
Optimal community detection in dense bipartite graphs

Julien Chhor*

Department of Mathematics and Statistics
Toulouse School of Economics
Toulouse, France
julien.chhor@tse-fr.eu

Parker Knight

Department of Biostatistics
Harvard University
Boston, MA, USA
pknight@g.harvard.edu

Abstract

We consider the problem of detecting a community of densely connected vertices in a high-dimensional bipartite graph of size $n_1 \times n_2$. Under the null hypothesis, the observed graph is drawn from a bipartite Erdős-Renyi distribution with connection probability p_0 . Under the alternative hypothesis, there exists an unknown bipartite subgraph of size $k_1 \times k_2$ in which edges appear with probability $p_1 = p_0 + \delta$ for some $\delta > 0$, while all other edges outside the subgraph appear with probability p_0 . Specifically, we provide non-asymptotic upper and lower bounds on the smallest signal strength δ^* that is both necessary and sufficient to ensure the existence of a test with small enough Type I and Type II errors. We also derive novel minimax-optimal tests achieving these fundamental limits when the underlying graph is sufficiently dense. Our proposed tests involve a combination of hard-thresholded nonlinear statistics of the adjacency matrix, the analysis of which may be of independent interest. In contrast with previous work, our non-asymptotic upper and lower bounds match for any configuration of n_1, n_2, k_1, k_2 .

1 Introduction

The analysis of community structure is a fundamental task in statistical network science [27, 24]. In the 2000's, a series of striking works including (but not limited to) [3, 9, 30, 53, 56] observe the ubiquity of community-based organization in real world network data. Inspired by these observations, a vast literature has emerged over the past two decades attempting to understand the statistical limits of detecting and recovering planted communities in random graphs [20, 48, 50, 59, 4, 8, 49], detecting geometry in graphs defined via latent metric spaces [14, 13, 44, 25], and establishing fundamental computational limits for the community detection and recovery problems [31, 16, 60, 46, 12]. Here, we draw a careful distinction between *community detection* and *community recovery*. In the community detection problem, also sometimes referred to as the Planted Dense Subgraph (PDS) problem, the statistician observes a random graph and aims to determine whether the graph was drawn from an Erdős-Renyi family, or if it contains a densely connected community of vertices (i.e., a dense subgraph). In contrast, the community recovery problem involves assigning each vertex in the observed graph to a latent community, under the assumption that such community structure exists. We focus on community detection in the present work, and we refer readers to [1] for a comprehensive review of the results pertaining to the recovery problem in the stochastic block model.

Existing work studying community detection can be divided into two categories. The first camp, including [4, 59, 57, 61], focuses on detecting the presence of a small community of k nodes in a high-dimensional graph of n nodes, typically in an asymptotic setting with $k = o(n)$. In these papers, the authors focus on deriving sufficient and necessary conditions on the difference in connection

*The authors contributed equally to this work and are ordered alphabetically.

probabilities between the planted subgraph and the rest of the vertices for which detection of the subgraph is possible. While these works are able to give asymptotically precise characterizations of the detection boundary for their problems of interest, their results are limited in generality, as they constrain themselves to fully symmetric settings (the adjacency matrices of the full graph and the subgraph, with the exception of [57], are square) and make asymptotic assumptions that prohibits the appearance of complex phenomena; see for instance equation (6) in [4].

A parallel line of work considers the community detection problem in stochastic block models (SBM's) [38, 48, 51, 2, 6, 7, 8, 28]. Here the focus is on the statistician's ability detect the presence of two or more communities in the observed graph, and deriving conditions on the difference in expected degree between the communities under which detection is possible. The early work [51] gives fundamental results related to the famed Kesten-Stigum threshold, although they are confined to the restricted case of two equally sized communities. While recent papers such as [8, 52] allow for a growing number of potentially asymmetric communities, these works still required the average expected degree across communities to be held constant, which prevents their results from applying to graphs with highly imbalanced communities. The restriction is lifted in [29], in which the authors study community detection in the degree-corrected block model. However, similar to the works of [4, 59, 57], the results of [29] hold only in a particular asymptotic regime.

Given the limitations of existing work, there remains a great interest in studying community detection in random graph models beyond those covered by the literature. In particular, previous results do not apply to the family of *bipartite graphs*, which are defined as sets of edges between two disjoint sets of vertices. In practice, the two sets of vertices may be of very different size, meaning that the results tailored for square adjacency matrices [4, 59, 2, 51] do not capture the subtleties of community detection in bipartite graphs. Furthermore, any planted subgraph in a bipartite graph is, by definition, itself bipartite. While the recent paper [57] considers the detection of bipartite communities, they assume that these communities are planted in a larger non-bipartite graph and hence, similarly to [4, 59, 2, 51], fail to fully characterize the detection boundary in the bipartite case. We also comment that bipartite graphs can be thought of as a special case of a multi-community SBM; however, as outlined above, the prior work studying this problem introduces its own limitations [8, 52, 29]. As bipartite graphs are increasingly used as modeling tools in the diverse fields of biology and medicine [54], information science [47], and game theory [55], this lack of understanding in the bipartite setting constitutes a significant gap in the statistics and machine learning literature.

In this paper, we aim to address this gap. Following [4, 59], we formalize the problem of detecting a community in a random bipartite graph as a hypothesis testing problem. We observe a random bipartite graph \mathbf{G} defined on disjoint sets of vertices (nodes) \mathcal{V}_1 and \mathcal{V}_2 of size n_1 and n_2 respectively. Under the null hypothesis, an edge appears between any two nodes in \mathbf{G} with equal probability $p_0 > 0$. Under the alternative hypothesis, there exist subsets of nodes $\mathcal{V}_{K_1} \subset \mathcal{V}_1$ and $\mathcal{V}_{K_2} \subset \mathcal{V}_2$ of size k_1 and k_2 such that any pair of nodes in $\mathcal{V}_{K_1} \times \mathcal{V}_{K_2}$ are connected with an elevated probability no less than $p_0 + \delta$, where $\delta > 0$. We derive the *minimax rate of separation* δ^* [35], which is defined as the smallest value of δ such that consistent community detection is possible (see Equation (4) for a formal definition). Our results are non-asymptotic in nature, and hold for any values of n_1, n_2, k_1 , and k_2 . This degree of precision allows us to describe subtle phase transitions in the separation rate δ^* which have not been observed in the existing community detection literature.

The paper is structured as follows. In Sections 1.1, 1.2, and 1.3 we review relevant literature, outline our contributions, and collect notation used through the paper. In Section 2 we formally state our problem of interest. Our main results and a discussion thereof are given in Section 3 and Section 4. We conclude with a summary of the limitations of our results and promising directions for future work in Section 5. All of the proofs are provided in the supplement.

