of the test at $\alpha = 0.05$. We chose the two Spectral methods proposed in Bickel and Sarkar (2016) as benchmarks since these are leading and well-established methods with formal guarantees. Even though the authors suggest only using the adjusted method, we will

https://doi.org/10.1017/nws.2024.1 Published online by Cambridge University Press

still compare both since, similar to our proposed framework, the authors propose a version of the test with an asymptotic threshold and an adjusted version of the test with a bootstrap correction.

4.2 Test against baseline value

First, we consider the baseline value test using Theorem 2.1, that is, we reject H_0 if

$$\tilde{T}(A) > \left(\gamma_0 + \frac{k_n}{K\hat{p}}\right)(1+\epsilon)$$

where $k_n = \{(\log K)/n\}^{1/2}$ and $\epsilon = 0.0001$. We let $n = 1000, 1200, \ldots, 3000$ and generate networks with K communities. When $n = 1000$, we let $K = 2$ where 60% and 40% of the nodes are in each community, respectively. When $1000 < n \leq 2000$, $K = 4$ with 40%, 20%, 20%, and 20% of the nodes distributed in each community. For $n > 2000$, we have $K = 6$ with 25%, 15%, . . ., 15% of nodes in each community. The edge probabilities P_{ij} are distributed such that

$$P_{ij} \overset{\text{indep.}}{\sim} \mathbb{1}(c_i^* = c_j^*)\mathsf{Uniform}(p_l, p_u) + \mathbb{1}(c_i^* \neq c_j^*)\mathsf{Uniform}(0.05, 0.07)$$

where $\boldsymbol{c}^*$ corresponds to the true community labels. We set $(p_l, p_u) = (0.075, 0.095)$, $(0.100, 0.120)$, and $(0.125, 0.145)$ when $K = 2$, 4 and 6, respectively, in order to fix $\mathsf{E}\{\tilde{\gamma}(P)\} \approx 0.20$ for all parameter combinations. This data-generating model has a block structure but with heterogeneous edge probabilities, which means P is not an SBM. We test against the null hypothesis $H_0 : \tilde{\gamma}(P) \leq \gamma_0 = 0.10$. The results are in Figure 1(a). Since $\mathsf{E}\{\tilde{\gamma}(P)\} > \gamma_0$, the test should reject. The cutoff is asymptotic, however, so the test has low power for small n. But when $n > 2000$, the test has a high power as n is large enough for the asymptotic results to apply.

For the next setting, we fix $n = 5000$ and let $K = 4$ with 40%, 20%, . . ., 20% of the nodes in each community. The edge probabilities P_{ij} are now distributed such that

$$P_{ij} \overset{\text{indep.}}{\sim} \mathbb{1}(c_i^* = c_j^*)\mathsf{Uniform}(0.125, 0.175) + \mathbb{1}(c_i^* \neq c_j^*)\mathsf{Uniform}(a, 0.125)$$

where $a = 0.10, 0.09, \ldots, 0.01$. We find that $\mathsf{E}\{\tilde{\gamma}(P)\} = 0.08, 0.09, \ldots, 0.23$ for different values of a and test against the null hypothesis $H_0 : \tilde{\gamma}(P) \leq \gamma_0 = 0.10$. The results are in Figure 1(b). When $a = 0.10$, $\tilde{\gamma}(P) = 0.08 < \gamma_0 = 0.10$ so we would expect the test to fail to reject which it does. The test should have a high rejection rate when $a = 0.07$ since $\tilde{\gamma}(P) = 0.12 > \gamma_0$. The power of the test, however, does not increase to one until $\tilde{\gamma}(P) = 0.15$. This small discrepancy is due to the fact that it is an asymptotic cutoff and we are generating networks with a finite number of nodes. When $\tilde{\gamma}(P) \geq 0.15$, the test consistently rejects as expected.

