Journal of Machine Learning Research 24 (2023) 1-34 Submitted 12/20; Published 1/23

Contextual Stochastic Block Model: Sharp Thresholds and Contiguity

Chen Lu CHENL819@MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology
Cambridge MA 02139, U.S.A.

Subhabrata Sen SUBHABRATASEN@FAS.HARVARD.EDU
Department of Statistics
Harvard University
Cambridge, MA 02138, U.S.A.

Editor: Edo Airoldi

Abstract

We study community detection in the *contextual stochastic block model* (Yan and Sarkar (2020); Deshpande et al. (2018)). Deshpande et al. (2018) studied this problem in the setting of sparse graphs with high-dimensional node-covariates. Using the non-rigorous *cavity method* from statistical physics (Mezard and Montanari (2009)), they calculated the sharp limit for community detection in this setting, and verified that the limit matches the information theoretic threshold when the average degree of the observed graph is large. They conjectured that the limit should hold as soon as the average degree exceeds one. We establish this conjecture, and characterize the sharp threshold for detection and weak recovery.

Keywords: community detection, weak-recovery, self-avoiding walks, contiguity.

1. Introduction

The community detection problem arises routinely in diverse applications, and has received significant attention recently in statistics and machine learning. In the simplest version of this problem, given access to a graph, one seeks to cluster the vertices into interpretable *communities* or groups of vertices, which are believed to reflect latent similarities among the nodes. From a theoretical standpoint, this problem has been extensively analyzed under specific generative assumptions on the observed graph; the most popular generative model in this context is the *stochastic block model* (SBM) (Holland et al. (1983)). Inspired by intriguing conjectures arising from the statistical physics community (Krzakala et al. (2013)), community detection under the stochastic block model has been studied extensively. As a consequence, the precise information theoretic limits for recovering the underlying communities have been derived, and optimal algorithms have been identified in this setting (for a survey of these recent breakthroughs, see Abbe (2017)).

In reality, the practitioner often has access to additional information in the form of node covariates, which complements the graph information. Statistically, it is natural to believe that clustering performance can be significantly improved by combining this covariate information with the graph structure. However, establishing this improvement in a formal context, and deriving procedures which combine the two are not straightforward. In Yan and Sarkar (2020), the authors formalize this question, and introduce a simple model for community detection with node covariates. We use the same framework in this paper.

We observe a graph $G = (V, E)$ on n vertices drawn from the sparse Stochastic Block Model $G(n, \frac{a}{n}, \frac{b}{n})$. Formally, we sample a community assignment vector $\sigma \in \{\pm 1\}^n$ uniformly; given σ, we draw edges with probability

$$\mathbb{P}[\{i,j\} \in E] = \begin{cases} \frac{a}{n} & \text{if } \ \sigma_i = \sigma_j, \\ \frac{b}{n} & \text{o.w.} \end{cases}$$

Let $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{n\times n}$ denote the adjacency matrix of the graph G. We define the average degree parameter $d = \frac{a+b}{2}$ and parametrize $a = d + \lambda\sqrt{d}$ and $b = d - \lambda\sqrt{d}$.

Further, at each node of the graph G, we observe a p-dimensional vector of covariates $\{B_i : 1 \le i \le n\}$. The covariates are also correlated with the underlying community assignment. Specifically, we observe

$$B_i = \sqrt{\frac{\mu}{n}}\sigma_i u + Z_i,$$

where $u \sim \mathcal{N}(0, I_p)$ is a latent gaussian vector, and $Z_i \sim \mathcal{N}(0, I_p)$. We construct the matrix $\mathbf{B} = [B_1, \cdots, B_n]^\top \in \mathbb{R}^{n\times p}$. In Yan and Sarkar (2020), the authors introduce a semidefinite programming (SDP) based algorithm for community detection in the above setting, which combines the graph with the node covariates. However, their results do not identify the information theoretic limits of community detection in this context, and do not identify the optimal community detection algorithm in this setting.

Note that when one has access to either the graph information or the covariate information, the information theoretic threshold is well known. Under the parametrization of the SBM, where only $\mathbf{A}$ is given, detection of the underlying community structure, as well as non-trivial recovery, are possible if and only if $\lambda > 1$. On the other hand, the case when only the covariate information $\mathbf{B}$ is available corresponds to a Gaussian mixture clustering problem. Under a high-dimensional asymptotic regime $\frac{n}{p} \to \gamma \in (0, \infty)$, random matrix considerations based on the BBP phase transition (Baik et al. (2005)), and contiguity arguments based on the second moment method (Perry et al. (2018)) imply that non-trivial detection and recovery are possible if and only if $\mu^2 > \gamma$.

In Deshpande et al. (2018), the second author and coauthors studied detection and recovery under this model in a high-dimensional asymptotic regime $\frac{n}{p} \to \gamma \in (0, \infty)$, and conjectured the sharp information theoretic limits in this problem: the underlying community structure can be detected if and only if $\lambda^2 + \frac{\mu^2}{\gamma} > 1$. In particular, this suggests that upon combining the graph information appropriately with the covariates, it is statistically possible to improve upon the optimal performance based on any single information source. The conjecture was derived using the non-rigorous *cavity method* from statistical physics (Mezard and Montanari (2009)), and rigorously established under an additional high-degree asymptotic $d \to \infty$ (after $n \to \infty$). However, numerical experiments in Deshpande et al. (2018) suggest that this high-degree asymptotic is unnecessary, and that the results are true as soon as $d > 1$. In this paper, we formally establish this conjecture. Throughout this article, we will work under the same high-dimensional asymptotic $\frac{n}{p} \to \gamma \in (0, \infty)$.

Our main contributions in this article are as follows.

(i) We first examine the detection problem (1), and establish the sharp threshold for mutual contiguity (see Theorem 1). The testing lower bound is derived using a traditional second moment argument. The upper bound is significantly more challenging—we devise a test statistic by counting certain appropriate self-avoiding walks in the graph.

(ii) Next, we turn to weak recovery, and establish that the threshold for weak recovery coincides with that for detection in this context. To establish the positive half of this result, we crucially utilize the self-avoiding walk based estimation idea introduced in Massoulié (2014); Hopkins and Steurer (2017); however, the presence of the graph data with the covariates makes this application significantly more challenging. We devise estimates for the pairwise correlations of the memberships by counting appropriate walk based statistics, and perform weak recovery by a subsequent projection and rounding step (we refer to Section 5 for further details). This is one of the main technical contributions of this article.

(iii) We then turn to the contiguity regime, and derive a precise expansion of the likelihood ratio in terms of appropriate cycle statistics. In turn, this identifies the precise statistics which distinguish the null from the alternative.

1.1 Main Results

Consider the hypothesis testing problem

$$\mathrm{H}_0 : (\lambda, \mu) = (0, 0) \quad \text{vs.} \quad \mathrm{H}_1 : (\lambda, \mu) \neq (0, 0). \tag{1}$$

We will denote the joint distributions of the data $(\mathbf{A}, \mathbf{B})$ by $\mathbb{P}_{\lambda,\mu}$, and keep the dependence on n, p implicit throughout. Further, we will assume that we are above the threshold for emergence of the giant component, i.e. $d > 1$ (Bollobás et al. (2007); Janson et al. (2011)).

Theorem 1 (Detection) *If $\lambda^2 + \frac{\mu^2}{\gamma} < 1$, $\mathbb{P}_{\lambda,\mu}$ is contiguous to $\mathbb{P}_{0,0}$. On the other hand, if $\lambda^2 + \frac{\mu^2}{\gamma} > 1$, the sequences $\mathbb{P}_{0,0}$ and $\mathbb{P}_{\lambda,\mu}$ are mutually asymptotically singular.*

Our next result addresses the threshold for weak recovery. We recall the relevant notion of weak recovery in this context.

Definition 2 (Weak Recovery) *An estimator $\hat{\sigma} := \hat{\sigma}(\mathbf{A}, \mathbf{B}) \in [-1, 1]^n$ achieves weak recovery if there exists $\varepsilon > 0$, independent of n, such that*

$$\frac{1}{n}\mathbb{E}_{\lambda,\mu}[|\langle \sigma, \hat{\sigma} \rangle|] \geq \varepsilon$$

as $n \to \infty$. We say that weak recovery is possible if there exists an estimator $\hat{\sigma}$ which achieves weak recovery.

Theorem 3 (Weak Recovery) *If $\lambda^2 + \frac{\mu^2}{\gamma} < 1$, then weak recovery is impossible. On the other hand, weak recovery is possible when $\lambda^2 + \frac{\mu^2}{\gamma} > 1$.*

Finally, we turn to the contiguity phase $\lambda^2 + \frac{\mu^2}{\gamma} < 1$, and derive an expansion for the likelihood ratio. We denote the likelihood ratio by

$$L_n = \frac{\mathrm{d}\mathbb{P}_{\lambda,\mu}}{\mathrm{d}\mathbb{P}_{0,0}}.$$

In this regime, $\mathbb{P}_{\lambda,\mu}$ and $\mathbb{P}_{0,0}$ cannot be distinguished with asymptotically negligible Type I and Type II errors. Statistically, the natural problem of interest concerns optimal detection, which is achieved by the likelihood ratio test (LRT). We will derive asymptotic expansions of the likelihood ratio under the null and the alternative. As a consequence, we will obtain the optimal power of the LRT. Along the way, we will obtain a family of statistics which "determine" the likelihood ratio. This will suggest computationally feasible statistics which attain optimal detection performance in this contiguous regime. Similar expansions for the likelihood ratio have been derived for *pure* spiked gaussian problems (the model B and its symmetric analogue) in the recent literature (Banerjee and Ma (2018); Alaoui and Jordan (2018); El Alaoui et al. (2020); Johnstone and Onatski (2020); Onatski et al. (2013, 2014)). Our approach in this regard will be motivated by the techniques introduced in Banerjee and Ma (2018). However, we note that in contrast to this literature, we have *both* a sparse random graph component, and a gaussian model. This necessitates crucial technical modifications—we emphasize the main differences in Section 3.

To this end, let us first introduce a class of cycle statistics. We will denote ω as a cycle on the factor graph corresponding to the posterior distribution (Mezard and Montanari (2009)), shown in Figure 1. Specifically, the factor graph is denoted as $G_{\mathsf{F}} = (V_{\mathsf{F}}, E_{\mathsf{F}})$. The vertices are split into two groups $V_{\mathsf{F}} = V_1 \cup V_2$, where V_1 denotes vertices from the adjacency matrix $\mathbf{A}$, with $|V_1| = n$, and V_2 denotes vertices from the covariate matrix $\mathbf{B}$, so $|V_2| = p$. In Figure 1, vertices in V_1 are shown by dots, and those in V_2 are shown by squares. The edges also split into two groups $E_{\mathsf{F}} = E_1 \cup E_2$, where $E_1 = \{\{i_1, i_2\} : i_1, i_2 \in V_1\}$, and $E_2 = \{\{i, j\} : i \in V_1, j \in V_2\}$. We will refer to edges in E_1 as A edges, and edges in E_2 as B edges. Because B edges must appear in consecutive pairs in a cycle, we refer to such pairs of B edges as B wedges. The graph of a cycle ω is denoted as $G_\omega = (V_\omega, E_\omega)$. We use k to denote the length of the cycle, and l to denote the number of B wedges in the cycle. For a cycle ω, we denote $G_{\omega,A} = (V_{\omega,A}, E_{\omega,A})$ the subgraph of the A edges, and $G_{\omega,B} = (V_{\omega,B}, E_{\omega,B})$ is the subgraph of B edges.

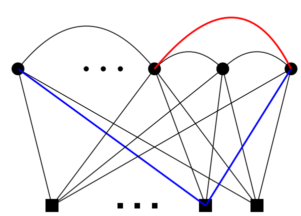

Figure 1: The factor graph corresponding to the posterior distribution. The dots represent the nodes in the adjacency graph $\mathbf{A}$, while the squares represent the variables corresponding to the Gaussian covariates $\mathbf{B}$. An A edge is highlighted in red, while a B-wedge is indicated in blue.

Definition 4 (Cycles) *For $k \geq l$, we define*

$$Y_{n,k,l} = \frac{1}{n^l} \sum_{\omega} \prod_{e_1 \in E_{\omega,A}} A_{e_1} \prod_{e_2 \in E_{\omega,B}} B_{e_2}$$

where the sum is over length k paths with l B-wedges, and the product is over the components of each path.