1.1 Prior work

The framework that we use to study optimal community detection in random graphs was pioneered by Arias-Castro and Verzelen [4, 59]. In [4], the authors provide matching asymptotic upper and lower bounds on the minimax risk of detecting a community of vanishing size in a dense graph. They also handle the cases when the size of the community and the baseline connection probability p_0 are unknown, and provide polynomial-time tests using a convex relaxation of the max degree test, following [11]. In [59], the authors extend their analysis to sparse graphs. While the papers [4, 59] broke significant ground, their results are still limited, as they only consider the restricted

fully symmetric setting of $n_1 = n_2$ and $k_1 = k_2$. Subsequent works derive analogous theoretical results for hypergraphs [62, 61], in which edges can be drawn between more than two vertices. The work [33] studies community detection under information constraints, namely that the statistician can only access small parts of the graph via non-adaptive edge queries. More recently, [57] considers the detection of a bipartite (i.e., imbalanced) community in a non-bipartite graph. We provide a detailed comparison of our results to the results of [57] in Section 3.

In the stochastic block model (SBM) literature, the community detection problem is posed as the task of distinguishing an SBM with a given set of parameters from an Erdős-Renyi graph. This problem was first considered by [38], whose results were later refined by [51, 48, 2], which established that detection is possible if and only if the signal strength is above the Kesten-Stigum threshold, as conjectured by [23]. In [6, 7], the authors prove analogous results for asymptotically growing expected degrees and provide optimal hypothesis tests based on signed cycle statistics. The work of [8] provides an impossibility result for distinguishability for SBM's with a possibly growing number of communities, and [28] analyzes a powerful test in this setting constructed using small subgraph statistics.

Under the framework of [4, 59], community detection is equivalent to detecting a submatrix in the observed adjacency matrix. The seminal work of [15] provides matching upper and lower bounds for detecting a sparse submatrix of elevated mean in a matrix of Gaussian random variables under a particular asymptotic regime. Following this, [46] initiates a rigorous study of statistical-computational gaps in the submatrix detection problem which was continued by [16, 32, 12]. In [45], the authors study the problem of detecting a planted sparse sub-tensor in a high dimensional Gaussian tensor. Finally, [22] derives upper and lower bounds for detecting the presence of multiple sparse submatrices in Gaussian noise that are tight up to log factors.

1.2 Our contributions

In this paper, we make the following contributions.

1. We fully settle the non-asymptotic expression of the minimax rate δ^* defined in (4), which holds provided the graph is sufficiently dense, as specified in Assumption 1. In contrast to the previous literature, our bounds always match up to multiplicative constants for any possible values of n_1, n_2, k_1 and k_2 , especially in the under-explored unbalanced regimes where $k_1 \ll k_2$ or $n_1 \ll n_2$. Our results reveal subtle phase transitions that, to the best of our knowledge, had not been documented in previous work.
2. Our lower bound on δ^* stated in Theorem 1 holds without requiring Assumption 1 and uses a very precise application of the second moment method [42]. It requires several non-trivial lemmas for controlling the moment-generating function of the product of binomial random variables, and substantially departs from the lower bound strategies proposed in [15] and [4].
3. Our upper bound stated in Theorem 2 is achieved by carefully combining three tests: A standard *total degree* test and two entirely novel tests referred to as the *truncated degree* test and the *max truncated degree* test. The two novel tests are defined using truncated non-linear functions of the adjacency matrix, building on the truncated χ^2 test recently studied in the Gaussian sequence model [21, 40, 43, 19]. However, substantial modifications of the truncated χ^2 test are needed to address two challenges:
 - (a) Moving from the vector case to the matrix case, which requires subtle Bonferroni corrections of the classical truncated χ^2 test.
 - (b) Handling matrices whose entries are Bernoulli rather than Gaussian random variables, which require significant adjustments of the truncated χ^2 test to carefully manage the complex sub-poissonian tails of the binomial distribution.

Our upper bound holds under Assumption 1, which we discuss thoroughly in later sections. We also prove a series of lemmas in the supplement that we use to control the Type I and Type II errors of these tests, and we anticipate that these results will be of interest to researchers studying related problems.

1.3 Notation

The following notation will be used throughout the paper. For $p \in \mathbb{N}$, let $[p] := 1, \dots, p$. We use $\mathcal{P}_k(n)$ to denote the set of subsets of $[n]$ of size k . For $a, b \in \mathbb{R}$, denote $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. We will use $a \lesssim b$ if there exists a constant $C > 0$ depending on η such that $a \leq Cb$. We say $a \asymp b$ if $a \lesssim b$ and $b \lesssim a$. For two sets A_1 and A_2 , we denote $A_1 \times A_2 = \{(i, j) : i \in A_1, j \in A_2\}$ as the Cartesian product of A_1 and A_2 . For a finite set A , we use $|A|$ to denote the cardinality of A . We use $\mathbf{1}_{\{\cdot\}}$ as the indicator function, meaning that $\mathbf{1}_A = 1$ if the event A occurs and $\mathbf{1}_A = 0$ otherwise. For a matrix \mathbf{X} , we use X_{ij} to denote its (i, j) th entry. Given two probability distributions \mathbb{P} and \mathbb{Q} , we use $\text{TV}(\mathbb{P}, \mathbb{Q}) = \sup_A |\mathbb{P}(A) - \mathbb{Q}(A)|$ to denote the total variation distance between \mathbb{P} and \mathbb{Q} . We use $c, C, \bar{C}, C', c_1, c_2, c_3$, and c_4 to denote constants whose value may change between instances of their usage in the paper. We will also use the convention that $[a, b] = \emptyset$ for any two real numbers a, b such that $a > b$. For any two integers $n, k \in \mathbb{N} \cup \{0\}$, we denote by $\binom{n}{k}$ the binomial coefficient, equal to $\frac{n!}{k!(n-k)!}$.

2 Problem statement

Let $n_1, n_2, k_1, k_2 \in \mathbb{N}$ be integers considered as fixed throughout the paper, and assume $k_1 \leq n_1$ and $k_2 \leq n_2$. Let $\mathbf{G} = (\mathcal{E}, \mathcal{V}_1, \mathcal{V}_2)$ be a random undirected bipartite graph defined on disjoint sets of vertices $\mathcal{V}_1 = \{V_{1,1}, \dots, V_{1,n_1}\}$ and $\mathcal{V}_2 = \{V_{2,1}, \dots, V_{2,n_2}\}$. We observe the adjacency matrix $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ with entries $A_{ij} = \mathbf{1}\{V_{1,i} \text{ is connected to } V_{2,j} \text{ by an edge}\}$ and assume that the edges are mutually independent Bernoulli random variables. For some parameter $p_0 \in [0, 1]$ assumed to be fixed throughout the paper, our objective is to determine whether each pair of vertices in $\mathcal{V}_1 \times \mathcal{V}_2$ is connected with probability p_0 , or if there exists a community (i.e., a bipartite subset of vertices) that is connected with greater probability $p_1 \geq p_0 + \delta$ for some $\delta > 0$.