4.3 Test against ER null

Next, we study the baseline model test using the bootstrap procedure described in Algorithm 2. First we test against the null hypothesis that the network was generated from the ER model. We showed that this test is also equivalent to the asymptotic test with $\gamma_0 = 0$ so we can also compare the rejection threshold from Theorem 2.1. A natural alternative model to the ER is the SBM (Holland et al., 1983) where $P_{ij} = B_{c_i,c_j}$ and $\boldsymbol{c}$ corresponds to the true community labels. We let $n = 250, 500, 1000, 1500, 2000$ and generate networks from an SBM with K communities. For $n < 1000$, we set $K = 2$ with 60% and 40% of the nodes in each community. $K = 4$ with a 40%, 20%, . . . , 20% split for $1000 \leq n < 2000$, and $K = 6$ for $n = 2000$ with 25%, 10%, . . . , 10% of nodes in each community. We fix the inter-community edge probability $B_{ij} = 0.025$ for $i \neq j$ and set $B_{ii} = 0.05, 0.075, 0.10$ when $K = 2, 4, 6$, respectively, for $i = 1, \ldots, K$. This ensures a fixed $\tilde{\gamma}(P) \approx 0.33$. The results are in Figure 1(c). Since the network is generated from an SBM and we are comparing with an ER null, we expect the test to yield a large rejection rate. We can see that

https://doi.org/10.1017/nws.2024.1 Published online by Cambridge University Press

