COMBINATORIAL-PROBABILISTIC TRADE-OFF: P-VALUES OF COMMUNITY PROPERTIES TEST IN THE STOCHASTIC BLOCK MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose an inferential framework testing the general community combinatorial properties of the stochastic block model. We aim to test the hypothesis on whether a certain community property is satisfied, e.g., whether a given set of nodes belong to the same community, and provide p-values for uncertainty quantification. Our framework is applicable to all symmetric community properties. To ease the challenges caused by the combinatorial nature of community properties, we develop a novel shadowing bootstrap method. Utilizing the symmetry, our method can find a shadowing representative of the true assignment and the number of tested assignments in the alternative is largely reduced. In theory, we introduce a combinatorial distance between two community classes and show a combinatorial-probabilistic trade-off phenomenon. Our test is honest as long as the product of the combinatorial distance between two communities and the probabilistic distance between two connection probabilities is sufficiently large. Besides, we show that such trade-off also exists in the information-theoretic lower bound. We also implement numerical experiments to show the validity of our method.

1 INTRODUCTION

Clustering is an important feature for network studies, which refers to the presence of node communities in the underlying graph. Community partitions the nodes into subgroups, within which a higher level of connectivity is perceived. Its broad spectrum of applications includes the fields of sociology (Wasserman & Faust, 1994), biology (Barabási & Oltvai, 2004), physics (Newman, 2003) and internet (Albert et al., 1999). Stochastic block model (SBM) (Holland et al., 1983) is one of the most widely studied statistical models depicting the network community structures. It is a random graph model which divides the nodes into disjoint communities and assigns the probability of connection between two nodes according to their community memberships. One of its central problems in previous studies is community detection. However, most of the existing research focused on estimating the community labeling without uncertainty quantification (Choi et al., 2012; Mossel et al., 2012; Airoldi et al., 2013; Massoulié, 2014; Abbe et al., 2016; Mossel et al., 2016). Some fundamental limits of the community recovery have also been established in previous studies. For example, Abbe et al. (2016) showed the optimal phase transition for the exact recovery of the community assignments using the maximum likelihood. The semi-definite relaxation methods (Abbe et al., 2016; Hajek et al., 2016; Agarwal et al., 2017; Bandeira, 2018) and the spectral methods (Yun & Proutiere, 2014; Abbe & Sandon, 2015; Gao et al., 2017; Abbe et al., 2020) are also shown to be optimal in exact recovery. Besides the exact recovery, Zhang & Zhou (2016) quantified the statistical rate of the community estimation via the mis-match ratio and showed the minimax rate of the mis-match ratio for community detection. In summary, the community estimation studies have two major limits: 1) it does not provide the p-values to evaluate the uncertainty of the estimation, which are essential in many scientific applications, and 2) it requires the recovery of community assignments for all nodes, while in many scientific applications we are interested in the community properties of a specific subset of nodes, e.g., whether two sets of nodes belong to the same community. We formulate the following examples of statistical hypotheses for illustration.

Example 1.1 (Same community test for m nodes)**.** We want to test whether m given nodes are in the same cluster or not. Without loss of generality, we have the hypothesis:

$$\mathrm{H}_0 : \text{Nodes } 1, \ldots, m \text{ belong to the same community,}$$
$$\mathrm{H}_1 : \text{There exists two nodes } 1 \leq j \neq k \leq m \text{ belonging to two different communities.}$$

Example 1.2 (Group community test)**.** For two groups of nodes, we know in prior that nodes within each group belong to the same community. We aim to further test whether these two groups belong to the same community. We denote one node set as $S_m = \{1, \ldots, m\}$ and the other as $S_{m'} = \{m+1, \ldots, m+m'\}$. The group community hypothesis is

$$\mathrm{H}_0 : \text{Nodes in } S_m \cup S_{m'} \text{ belong to the same community,}$$
$$\mathrm{H}_1 : \text{Nodes in } S_m \text{ belong to community } a\text{, but nodes in } S_{m'} \text{ belong to community } b \neq a.$$

Example 1.3 (Equal-sized communities test)**.** Given an SBM of n nodes and K communities, we aim to test whether each community has the same size. Namely, we aim to test the hypothesis:

$$\mathrm{H}_0 : \text{Each community has the size } n/K,$$
$$\mathrm{H}_1 : \text{At least one of the communities has size not equal to } n/K.$$

In order to conduct hypothesis tests including the above examples, we develop a general community property test. We consider the SBM with n nodes and K communities. Denote the community assignment of the nodes by $z = (z(1), ..., z(n)) \in \{1, \ldots, K\}^n$. The homogeneous SBM assumes that the edges of the random graph are independent Bernoulli random variables with connection probability p if $z(i) = z(j)$ and q if $z(i) \neq z(j)$. Let $\mathcal{C}_0, \mathcal{C}_1 \subset \{1, \ldots, K\}^n$ be two disjoint community assignment families. We are interested in the *general community property test*:

$$\mathrm{H}_0 : z \in \mathcal{C}_0 \text{ versus } \mathrm{H}_1 : z \in \mathcal{C}_1. \tag{1.1}$$

We characterize the hardness of the test by two kinds of "distances": the probabilistic distance between p and q, and the combinatorial distance between $\mathcal{C}_0$ and $\mathcal{C}_1$. The existing literature on community detection only focused on the probability distance, e.g., the Rényi divergence $I(p, q) = -2\log\left(\sqrt{pq} + \sqrt{(1-p)(1-q)}\right)$ (Zhang & Zhou, 2016). In comparison, our paper introduce a novel combinatorial distance between $\mathcal{C}_0$ and $\mathcal{C}_1$ denoted as $d(\mathcal{C}_0, \mathcal{C}_1)$ (see Definition 2.4) and propose a general testing method that is honest and powerful when

$$\text{Combinatorial-Probabilistic Trade-Off: } I(p, q)d(\mathcal{C}_0, \mathcal{C}_1) = \Omega(n^{\epsilon}) \tag{1.2}$$

for some arbitrarily small $\epsilon > 0$. On the other hand, we show the minimax lower bound of the test in the sense that H_0 and H_1 in (1.1) cannot be differentiated when $I(p, q)d(\mathcal{C}_0, \mathcal{C}_1) \leq c \log n$ for some constant $c > 0$.[1] The multiplication between $I(p, q)$ and $d(\mathcal{C}_0, \mathcal{C}_1)$ reveals the trade-off between the probabilistic distance and the combinatorial distance in the general community property test.