Formally, we formulate this testing problem in terms of the mean structure of the adjacency matrix \mathbf{A} . Define the matrix $\mathbf{P} = \mathbb{E}[\mathbf{A}] \in \mathbb{R}^{n_1 \times n_2}$, which represents the matrix of connection probabilities of the random graph \mathbf{G} . For any two subsets $K_1 \in \mathcal{P}_{k_1}(n_1)$, $K_2 \in \mathcal{P}_{k_2}(n_2)$ and any $\delta \geq 0$, we define

$$\Theta(K_1, K_2, \delta) = \left\{ \mathbf{P} \in \mathbb{R}^{n_1 \times n_2} \text{ s.t. } \begin{cases} \forall (i, j) \in K_1 \times K_2 : P_{ij} \geq p_0 + \delta \\ \forall (i, j) \notin K_1 \times K_2 : P_{ij} = p_0 \end{cases} \right\}.$$

A mean matrix \mathbf{P} belongs to $\Theta(K_1, K_2, \delta)$ if the bipartite community indexed by $K_1 \times K_2$ is more densely connected than the rest of the graph. Our focus in this paper is to optimally detect an unknown bipartite community of size $k_1 \times k_2$. To this end, for any $\delta > 0$, we let

$$\Theta(k_1, k_2, n_1, n_2, \delta) = \bigcup_{K_1 \in \mathcal{P}_{k_1}(n_1)} \bigcup_{K_2 \in \mathcal{P}_{k_2}(n_2)} \Theta(K_1, K_2, \delta).$$

We consider the following testing problem

$$H_0 : \forall (i, j) \in [n_1] \times [n_2], P_{ij} = p_0 \quad \text{against} \quad H_1 : \mathbf{P} \in \Theta(k_1, k_2, n_1, n_2, \delta). \quad (1)$$

The hypothesis H_0 is equivalent to observing a bipartite Erdős-Renyi random graph with parameter p_0 . The hypothesis H_1 is equivalent to the existence of an unknown bipartite community of size $k_1 \times k_2$ with connection probabilities at least $p_0 + \delta$.

A *test* is a measurable function of the observed data \mathbf{A} taking its values in $\{0, 1\}$. We measure the quality of a test Δ by its *risk*, defined as the sum of its Type I and worst-case Type II errors. Specifically, denoting by $\mathbb{P}_{\mathbf{P}}$ the probability distribution of \mathbf{A} for $\mathbf{P} \in \Theta(k_1, k_2, n_1, n_2, \delta)$, and letting \mathbb{P}_0 denote the distribution of \mathbf{A} under the null hypothesis, the risk of a test Δ is defined as

$$\mathcal{R}(\Delta, \delta) = \mathbb{P}_0(\Delta = 1) + \sup_{\mathbf{P} \in \Theta(k_1, k_2, n_1, n_2, \delta)} \mathbb{P}_{\mathbf{P}}(\Delta = 0). \quad (2)$$

The *minimax risk* associated with the problem (1) is defined as

$$\mathcal{R}^*(k_1, k_2, n_1, n_2, \delta) = \inf_{\Delta} \mathcal{R}(\Delta, \delta), \quad (3)$$

where the infimum is taken over all tests Δ . For a desired level of risk $\eta \in (0, 1)$, considered as a fixed constant throughout, we are interested in the *minimax separation rate* δ^* defined as

$$\delta^* = \inf \{ \delta > 0 : \mathcal{R}^*(k_1, k_2, n_1, n_2, \delta) \leq \eta \}. \quad (4)$$

The minimax separation rate δ^* encodes the difficulty of the testing problem. It is the infimal signal strength ensuring the existence of a test with Type I plus Type II errors controlled by the desired level of risk η . The goal of the paper is to derive the value of δ^* in terms of n_1, n_2, k_1, k_2 and p_0 up to absolute multiplicative constants, and to construct the minimax-optimal tests achieving a risk at most η for the testing problem (1) when the separation satisfies $\delta \geq C\delta^*$ for some constant $C > 0$ depending only on η .

3 Main results

For any $k_1, k_2, n_1, n_2 \in \mathbb{N}$ and for a constant $C > 0$ to be chosen later, we define

$$\psi(k_1, k_2, n_1, n_2) = \frac{1}{k_1} \log \left(1 + \frac{n_2}{k_2^2} \log \left(e \binom{n_1}{k_1} \right) \right) \quad (5)$$

$$\phi(k_1, k_2, n_1, n_2) = \begin{cases} \frac{n_1}{k_1^2} \log \left(1 + \frac{n_2}{k_2^2} \right) & \text{if } \frac{n_1}{k_1^2} \leq C \\ \infty & \text{otherwise.} \end{cases} \quad (6)$$

$$\nu(k_1, k_2, n_1, n_2) = \frac{1}{k_1} \log \left(\frac{n_2}{k_2} \right) \mathbf{1}_{\left\{ \frac{n_1 k_2}{k_1^2} \log \left(\frac{n_2}{k_2} \right) > 1 \right\}} \quad (7)$$

To alleviate the notation, we will write

$$\begin{aligned} \phi_{12} &= \phi(k_1, k_2, n_1, n_2) & \phi_{21} &= \phi(k_2, k_1, n_2, n_1) \\ \psi_{12} &= \psi(k_1, k_2, n_1, n_2) & \psi_{21} &= \psi(k_2, k_1, n_2, n_1) \\ \nu_{12} &= \nu(k_1, k_2, n_1, n_2) & \nu_{21} &= \nu(k_2, k_1, n_2, n_1). \end{aligned}$$

Finally, we define the quantities

$$R := R(k_1, k_2, n_1, n_2) = (\psi_{12} + \psi_{21}) \wedge \phi_{12} \wedge \phi_{21} \quad (8)$$

and

$$\tilde{R} = (\psi_{12} + \nu_{21}) \wedge (\psi_{21} + \nu_{12}) \wedge \phi_{12} \wedge \phi_{21}.$$

We will prove that the minimax separation rate δ^* satisfies

$$(\delta^*)^2 \asymp p_0(1 - p_0)R,$$

for any values of $k_1 k_2, n_1$ and n_2 , under Assumption 1 on p_0 given in Section 3.2.

3.1 Lower bound

The following theorem gives a lower bound on δ^* that holds for any values of n_1, n_2, k_1, k_2 and p_0 .

Theorem 1 *Let $\eta \in [0, 1]$ be given. There exist constants $\underline{c}, \bar{C} > 0$ that depend on η only, such that if $\delta^2 \leq \underline{c}p_0(1 - p_0)R$ with (6) defined with $C = \bar{C}$, then it holds*

$$\mathcal{R}^*(k_1, k_2, n_1, n_2, \delta) > \eta.$$

Theorem 1 yields the lower bound $(\delta^*)^2 \geq \underline{c}_\delta p_0(1 - p_0)R$ by definition of δ^* . The proof of Theorem 1 uses a careful application of the second moment argument [42, 37, 34, 36, 15, 4]. We refer the reader to the supplement for details.