Our first result establishes the limiting distribution of these cycle statistics under the null and alternative. To this end, it will be convenient to introduce some notation for the relevant index set in this problem. Let us define $\mathcal{J} \subset \mathbb{Z} \times \mathbb{Z}$ such that

$$\mathcal{J} = \{(k, 0) : k \geq 3\} \cup \{(k, l) : k \geq l \geq 1\}.$$

Proposition 5 *The collection*

$$\{Y_{n,k,l} : (k, l) \in \mathcal{J}\}$$

converges in distribution under both H_0 *and* H_1*. Further, the limiting random variables are independent under both* H_0 *and* H_1*. Finally,*

(i) Under H_0, $1 \leq l \leq k$,

$$Y_{n,k,0} \xrightarrow{d} \mathrm{Poi}\Big(\frac{1}{k} d^k\Big), \quad \frac{Y_{n,k,l} - p\mathbf{1}_{k=l=1}}{\sqrt{\frac{1}{2k}\binom{k}{l}\frac{d^{k-l}}{\gamma^l}}} \xrightarrow{d} \mathcal{N}(0, 1).$$

(ii) Under H_1, for any $1 \leq l \leq k$,

$$Y_{n,k,0} \xrightarrow{d} \mathrm{Poi}\Big(\frac{1}{k}(d^k + (\lambda\sqrt{d})^k)\Big), \quad \frac{Y_{n,k,l} - p\mathbf{1}_{k=l=1} - \frac{1}{2k}\binom{k}{l}\frac{(\lambda\sqrt{d})^{k-l}\mu^l}{\gamma^l}}{\sqrt{\frac{1}{2k}\binom{k}{l}\frac{d^{k-l}}{\gamma^l}}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Finally, if $l \geq 1$, the distributional limits continue to hold for k_n, l_n growing in n, as long as $l_n \leq k_n = o(\sqrt{\log n})$.

Let $\{\upsilon_{k,l,j} : j \in \{0, 1\}, (k, l) \in \mathcal{J}\}$ be collection of random variables with the desired limiting distributions. Specifically, $\upsilon_{k,0,0} \sim \mathrm{Poi}\Big(\frac{1}{k}d^k\Big)$ and for $l \geq 1$, $\upsilon_{k,l,0} \sim \mathcal{N}(\mu_{k,l,0}, \sigma^2_{k,l})$. Similarly, $\upsilon_{k,0,1} \sim \mathrm{Poi}\Big(\frac{1}{k}(d^k + (\lambda\sqrt{d})^k)\Big)$, and for $l \geq 1$, $\upsilon_{k,l,1} \sim \mathcal{N}(\mu_{k,l,1}, \sigma^2_{k,l})$. Here, the means and variances $\mu_{k,l,0}$ and $\sigma^2_{k,l}$ are as specified in Proposition 5 under the null, while $\mu_{k,l,1}$ denotes the mean under the alternative.

Theorem 6 *Consider (λ, μ) satisfying $\lambda^2 + \frac{\mu^2}{\gamma} < 1$. Then the following hold:*

1. *$\mathbb{P}_{\lambda,\mu}$ and $\mathbb{P}_{0,0}$ are asymptotically mutually contiguous.*
2. *Under H_0, we have that*

$$L_n \xrightarrow{d} \exp\Big(\sum_{k=1}^{\infty}\Big[\log(1 - \lambda^k d^{k/2})\upsilon_{k,0,0} - \frac{1}{k}(\lambda\sqrt{d})^k + \sum_{1 \leq l \leq k} \frac{\mu_{k,l,0}\upsilon_{k,l,1} - \frac{1}{2}\mu^2_{k,l,1}}{\sigma^2_{k,l}}\Big]\Big).$$

1.2 Related Literature

Covariate assisted clustering has been extensively studied across statistics, machine learning and computer science using diverse perspectives. The literature on this topic is quite diffuse, and its impossible to provide an exhaustive survey of this area. However, for the convenience of the reader, we survey the main methodological approaches, and discuss in-depth the main results relevant for our work.

From a methodological standpoint, generative model based approaches are very natural for this problem, and they have been extensively explored in this setting (Newman and Clauset (2016); Hoff (2003); Zanghi et al. (2010); Yang et al. (2009); Kim and Leskovec (2012); Leskovec and Mcauley (2012); Xu et al. (2012); Hoang and Lim (2014); Yang et al. (2013)). On the other hand, model free approaches, which cluster the nodes by optimizing a suitable loss function have also been popular (Binkiewicz et al. (2017); Zhang et al. (2016); Gibert et al. (2012); Zhou et al. (2009); Neville et al. (2003); Gunnemann et al. (2013); Dang and Viennet (2012); Cheng et al. (2011); Silva et al. (2012); Smith et al. (2016)). Bayesian methods (Chang and Blei (2010); Balasubramanyan and Cohen (2011)) provide another natural methodological approach for this problem. We refer the interested reader to Bothorel et al. (2015) for a survey of other approaches.

In a separate direction Aicher et al. (2014); Lelarge et al. (2015) study a version of community detection with informative edges. Lelarge et al. (2015) establishes only one side of the conjectured information theoretic threshold in this setting.

More recently, Binkiewicz et al. (2017); Zhang et al. (2016) analyze specific heuristic clustering algorithms under the block model formalism. However, consistency guarantees in this setting are derived for dense graphs, and under strong separability assumptions on the connection probabilities. Further, they do not identify the precise information theoretic thresholds for recovery. As a consequence, the precise information theoretic gains obtained from the additional covariates remains unclear.

Our work is closest in spirit to Yan and Sarkar (2020). They study an SDP based framework for community recovery. However, in contrast to our setting, they study low-dimensional covariates. They formally establish that clustering accuracy is improved upon combining the node information with the graph. In contrast, we study high-dimensional covariates, and establish that the information theoretic threshold is shifted in this setting.

Somewhat related to our inquiry, Kanade et al. (2016); Mossel and Xu (2016) study local algorithms for semisupervised clustering, i.e. when the true labels are given for a small fraction of nodes. While these algorithms are local, our analysis is global, and we capture the information theoretic limits in this problem.

1.3 Technical Contributions

The main contributions of this paper are the algorithm for weak recovery above the threshold $\lambda^2 + \frac{\mu^2}{\gamma} > 1$, and the asymptotic distribution of the cycle statistics in Proposition 5. Contiguity below the threshold follows from standard second moment arguments, and expansions of the likelihood ratio in the contiguity phase are based on a version of Janson's small-subgraph conditioning method. The small-subgraph conditioning argument follows the same template as Banerjee and Ma (2018), once the distribution of the cycle statistics Proposition 5 is given. In this section, we elaborate on our main technical contributions.

In prior work on community detection (see e.g. Mossel et al. (2015); Banerjee and Ma (2018); Hopkins and Steurer (2017); Massoulié (2014)), weak recovery has often been performed using statistics based on appropriate self-avoiding walks. In particular, the wide-applicability of this idea was emphasized in Hopkins and Steurer (2017). The general meta-algorithm introduced in Hopkins and Steurer (2017) has the following steps: (i) estimate the second moment $\sigma\sigma^{\mathrm{T}}$ using self-avoiding walk statistics, and (ii) use a generic projection and rounding procedure to derive the membership estimator. We follow the same strategy— the main technical challenge is to construct the appropriate estimator for the second moment. One might naturally suspect that an appropriate self-avoiding walk based statistic might be relevant in our setting. However, we have two data sources, encoded by the graph adjacency matrix and the matrix of gaussian covariates. As a consequence, the relevant statistics are not obvious in this setting. A first idea is to consider the factor graph for this problem (see Figure 1), and use self-avoiding walks of a fixed length. However, some thought reveals that this is sub-optimal; to see this, consider a setting where $\lambda^2 + \frac{\mu^2}{\gamma} > 1$, but $\lambda^2 < 1$. In this case, the block model alone does not contain any information regarding the underlying community assignment, so the walks based purely on the sparse graph will only contain noise.

Instead, we construct paths with edges from both the adjacency and covariate matrices, whose ratio of edges is given by $\lambda^2 : \frac{\mu^2}{\gamma}$. Conceptually, each path is constructed so that the contributions from the adjacency and covariate matrices reflect the amount of information from each source respectively. This approach can be potentially useful for other reconstruction problems which have multiple information sources.

On the other hand, the distribution of the likelihood ratio in the contiguity regime $\lambda^2 + \frac{\mu^2}{\gamma} < 1$ is determined by all cycles of finite length. For cycles with edges coming solely from the adjacency or covariance matrix, the distribution limits are Poisson (Mossel et al. (2015)) and Gaussian (Banerjee and Ma (2018)) respectively. In our setting, we also encounter mixed-cycles, comprising edges coming from both sources. Using a method of moments approach, we establish that the limiting distribution of these mixed cycles are all independent Gaussian random variables in the limit. Finally, we characterize the means and variances of these cycles under the null and alternative. We expect the general techniques to be useful in other settings as well.

1.4 Discussion:

(i) Network modeling with covariates: The contextual block model studied in this manuscript provides a formal framework to study community detection with auxiliary node covariates. The model is quite natural and succinct, and permits sharp theoretical analysis. In practice, one might add additional parameters to the model to improve empirical performance. For example, the marginal distribution of the node covariates in our model corresponds to a *spiked covariance model* with covariance matrix $I + \frac{\mu}{n}\sigma\sigma^{\top}$. Within this framework, a natural generalization would be to assume that the marginal distribution of the covariates is a spike covariance model with covariance matrix $\Sigma + \frac{\mu}{n}\sigma\sigma^{\top}$, for a positive semi-definite matrix Σ. In Section 6, we study the stability of our predictions under special classes of Σ. In particular, we study the empirical performance of a simple Belief Propagation (BP) Algorithm under this general model. One might also wonder whether the assumption of gaussianity on the node covariates is required in practice; Belief Propagation algorithms are typically *universal* in their behavior, as long as the covariates are not too dependent, and have sufficiently light tails. We also explore this phenomenon via numerical simulations in Section 6.

The contextual block model studied here samples the covariates given the true membership assignment. On the other hand, from a modeling perspective, it might be natural to model the distribution of the covariates first, and then specify a conditional model for the community memberships given the covariates. The model studied in Newman and Clauset (2016) is a prominent example in this regard. We note that these models are typically more involved, and are not amenable to sharp theoretical analysis.

(ii) Knowledge of λ, μ parameters: The community recovery algorithm introduced in this paper requires a priori knowledge of the parameters λ and μ. It is natural to wonder whether similar recovery algorithms exist if λ and μ are unknown. This question has been analyzed in the two corner cases $\lambda = 0$ and $\mu = 0$: if $\lambda = 0$, the entire statistical information comes from the node covariates. In this setting, if $\mu^2/\gamma > 1$, the sample covariance matrix $\mathbf{B}\mathbf{B}^\top$ has an outlying eigenvalue, which can be used to estimate μ. On the other hand, for $\mu = 0$, if $\lambda > 1$, the average degree d and λ can be estimated from the data using a walk based statistic (Mossel et al. (2015)). We conjecture that even in the general case, if $\lambda^2 + \frac{\mu^2}{\gamma} > 1$, the average degree $d > 1$, λ and μ should be estimable from the data, using appropriate walk based estimators. We leave this as a question for future inquiry.

Organization: The rest of the paper is structured as follows. We establish Theorem 1 in Section 2. In Section 3 we establish the asymptotic expansion of the likelihood ratio in the contiguity regime, and establish Theorem 6. We establish Theorem 3 in Sections 4 and 5. Finally, we supplement our theoretical results by some numerical experiments in Section 6.

2. Detection

We prove Theorem 1 in this section.

Proof We start with a proof of the information theoretic lower bound. Fix λ, μ such that $\lambda^2 + \frac{\mu^2}{\gamma} < 1$. We will use the traditional second moment approach. First, consider a complete data problem, where one observes the latent vectors $\sigma \in \{\pm 1\}^n$ and $u \in \mathbb{R}^p$. We denote the corresponding distribution as $\tilde{\mathbb{P}}_{\lambda,\mu}$. Thus we have,

$$L := \frac{\mathrm{d}\mathbb{P}_{\lambda,\mu}}{\mathrm{d}\mathbb{P}_{0,0}}(\mathbf{A}, \mathbf{B}) = \frac{\mathbb{E}_{\sigma,u}\left[\tilde{\mathbb{P}}_{\lambda,\mu}(\mathbf{A}, \mathbf{B}|\sigma, u)\right]}{\mathbb{P}_{0,0}(\mathbf{A}, \mathbf{B})},$$

where $\mathbb{E}_{\sigma,u}[\cdot]$ calculates the expectation with respect to the priors on σ and u. Consider the event $\mathcal{S} = \{u : \|u\|_2 \le (1+\delta)\sqrt{p}\}$, where $\delta > 0$ will be chosen appropriately. Define the truncated likelihood

$$\tilde{L} = \frac{\mathbb{E}_{\sigma,u}\Big[\tilde{\mathbb{P}}_{\lambda,\mu}(\mathbf{A},\mathbf{B}|\sigma,u)\mathbf{1}(u\in\mathcal{S})\Big]}{\mathbb{P}_{0,0}(\mathbf{A},\mathbf{B})}.$$

Finally, we claim that if $\lambda^2 + \frac{\mu^2}{\gamma} < 1$, then there exists a universal constant $C > 0$ such that $\mathbb{E}_{0,0}[\tilde{L}^2] \le C < \infty$. The desired contiguity follows from (Perry et al., 2018, Lemma 2.4).