2 Community Properties of the Stochastic Block Model

In our paper, we consider the fixed assignment stochastic block model, denoted by $\mathcal{M}(n, K, p, q, z)$. Denote by $[n] = \{1, \ldots, n\}$ for any integer n. For simplicity, we start with the scenario where the community sizes are even, and will generalize to the uneven case in Appendix A. We denote the even assignment class by $\mathcal{K}^n := \{z \in [K]^n : |\{i : z(i) = k\}| = n/K, \forall k \in [K]\}$. In our paper, we assume K to be bounded. Let $\mathbf{A} \in \{0, 1\}^{n \times n}$ be the symmetric adjacency matrix of the random graph generated from the SBM. In the following part of the paper, we will study the community property test with an observation of the adjacency matrix $\mathbf{A} \sim \mathcal{M}(n, K, p, q, z)$.

2.1 Symmetric Community Properties

In this section, we aim to define the community property and the distance between two community families. In general, we say a community property is a subset of $[K]^n$. However, such definition is too general and may include some ill-posed examples. For instance, if we can transfer one as-

[1] We refer to Theorem 3.2 and Theorem 4.1 for the rigorous arguments about the upper and lower bounds.

signment to another under certain permutation of the community labels, they are essentially the same assignment and should belong to the same community property. This motivates us to give the following definition of equivalent assignments.

Definition 2.1 (Equivalent community assignments)**.** Let S_K be the symmetric group containing all bijections from $[K]$ to itself. We say two assignments z and $z' \in [K]^n$ are equivalent, denoted as $z \simeq z'$, if there exists a permutation $\sigma \in S_K$ such that $\sigma(z) = z'$. Here $\sigma(z)$ means implementing the permutation σ to each entry of the vector z. More generally, given a node set $\mathcal{N} \subseteq [n]$, we denote by $z_{\mathcal{N}}$ the sub-vector of z with entries in $\mathcal{N}$. We say $z_{\mathcal{N}} \simeq z'_{\mathcal{N}}$ if there exists a permutation $\sigma \in S_K$ such that $\sigma(z_{\mathcal{N}}) = z'_{\mathcal{N}}$.

With Definition 2.1, we give the definition of symmetric community properties as follows.

Definition 2.2 (Symmetric community properties)**.** We say a community property $\mathcal{C}_0$ is symmetric, if there exist a node set $\mathcal{N} \subseteq [n]$ and an assignment $\widetilde{z} \in \mathcal{K}^n$, such that $\mathcal{C}_0 = \{z \in \mathcal{K}^n : z_{\mathcal{N}} \simeq \widetilde{z}_{\mathcal{N}}\}$. We say some $\mathcal{C}_1 \subseteq \mathcal{K}^n \backslash \mathcal{C}_0$ is an alternative property of $\mathcal{C}_0$ if $\mathcal{C}_1$ is closed under permutations on the support $\mathcal{N}$, i.e., for any $z \in \mathcal{C}_1 \subseteq \mathcal{K}^n \backslash \mathcal{C}_0$, if some z' satisfies $z'_{\mathcal{N}} \simeq z_{\mathcal{N}}$, then $z' \in \mathcal{C}_1$ as well.

Intuitively, the node set $\mathcal{N}$ and the assignment $\widetilde{z}$ in Definition 2.2 are the representative node set and assignment generating all possible assignments in the community property via permutation. The community property $\mathcal{C}_0$ is "symmetric" in the sense that all its assignments are equivalent on the support of node set $\mathcal{N}$. We impose the following assumption on testing symmetric properties.

Assumption 2.1 (Symmetric community property test)**.** In the hypothesis test $\mathrm{H}_0 : z \in \mathcal{C}_0$ v.s. $\mathrm{H}_1 : z \in \mathcal{C}_1$, we assume $\mathcal{C}_0, \mathcal{C}_1 \subseteq \mathcal{K}^n$ and $\mathcal{C}_0$ is symmetric and $\mathcal{C}_1$ is an alternative property of $\mathcal{C}_0$.

By Definition 2.2, $\mathcal{C}_1 = \mathcal{C}_0^c$ is an alternative property of $\mathcal{C}_0$. Meanwhile, $\mathcal{C}_1$ satisfying the assumption above could be a strict subset of $\mathcal{C}_0^c$, allowing more examples in practice. In fact, we can show that Examples 1.1 and 1.2 satisfies Assumption 2.1, whose concrete forms of $\mathcal{N}$ and $\widetilde{z}$ are given below.

• Example 1.1: Same community test for m nodes. We can define

$$\mathcal{C}_0 = \{z \in \mathcal{K}^n : z(1) = \cdots = z(m)\} \text{ and } \mathcal{C}_1 = \mathcal{K}^n \backslash \mathcal{C}_0, \tag{2.1}$$

while $\mathcal{N} = [m]$ and $\widetilde{z}$ is any assignment satisfying $\widetilde{z}_{\mathcal{N}} = (1, \ldots, 1) \in [K]^m$.

• Example 1.2: Same community test for groups. We have

$$\begin{aligned} \mathcal{C}_0 &= \{z \in \mathcal{K}^n : z(1) = \cdots = z(m) = z(m+1) = \cdots = z(m+m')\} \\ \mathcal{C}_1 &= \{z \in \mathcal{K}^n : z(1) = \cdots = z(m) \neq z(m+1) = \cdots = z(m+m')\}, \end{aligned} \tag{2.2}$$

where $\mathcal{C}_0$ is symmetric by choosing $\mathcal{N} = [m+m']$ and $\widetilde{z}_{\mathcal{N}} = (1, \ldots, 1) \in [K]^m$.

2.2 Combinatorial Distance Between Community Properties

In order to depict the relationship between $\mathcal{C}_0$ and $\mathcal{C}_1$, we propose a metric of distance in terms of the number of misaligned edges. We first define the set of misaligned edges between two assignments.