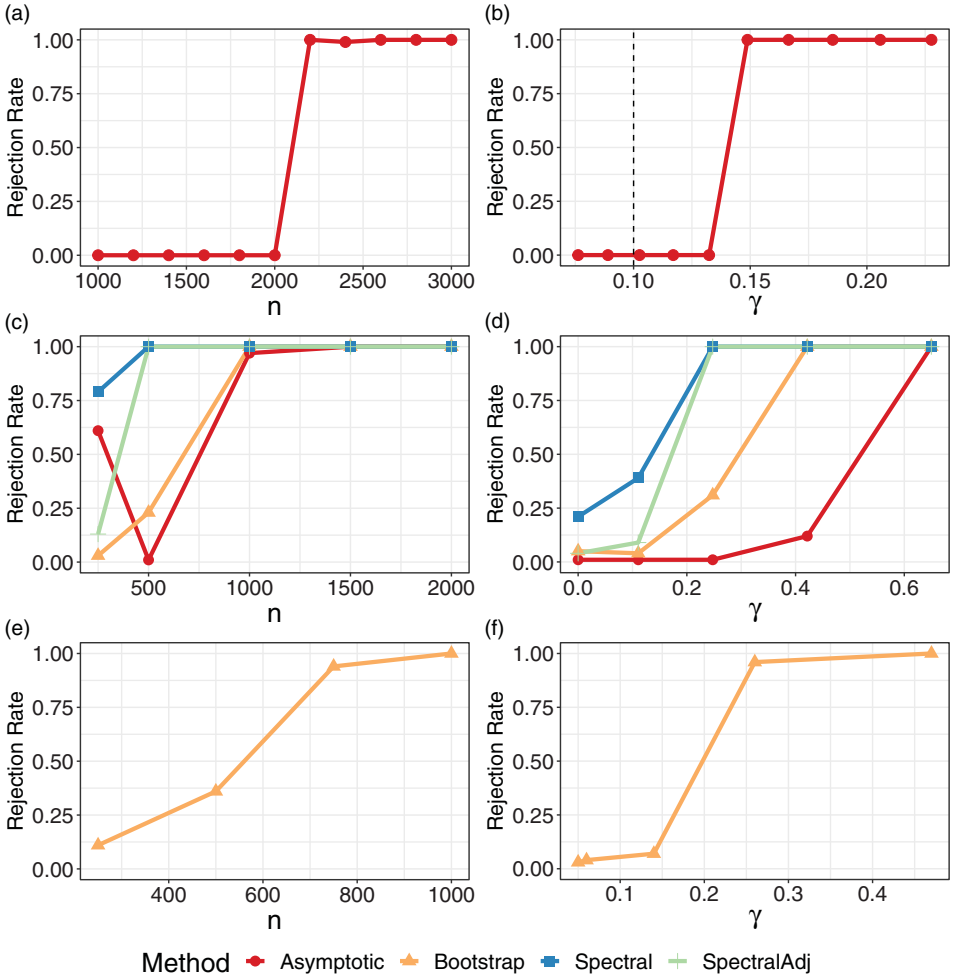

Figure 1. Rejection rates from simulation study. See Section 4 for complete details. (a) Baseline value null with fixed $\tilde{\gamma}(P)$; (b) baseline value null with fixed n; (c) Erdős–Rényi null with fixed $\tilde{\gamma}(P)$; (d) Erdős–Rényi null with fixed n; (e) Chung–Lu null with fixed $\tilde{\gamma}(P)$; (f) Chung–Lu null with fixed n.

both Spectral methods and the proposed bootstrap approach having an increasing rejection rate with increasing n and where the Spectral methods have a larger power. The asymptotic method has a large rejection rate for $n = 250$ but then drops to zero for $n = 500$ before increasing again at $n = 1000$. The reasons for this is that for $n = 250$, the number of communities K is being over-estimated. This causes the test statistic to inflate more than the cutoff, leading to a large rejection rate. For $n \geq 500$, K is more accurately estimated so we see trends that are expected.

In the second scenario, we fix $n = 1000$. Now, the intra-community edge probability is $B_{11} = B_{22} = 0.05$ and inter-community edge probability is $B_{12} = 0.05, 0.04, \ldots, 0.01$. This means that $\tilde{\gamma}(P) = 0$, 0.11, 0.25, 0.42, 0.65. The results are in Figure 1(d). When $\tilde{\gamma}(P) = 0$, the network is generated from an ER model meaning the null hypothesis is true so we expect a low rejection rate. The unadjusted Spectral method has large Type I error and the bootstrap method rejects slightly more than the α level of the test. When $\tilde{\gamma}(P) > 0$, the null hypothesis is false so we expect a large rejection rate. Both Spectral methods reach a power of one by $\tilde{\gamma}(P) = 0.25$ while the bootstrap method does not reach this power until $\tilde{\gamma}(P) = 0.42$. The asymptotic test, as expected, is the most

https://doi.org/10.1017/nws.2024.1 Published online by Cambridge University Press

conservative test. While the Spectral methods outperform the bootstrap test in both scenarios, this is to be expected because these methods were designed for this exact scenario and null. The proposed test is more general so it can be applied to many more settings, but is unlikely to beat a method designed for a specific scenario.

4.4 Test against CL null

Lastly, we study the baseline model test where now the CL model functions as the null, again using Algorithm 2. We drop both Spectral methods for these simulations as the ER null is hard-coded into them, thus making these methods inapplicable for testing against the CL null. We also drop the asymptotic test from the comparison because it is unclear how to set γ_0 for this null model. In fact, finding a closed-form expression for the rejection threshold for the CL null is an interesting avenue of future work. For the alternative model, we consider the degree-corrected block model (DCBM) (Karrer and Newman, 2011) where $P_{ij} = \theta_i\theta_j B_{c_i,c_j}$ and θ_i are node specific degree parameters. We generate networks from a DCBM with $n = 250, 500, \ldots, 1000$. For $n \leq 500$, we let $K = 2$ where 60% and 40% of the nodes are in each community. For $n > 500$, $K = 3$ with 40%, 30%, and 30% of the nodes distributed in each community. The degree parameters are generated $\theta_i \overset{\text{iid}}{\sim} \text{Uniform}(0.2, 0.3)$ and $B_{ii} = 1$ for $i = 1, \ldots, K$. In order to preserve $\tilde{\gamma}(P) \approx 0.33$, $B_{ij} = 0.5, 0.4$ when $K = 2, 3$, respectively, for $i \neq j$. The results are in Figure 1(e). Since the networks are generated from a DCBM, we expect a large rejection rate. We see that rejection rate increases monotonically with n, reaching a power of one by $n = 1000$.