It remains to establish that $\mathbb{E}_{0,0}[\tilde{L}^2] \le C < \infty$ for some universal $C > 0$. To this end, by Fubini's theorem, we note that

$$\mathbb{E}_{0,0}[\tilde{L}^2] = \mathbb{E}_{(\sigma,u),(\tau,v)}\Big[\mathbb{E}_{0,0}\Big[\frac{\tilde{\mathbb{P}}_{\lambda,\mu}(\mathbf{A},\mathbf{B}|\sigma,u)}{\mathbb{P}_{0,0}(\mathbf{A},\mathbf{B})}\frac{\tilde{\mathbb{P}}_{\lambda,\mu}(\mathbf{A},\mathbf{B}|\tau,v)}{\mathbb{P}_{0,0}(\mathbf{A},\mathbf{B})}\mathbf{1}(u,v\in\mathcal{S})\Big]\Big]. \tag{2}$$

Now, we have,

$$\frac{\tilde{\mathbb{P}}_{\lambda,\mu}(\mathbf{A},\mathbf{B}|\sigma,u)}{\mathbb{P}_{0,0}(\mathbf{A},\mathbf{B})} = \frac{\tilde{\mathbb{P}}_{\lambda,\mu}(\mathbf{A}|\sigma)}{\mathbb{P}_{0,0}(\mathbf{A})}\cdot\frac{\tilde{\mathbb{P}}_{\lambda,\mu}(\mathbf{B}|\sigma,u)}{\mathbb{P}_{0,0}(\mathbf{B})}.$$

We evaluate each term in turn. First, we have,

$$\frac{\tilde{\mathbb{P}}_{\lambda,\mu}(\mathbf{A}|\sigma)}{\mathbb{P}_{0,0}(\mathbf{A})} = \prod_{i<j} W_{ij}, W_{ij} = W_{ij}(\mathbf{A},\sigma) = \begin{cases} \frac{2a}{a+b} & \text{if } \sigma_i = \sigma_j, \quad A_{ij} = 1 \\ \frac{2b}{a+b} & \text{if } \sigma_i \neq \sigma_j, \quad A_{ij} = 1 \\ \frac{n-a}{n-(a+b)/2} & \text{if } \sigma_i = \sigma_j, \quad A_{ij} = 0 \\ \frac{n-b}{n-(a+b)/2} & \text{if } \sigma_i \neq \sigma_j, \quad A_{ij} = 0 \end{cases}$$

Second, direct computation yields

$$\frac{\tilde{\mathbb{P}}_{\lambda,\mu}(\mathbf{B}|\sigma,u)}{\mathbb{P}_{0,0}(\mathbf{B})} = \exp\Big(\sqrt{\frac{\mu}{n}}\sum_{i=1}^{n}\sigma_i Z_i^\top u - \frac{\mu}{2}\|u\|^2\Big).$$

Therefore,

$$\begin{aligned}&\mathbb{E}_{(\sigma,u),(\tau,v)}\Big[\mathbb{E}_{0,0}\Big[\frac{\tilde{\mathbb{P}}_{\lambda,\mu}(\mathbf{A},\mathbf{B}|\sigma,u)}{\mathbb{P}_{0,0}(\mathbf{A},\mathbf{B})}\frac{\tilde{\mathbb{P}}_{\lambda,\mu}(\mathbf{A},\mathbf{B}|\tau,v)}{\mathbb{P}_{0,0}(\mathbf{A},\mathbf{B})}\mathbf{1}(u,v\in\mathcal{S})\Big]\Big] \\ &= \mathbb{E}_{(\sigma,u),(\tau,v)}\Big[\mathbf{1}(u,v\in\mathcal{S})\mathbb{E}_{0,0}\Big[\prod_{i<j}W_{ij}V_{ij}\exp\Big(\sqrt{\frac{\mu}{n}}\sum_{i=1}^{n}Z_i^\top(\sigma_i u + \tau_i v) - \frac{\mu}{2}(\|u\|^2+\|v\|^2)\Big)\Big]\Big],\end{aligned}$$

where $V_{ij} = V_{ij}(\mathbf{A},\tau)$ is defined similarly to W_{ij}. Under $\mathbb{P}_{0,0}$, for any (σ,u) and (τ,v), $\mathbf{A}$ and $\mathbf{B}$ are independent. Setting $\rho = \rho(\sigma,\tau) = \frac{1}{n}\langle\sigma,\tau\rangle$, we have,

$$\mathbb{E}_{0,0}\Big[\exp\Big(\sqrt{\frac{\mu}{n}}\sum_{i=1}^{n}Z_i^\top(\sigma_i u + \tau_i v) - \frac{\mu}{2}(\|u\|^2+\|v\|^2)\Big)\Big] = \exp\Big(\frac{\mu}{n}\langle u,v\rangle\langle\sigma,\tau\rangle\Big).$$

Using (Mossel et al., 2015, Lemma 5.4), we have,

$$\mathbb{E}_{0,0}\Big[\prod_{i<j}W_{ij}V_{ij}\Big] = (1+o(1))e^{-\lambda^2/2-\lambda^4/4}\exp\Big(\frac{\rho^2\lambda^2}{2}(d+n)\Big).$$

Plugging these back into (2), we obtain,

$$\mathbb{E}_{0,0}[\tilde{L}^2] \le (1+o(1))e^{-\lambda^2/2-\lambda^4/4+\lambda^2 d/2}\mathbb{E}_{(\sigma,u),(\tau,v)}\Big[\exp\Big(n\Big(\frac{\rho^2\lambda^2}{2}+\frac{\mu}{\gamma}\rho\frac{\langle u,v\rangle}{p}\Big)\Big)\mathbf{1}(u,v\in\mathcal{S})\Big)\Big].,$$

We note that $\langle u,v\rangle = \|u\|\|v\|\langle\frac{u}{\|u\|},\frac{v}{\|v\|}\rangle$, $\|u\|,\|v\|\le(1+\delta)\sqrt{p}$ on the event $\mathcal{S}$, and $\langle\frac{u}{\|u\|},\frac{v}{\|v\|}\rangle \overset{d}{=} Y$, where Y is the first coordinate of a uniform vector on the unit sphere. Thus we have,

$$\mathbb{E}_{0,0}[\tilde{L}^2] \le (1+o(1))e^{-\lambda^2/2-\lambda^4/4+\lambda^2 d/2}\mathbb{E}\Big[\exp\Big(n\Big(\frac{\lambda^2}{2}X^2+(1+\delta)^2\frac{\mu}{\gamma}XY\Big)\Big)\Big],$$

where $X\overset{d}{=}\rho(\sigma,\tau)$ and Y is as described above. It is easy to see that $Y\in[-1,1]$ has density

$$f_Y(y)=\frac{\Gamma(p/2)}{\Gamma((p-1)/2)\Gamma(1/2)}(1-y^2)^{(p-3)/2}\le C\sqrt{n}(1-y^2)^{p/2},$$

for some universal constant $C>0$. Further, for $s\in(\frac{2}{n}\mathbb{Z})\cap[-1,1]$,

$$\mathbb{P}(X=s)=\frac{1}{2^n}\binom{n}{n(1+s)/2}\le\frac{C}{\sqrt{n}}\exp(nh(s)),$$

where $h(s)=-(1+s)/2\log(1+s)-(1-s)/2\log(1-s)$. Using $h(s)\le -s^2/2$, direct computation now yields that

$$\begin{aligned}&\mathbb{E}\Big[\exp\Big(n\Big(\frac{\lambda^2}{2}X^2+(1+\delta)^2\frac{\mu}{\gamma}XY\Big)\Big)\Big]\\ &\le Cn\int_{\mathbb{R}^2}\exp\Big[n\Big(\frac{\lambda^2}{2}s^2+\frac{\mu}{\gamma}(1+\delta)^2 sy-\frac{s^2}{2}-\frac{y^2}{2\gamma}\Big)\Big]dsdy\\ &< C'\end{aligned}$$

for some universal constant C', provided $\lambda^2+\frac{\mu^2}{\gamma}(1+\delta)^2<1$. This completes the proof.

Next, we turn to the regime $\lambda^2+\frac{\mu^2}{\gamma}>1$. We will devise a test based on the cycle statistic $Y_{n,k,l}$ with

$$\frac{l}{k}=\frac{\mu^2/\gamma}{\lambda^2+\mu^2/\gamma},$$

and some k, to be chosen appropriately. For k growing sufficiently slowly in n, Proposition 5 implies that $Y_{n,k,l}/\sigma_{k,l}$ is approximately $\mathcal{N}(0,1)$ under H_0, while it is distributed as $\mathcal{N}(\tilde{\mu},1)$ under H_1. The non-centrality parameter in this case is

$$\tilde{\mu}=\frac{1}{\sqrt{2k}}\Big(\binom{k}{l}(\lambda^2)^\alpha\Big(\mu^2/\gamma\Big)^\beta\Big)^{1/2}=\exp\Big(\frac{1}{2}k\log(\lambda^2+\mu^2/\gamma)+o(1)\Big).$$

Thus for k growing sufficiently slowly in n, we will get a sequence of consistent tests. This establishes the positive side of the detection threshold.

■

3. Likelihood ratio expansion

Armed with the distributional characterization of Proposition 5, we can characterlize the likelihood ratio expansion with a version of the small subgraph conditioning argument relevant to our setting. This argument was originally formalized by Robinson and Wormald (1992, 1994) in the context of random d-regular graphs, and was utilized by Mossel et al. (2015) in their study of community detection for the stochastic block model. On the other hand, inspired by a version of this argument developed by Janson (1995), Banerjee and Ma (2018) develop a Gaussian variant of this argument, and apply it to the study of contiguous regimes for Gaussian matrices with low-rank perturbations. Our setting naturally has a sparse graph, and a Gaussian component, and thus requires an extension. We expect this result to be useful in many other settings.

Proposition 7 (Small subgraph conditioning method) *Let $\mathbb{P}_n$ and $\mathbb{Q}_n$ be two sequences of probability measures, and let $\{Y_{n,k,l} : (k,l) \in \mathcal{J}\}$ be such that the following conditions hold:*

1. *$\mathbb{Q}_n$ is absolutely continuous w.r.t. $\mathbb{P}_n$.*
2. *All finite dimensional distributions of $\{Y_{n,k,l} : (k,l) \in \mathcal{J}\}$ converge to the null distribution under $\mathbb{P}_n$, and to the alternative distribution under $\mathbb{Q}_n$, as specified in Proposition 5.*
3. *The likelihood ratio $L_n = \frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n}$ satisfies:*

$$\limsup_{n\to\infty} \mathbb{E}_{\mathbb{P}_n}[L_n^2] \leq \exp\Big\{ -\frac{1}{2}\log(1-(\lambda^2+\frac{\mu^2}{\gamma})) - \frac{\lambda^2}{2} - \frac{\lambda^4}{4}\Big\} < \infty.$$

Then, we have the following consequences:

1. *$\mathbb{P}_n$ and $\mathbb{Q}_n$ are asymptotically mutually contiguous.*
2. *Under $\mathbb{P}_n$, we have that*

$$L_n \xrightarrow{d} \exp\Big(\sum_{k=1}^{\infty}\Big[\log(1-\lambda^k d^{k/2})\upsilon_{k,0,0} - \frac{1}{k}(\lambda\sqrt{d})^k + \sum_{1\leq l\leq k}\frac{\mu_{k,l,0}\upsilon_{k,l,0} - \frac{1}{2}\mu_{k,l,0}^2}{\sigma_{k,l}^2}\Big]\Big).$$

Given Proposition 5, the proof of this proposition is identical to the proof of Proposition 1 of Banerjee and Ma (2018), but with some Gaussian terms swapped out with Poisson terms. Thus we omit the proof.