Definition 2.3. For any two assignments $z_0 \in \mathcal{C}_0$ and $z_1 \in \mathcal{C}_1$, we define the two sets of misaligned edges as

$$\begin{aligned} &\mathcal{E}_1(z_0, z_1) = \{(i,j) : i < j, i, j \in [n], z_0(i) = z_0(j), z_1(i) \neq z_1(j)\} \text{ and} \\ &\mathcal{E}_2(z_0, z_1) = \{(i,j) : i < j, i, j \in [n], z_0(i) \neq z_0(j), z_1(i) = z_1(j)\}, \end{aligned}$$

where $\mathcal{E}_1(z_0, z_1)$ contains the edges whose corresponding nodes are assigned to the same community in z_0 but to two different communities by z_1, and $\mathcal{E}_2(z_0, z_1)$ is the opposite. See Figure 1 for illustration. We define $n_i(z_0, z_1) = |\mathcal{E}_i(z_0, z_1)|$, for $i = 1, 2$ as the cardinality of the two edge sets.

Now we are ready to propose the metric of assignment distance defined as follows.

Definition 2.4 (Community property distance)**.** We define the distance between two assignments z_0 and z_1 as $d(z_0, z_1) = n_1(z_0, z_1) \vee n_2(z_0, z_1)$. Correspondingly, we also define $d(z_0, \mathcal{C}_1) = \inf_{z_1 \in \mathcal{C}_1} d(z_0, z_1)$ and the distance between two community properties

$$d(\mathcal{C}_0, \mathcal{C}_1) = \inf_{z_0 \in \mathcal{C}_0, z_1 \in \mathcal{C}_1} d(z_0, z_1).$$

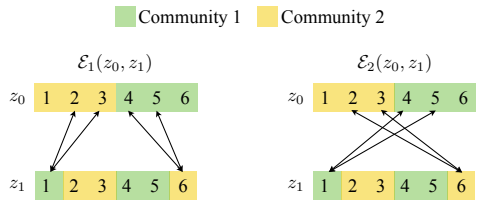


Figure 1: Example of misaligned edges in Definition 2.3.

By definition, the distance $d(\mathcal{C}_0, \mathcal{C}_1)$ is the minimal number of misaligned edges between $\mathcal{C}_0$ and $\mathcal{C}_1$. We refer the computation of $d(\mathcal{C}_0, \mathcal{C}_1)$ for more general examples to Section 2.4.

2.3 Likelihood-Ratio Test for Community Properties

Our method starts with defining a likelihood-ratio test statistic. We denote the observed adjacency matrix from the true model as $\mathbf{A} \sim \mathcal{M}(n, K, p, q, z^*)$, where z^* is the true assignment. The likelihood function of the stochastic block model is

$$f(\mathbf{A}; z, p, q) = \Pi_{i<j} p^{\mathbb{1}(z(i)=z(j))\mathbf{A}_{ij}} (1-p)^{\mathbb{1}(z(i)=z(j))(1-\mathbf{A}_{ij})} q^{\mathbb{1}(z(i)\neq z(j))\mathbf{A}_{ij}} (1-q)^{\mathbb{1}(z(i)\neq z(j))(1-\mathbf{A}_{ij})}.$$

We then denote the log-likelihood ratio statistic as

$$\text{LRT} = \log \frac{\sup_{z\in\mathcal{C}_1} f(\mathbf{A}; z, p, q)}{\sup_{z\in\mathcal{C}_0\cup\mathcal{C}_1} f(\mathbf{A}; z, p, q)}. \tag{2.3}$$

In order to conduct the property test, we study the limiting distribution of the likelihood ratio statistic. In specific, when $1/\rho_n = o(n^{1-c_2})$ for some constant $c_2 > 0$, where ρ_n is the signal strength, we are able to decompose the LRT as follows

$$\begin{aligned} \text{LRT} &= \sup_{z\in\mathcal{C}_1} \log f(\mathbf{A}; z, p, q) - \log f(\mathbf{A}; z^*, p, q) + o(1) \\ &= \sup_{z\in\mathcal{C}_1} g(p, q) \Bigg(\sum_{(i,j)\in\mathcal{E}_2(z^*,z)} \mathbf{A}_{ij} - \sum_{(i,j)\in\mathcal{E}_1(z^*,z)} \mathbf{A}_{ij} \Bigg) + o(1), \end{aligned} \tag{2.4}$$

where $g(p, q) = \log p(1-q)/\left(q(1-p)\right)$. We observe that the leading term in (2.4) is the difference of edges in two edge sets: $\mathcal{E}_2(z^*, z)$ and $\mathcal{E}_1(z^*, z)$. By Definition 2.4, the property distance $d(\mathcal{C}_0, \mathcal{C}_1)$ is larger when the two edge sets are larger, which makes the leading term larger as well. This implies why $d(\mathcal{C}_0, \mathcal{C}_1)$ characterizes the difficulty of the test.

Remark 2.1. The likelihood ratio statistic in (2.3) is similar to the one in Wang & Bickel (2017). They considered the hypothesis on a specific community property: the number of communities.

$$\mathcal{C}_0 = \{z | z \in [K-1]^n\} \text{ and } \mathcal{C}_1 = \{z | z \in [K]^n\} \backslash \mathcal{C}_0. \tag{2.5}$$

They showed that the suprema $\sup_{z\in\mathcal{C}_1} f(\mathbf{A}; z, p, q)$ in (2.3) is unique, as illustrated in Figure 2(a), which makes their LRT asymptotically normal for $\mathcal{C}_0, \mathcal{C}_1$ in (2.5). However, this is not always true for the general community properties. For some properties, there will be an exponential number of candidate assignments maximizing the likelihood in (2.3), as illustrated in Figure 2(b). Thus the LRT is no longer asymptotically normal for the general case. Therefore, despite the similar formality of the LRT, our testing procedure will be different from theirs.

To characterize the suprema in the LRT, we define the boundary of the alternative properties.