3.2 Upper bound

Now we present a matching upper bound on the minimax rate of separation δ^* . We do so by carefully combining three testing procedures, which are each constructed from the observed adjacency matrix \mathbf{A} . Our three hypothesis tests are defined as follows.

1. *Total degree test:* Let $\sigma_{\text{deg}} = \sqrt{n_1 n_2 p_0(1 - p_0)}$ and define the test statistic

$$t_{\text{deg}} = \frac{1}{\sigma_{\text{deg}}} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (A_{ij} - p_0) \quad (9)$$

The total degree test is defined as $\Delta_{\text{deg}}^h = \mathbf{1}(t_{\text{deg}} > h)$ for a choice of threshold $h > 0$.

2. *Truncated degree test:* To define this test, we introduce the Bennett function, defined as

$$h_B(x) = (1+x)\log(1+x) - x, \quad \forall x > -1. \quad (10)$$

$$h_B(-1) = 1. \quad (11)$$

For any $j \in [n_2]$, and $a \geq 1$, let $\bar{\sigma} = \sqrt{n_1 p_0 (1-p_0)}$ and define

$$\begin{aligned} \bar{A}_j &= \frac{1}{\bar{\sigma}} \sum_{i=1}^{n_1} (A_{ij} - p_0) \\ W_j &= n_1(1-p_0) h_B\left(-\frac{\bar{\sigma} \bar{A}_j}{n_1(1-p_0)}\right) + n_1 p_0 h_B\left(\frac{\bar{\sigma} \bar{A}_j}{n_1 p_0}\right). \end{aligned} \quad (12)$$

Letting $\nu_a^{n_1} = \mathbb{E}_0 [W_1 | \bar{A}_1 \geq a]$ and $\tau > 0$ denote a parameter to be chosen later, define

$$t_{\text{trunc-deg},1} = \sum_{j=1}^{n_2} (W_j - \nu_\tau^{n_1}) \mathbf{1}(\bar{A}_j \geq \tau). \quad (13)$$

The truncated degree test is as $\Delta_{\text{trunc-deg},1}^h = \mathbf{1}(t_{\text{trunc-deg},1} > h)$ for some threshold $h > 0$.

3. *Max truncated degree test:* For any $J_1 \in \mathcal{P}_{k_1}(n_1)$ and any $j \in [n_2]$, we let $\sigma_{J_1,j} = \sqrt{k_1 p_0 (1-p_0)}$ and define

$$\begin{aligned} t_{J_1,j} &= \frac{1}{\sigma_{J_1,j}} \sum_{i \in J_1} (A_{ij} - p_0) \\ W_{J_1,j} &= k_1(1-p_0) h_B\left(-\frac{\sigma_{J_1,j} t_{J_1,j}}{k_1(1-p_0)}\right) + k_1 p_0 h_B\left(\frac{\sigma_{J_1,j} t_{J_1,j}}{k_1 p_0}\right). \end{aligned} \quad (14)$$

Letting $\nu_a^{k_1} = \mathbb{E}_0 [W_{J_1} | t_{J_1,j} \geq a]$ and $\tau > 0$ be a parameter to be chosen later, define

$$t_{\text{max-trunc-deg},1} = \max \left\{ \sum_{j=1}^{n_2} (W_{J_1,j} - \nu_\tau^{k_1}) \mathbf{1}(t_{J_1,j} > \tau) \mid J_1 \in \mathcal{P}_{k_1}(n_1) \right\}. \quad (15)$$

Finally, for a choice of threshold $h > 0$, we define the maximum truncated degree test as $\Delta_{\text{max-trunc-deg},1}^h = \mathbf{1}(t_{\text{max-trunc-deg},1} > h)$.

Analogously, we define the tests $\Delta_{\text{max-trunc-deg},2}^h$ and $\Delta_{\text{trunc-deg},2}^h$ by swapping the roles of k_1, k_2 , and n_1, n_2 in the definitions of $\Delta_{\text{max-trunc-deg},1}^h$ and $\Delta_{\text{trunc-deg},1}^h$ respectively.

For some constant $c_1 > 0$ depending only on the desired level of risk η , we further combine the degree and truncated degree tests as follows. Let

$$\Delta_a^{h_1, h_2} = \begin{cases} \Delta_{\text{trunc-deg},1}^{h_1} & \text{if } \frac{n_2}{k_2^2} \geq c_1, \\ \Delta_{\text{deg}}^{h_2} & \text{otherwise} \end{cases} \quad \text{and} \quad \Delta_b^{h'_1, h'_2} = \begin{cases} \Delta_{\text{trunc-deg},2}^{h'_1} & \text{if } \frac{n_1}{k_1^2} \geq c_1, \\ \Delta_{\text{deg}}^{h'_2} & \text{otherwise.} \end{cases}$$

Using these procedures as our building blocks, we construct our final test Δ^* by applying the relevant test depending on what part dominates in the expression of the rate R . Specifically, our optimal test is defined as

$$\Delta^* = \begin{cases} \Delta_{\text{max-trunc-deg},1}^{h_3} & \text{if } \tilde{R} = \psi_{12} + \nu_{21} \\ \Delta_{\text{max-trunc-deg},2}^{h_3} & \text{if } \tilde{R} = \psi_{21} + \nu_{12} \\ \Delta_a^{h_1, h_2} & \text{if } \tilde{R} = \phi_{12} \\ \Delta_b^{h'_1, h'_2} & \text{if } \tilde{R} = \phi_{21}. \end{cases}$$

Our upper bound on the risk of Δ^* holds under the following assumption, which guarantees that the observed graph \mathbf{G} is sufficiently dense.

Assumption 1 (Graph density.) *There exists a constant $C_\eta > 0$ depending only on η such that the baseline connection probability p_0 satisfies $p_0 \leq 1/4$ and*

$$p_0 \geq \begin{cases} \frac{C_\eta}{k_1 k_2} \log \left(e \binom{n_1}{k_1} \binom{n_2}{k_2} \right) & \text{if } \tilde{R} = (\psi_{12} + \nu_{21}) \wedge (\psi_{21} + \nu_{12}) \\ \frac{C_\eta}{n_1} \log \left(1 + \frac{n_2}{k_2^2} \right) & \text{if } \tilde{R} = \phi_{12} \text{ and } n_2 > k_2^2 \\ \frac{C_\eta}{n_2} \log \left(1 + \frac{n_1}{k_1^2} \right) & \text{if } \tilde{R} = \phi_{21} \text{ and } n_1 > k_1^2 \\ \frac{C_\eta}{n_1 n_2} & \text{otherwise.} \end{cases}$$

The upper bound $p_0 \leq \frac{1}{4}$ in Assumption 1 is used to apply Slud's inequality for binomial anti-concentration [58]. We invoke the different cases of Assumption 1 depending on the exact construction of Δ^* . Assumption 1 places us in a quasi-normal moderate deviation regime, which enables us to control the tails of the test statistics (9), (13), and (15) with sufficient precision. These tests are based on binomial statistics, which exhibit sub-Gaussian concentration properties in the moderate deviation regime and sub-poissonian concentration in the large deviation regimes (see, e.g., [5, 17, 39, 18]). This dichotomy has also been noted in the two companion papers [4] and [59], which split the analysis between the case of *dense* graphs, where a quasi-normal regime emerges (see the first row of Table 1 in [4]), and *sparse* graphs, where fundamentally different behaviors occur due to lower connectivity. In this paper, we follow the same approach by restricting to dense graphs as specified in Assumption 1, and leave the case of sparse graphs for future work. Furthermore, Assumption 1 has a natural probabilistic interpretation, as codified in the following propositions.