3.1 Proof of Proposition 5

We prove Proposition 5 in this section. Formally, fix $(k_1,l_1),\cdots,(k_r,l_r) \in \mathcal{J}$, and $m_1,\cdots,m_r \geq 1$. Without loss of generality, we assume that there exists $r_1 \leq r$ such that $l_1 = \cdots = l_{r_1} = 0$ and $l_j > 0$ for $j > r_1$. Further, assume that $k_1 < k_2 < \cdots < k_{r_1}$ and $k_{r_1+1} < k_{r_1+2} < \cdots < k_r$. For convenience, we will denote

$$Z_{n,k_j,0} = Y_{n,k_j,0}, \quad 1 \leq j \leq r_1,$$
$$Z_{n,k_{r_1+1},l_{r_1+1}} = Y_{n,k_{r_1+1},l_{r_1+1}} - p\mathbf{1}_{k_{r_1+1}=l_{r_1+1}=1}, Z_{k_j,l_j} = Y_{n,k_j,l_j}, \quad j > r_1.$$

We will show that for $j \le r_1$, Z_{n,k_j,l_j} have Poisson limits, and for $j > r_1$, Z_{n,k_j,l_j} have Gaussian limits. This is done with the method of moments: we will establish that as $n \to \infty$,

$$\begin{aligned}
&\mathbb{E}_{0,0}\Big[\prod_{j=1}^{r} Z_{n,k_j,l_j}^{m_j}\Big] \to \prod_{j=1}^{r_1} \mathbb{E}[v_{k_j,l_j,0}^{m_j}] \cdot \prod_{j=r_1+1}^{r} \sigma_{k_j,l_j}\mathbb{E}[\xi_j^{m_j}], \\
&\mathbb{E}_{\lambda,\mu}\Big[\prod_{j=1}^{r} Z_{n,k_j,l_j}^{m_j}\Big] \to \prod_{j=1}^{r_1} \mathbb{E}[v_{k_j,l_j,1}^{m_j}] \cdot \prod_{j=r_1+1}^{r} \sigma_{k_j,l_j}\mathbb{E}[\xi_j^{m_j}],
\end{aligned} \tag{3}$$

where $\{\xi_1, \xi_2, \cdots\}$ is a sequence of iid $\mathcal{N}(0,1)$ random variables. It is easy to verify that the random variables Z_{n,k_j,l_j} satisfy Carleman's condition, so (3) implies the desired convergence in distribution in Proposition 5.

We first establish a *decoupling* lemma, which will allow us to separate the analysis of terms with $l = 0$, which have Poisson limits, from terms with $l > 0$, which have Gaussian limits.

Lemma 8 *Fix $r \ge 1$ and $r_1 \le r$. Fix $(k_1, l_1), \cdots, (k_r, l_r) \in \mathcal{J}$, and $m_1, \cdots, m_r \ge 1$. Further assume that $l_j = 0$ for $j \le r_1$, and $l_j > 0$ for $r_1 < j \le r$. Then we have, as $n \to \infty$,*

$$\begin{aligned}
\Big|\mathbb{E}_{0,0}\Big[\prod_{j=1}^{r} Z_{n,k_j,l_j}^{m_j}\Big] - \mathbb{E}_{0,0}\Big[\prod_{j=1}^{r_1} Z_{n,k_j,l_j}^{m_j}\Big]\mathbb{E}_{0,0}\Big[\prod_{j=r_1+1}^{r} Z_{n,k_j,l_j}^{m_j}\Big]\Big| \to 0, \\
\Big|\mathbb{E}_{\lambda,\mu}\Big[\prod_{j=1}^{r} Z_{n,k_j,l_j}^{m_j}\Big] - \mathbb{E}_{\lambda,\mu}\Big[\prod_{j=1}^{r_1} Z_{n,k_j,l_j}^{m_j}\Big]\mathbb{E}_{\lambda,\mu}\Big[\prod_{j=r_1+1}^{r} Z_{n,k_j,l_j}^{m_j}\Big]\Big| \to 0.
\end{aligned}$$

Proof We first establish the assertion under H_0. We have

$$\begin{aligned}
&\mathbb{E}_{0,0}\Big[\prod_{j=1}^{r} Z_{n,k_j,l_j}^{m_j}\Big] = \mathbb{E}_{0,0}[T_1 \cdot T_2 \cdot T_3], \\
&T_1 = \prod_{j=1}^{r_1}\Big(\sum_{\omega_j} \prod_{e \in E_{\omega_j,A}} A_e\Big)^{m_j}, \\
&T_2 = \Big(\frac{1}{n^{l_{r_1+1}}} \sum_{\omega_{r_1+1}} \prod_{e \in E_{\omega_{r_1+1},A}} A_e \prod_{e \in E_{\omega_{r_1+1},B}} B_e - p \cdot \mathbf{1}(k_{r_1+1} = l_{r_1+1} = 1)\Big)^{m_{r_1+1}}, \\
&T_3 = \prod_{j=r_1+2}^{r}\Big(\frac{1}{n^{l_j}} \sum_{\omega_j} \prod_{e \in E_{\omega_j,A}} A_e \prod_{e \in E_{\omega_j,B}} B_e\Big)^{m_j}.
\end{aligned}$$

Expanding, these terms may be expressed as

$$T_1 = \prod_{j=1}^{r_1} \Big(\sum_{\omega_{j,1},\cdots,\omega_{j,m_j}} \prod_{q=1}^{m_j} \prod_{e \in E_{\omega_{j,q},A}} A_e \Big),$$

$$T_2 = \frac{1}{n^{l_{r_1+1} m_{r_1+1}}} \sum_{\omega_{r_1+1,1},\cdots,\omega_{r_1+1,m_{r_1+1}}} \prod_{q=1}^{m_{r_1+1}} \Big(\prod_{e \in E_{\omega_{r_1+1,q},A}} A_e \prod_{e \in E_{\omega_{r_1+1,q},B}} B_e - p \cdot \mathbf{1}(k_{r_1+1} = l_{r_1+1} = 1) \Big),$$

$$T_3 = \prod_{j=r_1+2}^{r} \Big(\frac{1}{n^{l_j m_j}} \sum_{\omega_{j,1},\cdots,\omega_{j,m_j}} \prod_{q=1}^{m_j} \prod_{e \in E_{\omega_{j,q},A}} A_e \prod_{e \in E_{\omega_{j,q},B}} B_e \Big).$$

Thus we have,

$$\begin{aligned} &\mathbb{E}_{0,0}[T_1 T_2 T_3 | \mathbf{A}] \\ &= \frac{1}{n^{\sum_{j=r_1+1}^{r} l_j m_j}} \sum_{j \in [r], 1 \le q_j \le m_j, \omega_{j,q_j}} \\ &\qquad \mathbb{E}_{0,0} \left[\begin{array}{c} \Big(\prod_{j=1}^{r_1} \prod_{q_j \le m_j} \prod_{e \in E_{\omega_{j,q_j},A}} A_e \Big) \times \\ \prod_{q=1}^{m_{r_1+1}} \Big(\prod_{e \in E_{\omega_{r_1+1,q},A}} A_e \prod_{e \in E_{\omega_{r_1+1,q},B}} B_e - p \cdot \mathbf{1}(k_{r_1+1} = l_{r_1+1} = 1) \Big) \times \Big| A \\ \prod_{j=r_1+2}^{r} \prod_{q_j=1}^{m_j} \prod_{e \in E_{\omega_{j,q_j},A}} A_e \prod_{e \in E_{\omega_{j,q_j},B}} B_e \end{array} \right]. \end{aligned}$$

Consider first the case $(k_{r_1+1}, l_{r_1+1}) \neq (1,1)$. In this case,

$$\begin{aligned} \mathbb{E}_{0,0}[T_1 T_2 T_3] &= \frac{1}{n^{\sum_{j=r_1+1}^{r} l_j m_j}} \times \\ &\sum_{j \in [r], 1 \le q_j \le m_j, \omega_{j,q_j}} \mathbb{E}_{0,0}\Big[\Big(\prod_{j=1}^{r_1} \prod_{q_j \le m_j} \prod_{e \in E_{\omega_{j,q_j},A}} A_e \Big) \Big(\prod_{j=r_1+1}^{r} \prod_{q_j \le m_j} \prod_{e \in E_{\omega_{j,q_j},A}} A_e \Big) \Big] \times \\ &\qquad \mathbb{E}_{0,0}\Big[\prod_{j=r_1+1}^{r} \prod_{q_j \le m_j} \prod_{e \in E_{\omega_{j,q_j},B}} B_e \Big]. \end{aligned}$$

On the other hand, a similar calculation yields

$$\begin{aligned} \mathbb{E}_{0,0}\Big[\prod_{j=1}^{r_1} Z_{n,k_j,l_j}^{m_j} \Big] \mathbb{E}_{0,0}\Big[\prod_{j=r_1+1}^{r} Z_{n,k_j,l_j}^{m_j} \Big] &= \frac{1}{n^{\sum_{j=r_1+1}^{r} l_j m_j}} \times \\ \sum_{j \in [r], 1 \le q_j \le m_j, \omega_{j,q_j}} \mathbb{E}_{0,0}\Big[\Big(\prod_{j=1}^{r_1} \prod_{q_j \le m_j} \prod_{e \in E_{\omega_{j,q_j},A}} A_e \Big) \Big] &\mathbb{E}_{0,0}\Big[\Big(\prod_{j=r_1+1}^{r} \prod_{q_j \le m_j} \prod_{e \in E_{\omega_{j,q_j},A}} A_e \Big) \Big] \times \qquad (4) \\ \mathbb{E}_{0,0}\Big[\prod_{j=r_1+1}^{r} \prod_{q_j \le m_j} \prod_{e \in E_{\omega_{j,q_j},B}} B_e \Big]. \end{aligned}$$

To establish that the difference between the two quantities is $o(1)$, note that if two cycles overlap on m-edges, we gain a factor $O(n^m)$. However, this naturally implies they overlap on at least $(m+1)$ vertices. As a result, we lose a factor n^{m+1} in choosing these cycles. As a result, the dominant contributions arise from non-overlapping cycles, thus establishing the desired claim in this case. The proof for $(k_{r_1+1}, l_{r_1+1}) = (1,1)$ is exactly analogous.

Under $\mathbb{P}_{\lambda,\mu}$, if $(k_{r_1+1}, l_{r_1+1}) \neq (1,1)$, we have,

$$\mathbb{E}_{\lambda,\mu}\Big[\prod_{j=1}^{r} Z_{n,k_j,l_j}^{m_j}\Big] = \frac{1}{n^{\sum_{j=r_1+1}^{r} l_j m_j}} \times$$
$$\mathbb{E}_{\sigma}\Big[\sum_{j\in[r],1\leq q_j\leq m_j,\omega_{j,q_j}} \mathbb{E}_{\lambda,\mu}\Big[\Big(\prod_{j=1}^{r_1}\prod_{q_j\leq m_j}\prod_{e\in E_{\omega_{j,q_j},A}} A_e\Big)\Big(\prod_{j=r_1+1}^{r}\prod_{q_j\leq m_j}\prod_{e\in E_{\omega_{j,q_j},A}} A_e\Big) \mid \sigma\Big] \times$$
$$\mathbb{E}_{\lambda,\mu}\Big[\prod_{j=r_1+1}^{r}\prod_{q_j\leq m_j}\prod_{e\in E_{\omega_{j,q_j},B}} B_e \mid \sigma\Big]\Big].$$

On the other hand,

$$\mathbb{E}_{\lambda,\mu}\Big[\prod_{j=1}^{r_1} Z_{n,k_j,l_j}^{m_j}\Big]\mathbb{E}_{\lambda,\mu}\Big[\prod_{j=r_1+1}^{r} (Z_{n,k_j,l_j})^{m_j}\Big] = \frac{1}{n^{\sum_{j=r_1+1}^{r} l_j m_j}} \times$$
$$\mathbb{E}_{\sigma}\Big[\sum_{j\in[r],1\leq q_j\leq m_j,\omega_{j,q_j}} \mathbb{E}_{\lambda,\mu}\Big[\Big(\prod_{j=1}^{r_1}\prod_{q_j\leq m_j}\prod_{e\in E_{\omega_{j,q_j},A}} A_e\Big) \mid \sigma\Big] \times$$
$$\mathbb{E}_{\lambda,\mu}\Big[\Big(\prod_{j=r_1+1}^{r}\prod_{q_j\leq m_j}\prod_{e\in E_{\omega_{j,q_j},A}} A_e\Big) \mid \sigma\Big] \times \mathbb{E}_{\lambda,\mu}\Big[\prod_{j=r_1+1}^{r}\prod_{q_j\leq m_j}\prod_{e\in E_{\omega_{j,q_j},B}} B_e \mid \sigma\Big]\Big].$$