For the second scenario, we fix $K = 4$ with 40%, 20%, ..., 20% of the nodes in each community and $n = 1000$. Let $B_{ii} = 1$, $B_{ij} = 1,\ 0.8,\ \ldots,\ 0.2$ for $i \neq j = 1, \ldots, 4$. Thus, $\tilde{\gamma}(P) = 0.05, 0.06, 0.14, 0.26, 0.47$. The results are in Figure 1(f). When $B_{12} = 1$ ($\tilde{\gamma}(P) = 0.03$), the networks are generated from the CL (null) model so we expect a low rejection rate and the bootstrap test has a low Type I error. When $B_{12} < 1$ ($\tilde{\gamma}(P) > 0.05$), the null hypothesis is false so we expect a large rejection rate. The bootstrap reaches a power of one by $\tilde{\gamma}(P) = 0.47$.

5. Real data analysis

We now study the proposed method on two networks: DBLP and hospital interactions. The DBLP is a computer science bibliography website and this network was extracted by Gao et al. (2009) and Ji et al. (2010). Here, each node represents an author and an edge signifies that the two authors attended the same conference. Additionally, we only consider two author's research areas, databases and information retrieval. The hospital network (Vanhems et al., 2013) captures the interacts between patients and healthcare providers at a hospital in France. Each person is represented by a node and an edge signifies that they were in close proximity.

For each dataset, we consider several metrics. First, we compute $\tilde{T}(A)$ and find the largest value of γ_0 such that the test in (9) is still rejected. We also compute the p-value and find the bootstrap histogram of the test statistic for the ER and CL null models (with $B = 1,000$). Considering both null hypotheses together allows us to gain a richer understanding of the network. We compare the proposed method to the p-value from the adjusted Spectral method (Bickel and Sarkar, 2016). See Table 1 for numeric results and Figure 2 for histograms from the bootstrap method. While Table 1 provides a succinct summary, the plots in Figure 2 provide more details and insights. In these plots, the observed test statistic computed from the dataset is plotted as a vertical line along with histograms representing bootstrap distributions from various benchmark models. These simple but informative plots give practitioners a reference of how the observed community structure compares to the range of community structure in various benchmarks.

https://doi.org/10.1017/nws.2024.1 Published online by Cambridge University Press

Table 1. The number of nodes n and edges m for real-world networks. $\tilde{T}(A)$ is the observed value of the *E2D2* parameter and γ_0 is the largest null value such that the baseline value test would be rejected. Additionally, we report the p-values for the adjusted Spectral method and Bootstrap method against different null hypotheses (ER = Erdős–Rényi, CL = Chung–Lu)

					Spectral adj.	Bootstrap	
Network	n	m	$\tilde{T}(A)$	γ_0	ER	ER	CL
DBLP	2,203	1,148,044	0.75	0.73	< 0.001	0.000	0.000
Hospital	75	1,139	0.32	0.22	< 0.001	0.000	0.104