Proposition 1 *Let $c > 0$. There exists a constant $C > 0$ such that if*

$$\frac{C}{k_1 k_2} \log \left(e \binom{n_1}{k_1} \binom{n_2}{k_2} \right) \leq p_0 \leq \frac{1}{4},$$

then it holds $\mathbb{P}_0(\mathbf{G} \text{ has an empty } k_1 \times k_2 \text{ bipartite subgraph}) \leq c$.

Here, we say that a subgraph \mathbf{G}_1 of \mathbf{G} is empty if all the vertices in \mathbf{G}_1 has degree equal to 0.

Proposition 2 *Let $c > 0$. There exists a constant $C > 0$ such that if*

$$\frac{C}{n_2 k_1} \log \left(e \binom{n_1}{k_1} \right) \leq p_0 \leq \frac{1}{4},$$

then it holds $\mathbb{P}_0(\exists I \in \mathcal{P}_{k_1}(n_1) : \sum_{i \in I} \sum_{j=1}^{n_2} A_{ij} = 0) \leq c$.

Propositions 1 and 2 shows that Assumption 1 prevents the occurrence of extreme events under \mathbb{P}_0 that would otherwise inflate the Type I error of our testing procedures, especially the max truncated degree test in (15). We remark that the lower bound on p_0 assumed in Proposition 2 is stronger than what we need in Assumption 1; we include this result to provide intuition.

The following theorem guarantees that the minimax risk of Δ^* is small whenever δ is large enough.

Theorem 2 *Let $\eta \in [0, 1]$ be given. Suppose that Assumption 1 holds. Then there exist constants $C_\delta, \bar{C} > 0$ and thresholds $h_1, h_2, h_3, h_4 > 0$ such that if $\delta^2 \geq C_\delta p_0 (1 - p_0) R$ with (6) defined with $C = \bar{C}$, then the test Δ^* satisfies*

$$\mathcal{R}(\Delta^*, \delta) < \eta.$$

Theorem 2 demonstrates that our proposed test Δ^* is able to match the lower bound on δ^* given in Theorem 1 under Assumption 1. Combining Theorems 1 and 2, we have therefore identified the minimax rate of separation δ^* up to constants.

4 Discussion

4.1 Discussion on the proposed tests

Our test Δ^* is a careful combination of three testing procedures. The total degree test has been commonly applied in previous works [4, 57], along with the *scan test*, which consists of scanning

over all possible subgraphs of size $k_1 \times k_2$ and rejecting if one of them contains an unusually large number of edges. Notably, our results do not make use of the scan test, and highlight that it can always be successfully replaced by a max truncated degree test, with lower time complexity².

Our two novel tests $\Delta_{\text{trunc-deg},1}^h$ and $\Delta_{\text{max-trunc-deg},1}^{h_3}$ build on the truncated χ^2 test recently developed in the Gaussian sequence model. Given a vector $X \in \mathbb{R}^d$, the truncated χ^2 test statistic is given by

$$T = \sum_{j=1}^d (X_j^2 - \nu_a) \mathbf{1}(|X_j| \geq a)$$

for some $a > 0$ and where $\nu_a = \mathbb{E}(Z^2 | |Z| \geq a)$. On top of being fast to compute, this test has been shown to be optimal in the Gaussian sequence model for detecting s -sparse alternatives separated in ℓ_2 norm when $s \leq \sqrt{d}$ [21], whereas the max test, which rejects for large enough values of $\max \{ \sum_{j \in J} X_j^2 | J \in \mathcal{P}_s(d) \}$, is known to be statistically suboptimal [10] while exhibiting exponential time complexity.

We aim to adapt this test to the matrix case where $X \in \mathbb{R}^{d_1 \times d_2}$. For the sake of discussion, we will focus on the truncated degree test (13), the ideas underlying the max truncated degree test (15) being analogous. An idea for adapting the truncated χ^2 test would be to use

$$\sum_{i=1}^{n_1} \left(\left(\sum_{j=1}^{n_2} X_{ij} \right)^2 - \nu_a \right) \mathbf{1} \left(\sum_{j=1}^{n_2} X_{ij} \geq a \right)$$

for some suitable re-centering parameters ν_a and ν'_a . Whereas squares of Gaussian random variables concentrate sub-exponentially (see Lemma 5 in [43]), squares of binomial distributions do not exhibit such favorable concentration properties, due to their sub-poissonian tails. In the definition of the test statistics (12) and (14), we therefore replace the square function with a carefully chosen function with slower growth. Focusing on the truncated degree test for illustration, and recalling (12), we can check that

$$W_j = n_1(1-p_0) h_B \left(-\frac{\bar{\sigma} \bar{A}_j}{n_1(1-p_0)} \right) + n_1 p_0 h_B \left(\frac{\bar{\sigma} \bar{A}_j}{n_1 p_0} \right) \asymp \bar{\sigma} \bar{A}_j \log \left(1 + \frac{\bar{A}_j}{\bar{\sigma}} \right).$$

When $\bar{A}_j \leq \bar{\sigma}$, which exactly corresponds to the moderate deviation regime for binomial distributions, we obtain $W_j \asymp \bar{A}_j^2$ and our test statistic $t_{\text{trunc-deg},1} = \sum_{j=1}^{n_2} (W_j - \nu_{\tau}^{n_1}) \mathbf{1}(\bar{A}_j \geq \tau)$ becomes analogous to the truncated χ^2 test described above. In the large deviation regime where $\bar{A}_j \gg \bar{\sigma}$, we have $W_j \asymp \bar{\sigma} \bar{A}_j \log(\frac{\bar{A}_j}{\bar{\sigma}})$, which grows less fast than \bar{A}_j^2 and mitigates the extreme values of the binomial distribution.