To see that the difference between the quantities is again $o(1)$, first condition on the choice of σ, and consider the difference in the inner sums. Again, if two cycles overlap on m edges, we will gain a factor of $O(n^m)$, but we lose a factor of n^{m+1} for the number of such choices. Thus the dominant contribution again comes from the non-overlapping cycles. The proof from $(k_{r_1+1}, l_{r_1+1}) = (1,1)$ follows analogously. ∎

Armed with Lemma 8, we turn to a proof of Proposition 5. Note that under $\mathbb{P}_{\lambda,\mu}$, $B_{ij} = X_{ij} + Z_{ij}$, where $X_{ij} = \sqrt{\frac{\mu}{n}} \cdot \sigma_i u_j$. Thus we have,

$$\begin{aligned}
Y_{n,k,l} &= \frac{1}{n^l}\sum_{\omega}\prod_{e_1\in E_{\omega,A}} A_{e_1}\prod_{e_2\in E_{\omega,B}} B_{e_2} \\
&= \frac{1}{n^l}\sum_{\omega}\prod_{e_1\in E_{\omega,A}} A_{e_1}\prod_{e_2\in E_{\omega,B}} (X_{e_2}+Z_{e_2}) \\
&= T_{1,k,l} + T_{2,k,l} + T_{3,k,l},
\end{aligned}$$

where

$$T_{1,k,l} = \frac{1}{n^l} \sum_{\omega} \prod_{e_1 \in E_{\omega,A}} A_{e_1} \prod_{e_2 \in E_{\omega,B}} Z_{e_2}, \tag{5}$$
$$T_{2,k,l} = \frac{1}{n^l} \sum_{\omega} \prod_{e_1 \in E_{\omega,A}} A_{e_1} \prod_{e_2 \in E_{\omega,B}} X_{e_2},$$
$$T_{3,k,l} = Y_{n,k,l} - T_{1,k,l} - T_{2,k,l}.$$

We will establish the following lemmas.

Lemma 9 *In the setting of Proposition 5, for $k \geq l \geq 1$,*

$$T_{1,k,l} - p\mathbf{1}_{k=l=1} \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{2k}\binom{k}{l}\frac{d^{k-l}}{\gamma^l}\Big).$$

under $\mathbb{P}_{\lambda,\mu}$. Further, the limiting random variables are asymptotically independent.

Lemma 10 *In the setting of Proposition 5, for $k \geq l \geq 1$,*

$$T_{2,k,l} \xrightarrow{P} \frac{1}{2k}\binom{k}{l}\Big(\frac{\mu}{\gamma}\Big)^l (\lambda\sqrt{d})^{k-l}.$$

Lemma 11 *In the setting of Proposition 5, for $k \geq l \geq 1$, $T_{3,k,l} \xrightarrow{P} 0$.*

The proof of Proposition 5 is straightforward from these intermediate lemmas.
Proof First we note that Lemma 8 immediately implies that the terms Z_{n,k_i,l_i} with $i \leq r_1$ are asymptotically independent to the terms Z_{n,k_j,l_j}, with $j > r_1$. Moreover, the limiting distributions of $\{Y_{n,k,0} : k \geq 3\}$ under both H_0 and H_1 follows directly from (Mossel et al., 2015, Theorem 3.1). Thus, we have, as $n \to \infty$,

$$\mathbb{E}_{0,0}\Big[\prod_{j=1}^{r_1} Z_{n,k_j,0}^{m_j}\Big] \to \prod_{j=1}^{r_1} \mathbb{E}[\upsilon_{k_j,0,0}^{m_j}], \quad \mathbb{E}_{\lambda,\mu}\Big[\prod_{j=1}^{r_1} Z_{n,k_j,0}^{m_j}\Big] \to \prod_{j=1}^{r_1} \mathbb{E}[\upsilon_{k_j,0,1}^{m_j}].$$

Consequently, it suffices to analyze the terms with $l > 0$, and we show that in this case $Z_{n,k,l}$ will have asymptotically Gaussian limits under H_0 and H_1. We complete the proof under $\mathbb{P}_{\lambda,\mu}$. The null distribution follows by setting $\lambda = \mu = 0$.

Considering the cycles with $k \geq l \geq 1$, we use Lemma 9, 10 and 11 to conclude that the cycle statistics $Y_{n,k,l}$ have the desired gaussian limits under $\mathbb{P}_{\lambda,\mu}$. Further, Lemma 9 implies that these variables are asymptotically independent. This completes the proof. ∎

It remains to prove Lemmas 9, 10 and 11. We start with the proof of Lemma 9.
Proof We complete the proof via the following steps.

(i) Calculation of the mean and variance of $T_{1,k,l}$ under $\mathbb{P}_{\lambda,\mu}$.

(ii) Verification of Wick's formula.

(iii) Verification of asymptotic independence.

We address (i), and calculate the means and variances of the $T_{1,k,l}$ statistics. Note that the case $l = k$ corresponds to the cycle statistics in Banerjee and Ma (2018), and we can read off the expectations and variances directly. Consider the case $0 < l < k$. We have, using (5),

$$\mathbb{E}_{\lambda,\mu}[T_{1,k,l}] = 0.$$

Moving onto the variance, we have,

$$\begin{aligned}\mathbb{E}_{\lambda,\mu}[T_{1,k,l}^2] &= \frac{1}{n^{2l}} \sum_{\omega_1,\omega_2} \mathbb{E}_{\lambda,\mu}\Big[\Big(\prod_{e_1\in E_{\omega_1,A}} A_{e_1} \prod_{e_2\in E_{\omega_1,B}} Z_{e_2}\Big)\Big(\prod_{e_1\in E_{\omega_2,A}} A_{e_1} \prod_{e_2\in E_{\omega_2,B}} Z_{e_2}\Big)\Big] \\ &= \frac{1}{n^{2l}} \sum_{\omega} \mathbb{E}_{\lambda,\mu}\Big[\Big(\prod_{e_1\in E_{\omega,A}} A_{e_1} \prod_{e_2\in E_{\omega,B}} Z_{e_2}\Big)^2\Big] + \tilde{T}_1,\end{aligned} \tag{6}$$

where $\tilde{T}_1$ tracks the contribution from the pairs (ω_1, ω_2) with $\omega_1 \neq \omega_2$. First, observe that

$$\begin{aligned}&\frac{1}{n^{2l}} \sum_{\omega} \mathbb{E}_{\lambda,\mu}\Big[\Big(\prod_{e_1\in E_{\omega,A}} A_{e_1} \prod_{e_2\in E_{\omega,B}} Z_{e_2}\Big)^2\Big] \\ &= \frac{1}{n^{2l}} \sum_{\omega} \mathbb{E}_{\lambda,\mu}\Big[\prod_{e_1\in E_{\omega,A}} A_{e_1} \prod_{e_2\in E_{\omega,B}} Z_{e_2}^2\Big] \\ &= \frac{1}{n^{2l}} \sum_{\omega} \mathbb{E}_{\sigma}\Big[\prod_{e_1\in E_{\omega,A}} \Big(\frac{d + \lambda\sqrt{d}\,\sigma_{e_1^-}\sigma_{e_1^+}}{n}\Big)\Big] \\ &= \frac{1}{n^{l+k}} \sum_{\omega} \mathbb{E}_{\sigma}\Big[\sum_{z_e\in\{d,\lambda\sqrt{d}\sigma_{e^-}\sigma_{e^+}\}:e\in E_{\omega,A}} \prod_{e\in E_{\omega,A}} z_e\Big] \\ &= \frac{1}{n^{2l}} \sum_{\omega} \Big(\frac{d}{n}\Big)^{k-l}.\end{aligned}$$

where the last step follows from the obervation that ω has $k - l$ A-edges, and that only terms which contribute d for each edge survive in the limit. Recall that the number of k cycles with l B-wedges is $\frac{1}{2k}\binom{k}{l} n^k p^l$. This implies

$$\begin{aligned}&\frac{1}{n^{2l}} \sum_{\omega} \mathbb{E}_{\lambda,\mu}\Big[\Big(\prod_{e_1\in E_{\omega,A}} A_{e_1} \prod_{e_2\in E_{\omega,B}} Z_{e_2}\Big)^2\Big] = \frac{1}{n^{2l}} \sum_{\omega} \Big(\frac{d}{n}\Big)^{k-l} = \frac{1}{n^{2l}} \cdot \Big(\frac{d}{n}\Big)^{k-l} \cdot \frac{1}{2k} n^k \binom{k}{l} p^l \\ &= \frac{1}{2k}\binom{k}{l} \frac{d^{k-l}}{\gamma^l}(1 + o(1)).\end{aligned}$$

We will next establish that this is the dominant term in the asymptotic variance, and that $\tilde{T}_1 \to 0$ as $n \to \infty$. Note that a product term corresponding to (ω_1, ω_2) with $\omega_1 \neq \omega_2$ has a non-zero contribution provided they share exactly the same B-wedges. For any two such cycles (ω_1, ω_2), suppose they share α_1 A edges. Note that $\omega_1 \neq \omega_2$, and thus $0 \leq \alpha_1 < k-l$.

As ω_1 and ω_2 cannot differ on exactly one edge, in fact, this implies $\alpha_1 \le k-l-2$. Setting $\alpha_2 = k-l-\alpha_1$, we have,

$$\begin{aligned}\tilde{T}_1 &= \frac{1}{n^{2l}} \sum_{\omega_1 \neq \omega_2} \mathbb{E}_{\lambda,\mu}\Big[\Big(\prod_{e_1 \in E_{\omega_1,A}} A_{e_1} \prod_{e_2 \in E_{\omega_1,B}} Z_{e_2}\Big)\Big(\prod_{e_1 \in E_{\omega_2,A}} A_{e_1} \prod_{e_2 \in E_{\omega_2,B}} Z_{e_2}\Big)\Big] \\ &= \frac{1}{n^{2l}} \sum_{\alpha_2=2}^{k-l} \sum_{|E_{\omega_1,A} \cap E_{\omega_2,A}| = k-l-\alpha_2} \Big(\frac{d}{n}\Big)^{k-l+\alpha_2}.\end{aligned}$$

It remains to count the number of pairs (ω_1, ω_2) with $|E_{\omega_1,A} \cap E_{\omega_2,A}| = k-l-\alpha_2$ for all $2 \le \alpha_2 \le k-l$. We derive a rough upper bound to the number of such pairs as follows: there are $O(n^k p^l)$ choices for the first cycle, and $O(n^{\alpha_2 - 1})$ choices for the second cycle ω_2, given ω_1. Thus we bound the number of such pairs as $C(k,\gamma) n^{k+l+\alpha_2-1}$, where $C(k,\gamma) > 0$ is independent of n. Plugging in this bound, we obtain

$$\tilde{T}_1 \le C(k,\gamma) \frac{1}{n^{2l}} \sum_{\alpha_2=2}^{k-l} n^{k+l+\alpha_2-1} \cdot \Big(\frac{d}{n}\Big)^{k-l+\alpha_2} = O\Big(\frac{1}{n}\Big).$$

This controls $\tilde{T}_1$, and establishes the variance under $\mathbb{P}_{\lambda,\mu}$.

We next turn to (ii). We want to show that $T_{1,k,l}$ are asymptotically Gaussian by analyzing their moments (3). This is done by checking that the limits of the moments satisfy Wick's formula. We will perform the calculations assuming that $(k_{r_1+1}, l_{r_1+1}) \neq (1,1)$; the same calculations hold in the case of equality and hence are omitted. Formally, we show that for $W_{ni} \in \{T_{1,k_{r_1+1},l_{r_1+1}}, \cdots, T_{1,k_r,l_r}\}$, $i \in [m]$, we have that

$$\lim_{n\to\infty} \mathbb{E}[W_{n1} \cdots W_{nm}] = \begin{cases} \sum_\eta \prod_{i=1}^{m/2} \mathbb{E}[W_{n\eta(i,1)} W_{n\eta(i,2)}] + o(1) & \text{if } m \text{ even} \\ o(1) & \text{o.w.} \end{cases} \tag{7}$$

where η is a partition of $[m]$ into $\frac{m}{2}$ blocks of size two, and $\eta(i,j)$ denotes the j-th element of the i-th block, where $j \in \{1,2\}$. Wick's formula (Wick (1950)) then implies that the limiting distribution must be Gaussian, as long as the limits of $\mathbb{E}[W_{n\eta(i,1)} W_{n\eta(i,2)}]$ exist.