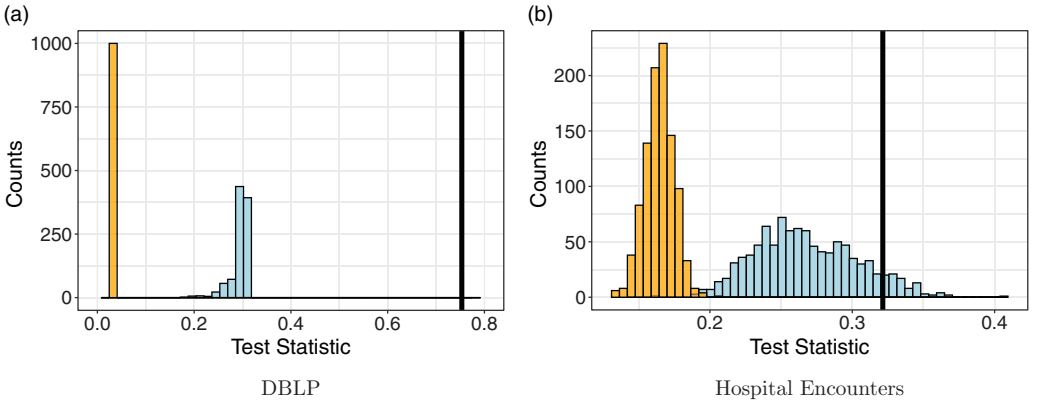

Figure 2. Histograms of bootstrap samples from the proposed method for the two real datasets. The orange histogram is with the Erdős–Rényi null, and the blue histogram is with the Chung–Lu null. The vertical line (black) indicates the value of the test statistic.

For the DBLP network, since we only selected two research areas, we set $K = 2$ to find $\tilde{T}(A)$. The observed value of $\tilde{T}(A) = 0.75$ is quite large and, due to the network's size and density, would reject the baseline value test up to $\gamma_0 = 0.73$. This means that if P generated this network, then we can assert that $\tilde{\gamma}(P) \geq 0.73$. Additionally, all p-values for the model-based tests are effectively zero and, moreover, the observed test statistic is in the far right tail of both bootstrap null distributions. Since both tests are rejected, then it is very unlikely that the network was generated from an ER or CL model and, instead, implies that there is highly significant community structure in this network. This finding accords with existing literature (e.g., (Sengupta and Chen, 2018)).

For the Hospital network, the test of $H_0 : \gamma_0 = 0$ is rejected, as are both tests with ER null. This is a sensible result since we showed that $\gamma_0 = 0$ is equivalent to testing against the ER null. The p-value for the CL null, however, is not significant at the $\alpha = 0.05$ level, meaning that this network does not have more community structure than we would expect to occur from a CL network by chance. So while there is strong evidence that the network diverges from an ER model, these findings indicate that perhaps the low ER-null p-values are due to degree heterogeneity rather than community structure. Indeed, the histogram shows how the test statistic is very unlikely to have been drawn from the ER distribution, but is reasonably likely to have come from the CL distribution since this distribution has a greater mean and variance. Using an ER null alone would have led to the conclusion that there is community structure in this network. By using multiple nulls together with the proposed method, however, we gain a fuller understanding of the network by concluding that degree heterogeneity may be masquerading as community structure.

https://doi.org/10.1017/nws.2024.1 Published online by Cambridge University Press

6. Discussion

In this work, we proposed two methods to test for community structure in networks. These tests are rooted in a formal and general definition of the E2D2 parameter. This metric is simple, flexible, and well connected to the conceptual notion of community structure which we argue makes it a more principled approach. In fact, the test statistic can even be used as a descriptive statistic to quantify the strength of community structure in a network. Existing methods are based on specific random graph models, such as the ER model, which are implicitly presumed to be the only models that do not have community structure. While our second testing approach fits into this framework as well, the general nature of the E2D2 parameter means that we can test against nearly any null model (subject to conditional edge independence) to obtain a richer set of practical insights compared to existing methods. Given a network, we recommend that practitioners first carry out the test against the ER null. If this test is rejected, further tests should be carried out to check whether the it could be due to some other network feature like degree heterogeneity. Thus, the method not only helps decide whether the network appears to exhibit community structure but also helps understand the source of this ostensible community structure.