Heuristically, the truncated degree test is optimal when one sub-graph of the elevated community is large and the other is small. The max-truncated degree test is optimal when both sub-graphs are small. To demonstrate this: Suppose that one of the communities is very large, for example $k_2 = n_2$ to fix ideas, and the other one is small ($k_1 \ll n_1$). In the corresponding adjacency matrix, all entries within a row are i.i.d. Bernoulli random variables with the same probability parameter. It is well-known that a sufficient statistic of m i.i.d. data points $X_1, \dots, X_m \sim \text{Bernoulli}(p)$ is the sum $\sum_{i=1}^m X_i \sim \text{Binomial}(m, p)$. Therefore, the column vector obtained as the row-wise sum of the adjacency matrix is a sufficient statistic of the data. This vector is simply the vector of node degrees in the first sub-graph of the bipartite graph. Therefore, when one of the two sub-graphs is large enough, the problem can be reduced to a vector-based problem. A classical test statistic in such cases is the truncated chi-square test, which we apply to the degree vector with a slight refinement to account for the sub-Poissonian tails of the binomial distribution. This yields the truncated degree test.

When both sub-graphs of the elevated community are small, collapsing the matrix into a single vector (by summing over rows or columns) discards too much information from the data. In this case, a natural idea would rather be to scan over all possible bipartite subgraphs of size $k_1 \times k_2$ and reject the null hypothesis when one such subgraph contains an unusually large number of edges—an approach proposed in [57]. Unfortunately, this procedure is not optimal. Instead, we propose a non-trivial refinement that builds on the truncated chi-square test, adapted to tackle two key challenges: 1) the

²Computing $\Delta_{\text{max-trunc-deg},1}^{h_3}$ requires $O(k_1 n_2 \binom{n_1}{k_1})$ operations, rather than $O(k_1 k_2 \binom{n_1}{k_1} \binom{n_2}{k_2})$ for the scan test.

data are matrix-valued, not a vector-valued, and 2) the entries are Bernoulli rather than Gaussian random variables. Our procedure, the max-truncated degree test as formally defined in equation (14), is obtained by scanning and summing over subsets along one dimension of the matrix, applying a non-linear transformation based on the Bennett function, and truncating along the other dimension – analogous to the truncated chi-square test statistic. This procedure captures subtle concentration effects of bipartite graphs when both communities are small and achieves the minimax optimal risk in this regime.

4.2 Comparison with existing results

Through Theorems 1 and 2, we have shown

$$(\delta^*)^2 \asymp p_0(1-p_0)R, \quad (16)$$

with R defined in (8). We can understand the nuances of (16) by first considering our results when restricted to the fully balanced (effectively non-bipartite) setting. Suppose that $k_1 = k_2 = k$ and $n_1 = n_2 = n$, and that $n \leq Ck^2$. From the definition of R , we can rewrite (16) as

$$\frac{(\delta^*)^2}{p_0(1-p_0)} \asymp \frac{1}{k} \log\left(\frac{n}{k}\right) \wedge \frac{n}{k^2} \log\left(1 + \frac{n}{k^2}\right). \quad (17)$$

If $\frac{n}{k^2} \gg 1$ or $\frac{n}{k^2} \ll 1$, the form of δ^* in (17) matches that given by Theorem 2 of [4] in the moderate deviation regime imposed by Assumption 1. Our result is, in fact, more refined than that of [4], as we are able to precisely capture the phase transition around $\frac{n}{k^2} \asymp 1$ thanks to our truncated degree test.

The detection boundary for imbalanced bipartite graphs is more complex. The following proposition describes a new phase transition that emerges in the imbalanced bipartite setting, but not in the non-bipartite setting.

Proposition 3 *Suppose that $k_1^2 \geq \bar{c}n_1k_2$ for a constant $\bar{c} > 0$ and $k_j \leq c_jn_j$ for $j \in \{1, 2\}$ where $c_1, c_2 > 0$ are sufficiently small constants. Additionally, suppose that there exists a constant $\alpha > 0$ such that $n_2 \geq k_2^{2+\alpha}$ and that $\frac{n_1}{k_1} \geq e \log\left(\frac{n_2}{k_2}\right)$. Then it holds*

$$\frac{(\delta^*)^2}{p_0(1-p_0)} \asymp \frac{1}{k_2} \log\left(1 + \frac{n_1k_2}{k_1^2} \log(n_2)\right). \quad (18)$$

In particular, this reveals a phase transition at $\frac{n_1k_2}{k_1^2} \log(n_2) \asymp 1$.

To our knowledge, (18) is described nowhere in the existing community detection literature, as previous works either impose restrictions on the shape of the observed graph and the community of interest (i.e. enforcing $n_1 = n_2$ and/or $k_1 = k_2$) or provide results that are not sharp for all configurations of n_1, n_2, k_1 , and k_2 . The recent paper of [57] presents the results that are most similar to ours in the literature. They consider the case $n_1 = n_2 = n$ with $k_1 \neq k_2$ in the asymptotic regime $k_1 + k_2 = o(n)$. We again emphasize that our results hold for $n_1 \neq n_2$ and any k_1 and k_2 , and as such are immediately more general than those of [57]. However, we argue that our results are more precise even in the restricted case of $n_1 = n_2 = n$ and $k_1 + k_2 = o(n)$. Under the assumption that $p_0, \delta^* = \Theta(n^{-\alpha})$ for $\alpha \in (0, 2]$, in Theorem 1 of [57] the authors show that

$$\frac{(\delta^*)^2}{p_0(1-p_0)} = \Omega\left(\frac{1}{k_1 \wedge k_2} \wedge \frac{n^2}{k_1^2 k_2^2} \wedge \frac{n}{k_1^2 \vee k_2^2}\right),$$

and in Theorem 2, they show

$$\frac{(\delta^*)^2}{p_0(1-p_0)} = O\left(\frac{\log(n)}{k_1 \wedge k_2} \wedge \frac{n^2}{k_1^2 k_2^2} \wedge \frac{n \log(n)}{k_1^2 \vee k_2^2}\right).$$

We refer readers to equations (14) and (15) in [57] and the surrounding discussion for details. Their lower bound on δ^* is loose by a logarithmic factor in n , and the asymptotic assumption $\delta^* = \Theta(n^{-\alpha})$ excludes the regime in which the complex rate that we derive in (16) emerges. As such, the results of [57] do not describe, for instance, the phase transition (18). We also comment on the assumptions on p_0 made in [57] and in our work. Recall that the results of [57] require $p_0 \gtrsim n^{-\alpha}$ for $\alpha \in (0, 2]$, where under our Assumption 1, we require at least $p_0 \gtrsim \frac{1}{n^2}$ and, in certain regimes, we need

$p_0 \gtrsim \frac{1}{k_1 k_2} \log\left(\binom{n}{k_1} \binom{n}{k_2}\right)$. This assumption is stronger than that of [57], and hence excludes our results from holding for all values of p_0 considered by [57]. However, as their results are not sharp, they are not able to describe the entirety of the detection boundary in the case when our Assumption 1 does not hold. We anticipate that the minimax rate of separation δ^* may actually differ non-trivially from the rate derived in [57] in this setting, and we view the extension of our results beyond Assumption 1 as a crucial piece of future work. Finally, we remark that [26] extends the results of [57] beyond planted bipartite subgraphs, but they do so in the same asymptotic setting and thus suffer from the same limitations.