For a choice of $W_{n1}, \cdots, W_{nm}$, let $\omega_{1:m}$ be a collection of cycles $\omega_1, ..., \omega_m$, such that ω_i is of length k_i, with l_i B wedges and x_i contiguous blocks of A type edges and B type wedges. Note then that

$$\begin{aligned}&\mathbb{E}_{\lambda,\mu}[W_{n1} \cdots W_{nm}] \\ &= \mathbb{E}_{\lambda,\mu}\Big[n^{-\sum_i l_i} \sum_{\omega_{1:m}} \prod_{i \le m} \prod_{e_1 \in E_{\omega_i,A}} A_{e_1} \prod_{e_2 \in E_{\omega_i,B}} Z_{e_2}\Big] \\ &= n^{-\sum_i l_i} \sum_{\omega_{1:m}} \mathbb{E}_{\lambda,\mu}\Big[\prod_{i\le m} \prod_{e_1 \in E_{\omega_i,A}} A_{e_1}\Big] \mathbb{E}_{\lambda,\mu}\Big[\prod_{i \le m} \prod_{e_2 \in \tilde{E}_{\omega_i}} Z_{e_2}\Big]\end{aligned} \tag{8}$$

where $\tilde{E}_{\omega_i}$ are the edges of $\tilde{G}_{\omega_i} = G_{\omega_i}/G_{\omega_i,A}$, the quotient graph where for each $j \le x_i$, G_{ω_i,α_j}, the graph of the jth A block in ω_i, is identified as a vertex of $\tilde{G}_{\omega_i}$. We denote the vertices of $\tilde{G}_{\omega_i}$ as $\tilde{V}_{\omega_i} = \tilde{V}^1_{\omega_i} \cup \tilde{V}^2_{\omega_i}$, where $\tilde{V}^1$ are the vertices inherited from G_{ω_i}, and $\tilde{V}^2$

are the vertices produced by the quotient operator. Among $\tilde{V}^2$, we define the following equivalence relationship: if the first and last vertices of G_{ω_i,α^i_j} and G_{ω_h,α^h_p} are the same, then we consider the vertices in $\tilde{G}_{\omega_i}$ and $\tilde{G}_{\omega_h}$, which correspond to the quotient image of G_{ω_i,α^i_j} and G_{ω_h,α^h_p} respectively, to be the same. In order for the contribution of $\omega_{1:m}$ to be non-zero, the B edges have to be included at least twice, and we will call such a collection $\tilde{G}_{\omega_{1:m}}$ a weak CLT sentence. Given a weak CLT sentence, we define a partition $\eta(\tilde{G}_{\omega_{1:m}})$ of $[m]$ as follows: i and j are in the same partition if $\tilde{G}_{\omega_i}$ and $\tilde{G}_{\omega_j}$ share at least one edge. As a result, we can express (8) as follows:

$$n^{-\sum_i l_i}\sum_{G_{\omega_{1:m},A}}\mathbb{E}_{\lambda,\mu}\Big[\prod_{e_1\in E_{\omega_{1:m},A}}A_{e_1}\Big]\sum_{\eta}\sum_{\substack{\tilde{G}_{\omega_{1:m}}\\ \eta(\tilde{G}_{\omega_{1:m}})=\eta}}\mathbb{E}_{\lambda,\mu}\Big[\prod_{e_2\in\tilde{E}_{\omega_{1:m}}}Z_{e_2}\Big].$$

Let t be the total number of vertices of $\tilde{G}_{\omega_{1:m}}$. Let us consider the case where $\eta(\tilde{G}_{\omega_{1:m}})$ contains strictly less than $\frac{m}{2}$ blocks, which includes all cases when m is odd. In this case (Anderson and Zeitouni, 2006, Lemma 4.10) implies that $t<\sum_i l_i$. Following the proof of (Banerjee and Ma, 2018, Lemma 3), we note that the number of weak CLT sentences summed over is bounded by

$$O\Big(\sum_i l_i\Big)^{O(\sum_i l_i)}n^{t-\sum_i x_i}.$$

This comes from the fact that $\sum_i x_i$ of the vertices are automatically fixed from the quotient operator. As a result, for a particular partition η, where $|\eta|<\frac{m}{2}$, we note that:

$$\begin{aligned}
&n^{-\sum_i l_i}\sum_{G_{\omega_{1:m},A}}\mathbb{E}_{\lambda,\mu}\Big[\prod_{e_1\in E_{\omega_{1:m},A}}A_{e_1}\Big]\sum_{\substack{\tilde{G}_{\omega_{1:m}}\\ \eta(\tilde{G}_{\omega_{1:m}}=\eta)}}\mathbb{E}_{\mu,\nu}\Big[\prod_{e_2\in\tilde{E}_{\omega_{1:m}}}Z_{e_2}\Big]\\
&=n^{-\sum_i l_i}\sum_{G_{\omega_{1:m},A}}\mathbb{E}_{\lambda,\mu}\Big[\prod_{e_1\in E_{\omega_{1:m},A}}A_{e_1}\Big]\sum_{\substack{\tilde{G}_{\omega_{1:m}}\\ \eta(\tilde{G}_{\omega_{1:m}}=\eta)}}O(1)^{O(\sum_i l_i)}\\
&=n^{-\sum_i l_i}O\Big(\sum_i l_i\Big)^{O(\sum_i l_i)}n^{t-\sum_i x_i}\sum_{G_{\omega_{1:m},A}}\mathbb{E}_{\lambda,\mu}\Big[\prod_{e_1\in E_{\omega_{1:m},A}}A_{e_1}\Big]\\
&=n^{-\sum_i l_i}O\Big(\sum_i l_i\Big)^{O(\sum_i l_i)}n^{t-\sum_i x_i}O\Big(\frac{1}{n}\Big)^{\sum_i\sum_{j\le x_i}2\alpha^i_j}O(n)^{\sum_i\sum_{j\le x_i}(2\alpha^i_j+1)}\\
&=O\Big(\sum_i l_i\Big)^{O(\sum_i l_i)}O(n)^{t-\sum_i l_i}.
\end{aligned}$$

The penultimate line holds because for an A block of length $2\alpha^i_j$, there are $2\alpha^i_j+1$ vertices, for which there are $O(n)$ options each. Thus the total number of choices for $E_{\omega_{1:m},A}$ is $O(n)^{\sum_i\sum_{j\le x_i}(2\alpha^i_j+1)}$. Since $t<\sum_i l_i$, we see that the contribution of such a term is $o(1)$.

We have thus shown that the leading order term of (8) consists of weak CLT sentences whose partition $\eta(\tilde{G}_{\omega_{1:m}})$ has exactly $\frac{m}{2}$ blocks. Note that this automatically implies that

(8) is $o(1)$ if m is odd, and that for m even, the leading order weak CLT sentences have partitions η with only blocks of size two, which is exactly what we need for (7). From our calculations in step (i), we see that the variances of $Z_{n,k,l}$ under the null are as we claimed.

Finally, we verify step (iii), that for $l \geq 1$, $Z_{n,k,l}$ are asymptotically independent. From the discussion above (7), this amounts to checking that for $W_{ni} \neq W_{nj}$, $\mathbb{E}_{\lambda,\mu}[W_{ni}W_{nj}] \xrightarrow{P} 0$. Note that this expectation equals 0 if $l_i \neq l_j$. Thus it suffices to consider the case $l_i = l_j$, $k_i < k_j$ and we have:

$$\mathbb{E}_{\lambda,\mu}[W_{ni}W_{nj}] = n^{-2l_i} \sum_{\omega_i,\omega_j} \mathbb{E}_{\lambda,\mu}\left[\prod_{e_1 \in E_{\omega_i}} A_{e_1} \prod_{e_1 \in E_{\omega_j}} A_{e_1}\right]$$

where the sum is taken over ω_i, ω_j that intersect on all l_i B-wedges, which gives

$$\mathbb{E}_{\lambda,\mu}[W_{ni}W_{nj}] \leq C(k_i, d, \gamma) n^{-2l_i} \left(\frac{1}{n}\right)^{k_i+k_j-2l_i} n^{k_i+k_j-l_i-1} = o(1).$$

This concludes the demonstration of Wick's formula under $\mathbb{P}_{\lambda,\mu}$, and hence proves the desired convergence in distribution.

■

We turn next to the proof of Lemma 10.

Proof For any $1 \leq i_1 < \cdots < i_k \leq n$, we let $\mathbf{i}_{1:k} = (i_1, \cdots, i_k)$. Similarly, for $1 \leq j_1 < \cdots < j_l \leq p$, set $\mathbf{j}_{1:l} = (j_1, \cdots, j_p)$. Finally, given $\mathbf{i}_{1:k}$ and $\mathbf{j}_{1:l}$, let $\mathcal{C}(\mathbf{i}_{1:k}, \mathbf{j}_{1:l})$ denote the set of cycles with $k-l$ A-edges and l B-wedges on the chosen vertices. Armed with this notation, we observe that

$$\begin{aligned} T_{2,k,l} &= \frac{1}{n^l} \sum_{\omega} \prod_{e_1 \in E_{\omega,A}} A_{e_1} \prod_{e_2 \in E_{\omega,B}} X_{e_2} \\ &= \frac{1}{n^l} \sum_{\mathbf{i}_{1:k}, \mathbf{j}_{1:l}} \sum_{\omega \in \mathcal{C}(\mathbf{i}_{1:k}, \mathbf{j}_{1:l})} \prod_{e_1 \in E_{\omega,A}} A_{e_1} \prod_{e_2 \in E_{\omega,B}} X_{e_2}. \end{aligned}$$

Given $\mathbf{i}_{1:k}$, let $\mathcal{C}(\mathbf{i}_{1:k})$ denote all length k cycles on the vertices $i_1, i_2, \cdots, i_k$, with $(k-l)$ edges colored to be of type A, and the remaining edges colored to be of type B. For any edge in the cycle, let $t(e) \in \{A, B\}$ denote its type. This implies

$$T_{2,k,l} = \frac{1}{n^l} \Big(\sum_{\mathbf{i}_{1:k}} \sum_{\omega \in \mathcal{C}(\mathbf{i}_{1:k})} \prod_{t(e)=A} A_e \prod_{t(e)=B} \sigma_{e^-}\sigma_{e^+} \Big) \Big(\Big(\frac{\mu}{n}\Big)^l \sum_{\mathbf{j}_{1:l}} \prod_{h=1}^{l} u_{j_h}^2 \Big).$$

Note that for fixed $\mathbf{j}_{1:l}$, u_{jh}^2 are independent with mean 1, so law of large numbers gives

$$\frac{1}{n^l} \sum_{\mathbf{j}_{1:l}} \prod_{h=1}^{l} u_{j_h}^2 \xrightarrow{P} 1.$$

Thus it suffices to control the other term. Observe that

$$\begin{aligned}
&\mathbb{E}_{\lambda,\mu}\Big[\frac{1}{n^l}\Big(\sum_{\mathbf{i}_{1:k}}\sum_{\omega\in\mathcal{C}(\mathbf{i}_{1:k})}\prod_{t(e)=A}A_e\prod_{t(e)=B}\sigma_{e^-}\sigma_{e^+}\Big)\Big]\\
=&\mathbb{E}_{\sigma}\Big[\frac{1}{n^l}\Big(\sum_{\mathbf{i}_{1:k}}\sum_{\omega\in\mathcal{C}(\mathbf{i}_{1:k})}\prod_{t(e)=A}\Big(\frac{d+\lambda\sqrt{d}\,\sigma_{e^-}\sigma_{e^+}}{n}\Big)\prod_{t(e)=B}\sigma_{e^-}\sigma_{e^+}\Big)\Big]\\
=&\mathbb{E}_{\sigma}\Big[\frac{1}{n^k}\Big(\sum_{\mathbf{i}_{1:k}}\sum_{\omega\in\mathcal{C}(\mathbf{i}_{1:k})}\sum_{z_e\in\{d,\lambda\sqrt{d}\,\sigma_{e^-}\sigma_{e^+}\}:t(e)=A}\prod_{t(e)=A}z_e\prod_{t(e)=B}\sigma_{e^-}\sigma_{e^+}\Big)\Big].
\end{aligned}$$