Definition 2.5 (Boundary of community class)**.** For a given null assignment $z_0 \in \mathcal{C}_0$, we define the boundary of $\mathcal{C}_1$ as:

$$B_{z_0} = \left\{ z \in \mathcal{C}_1 : d(z_0, z) = d(z_0, \mathcal{C}_1) \right\}.$$

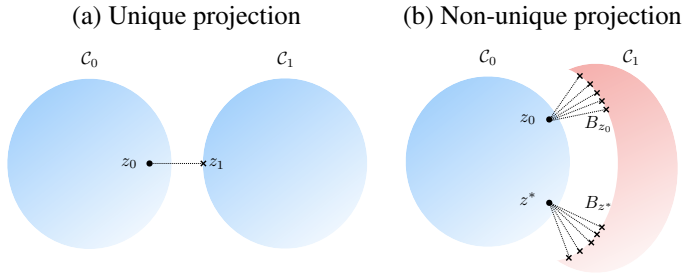


Figure 2: The boundary of $\mathcal{C}_1$ given z_0. On the left panel, the projection is unique and $B_{z_0} = z_1$. On the right panel, the projection is not unique, where z_0 is the shadowing assignment of z^*, and B_{z^*} and B_{z_0} have similar structures.

Namely, B_{z_0} is the projection of z_0 onto $\mathcal{C}_1$. Our analysis shows that the likelihood maximizer is asymptotically equivalent to the boundary B_{z*}.[2] In the next section, we will study the asymptotic property of the LRT under such case by studying the structure of B_{z_0}.

2.4 Shadowing Bootstrap for the Property Test

There are two major challenges to estimate the p-value of LRT. First, the suprema of the likelihood is not unique and therefore the limiting distribution of LRT is not necessarily normal. By (2.4), we can in turn study the limiting distribution of the leading term

$$L := \sup_{z \in \mathcal{C}_1} g(p,q) \Bigg(\sum_{(i,j) \in \mathcal{E}_2(z^*,z)} \mathbf{A}_{ij} - \sum_{(i,j) \in \mathcal{E}_1(z^*,z)} \mathbf{A}_{ij} \Bigg). \tag{2.6}$$

Chernozhukov et al. (2013) studied the limiting distribution of the maximal of high dimensional empirical process and proposed to estimate its quantile by multiplier bootstrap. However, their method imposes that the dimension d of the empirical process and the sample size n satisfies $\log d / n^{1/5} = o(1)$, whereas in (2.6), the dimension $d = |\mathcal{C}_1|$ could be of the order K^n and violates the scaling condition. To handle such problem, our key observation is that the supreme over the alternative $\mathcal{C}_1$ can be represented by the supreme over its boundary B_{z*}. In particular, we show that the leading term L is asymptotically the same as the following statistic:

$$L_0 := \sup_{z \in B_{z^*}} g(p,q) \Bigg(\sum_{(i,j) \in \mathcal{E}_2(z^*,z)} \mathbf{A}_{ij} - \sum_{(i,j) \in \mathcal{E}_1(z^*,z)} \mathbf{A}_{ij} \Bigg). \tag{2.7}$$

The cardinality of B_{z^*} is much smaller than that of $\mathcal{C}_1$ (usually polynomial to n), and therefore satisfies the scaling condition of high dimensional multiplier bootstrap. However, we cannot construct B_{z^*} in practice as z^* is unknown. This leads to the second challenge: how to find B_{z^*} in practice? Our key insight to solve the second challenge is to utilize the symmetry property in Definition 2.2. This insight relies on the following lemma characterizing the covariance of two processes.

Lemma 2.2 (Shadowing symmetry)**.** For a given $z \in \mathcal{C}_0$, we list the assignments in the boundary B_z as $z_1, z_2, \ldots, z_{|B_z|}$. Define a $|B_z|$-dimensional vector $\boldsymbol{L}_z$ as

$$(\boldsymbol{L}_z)_k = g(p,q) \Bigg(\sum_{(i,j) \in \mathcal{E}_2(z,z_k)} \mathbf{A}_{ij} - \sum_{(i,j) \in \mathcal{E}_1(z,z_k)} \mathbf{A}_{ij} \Bigg), \text{ for } k = 1, 2, \ldots, |B_z|.$$

Suppose Assumption 2.1 holds. For any $z_0, z_0' \in \mathcal{C}_0$, we have $|B_{z_0}| = |B_{z_0'}|$ and $\mathrm{Cov}(\boldsymbol{L}_{z_0})$ equals to $\mathrm{Cov}(\boldsymbol{L}_{z_0'})$ up to permutation, i.e., there existing a permutation $\tau \in S_{|B_{z_0}|}$ such that $\mathrm{Cov}(\boldsymbol{L}_{z_0})_{kl} = \mathrm{Cov}(\boldsymbol{L}_{z_0'})_{\tau(k)\tau(l)}$ for all $k, l = 1, \ldots, |B_{z_0}|$.

[2] See Lemma C.2 in the Appendix for the rigorous argument.

We refer to Appendix E.1 for the proof of Lemma 2.2. Therefore, we can avoid directly constructing B_{z^*} and choose instead any $z \in \mathcal{C}_0$ as a "shadowing assignment" to construct the shadowing statistic

$$L_0(z_0) := \sup_{z \in B_{z_0}} g(p,q)\Big(\sum_{(i,j)\in\mathcal{E}_2(z_0,z)} \mathbf{A}_{ij} - \sum_{(i,j)\in\mathcal{E}_1(z_0,z)} \mathbf{A}_{ij}\Big). \tag{2.8}$$

The following proposition shows that the quantiles of $L_0(z_0)$ and L_0 are asymptotically the same.