5 Conclusion, limitations, and open problems

We have presented a rigorous study of the community detection problem in bipartite graphs. We present a lower bound on the minimax rate of separation, and describe a novel optimal testing strategy that achieves the minimax rate when the observed graph is dense. Our non-asymptotic results hold for all possible dimensions of the observed graph and the planted subgraph, which reveals new behavior in the detection boundary. Here, we outline some limitations and opportunities for future work.

Adaptivity to unknown p_0 , k_1 , and k_2 . All three of our tests require knowledge of p_0 , the connection probability under the null hypothesis, which is typically unavailable in practice. In [4], the authors consider this problem in the case of non-bipartite graphs, and find that community detection without knowledge of p_0 exhibits substantially different behavior. We anticipate that similar phenomena will arise in the bipartite setting.

Furthermore, our max truncated degree test requires knowledge of the size of the community as encoded in k_1 and k_2 . Optimal adaptivity to unknown k_1 and k_2 is achieved in previous works such as [15, 4] by scanning over possible values of k_1 and k_2 and performing a Bonferroni correction. However, this is not an option for us, since the tails of the test statistic (15) are too heavy to still be optimal after a union bound. One possible strategy is to use Lepski-style adaption as in [41]; we leave this to future work.

Beyond dense bipartite graphs. The tightness of our upper bound in Theorem 2 relies on Assumption 1. Prior work in the non-bipartite community detection literature [59] suggests that the detection landscape is very different for sparse graphs. Understanding optimal detection in bipartite graphs when p_0 is extremely small will potentially require testing procedures and new techniques for proving lower bounds, likely based on the truncated second moment method [35, 15, 4, 59].

Computational considerations. Our max truncated degree test is unfeasible to compute on datasets of even moderate size. It is well established that the community detection problem in non-bipartite graphs exhibits a statistical-computational gap [46, 4, 31, 12, 22, 57, 26], meaning that it is not, in general, possible to optimally detect a planted community with a test that runs in polynomial time. We conjecture that this gap persists in the bipartite case, and establishing this phenomenon formally is an open problem of great interest.

References

- [1] Emmanuel Abbe. “Community detection and stochastic block models: recent developments”. In: *Journal of Machine Learning Research* 18.177 (2018), pp. 1–86.
- [2] 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*. IEEE. 2015, pp. 670–688.
- [3] Réka Albert and Albert-László Barabási. “Statistical mechanics of complex networks”. In: *Reviews of modern physics* 74.1 (2002), p. 47.
- [4] Ery Arias-Castro and Nicolas Verzelen. “Community detection in dense random networks”. In: (2014).
- [5] Ery Arias-Castro and Meng Wang. “The sparse Poisson means model”. In: (2015).
- [6] Debapratim Banerjee. “Contiguity and non-reconstruction results for planted partition models: the dense case”. In: *Electronic Journal of Probability* (2018).
- [7] Debapratim Banerjee and Zongming Ma. “Optimal hypothesis testing for stochastic block models with growing degrees”. In: *arXiv preprint arXiv:1705.05305* (2017).
- [8] Jess Banks et al. “Information-theoretic thresholds for community detection in sparse networks”. In: *Conference on Learning Theory*. PMLR. 2016, pp. 383–416.
- [9] Albert-Laszlo Barabasi and Zoltan N Oltvai. “Network biology: understanding the cell’s functional organization”. In: *Nature reviews genetics* 5.2 (2004), pp. 101–113.
- [10] Yannick Baraud. “Non-asymptotic minimax rates of testing in signal detection”. In: (2002).
- [11] Quentin Berthet and Philippe Rigollet. “Optimal detection of sparse principal components in high dimension”. In: *Annals of Statistics* (2013).
- [12] Matthew Brennan, Guy Bresler, and Wasim Huleihel. “Reducibility and computational lower bounds for problems with planted sparse structure”. In: *Conference On Learning Theory*. PMLR. 2018, pp. 48–166.
- [13] Matthew Brennan, Guy Bresler, and Dheeraj Nagaraj. “Phase transitions for detecting latent geometry in random graphs”. In: *Probability Theory and Related Fields* 178.3 (2020), pp. 1215–1289.
- [14] Sébastien Bubeck et al. “Testing for high-dimensional geometry in random graphs”. In: *Random Structures & Algorithms* 49.3 (2016), pp. 503–532.
- [15] Cristina Butucea and Yuri I Ingster. “Detection of a sparse submatrix of a high-dimensional noisy matrix”. In: *Bernoulli* (2013).
- [16] Yudong Chen and Jiaming Xu. “Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices”. In: *Journal of Machine Learning Research* 17.27 (2016), pp. 1–57.
- [17] Julien Chhor and Alexandra Carpentier. “Goodness-of-fit testing for Hölder-continuous densities: Sharp local minimax rates”. In: *Bernoulli* (2025).
- [18] Julien Chhor and Alexandra Carpentier. “Sharp local minimax rates for goodness-of-fit testing in multivariate binomial and Poisson families and in multinomials.” In: *Mathematical Statistics & Learning* 5 (2022).
- [19] Julien Chhor, Rajarshi Mukherjee, and Subhabrata Sen. “Sparse signal detection in heteroscedastic Gaussian sequence models: sharp minimax rates”. In: *Bernoulli* 30.3 (2024), pp. 2127–2153.
- [20] Amin Coja-Oghlan. “Graph partitioning via adaptive spectral techniques”. In: *Combinatorics, Probability and Computing* 19.2 (2010), pp. 227–284.
- [21] Olivier Collier, Laëtitia Comminges, and Alexandre B Tsybakov. “Minimax estimation of linear and quadratic functionals on sparsity classes”. In: (2017).
- [22] Marom Dadon, Wasim Huleihel, and Tamir Bendory. “Detection and recovery of hidden submatrices”. In: *IEEE Transactions on Signal and Information Processing over Networks* 10 (2024), pp. 69–82.
- [23] Aurelien Decelle et al. “Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications”. In: *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 84.6 (2011), p. 066106.