Observe that we have a non-zero contribution in the sum above if and only if $z_e = \lambda\sqrt{d}\sigma_{e^-}\sigma_{e^+}$ for all e such that $t(e) = A$; in this case, each term contributes $(\lambda\sqrt{d})^{k-l}$. It remains to count the number of length k cycles with l B-wedges. Any such cycle has k-vertices of type A, and this choice can be done in n^k ways. The positions of the B-wedges can be chosen in $\binom{k}{l}$ ways. The B-vertices on the B-wedges can be chosen in p^l ways. Finally, we divide this count by $2k$ to account for overcounting due to cyclic shifts. Thus, upon simplification,

$$\mathbb{E}_{\lambda,\mu}\Big[\frac{1}{n^l}\Big(\sum_{\mathbf{i}_{1:k}}\sum_{\omega\in\mathcal{C}(\mathbf{i}_{1:k})}\prod_{t(e)=A}A_e\prod_{t(e)=B}\sigma_{e^-}\sigma_{e^+}\Big)\Big] = (1+o(1))\frac{1}{2k}\binom{k}{l}\Big(\frac{\mu}{\gamma}\Big)^l(\lambda\sqrt{d})^{k-l}.$$

Next, we show that

$$\mathrm{Var}_{\lambda,\mu}\Big(\frac{1}{n^l}\Big(\sum_{\mathbf{i}_{1:k}}\sum_{\omega\in\mathcal{C}(\mathbf{i}_{1:k})}\prod_{t(e)=A}A_e\prod_{t(e)=B}\sigma_{e^-}\sigma_{e^+}\Big)\Big) = o(1)$$

as $n \to \infty$. This amounts to checking that

$$\begin{aligned}
&\mathbb{E}_{\lambda,\mu}\left[\frac{1}{n^{2l}}\Big(\sum_{\mathbf{i}_{1:k}}\sum_{\omega\in\mathcal{C}(\mathbf{i}_{1:k})}\prod_{t(e)=A}A_e\prod_{t(e)=B}\sigma_{e^-}\sigma_{e^+}\Big)^2\right]\\
&=\mathbb{E}_{\lambda,\mu}\left[\frac{1}{n^{l}}\Big(\sum_{\mathbf{i}_{1:k}}\sum_{\omega\in\mathcal{C}(\mathbf{i}_{1:k})}\prod_{t(e)=A}A_e\prod_{t(e)=B}\sigma_{e^-}\sigma_{e^+}\Big)\right]^2 + o(1).
\end{aligned}$$

This turns out to be true because of the same argument for Lemma 8 (specifically the discussion after (4)). In conclusion, this establishes that

$$T_{2,k,l} \xrightarrow{P} \frac{1}{2k}\binom{k}{l}\Big(\frac{\mu}{\gamma}\Big)^l(\lambda\sqrt{d})^{k-l}.$$

This completes the proof. ■

Finally, we complete the proof of Lemma 11.

Proof Note that $\mathbb{E}_{\lambda,\mu}[T_{3,k,l}] = 0$, and thus it suffices to establish that $\mathbb{E}_{\lambda,\mu}[T_{3,k,l}^2] = o(1)$ as $n \to \infty$. Now, we can express $T_{3,k,l} = \sum_\omega V_{n,k,l,\omega}$, where

$$V_{n,k,l,\omega} = \frac{1}{n^l} \prod_{e \in E_{\omega,A}} A_e \sum_{E_{\omega,f} \subsetneq E_{\omega,B}} \prod_{e \in E_{\omega,f}} X_e \prod_{e \in E_{\omega,B} \setminus E_{\omega,f}} Z_e.$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\lambda,\mu}[T_{3,k,l}^2] &= \sum_{\omega_1,\omega_2} \mathbb{E}_{\lambda,\mu}[V_{n,k,l,\omega_1} V_{n,k,l,\omega_2}] \\ &:= \sum_{\omega_1,\omega_2} \sum_{E_{\omega_1,f} \subsetneq E_{\omega_1,B}, E_{\omega_2,f} \subsetneq E_{\omega_2,B}} \mathbb{E}_{\lambda,\mu}[V_{n,k,l,\omega_1,E_{f,\omega_1}} V_{n,k,l,\omega_2,E_{f,\omega_2}}],\end{aligned}$$

where we define

$$V_{n,k,l,\omega,E_{f,\omega}} := \frac{1}{n^l} \prod_{e \in E_{\omega,A}} A_e \prod_{e \in E_{\omega,f}} X_e \prod_{e \in E_{\omega,B} \setminus E_{\omega,f}} Z_e.$$

This implies $\mathbb{E}_{\lambda,\mu}[V_{n,k,l,\omega_1,E_{f,\omega_1}} V_{n,k,l,\omega_2,E_{f,\omega_2}}]$ is zero unless $E_{\omega_1,B} \setminus E_{\omega_1,f} = E_{\omega_2,B} \setminus E_{\omega_2,f}$. Given ω_1, ω_2, the terms which affect the contribution by powers of n are the edges in $E_{\omega_1,A} \cap E_{\omega_2,A}$. The dominant contribution arises from ω_1, ω_2 such that $|E_{\omega_1,A} \cap E_{\omega_2,A}| = 0$—this follows using the same reasoning used to identify the dominant order of the variance in the proof of Lemma 9. Further, note that overlaps in the edges in $E_{\omega_1,f}$ and $E_{\omega_2,f}$ affect the expectation, but only to constant order. For a pair (ω_1, ω_2) satisfying these conditions

$$\mathbb{E}_{\lambda,\mu}[V_{n,k,l,\omega_1,E_{f,\omega_1}} V_{n,k,l,\omega_2,E_{f,\omega_2}}] = \frac{1}{n^{2l}} \mathbb{E}_{\lambda,\mu}\Big[\prod_{e \in E_{\omega_1,A}} A_e \prod_{e \in E_{\omega_2,A}} A_e \prod_{e \in E_{\omega_1,f}} X_e \prod_{e \in E_{\omega_2,f}} X_e \Big].$$

As ω_1, ω_2 are both length k cycles with $k - l$ A edges and l B-edges, and $E_{\omega_1,B} \setminus E_{\omega_1,f} = E_{\omega_2,B} \setminus E_{\omega_2,f}$, we have $|E_{\omega_1,f}| = |E_{\omega_2,f}| := x$. In turn, this implies that there exists $C := C(k,l) > 0$ such that

$$\mathbb{E}_{\lambda,\mu}[V_{n,k,l,\omega_1,E_{f,\omega_1}} V_{n,k,l,\omega_2,E_{f,\omega_2}}] \le C \frac{1}{n^{2l+x}} \mathbb{E}_{\lambda,\mu}\Big[\prod_{e \in E_{\omega_1,A}} A_e \prod_{e \in E_{\omega_2,A}} A_e \Big] \le C' \frac{1}{n^{2k+x}},$$

where $C' := C'(k,l,\lambda,d) > 0$ is a constant independent of n. There are only finitely many choices of $E_{\omega_1,f}$ and $E_{\omega_2,f}$, and therefore, for each ω_1, ω_2,

$$\sum_{E_{\omega_1,f} \subsetneq E_{\omega_1,B}, E_{\omega_2,f} \subsetneq E_{\omega_2,B}} \mathbb{E}_{\lambda,\mu}[V_{n,k,l,\omega_1,E_{f,\omega_1}} V_{n,k,l,\omega_2,E_{f,\omega_2}}] \le C'' \frac{1}{n^{2k+x}},$$

for a larger constant C''. Finally, we sum over ω_1, ω_2. If the two cycles intersect on x edges, they have $x+1$ vertices in common. Thus the number of pairs ω_1, ω_2 with x common edges is $O(n^{2k+2l-x-1})$. Summing, we have the conclusion that $\mathbb{E}_{\lambda,\mu}[T_{3,k,l}^2] = o(1)$.

■

4. Weak Recovery under the Threshold

In the remaining two sections, we will prove Theorem 3. In this section, we show the first part, that weak recovery is impossible when $\lambda^2+\frac{\mu^2}{\gamma}<1$. We follow the general proof scheme in Banerjee (2018). The proof is information theoretic, and the main idea is contained in the following proposition.

Proposition 12 *When $\lambda^2+\frac{\mu^2}{\gamma}<1$, then for any fixed r, and any two configurations $(\sigma_1,...,\sigma_r),(\tau_1,...,\tau_r)\in\{\pm1\}^r$, we have that as $n\to\infty$,*

$$\|\mathbb{P}_{\lambda,\mu}(\cdot\mid\sigma_{1:r})-\mathbb{P}_{\lambda,\mu}(\cdot\mid\tau_{1:r})\|_{\mathrm{TV}}\to 0.$$

Proof The idea is to bound the total variation with a function of the second moment of the likelihood ratios,

$$L_{\sigma,n}=\frac{\mathrm{d}\mathbb{P}_{\lambda,\mu}}{\mathrm{d}\mathbb{P}_{0,0}}(\cdot\mid\sigma_{1:r})\ ,\ L_{\tau,n}=\frac{\mathrm{d}\mathbb{P}_{\lambda,\mu}}{\mathrm{d}\mathbb{P}_{0,0}}(\cdot\mid\tau_{1:r}),$$

and then noting that because r is fixed, the second moments of these likelihood ratios will converge to a limit independent of $\sigma_{1:r}$ and $\tau_{1:r}$, making the upper bound converge to 0. However, as we have seen in the other arguments, we need some truncation. Let us define the truncated likelihood ratios

$$\tilde{L}_{\sigma,n}=\frac{\mathbb{E}_{\sigma_{-r},u}\left[\mathbb{P}_{\lambda,\mu}\left(\mathbf{A},\mathbf{B}\mid\sigma_{1:r},\sigma_{-r},u\right)\mathbf{1}(u\in\mathcal{S})\right]}{\mathbb{P}_{0,0}\left(\mathbf{A},\mathbf{B}\right)}$$
$$\tilde{L}_{\tau,n}=\frac{\mathbb{E}_{\tau_{-r},u}\left[\mathbb{P}_{\lambda,\mu}\left(\mathbf{A},\mathbf{B}\mid\tau_{1:r},\tau_{-r},u\right)\mathbf{1}(u\in\mathcal{S})\right]}{\mathbb{P}_{0,0}\left(\mathbf{A},\mathbf{B}\right)},$$

where $\sigma_{-r},\tau_{-r}\in\{\pm1\}^{n-r}$ are the other coordinates, and $\mathcal{S}=\{\|u\|\leq 2\sqrt{p}\}$. Define distributions given by these truncated likelihood ratios:

$$\mathbb{Q}_{\sigma,n}(\Omega\mid\sigma_{1:r})=\frac{1}{\mathbb{P}_n\left(\mathcal{S}\right)}\mathbb{E}_{0,0}\left[\tilde{L}_{\sigma,n}\,\mathbb{1}(\Omega)\mid\sigma_{1:r}\right]$$
$$\mathbb{Q}_{\tau,n}(\Omega\mid\tau_{1:r})=\frac{1}{\mathbb{P}_n\left(\mathcal{S}\right)}\mathbb{E}_{0,0}\left[\tilde{L}_{\tau,n}\,\mathbb{1}(\Omega)\mid\tau_{1:r}\right].$$