Proposition 2.3. Suppose Assumption 2.1 holds, $\log|B_{z^*}| = O(\log n)$ and $1/\rho_n = o(n^{1-c_2})$ for some constant $c_2 > 0$. For any $z_0 \in \mathcal{C}_0$, we have

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} |\mathbb{P}(L_0 < t) - \mathbb{P}(L_0(z_0) < t)| = 0.$$

We defer the proof to Appendix C.1. Now we are ready to present the shadowing bootstrap procedure. Based on the previous discussion, we aim to estimate the p-value based on $L_0(z_0)$. To achieve this, we take an arbitrary $z_0 \in \mathcal{C}_0$, and generate one realization of the adjacency matrix $\widehat{\mathbf{A}} \sim \mathcal{M}(n, K, \widehat{p}, \widehat{q}, z_0)$. Here $\widehat{p}$ and $\widehat{q}$ are the maximum likelihood estimator

$$(\widehat{p}, \widehat{q}) = \mathrm{argmax}_{(p,q)} \sup_{z\in\mathcal{C}_0\cup\mathcal{C}_1} f(\mathbf{A}; z, p, q) \tag{2.9}$$

The likelihood ratio statistic is

$$\widehat{\mathrm{LRT}} = \log \sup_{z\in\mathcal{C}_1} f(\mathbf{A}; z, \widehat{p}, \widehat{q}) - \log \sup_{z\in\mathcal{C}_0\cup\mathcal{C}_1} f(\mathbf{A}; z, \widehat{p}, \widehat{q}). \tag{2.10}$$

The next step is to find the assignments in B_{z_0}. We can construct B_{z_0} by Definition 2.5 in general. We will show later how to construct B_{z_0} for the concrete examples. To estimate the p-value, we apply the Gaussian multiplier bootstrap. Let $\{e_{ij}\}_{1\le i<j\le n}$ be independent standard Gaussian random variables and define

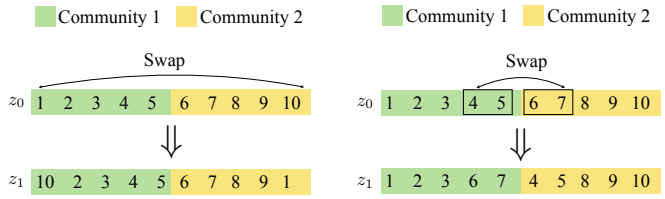
$$W_n = \sup_{z\in B_{z_0}} \sum_{1\le i<j\le n} \big(\widehat{\mathbf{A}}_{ij} - \mathbb{E}_{\widehat{p},\widehat{q}}(\widehat{\mathbf{A}}_{ij})\big)\big(\mathbb{1}[(i,j)\in\mathcal{E}_2(z_0,z)] - \mathbb{1}[(i,j)\in\mathcal{E}_1(z_0,z)]\big)e_{ij}, \tag{2.11}$$

where $\mathbb{E}_{\widehat{p},\widehat{q}}(\widehat{\mathbf{A}}_{ij}) = \widehat{p}$ if $z_0(i) = z_0(j)$ and $\mathbb{E}_{\widehat{p},\widehat{q}}(\widehat{\mathbf{A}}_{ij}) = \widehat{q}$ otherwise. Then we estimate the p-value via the multiplier bootstrap

$$p_W = \mathbb{P}\big(g(\widehat{p},\widehat{q})W_n + g(\widehat{p},\widehat{q})\widehat{\mu}_0 \ge \widehat{\mathrm{LRT}} | \widehat{\mathbf{A}}, \mathbf{A}\big), \tag{2.12}$$

where $\widehat{\mu}_0 = d(\mathcal{C}_0, \mathcal{C}_1)(\widehat{q} - \widehat{p})$ is the estimator of the mean of the process in (2.8). Finally, we reject the null $\mathrm{H}_0 : z^* \in \mathcal{C}_0$ if $p_W \le \alpha$ and do not reject H_0 otherwise.

Now we provide concrete steps to construct B_{z_0} and calculate $d(\mathcal{C}_0, \mathcal{C}_1)$ for Examples 1.1 and 1.2.



(a) Example 1.1 (b) Example 1.2

Figure 3: Procedure to construct the assignment in the boundary B_{z_0}. Panel (a) is to test whether the first 3 nodes belong to the same community, and Panel (b) is to test whether node sets $\{1, 2, 3\}$ and $\{4, 5\}$ belong to the same community.

• Example 1.1: For any $z_0 \in \mathcal{C}_0$, to construct B_{z_0}, we aim to find assignments whose distance to z_0 is $d(z_0, \mathcal{C}_1)$. The simplest way is to exchange the community assignment of one node $s \in [m]$ with another node s' from a different community (see z_1 in Figure 3(a) for an example when $n = 10$, $m = 3$, and $K = 2$). It is easy to check all such assignments belong to B_{z_0}. On the other hand, any

other operation will incur more node-wise misclassification and the edge-wise misalignment will be much larger. As for $d(\mathcal{C}_0, \mathcal{C}_1)$, we start with evaluating $n_1(z_0, z_1)$ for some $z_1 \in B_{z_0}$. The edges whose connection probability is changed from p to q will be the edges between the swapped nodes and the rest of the nodes in their original communities. Therefore, we have $n_1(z_0, z_1) = 2(n/K-1)$. Similarly, $n_2(z_0, z_1) = 2(n/K - 1)$ and thus $d(\mathcal{C}_0, \mathcal{C}_1) = 2(n/K - 1)$. In summary, B_{z_0} is composed of all the assignments obtained from swapping two nodes from different communities in z_0. The distance between two classes is $d(\mathcal{C}_0, \mathcal{C}_1) = 2(n/K - 1)$.

• Example 1.2: In this example, without loss of generality we can assume that $m' \leq m$. Then to project an arbitrary $z_0 \in \mathcal{C}_0$ onto $\mathcal{C}_1$, we will exchange the cluster assignment of the set $\mathcal{S}_{m'} = \{m+1, \ldots, m+m'\}$ with another set $\mathcal{S}'_{m'}$ from a different cluster of cardinality m' to obtain the smallest number of misaligned edges. See Figure 3(b). Correspondingly, for $z_1 \in B_{z_0}$, we have $d(z_0, z_1) = n_1(z_0, z_1) = n_2(z_0, z_1) = 2m'(n/K - m')$, and thus $d(\mathcal{C}_0, \mathcal{C}_1) = 2m'(n/K - m')$. In summary, B_{z_0} is composed of all the assignments which can be obtained from reassigning the label of nodes $m+1, \ldots, m+m'$ in z_0. The distance between two classes is $d(\mathcal{C}_0, \mathcal{C}_1) = 2m'(n/K - m')$.

3 Validity of Community Property Test

In this section, we show that our testing method is honest and powerful. Before presenting our theorems, we first give the following assumption for the alternative class $\mathcal{C}_1$.