- [24] Asim K Dey, Yahui Tian, and Yulia R Gel. “Community detection in complex networks: From statistical foundations to data science applications”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 14.2 (2022), e1566.
- [25] Quentin Duchemin and Yohann De Castro. “Random geometric graph: Some recent developments and perspectives”. In: *High Dimensional Probability IX: The Ethereal Volume* (2023), pp. 347–392.
- [26] Dor Elimelech and Wasim Huleihel. “Detecting Arbitrary Planted Subgraphs in Random Graphs”. In: *arXiv preprint arXiv:2503.19069* (2025).
- [27] Santo Fortunato and Mark EJ Newman. “20 years of network community detection”. In: *Nature Physics* 18.8 (2022), pp. 848–850.
- [28] Chao Gao and John Lafferty. “Testing network structure using relations between small subgraph probabilities”. In: *arXiv preprint arXiv:1704.06742* (2017).
- [29] Chao Gao et al. “Community detection in degree-corrected block models”. In: *The Annals of Statistics* 46.5 (2018), pp. 2153–2185.
- [30] Michelle Girvan and Mark EJ Newman. “Community structure in social and biological networks”. In: *Proceedings of the national academy of sciences* 99.12 (2002), pp. 7821–7826.
- [31] Bruce Hajek, Yihong Wu, and Jiaming Xu. “Computational lower bounds for community detection on random graphs”. In: *Conference on Learning Theory*. PMLR, 2015, pp. 899–928.
- [32] Bruce Hajek, Yihong Wu, and Jiaming Xu. “Information limits for recovering a hidden community”. In: *IEEE Transactions on Information Theory* 63.8 (2017), pp. 4729–4745.
- [33] Wasim Huleihel, Arya Mazumdar, and Soumyabrata Pal. “Random subgraph detection using queries”. In: *Journal of Machine Learning Research* 25.126 (2024), pp. 1–25.
- [34] Yu I Ingster. “Minimax testing of nonparametric hypotheses on a distribution density in the L_p metrics”. In: *Theory of Probability & Its Applications* 31.2 (1987), pp. 333–337.
- [35] Yuri Ingster and Irina A Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Vol. 169. Springer Science & Business Media, 2012.
- [36] Yuri I Ingster, Alexandre B Tsybakov, and Nicolas Verzelen. “Detection boundary in sparse regression”. In: (2010).
- [37] Yuri Izmailovich Ingster. “On the minimax nonparametric detection of signals in white gaussian noise”. In: *Problemy Peredachi Informatsii* 18.2 (1982), pp. 61–73.
- [38] Svante Janson. “Asymptotic equivalence and contiguity of some random graphs”. In: *Random Structures & Algorithms* 36.1 (2010), pp. 26–45.
- [39] Subhodh Kotekal, Julien Chhor, and Chao Gao. “Locally sharp goodness-of-fit testing in sup norm for high-dimensional counts”. In: *arXiv preprint arXiv:2409.08871* (2024).
- [40] Subhodh Kotekal and Chao Gao. “Minimax rates for sparse signal detection under correlation”. In: *Information and Inference: A Journal of the IMA* 12.4 (2023), pp. 2873–2969.
- [41] Subhodh Kotekal and Chao Gao. “Sparsity meets correlation in Gaussian sequence model”. In: *arXiv preprint arXiv:2312.09356* (2023).
- [42] Lucien LeCam. “Convergence of estimates under dimensionality restrictions”. In: *The Annals of Statistics* (1973), pp. 38–53.
- [43] Haoyang Liu, Chao Gao, and Richard J. Samworth. “Minimax rates in sparse, high-dimensional change point detection”. In: *The Annals of Statistics* 49.2 (2021), pp. 1081–1112. DOI: [10.1214/20-AOS1994](https://doi.org/10.1214/20-AOS1994).
- [44] Suqi Liu and Miklós Z Rácz. “Phase transition in noisy high-dimensional random geometric graphs”. In: *Electronic Journal of Statistics* 17.2 (2023), pp. 3512–3574.
- [45] Yuetian Luo and Anru R Zhang. “Tensor clustering with planted structures: Statistical optimality and computational limits”. In: *The Annals of Statistics* 50.1 (2022), pp. 584–613.
- [46] Zongming Ma and Yihong Wu. “COMPUTATIONAL BARRIERS IN MINIMAX SUBMATRIX DETECTION”. In: *The Annals of Statistics* (2015), pp. 1089–1116.
- [47] Cristina Maier and Dan Simovici. “Bipartite graphs and recommendation systems”. In: *Journal of Advances in Information Technology-in print* (2022).
- [48] Laurent Massoulié. “Community detection thresholds and the weak Ramanujan property”. In: *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. 2014, pp. 694–703.

- [49] Elchanan Mossel, Joe Neeman, and Allan Sly. “A proof of the block model threshold conjecture”. In: *Combinatorica* 38.3 (2018), pp. 665–708.
- [50] Elchanan Mossel, Joe Neeman, and Allan Sly. “Belief propagation, robust reconstruction and optimal recovery of block models”. In: *Conference on Learning Theory*. PMLR. 2014, pp. 356–370.
- [51] Elchanan Mossel, Joe Neeman, and Allan Sly. “Reconstruction and estimation in the planted partition model”. In: *Probability Theory and Related Fields* 162 (2015), pp. 431–461.
- [52] Elchanan Mossel, Allan Sly, and Youngtak Sohn. “Weak recovery, hypothesis testing, and mutual information in stochastic block models and planted factor graphs”. In: *arXiv preprint arXiv:2406.15957* (2024).
- [53] Gergely Palla et al. “Uncovering the overlapping community structure of complex networks in nature and society”. In: *Nature* 435.7043 (2005), pp. 814–818.
- [54] Georgios A Pavlopoulos et al. “Bipartite graphs in systems biology and medicine: a survey of methods and applications”. In: *GigaScience* 7.4 (2018), giy014.
- [55] Jorge Pena and Yannick Rochat. “Bipartite graphs as models of population structures in evolutionary multiplayer games”. In: (2012).
- [56] Martin Rosvall and Carl T Bergstrom. “Maps of random walks on complex networks reveal community structure”. In: *Proceedings of the national academy of sciences* 105.4 (2008), pp. 1118–1123.
- [57] Asaf Rotenberg, Wasim Huleihel, and Ofer Shayevitz. “Planted Bipartite Graph Detection”. In: *IEEE Transactions on Information Theory* (2024).
- [58] Eric V Slud. “Distribution inequalities for the binomial law”. In: *The Annals of Probability* 5.3 (1977), pp. 404–412.
- [59] Nicolas Verzelen and Ery Arias-Castro. “Community detection in sparse random networks”. In: (2015).
- [60] Yihong Wu and Jiaming Xu. “Statistical problems with planted structures: Information-theoretical and computational limits”. In: *arXiv preprint arXiv:1806.00118* (2018).
- [61] Mingao Yuan and Zuofeng Shang. “Sharp detection boundaries on testing dense subhypergraph”. In: *Bernoulli* 28.4 (2022), pp. 2459–2491.
- [62] Mingao Yuan et al. “Testing community structure for hypergraphs”. In: *The Annals of Statistics* 50.1 (2022), pp. 147–169.

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We substantiate all claims made in the abstract and introduction in Section 3.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We address all serious limitations in Section 5, and outline many possible avenues for future work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide all proofs in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We affirm that we comply with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Our work is entirely theoretical, and we anticipate no societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: We do not use any existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not release any assets in the present work.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.