There are several interesting future research directions. First, the proposed E2D2 parameter and bootstrap testing framework could be adapted for sequential testing. In Ghosh and Barnett (2023), the authors propose a general framework for sequentially testing for $H_0 : K$ vs. H_1:$K+1$ communities in the network. Using the proposed parameter, we would first find $\tilde{T}(A)$ setting $K = 2$, and then generate bootstrap samples with $K = 1$ (ER model). If this test is rejected, we find $\tilde{T}(A)$ with $K = 3$ and generate bootstrap samples from an SBM with $K = 2$. Note that scaling the E2D2 parameter by K ensures a fair comparison of the metric across different values of K. This process continues until we fail to reject the null hypothesis. Since our proposed bootstrap procedure yields a valid p-value, we can directly apply the results in Ghosh and Barnett (2023) to ensure a prespecified error tolerance.

Additionally, the proposed E2D2 parameter is currently limited to quantifying assortative community structure. Extending the method to handle disassortative and/or bi-partite networks would be a interesting contribution. Next, while the asymptotic test is more adept to scale to large networks, the bootstrap test is limited to networks of up to (roughly) $n = 10,000$ nodes for computational feasibility. There are also open theoretical questions including a more precise asymptotic cutoff that accounts for correlation between the random variables as well as bootstrap theory for the maximum test statistic. Moreover, the ideas from this work could be extended to test for other network properties like core–periphery structure (Borgatti and Everett, 2000).

Acknowledgments. We would like to thank the anonymous referees whose comments greatly improved the quality of the manuscript. We would also like to thank Alvin Sheng for help with developing the initial ideas.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/nws.2024.1.

Competing interests. None.

Data availability. All data used in this paper are readily available online.

Funding statement. The authors have no funding to report.

Notes

1 Mukherjee and Sen (2021) test for the significance of *degrees* in a DCBM, that is, H_0: SBM vs. H_1: DCBM whereas the proposed test considers H_0: CL vs. H_1: DCBM.

2 Subject to the modeling constraints mentioned in Section 2.2.

3 Note that for the ER data-generating model, there is effectively only one community. During the inference procedure, however, K is unknown and must be estimated from the network. Thus, in this situation, K is better considered as the number of groups a node can be assigned to, rather than the number of communities.

References

Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T. . . . Qin, Y. (2017). Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research*, *18*(1), 8393–8484.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.

Bickel, P. J., & Chen, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, *106*(50), 21068–21073.

Bickel, P. J., & Sarkar, P. (2016). Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, *78*(1), 253–273.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: theory and experiment*, *2008*(10), P10008.

Borgatti, S. P., & Everett, M. G. (2000). Models of core/periphery structures. *Social Networks*, *21*(4), 375–395.

Chung, F., & Lu, L. (2002). The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, *99*(25), 15879–15882.

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*(6), 066111.

Erdős, P., & Rényi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen*, *6*, 260–297.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3-5), 75–174.

Fosdick, B. K., Larremore, D. B., Nishimura, J., & Ugander, J. (2018). Configuring random graph models with fixed degree sequences. *Siam Review*, *60*(2), 315–355.

Fotouhi, B., Momeni, N., Allen, B., & Nowak, M. A. (2019). Evolution of cooperation on large networks with community structure. *Journal of the Royal Society Interface*, *16*(152), 20180677.

Gao, J., Liang, F., Fan, W., Sun, Y., & Han, J. (2009). Graph-based consensus maximization among multiple supervised and unsupervised models. *Advances in Neural Information Processing Systems*, *22*, 585–593.

Ghosh, R. P., & Barnett, I. (2023). Selecting a significance level in sequential testing procedures for community detection. *Applied Network Science*, *8*(1), 49.

Guo, Z., Cho, J.-H., Chen, R., Sengupta, S., Hong, M., & Mitra, T. (2020). Online social deception and its countermeasures: a survey. *IEEE Access*, *9*, 1770–1806.

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098.

Holland, P., Laskey, K., & Leinhardt, S. (1983). Stochastic blockmodels: first steps. *Social Networks*, *5*(2), 109–137.