Assumption 3.1 (Scattering of $\mathcal{C}_1$)**.** For any $z_0 \in \mathcal{C}_0$, we have $|B_{z_0}| = O(n^{c_0})$ for some constant $c_0 > 0$.

We call this assumption the *scattering assumption* as it ensures that the assignments in $\mathcal{C}_1$ are scattered and not too concentrated on the boundary. We reparameterize p, q as $p = \rho_n \lambda_1$ and $q = \rho_n \lambda_2$, where we assume λ_1 and λ_2 are constants independent of n. The following main theorem shows that our test is honest and powerful for general symmetric community properties.

Theorem 3.2. Suppose Assumptions 2.1 and 3.1 hold, $d(\mathcal{C}_0, \mathcal{C}_1) = o(n^{c_1})$ for some constant $c_1 < 2$, and $1/\rho_n = o(n^{1-c_2})$ for some constant $c_2 > 0$. We have

$$\lim_{n\to\infty} \sup_{z^* \in \mathcal{C}_0} \mathbb{P}(p_W \leq \alpha) = \alpha \text{ and } \lim_{n\to\infty} \sup_{z^* \in \mathcal{C}_0} \mathbb{P}(\text{reject H}_0) = \alpha.$$

Moreover, if $d(\mathcal{C}_0, \mathcal{C}_1) I(p, q) = \Omega(n^{\varepsilon})$ for some arbitrarily small constant $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \inf_{z^* \in \mathcal{C}_1} \mathbb{P}(\text{reject H}_0) = 1.$$

We defer the proof to Appendix C.2. One may note that under the regime of Theorem 3.2, exact recovery can be achieved. However, previous exact recovery methods cannot provide uncertainty quantification. On the other hand, our approach can provide the p-values for the general community properties which is more challenging compared to estimating the community labeling. We now apply Theorem 3.2 to Examples 1.1 and 1.2 by checking Assumption 3.1. The following proposition shows that Assumption 3.1 is satisfied for Examples 1.1 and 1.2.

Proposition 3.3. For Example 1.1, we have $|B_{z_0}| = O(mn)$ and $d(\mathcal{C}_0, \mathcal{C}_1) = 2(n/K - 1) = O(n/K)$. For Example 1.2, $|B_{z_0}| = O(K(n/K)^{m \wedge m'})$ and $d(\mathcal{C}_0, \mathcal{C}_1) = 2(m \wedge m')(n/K - m \wedge m')$.

Plugging these results to the general Theorem 3.2, we have the following corollary.

Corollary 3.4 (Examples 1.1 and 1.2)**.** Suppose $1/\rho_n = o(n^{1-c_2})$ for some constant $c_2 > 0$. If $m \leq n/K$ for Example 1.1 or $m \wedge m' = O(1)$ for Example 1.2, we have $\lim_{n\to\infty} \sup_{z^* \in \mathcal{C}_0} \mathbb{P}(\text{reject H}_0) = \alpha$. Moreover, if $I(p, q)n/K = \Omega(n^{\varepsilon})$ for some small positive constant ε, we have $\lim_{n\to\infty} \sup_{z^* \in \mathcal{C}_1} \mathbb{P}(\text{reject H}_0) = 1$.

4 Information-Theoretic Lower Bound

In this section, we discuss the information-theoretic lower bound of community property test. We will give the lower bound of the minimax risk of all possible test ψ defined as

$$r(\mathcal{C}_0, \mathcal{C}_1) = \inf_{\psi} \Big\{ \sup_{z \in \mathcal{C}_0} \mathbb{P}_z(\psi = 1) + \sup_{z \in \mathcal{C}_1} \mathbb{P}_z(\psi = 0) \Big\}.$$

We will show that the combinatorial-probabilistic trade-off phenomenon appears in the lower bound as well, thus it essentially characterizes the hardness of the community property test. In order to establish the lower bound, we first introduce the concept of packing number of community class $\mathcal{C}_1$.

Definition 4.1 (ε-packing of B_{z_0})**.** For any $z_0 \in \mathcal{C}_0$, we say $\{z_1, z_2, ..., z_N\} \subseteq B_{z_0}$ is an ε-packing of B_{z_0}, if for any $z_j \neq z_k$ we have $|\mathcal{E}_{1,2}(z_0, z_j) \cap \mathcal{E}_{1,2}(z_0, z_k)| \leq \varepsilon$. The ε-packing number of B_{z_0}, denoted as $N(B_{z_0}, \varepsilon)$, is the maximum cardinality of any ε-packing of B_{z_0}.

By Definition 2.5, B_{z_0} collects the alignments in $\mathcal{C}_1$ which are closest to $z_0 \in \mathcal{C}_0$, which are the hardest cases to test. The following theorem shows the lower bound of the community property test can be characterized by the packing number of these hardest cases.

Theorem 4.1. Suppose $\mathcal{C}_0, \mathcal{C}_1 \subseteq \mathcal{K}^n$, $1/\rho_n = o(n^{1-c_2})$ for some constant $c_2 > 0$ and $p \leq 1 - \delta$ for some constant $\delta > 0$. If there exists a $z_0 \in \mathcal{C}_0$ such that $\log N\big(B_{z_0}, \sqrt{d(z_0, \mathcal{C}_1)}\big) = O(\log n)$ and

$$\limsup_{n\to\infty} \frac{d(z_0, \mathcal{C}_1) I(p, q)}{\log N\big(B_{z_0}, \sqrt{d(z_0, \mathcal{C}_1)}\big)} < 1, \tag{4.1}$$

then $\liminf_{n\to\infty} r(\mathcal{C}_0, \mathcal{C}_1) \geq 1/2$.

We defer the proof to Appendix D.1. From (4.1), we see that the packing entropy $\log N(B_{z_0}, \sqrt{d(z_0, \mathcal{C}_1)})$ is the lower bound of the signal strength. In general, the packing entropy is $O(\log n)$. Comparing to the upper bound $d(\mathcal{C}_0, \mathcal{C}_1) I(p, q) = \Omega(n^{\varepsilon})$ for some arbitrarily small constant $\varepsilon > 0$ in Theorem 3.2, there is a gap to $O(\log n)$ in the lower bound. We conjecture that this gap exists as both our upper and lower bounds are for general community property test. We will find a finer analysis in future research. The following theorem gives an alternative lower bound result relaxing the scaling conditions in Theorem 4.1.

Theorem 4.2. Suppose $\mathcal{C}_0, \mathcal{C}_1 \subseteq \mathcal{K}^n$, $0 < q < p \leq 1 - \delta$ for some constant $\delta > 0$ and $\lim_{n\to\infty} d(\mathcal{C}_0, \mathcal{C}_1) p = \infty$. If one of the following conditions:

(1) $d(\mathcal{C}_0, \mathcal{C}_1) I(p, q) \leq c$ for some sufficiently small constant c;

(2) $\lim_{n\to\infty} d(\mathcal{C}_0, \mathcal{C}_1) I(p, q) = \infty$, but there exists a $z_0 \in \mathcal{C}_0$ such that

$$\limsup_{n\to\infty} \frac{d(z_0, \mathcal{C}_1) I(p, q)}{\log N(B_{z_0}, 0)} < 1,$$

is satisfied, then $\liminf_{n\to\infty} r(\mathcal{C}_0, \mathcal{C}_1) \geq 1/2$.

We defer the proof to Appendix D.1. Theorems 4.1 and 4.2 both show the lower bound with the combinatorial-probabilistic trade-off. Theorem 4.1 has a sharper lower bound on $d(z_0, \mathcal{C}_1) I(p, q)$ under a stronger scaling condition. In comparison, Theorem 4.2 has a less sharp lower bound with weaker scaling conditions. When $d(\mathcal{C}_0, \mathcal{C}_1) I(p, q)$ is bounded, we cannot differentiate two hypotheses. When $d(\mathcal{C}_0, \mathcal{C}_1)$ goes to infinity, Theorem 4.2 condition (2) gives the lower bound $d(\mathcal{C}_0, \mathcal{C}_1) I(p, q) < \log N(B_{z_0}, 0)$. If we have stronger scaling conditions in Theorem 4.1, we get a sharper lower bound $d(\mathcal{C}_0, \mathcal{C}_1) I(p, q) < \log N(B_{z_0}, \sqrt{d(z_0, \mathcal{C}_1)})$. Now we apply the general theorems for the lower bound to Examples 1.1 and 1.2. By (4.1) in Theorem 4.1, a key quantity for the lower bound is the packing number $N(B_{z_0}, \sqrt{d(z_0, \mathcal{C}_1)})$. The following proposition gives concrete results for the two examples.

Proposition 4.3. The packing number for Example 1.1 is $N(B_{z_0}, \sqrt{d(z_0, \mathcal{C}_1)}) = m$, and the packing number for Example 1.2 is $N(B_{z_0}, \sqrt{d(z_0, \mathcal{C}_1)}) = N(B_{z_0}, 0) = 1$.

Applying Proposition 4.3 to Theorem 4.1 and Theorem 4.2, we provide the lower bound for Examples 1.1 and 1.2 in Corollary 4.4 and Corollary 4.5 respectively.

Corollary 4.4. For $\mathcal{C}_0$ and $\mathcal{C}_1$ in (2.1), if $1/\rho_n = o(n^{1-c_2})$ for some constant $c_2 > 0$, $p < 1 - \delta$ for some constant $\delta > 0$ and $\limsup_{n\to\infty} 2nI(p, q)/(K \log m) < 1$, then $\liminf_{n\to\infty} r(\mathcal{C}_0, \mathcal{C}_1) \geq 1/2$.

Corollary 4.5. For $\mathcal{C}_0$ and $\mathcal{C}_1$ in (2.2), if $np \to \infty$, $0 < q < p \leq 1 - \delta$ for some constant $\delta > 0$ and $\limsup_{n\to\infty} nI(p, q) < c$ for some sufficiently small constant $c > 0$, then $\liminf_{n\to\infty} r(\mathcal{C}_0, \mathcal{C}_1) \geq 1/2$.

5 NUMERICAL RESULTS ON SYNTHETIC DATA

We conduct the shadowing bootstrap on Examples 1.1 and 1.3. We test both hypotheses at the significance level $\alpha = 0.05$, and consider $K = 2$ and $n = 200, 600, 1000$. The connection probabilities are set to be $p = (1+\Delta)\rho_n$ and $q = (1-\Delta)\rho_n$, where $\rho_n = (n/K)^{-0.3}$ and Δ is the parameter controlling the difference between p and q. We set Δ from 0 to 0.8. The maximum likelihood estimator is initialized by singular value decomposition estimator to boost the computation. For Example 1.1, we set $m = \lceil (n/K)^{\delta}/2 \rceil$ for $\delta = 0.3, 0.5, 0.7$ to explore the influence of m on type-I and II errors.

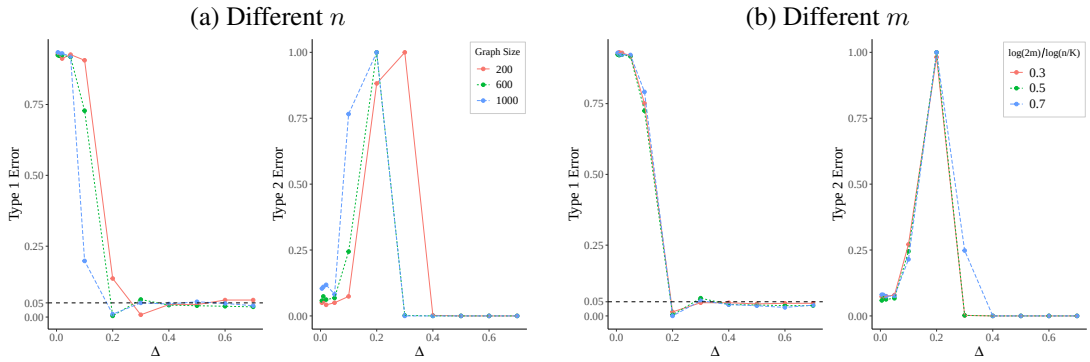


Figure 4: Type-I and type-II errors for Example 1.1. Panel (a) shows the results for different graph sizes $n = 200, 600, 1000$ with m set as $\lceil (n/K)^{0.5}/2 \rceil$. Panel (b) shows the results $m = \lceil (n/K)^{\delta}/2 \rceil$ with $\delta = 0.3, 0.5, 0.7$ and the graph size $n = 600$.