Ji, M., Sun, Y., Danilevsky, M., Han, J., & Gao, J. (2010). Graph regularized transductive classification on heterogeneous information networks. In *Machine Learning and Knowledge Discovery in Databases* (pp. 570–586). Springer.

Jin, J. (2015). Fast community detection by SCORE. *The Annals of Statistics*, *43*(1), 57–89.

Kane, G. C., Alavi, M., Labianca, G., & Borgatti, S. P. (2014). What's different about social media networks? a framework and research agenda. *MIS Quarterly*, *38*(1), 275–304.

Karrer, B., & Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, *83*(1), 016107.

Kempe, D., Kleinberg, J., & Tardos, É. (2003). Maximizing the spread of influence through a social network. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146).

Lancichinetti, A., Radicchi, F., & Ramasco, J. J. (2010). Statistical significance of communities in networks. *Physical Review E*, *81*(4), 046110.

Leitch, J., Alexander, K. A., & Sengupta, S. (2019). Toward epidemic thresholds on temporal networks: a review and open questions. *Applied Network Science*, *4*(1), 105.

Levin, K., & Levina, E. (2019). Bootstrapping networks with latent space structure. arXiv preprint arXiv: 1907.10821.

Li, Y., & Qi, Y. (2020). Asymptotic distribution of modularity in networks. *Metrika*, *83*(4), 467–484.

Mancoridis, S., Mitchell, B. S., Rorres, C., Chen, Y., & Gansner, E. R. (1998). Using automatic clustering to produce high-level system organizations of source code. In *Proceedings of the 6th International Workshop on Program Comprehension* (pp. 45–52). IEEE.

Mason, O., & Verwoerd, M. (2007). Graph theory and networks in biology. *IET Systems Biology*, *1*(2), 89–119.

Mukherjee, R., & Sen, S. (2021). Testing degree corrections in stochastic block models. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, Vol. *57*, (pp. 1583–1635). Institut Henri Poincaré.

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review*, *74*(3 Pt 2), 036104.

https://doi.org/10.1017/nws.2024.1 Published online by Cambridge University Press

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. In *Advances in neural information processing systems* (pp. 849–856).
Osawa, S., & Murata, T. (2015). Selecting seed nodes for influence maximization in dynamic networks. In: Mangioni, G., Simini, F., Uzzo, S. M., & Wang, D., ed. *Complex Networks VI* (pp. 91–98). Springer.
Palowitch, J., Bhamidi, S., & Nobel, A. B. (2018). Significance-based community detection in weighted networks. *Journal of Machine Learning Research*, *18*, 1–48.
Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p∗) models for social networks. *Social Networks*, *29*(2), 173–191.
Rohe, K., Chatterjee, S., & Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, *39*(4), 1878–1915.
Sengupta, S., & Chen, Y. (2015). Spectral clustering in heterogeneous networks. *Statistica Sinica*, *25*, 1081–1106.
Sengupta, S., & Chen, Y. (2018). A block model for node popularity in networks with community structure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology*, *80*(2), 365–386.
Sussman, D. L., Tang, M., Fishkind, D. E., & Priebe, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, *107*(499), 1119–1128.
Vanhems, P., Barrat, A., Cattuto, C., Pinton, J.-F., Khanafer, N., Regis, C. . . . Voirin, N. (2013). Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PloS One*, *8*(9), e73970.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, *393*(6684), 440–442.
Yanchenko, E., Murata, T., & Holme, P. (2023). Influence maximization on temporal networks: a review. arXiv preprint arXiv: 2307.00181.
Yuan, M., Liu, R., Feng, Y., & Shang, Z. (2022). Testing community structure for hypergraphs. *The Annals of Statistics*, *50*(1), 147–169.

Cite this article: Yanchenko E. and Sengupta S. A generalized hypothesis test for community structure in networks. *Network Science*. https://doi.org/10.1017/nws.2024.1