In Figure 4, we show how the type-I and type-II errors vary with Δ, n and m for Example 1.1 (via 500 Monte Carlos). As Δ increases, the type-I error converges to the nominal level 0.05, which shows that our method is honest. Type-II error is small when Δ is around zero as the test will always reject the null when the signal strength is too small, while it increases drastically as the type-I error drops to 0. When Δ is large enough, the type-II error converges to 0, showing that our test is powerful. In Figure 4(b), we can see that the type-I and type-II errors for different m's converge similarly as Δ increases. This is consistent with Corollary 3.4 as the scaling condition is irrelevant to m. Simulation results for Example 1.3 are shown in Figure 5. The type 1 and type 2 error rates vary similarly as in Example 1.1.

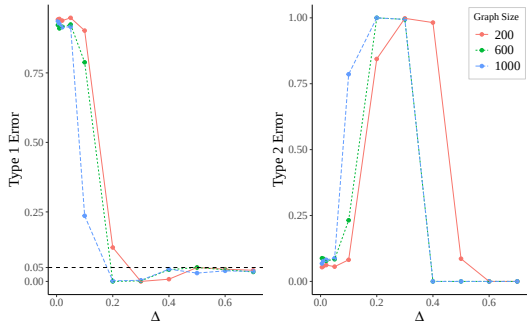


Figure 5: Type-I and type-II errors for Example 1.3 as Δ varies in $[0, 0.8]$ and the graph size $n = 200, 600, 1000$.

6 CONCLUSION

We propose a likelihood ratio test for the community property test with uncertainty quantification, implemented via a "shadowing bootstrap" method by utilizing the symmetry of $\mathcal{C}_0$ and $\mathcal{C}_1$. We show that our test is generally honest and powerful as long as $\mathcal{C}_0$ and $\mathcal{C}_1$ are symmetric. We also prove the minimax lower bound of the general community property test.

REFERENCES

Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: fundamental limits and efficient algorithms for recovery. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS 2015*, pp. 670–688. IEEE Computer Soc., Los Alamitos, CA, 2015. doi: 10.1109/FOCS.2015.47. URL https://doi.org/10.1109/FOCS.2015.47.

Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Trans. Inform. Theory*, 62(1):471–487, 2016. ISSN 0018-9448. doi: 10.1109/TIT.2015.2490670. URL https://doi.org/10.1109/TIT.2015.2490670.

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Statist.*, 48(3):1452–1474, 2020. ISSN 0090-5364. doi: 10.1214/19-AOS1854. URL https://doi.org/10.1214/19-AOS1854.

Naman Agarwal, Afonso S. Bandeira, Konstantinos Koiliaris, and Alexandra Kolla. Multisection in the stochastic block model using semidefinite programming. In *Compressed sensing and its applications*, Appl. Numer. Harmon. Anal., pp. 125–162. Birkhäuser/Springer, Cham, 2017.

Edo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 692–700. Curran Associates, Inc., 2013.

Réka Albert, Hawoong Jeong, and Albert-László Barabási. Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999. doi: 10.1038/43601. URL https://doi.org/10.1038/43601.

Afonso S. Bandeira. Random Laplacian matrices and convex relaxations. *Found. Comput. Math.*, 18(2):345–379, 2018. ISSN 1615-3375. doi: 10.1007/s10208-016-9341-9. URL https://doi.org/10.1007/s10208-016-9341-9.

Albert-László Barabási and Zoltán N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004. doi: 10.1038/nrg1272. URL https://doi.org/10.1038/nrg1272.

Peter Bickel, David Choi, Xiangyu Chang, and Hai Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.*, 41(4): 1922–1943, 2013. ISSN 0090-5364. doi: 10.1214/13-AOS1124. URL https://doi.org/10.1214/13-AOS1124.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819, 2013. ISSN 0090-5364. doi: 10.1214/13-AOS1161. URL https://doi.org/10.1214/13-AOS1161.

D. S. Choi, P. J. Wolfe, and E. M. Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284, 2012. ISSN 0006-3444. doi: 10.1093/biomet/asr053. URL https://doi.org/10.1093/biomet/asr053.

Chao Gao, Zongming Ma, Anderson Y. Zhang, and Harrison H. Zhou. Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.*, 18:Paper No. 60, 45, 2017. ISSN 1532-4435.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming: extensions. *IEEE Trans. Inform. Theory*, 62(10):5918–5937, 2016. ISSN 0018-9448. doi: 10.1109/TIT.2016.2594812. URL https://doi.org/10.1109/TIT.2016.2594812.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: first steps. *Social Networks*, 5(2):109–137, 1983. ISSN 0378-8733. doi: 10.1016/0378-8733(83)90021-7. URL https://doi.org/10.1016/0378-8733(83)90021-7.

Laurent Massoulié. Community detection thresholds and the weak Ramanujan property. In *STOC'14—Proceedings of the 2014 ACM Symposium on Theory of Computing*, pp. 694–703. ACM, New York, 2014.

Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction, 2012.

Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. *Electronic Journal of Probability*, 21(0), 2016. ISSN 1083-6489. doi: 10.1214/16-ejp4185. URL http://dx.doi.org/10.1214/16-EJP4185.

M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 45(2):167–256, 2003. ISSN 0036-1445. doi: 10.1137/S003614450342480. URL https://doi.org/10.1137/S003614450342480.

Y. X. Rachel Wang and Peter J. Bickel. Likelihood-based model selection for stochastic block models. *Ann. Statist.*, 45(2):500–528, 2017. ISSN 0090-5364. doi: 10.1214/16-AOS1457. URL https://doi.org/10.1214/16-AOS1457.

Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

Se-Young Yun and Alexandre Proutiere. Accurate community detection in the stochastic block model via spectral algorithms, 2014.

Anderson Y. Zhang and Harrison H. Zhou. Minimax rates of community detection in stochastic block models. *Ann. Statist.*, 44(5):2252–2280, 2016. ISSN 0090-5364. doi: 10.1214/15-AOS1428. URL https://doi.org/10.1214/15-AOS1428.