From the proof of (Banerjee and Ma, 2018, Proposition 1), we know that $\|\mathbb{P}_{\lambda,\mu}\left(\cdot\mid\sigma_{1:r}\right)-\mathbb{Q}_{\sigma,n}(\cdot\mid\sigma_{1:r})\|_{\mathrm{TV}}$ and $\|\mathbb{P}_{\lambda,\mu}\left(\cdot\mid\tau_{1:r}\right)-\mathbb{Q}_{\tau,n}(\cdot\mid\tau_{1:r})\|_{\mathrm{TV}}$ both vanish as $n\to\infty$. Thus, to prove the proposition, it suffices to check that $\|\mathbb{Q}_{\sigma,n}(\cdot\mid\sigma_{1:r})-\mathbb{Q}_{\tau,n}(\cdot\mid\tau_{1:r})\|_{\mathrm{TV}}\to 0$. Note that

$$\|\mathbb{Q}_{\sigma,n}(\cdot\mid\sigma_{1:r})-\mathbb{Q}_{\tau,n}(\cdot\mid\tau_{1:r})\|_{TV}=\frac{1}{\mathbb{P}_n\left(\mathcal{S}\right)}\mathbb{E}_{0,0}\left[|\tilde{L}_{\sigma,n}-\tilde{L}_{\tau,n}|\right]\leq\frac{1}{\mathbb{P}_n\left(\mathcal{S}\right)}\mathbb{E}_{0,0}\left[(\tilde{L}_{\sigma,n}-\tilde{L}_{\tau,n})^2\right]^{\frac{1}{2}},$$

where we have used the Cauchy Schwarz inequality. Now, we have,

$$\begin{aligned}
&\mathbb{E}_{0,0}\left[(\tilde{L}_{\sigma,n}-\tilde{L}_{\tau,n})^2\right]\\
&=\mathbb{E}_{0,0}\Bigg[\frac{1}{\mathbb{P}_{0,0}\left(\mathbf{A},\mathbf{B}\right)^2}\,\mathbb{E}_{\sigma_{-r},\tau_{-r},u,v}\Big\{\big(\mathbb{P}_{\lambda,\mu}\left(\mathbf{A},\mathbf{B}\mid\sigma_{1:r},\sigma_{-r},u\right)\mathbb{P}_{\lambda,\mu}\left(\mathbf{A},\mathbf{B}\mid\sigma_{1:r},\tau_{-r},v\right)\\
&\qquad+\mathbb{P}_{\lambda,\mu}\left(\mathbf{A},\mathbf{B}\mid\tau_{1:r},\sigma_{-r},u\right)\mathbb{P}_{\lambda,\mu}\left(\mathbf{A},\mathbf{B}\mid\tau_{1:r},\tau_{-r},v\right)\\
&\qquad-2\mathbb{P}_{\lambda,\mu}\left(\mathbf{A},\mathbf{B}\mid\sigma_{1:r},\sigma_{-r},u\right)\mathbb{P}_{\lambda,\mu}\left(\mathbf{A},\mathbf{B}\mid\tau_{1:r},\tau_{-r},v\right)\big)\mathbf{1}(u,v\in\mathcal{S})\Big\}\Bigg].
\end{aligned}$$

Thus we just have to prove that the quantity

$$\mathbb{E}_{0,0}\left[\frac{\mathbb{P}_{\lambda,\mu}\left(\mathbf{A},\mathbf{B}\mid\sigma_{1:r},\sigma_{-r},u\right)\mathbb{P}_{\lambda,\mu}\left(\mathbf{A},\mathbf{B}\mid\tau_{1:r},\tau_{-r},v\right)}{\mathbb{P}_{0,0}\left(\mathbf{A},\mathbf{B}\right)^2}\mathbf{1}(u,v\in\mathcal{S})\right]$$

has a limit which is independent of $\sigma_{1:r}$ and $\tau_{1:r}$. But we know that this is true because of the second moment calculations in Section 2, so we are done. ■

Then the impossibility of reconstruction follows from some technical calculations. The proof of the next two results follow directly from Proposition 6.2 and Theorem 2.2 of Banerjee (2018) respectively.

Proposition 13 *Let $\lambda^2+\frac{\mu^2}{\gamma}<1$. Let $S\subset[n]$ be such that $|S|=r$, with r finite and fixed, and let $u\in[n]$ be a single index such that $u\notin S$. Then, as $n\to\infty$, we have that*

$$\mathbb{E}_{\lambda,\mu}[\|\mathbb{P}_{\lambda,\mu}(\sigma_u\mid\mathbf{A},\mathbf{B},\sigma_S)-\mathbb{P}_{0,0}(\sigma_u)\|_{\mathrm{TV}}\mid\sigma_S]\to 0.$$

Theorem 14 *If $\lambda^2+\frac{\mu^2}{\gamma}<1$, then reconstruction is impossible, i.e. let the overlap be defined as*

$$ov(\sigma,\tau)=\frac{1}{n}\sum_{i=1}^{n}\sigma_i\tau_i-\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}\tau_i\right),$$

then for any estimator $\hat\sigma(\mathbf{A},\mathbf{B})\in\{\pm1\}^n$, we have that

$$ov(\sigma,\hat\sigma)\xrightarrow{p}0.$$

Because $\frac{1}{n}\sum_{i=1}^{n}\sigma_i\xrightarrow{p}0$, and $\frac{1}{n}\sum_{i=1}^{n}\hat\sigma_i$ is bounded, we see that $\frac{1}{n}\langle\sigma,\hat\sigma\rangle\xrightarrow{P}0$, so weak recovery is impossible.

5. Weak Recovery with Self Avoiding Walks

In this section we finish the proof of Theorem 3, by showing that weak recovery is possible whenever $\lambda^2+\frac{\mu^2}{\gamma}>1$. Because weak recovery is possible as soon as either $\lambda^2>1$ (Mossel et al. (2013); Massoulié (2014)) or $\frac{\mu^2}{\gamma}>1$ (Baik et al. (2005)), we only need to consider the case that λ^2, $\frac{\mu^2}{\gamma}<1$. We will construct an estimator $\hat\sigma$ that is computable in quasi-polynomial time. We use a strategy introduced in Hopkins and Steurer (2017) in the context of community detection in the block model. Under this approach, one seeks to design an appropriate set of "low-degree" polynomials in the data $(\mathbf{A},\mathbf{B})$, and recover the signals based on these polynomials. We note that in the specific context of community detection, this approach was already latent in the approach of Massoulié (2014) and Bordenave et al. (2015), based on self-avoiding/non-backtracking walks.

We seek to calculate a polynomial $P(\mathbf{A},\mathbf{B})$, which estimates $\sigma\sigma^{\mathrm{T}}$. Formally, suppose we had an estimator satisfying

$$\mathbb{E}_{\lambda,\mu}\Big[\langle P(\mathbf{A},\mathbf{B}),\sigma\sigma^{\mathrm{T}}\rangle\Big]\geq n\delta\mathbb{E}_{\lambda,\mu}\Big[\|P(\mathbf{A},\mathbf{B})\|_F^2\Big]^{\frac{1}{2}} \tag{9}$$

for some universal constant $\delta > 0$. Then (Hopkins and Steurer, 2017, Theorem 1) implies that there exists $\delta' = \delta'(\delta)$, and an estimator $\hat{\sigma}$ such that

$$\frac{1}{n^2}\mathbb{E}_{\lambda,\mu}[\langle \sigma, \hat{\sigma}\rangle^2] \geq \delta'.$$

This ensures weak recovery in our setting. To construct the estimator $\hat{\sigma}$, from the matrix $P(\mathbf{A}, \mathbf{B})$ constructed above, we compute a matrix Σ with minimum Frobenius norm that satisfies the following constraints:

$$\begin{aligned} \operatorname{diag}(\Sigma) &= \mathbb{1} \\ \frac{\langle P(\mathbf{A}, \mathbf{B}), \Sigma\rangle}{\|P(\mathbf{A}, \mathbf{B})\|_F \cdot n} &\geq \delta' \\ \Sigma &\succeq 0 \end{aligned}$$

and then output the vector $\hat{\sigma} \in \{\pm 1\}^n$ obtained by taking coordinate-wise signs of a centered Gaussian vector with covariance Σ.

Lemma 15 *The estimator $\hat{\sigma}$ achieves weak recovery whenever $\lambda^2 + \frac{\mu^2}{\gamma} > 1$.*

Proof The proof of weak recovery follows immediately from the proof of (Hopkins and Steurer, 2017, Lemma 3.5). ■

Remark 16 *This estimator takes $n^{O(\log n)/poly(\delta)}$ time to compute, as we use certain Self Avoiding Walks (SAWs) of length $\Theta(\log n)$. The running time can be improved to $n^{poly(1/\delta)}$ by the idea of color coding, and more discussions on this improvement can be found in (Hopkins and Steurer, 2017, Section 2.5).*

In the remainder of the section, we will construct an estimator and establish (9).

5.1 Self Avoiding Walks

It remains to construct the polynomial $P(\mathbf{A}, \mathbf{B})$. To this end, we will use self avoiding walks on the underlying factor-graph. First, for $i_1, i_2 \in [n]$ and $j \in [p]$, we define

$$\begin{aligned} \hat{A}_{i_1,i_2} &= \frac{2n}{a-b}\Big(A_{ij} - \frac{a+b}{2n}\Big), \\ \hat{B}^j_{i_1,i_2} &= \frac{n}{\mu}B_{i_1,j}B_{i_2,j} = \frac{n}{\mu}\Big(\sqrt{\frac{\mu}{n}}\sigma_{i_1}u_j + Z_{i_1,j}\Big)\Big(\sqrt{\frac{\mu}{n}}\sigma_{i_2}u_j + Z_{i_2,j}\Big). \end{aligned} \tag{10}$$

Direct computation yields that

$$\begin{aligned} \mathbb{E}_{\lambda,\mu}[\hat{A}_{i_1,i_2}|\sigma] = \sigma_{i_1}\sigma_{i_2}, \quad \operatorname{Var}_{\lambda,\mu}(A_{i_1,i_2}) = \frac{n}{\lambda^2}, \\ \mathbb{E}_{\lambda,\mu}[\hat{B}^j_{i_1,i_2}|\sigma] = \sigma_{i_1}\sigma_{i_2}, \quad \operatorname{Var}_{\lambda,\mu}(\hat{B}^j_{i_1,i_2}) = \frac{np}{\mu^2/\gamma}. \end{aligned}$$

Recall the factor graph, as shown in Figure 1. For $i_1, i_2 \in [n]$, we will associate the weight $\hat{A}_{i_1,i_2}$ to the A edge $\{i_1, i_2\}$. Similarly, for $j \in [p]$, we associate the weight $\hat{B}^j_{i_1} = \sqrt{\frac{\mu}{n}} B_{i_1,j}$ to the B edge $\{i_1, j\}$, and the weight $\hat{B}^j_{i_1,i_2}$ to the B wedge $\{i_1, j, i_2\}$.

Fix $i_1, i_2 \in [n]$, and let $k \geq l \geq 1$ be integers that we will specify later. Consider a path α on the factor graph that starts at i_1 and ends at i_2, which contains $k - l$ A type edges, and l B type wedges. We will require that the A-type edges on the path are "self-avoiding", i.e., no edge of type A occurs more than once. Further, if $j_1, \cdots, j_l$ denote the vertices in V_2 which lie on the path α, we will require that these vertices are distinct. Let $\mathcal{L}(i_1, i_2, k, l)$ denote the set of all such paths α, for any given $i_1, i_2 \in [n]$, and $k \geq l$. Given any path α, construct a polynomial on entries of $(\mathbf{A}, \mathbf{B})$ by

$$p_\alpha = \prod_{e \in \alpha} \text{weight}(e),$$

where weight(i_1, i_2) is $\hat{A}_{i_1,i_2}$ and $\hat{B}^j_{i_1.i_2}$ when the edge (i_2, i_2) is of type A and B respectively. Direct computation yields that

$$\mathbb{E}_{\lambda,\mu}[p_\alpha|\sigma] = \sigma_{i_1}\sigma_{i_2}, \quad \text{Var}_{\lambda,\mu}(p_\alpha) = \Big(\frac{n}{\lambda^2}\Big)^{k-l}\Big(\frac{np}{\mu^2/\gamma}\Big)^l (1 + o(1)). \tag{11}$$

Thus we see that p_α is an unbiased estimator for $\sigma_{i_1}\sigma_{i_2}$, but its large variance renders it useless on its own. Fortunately, there are many paths $\alpha \in \mathcal{L}(i_1, i_2, k, l)$, and we might hope that we can reduce the variance by averaging over polynomials from different paths as follows

$$P_{i_1,i_2}(\mathbf{A}, \mathbf{B}) = \frac{1}{|\mathcal{L}(i_1, i_2, k, l)|} \sum_{\mathcal{L}(i_1,i_2,k,l)} p_\alpha.$$

Note that P_{i_1,i_2} is still an unbiased estimator for $\sigma_{i_1}\sigma_{i_2}$. Finally, we set $P(\mathbf{A}, \mathbf{B}) = \{P_{i_1,i_2}(\mathbf{A}, \mathbf{B}) : 1 \leq i_1 < i_2 \leq n\}$ to be the estimator for the matrix $\sigma\sigma^\top$.
It remains to check (9) whenever $\lambda^2 + \frac{\mu^2}{\gamma} > 1$. We establish this in the next lemma.

Lemma 17 *Assume $d > 1$ and $\lambda^2 + \frac{\mu^2}{\gamma} > 1 + \varepsilon$, for some $\varepsilon > 0$. Then there exists universal constants $C > 0$ and $c > 0$ such that setting $k = C \log n/\varepsilon^c$ and $l := l(k, \lambda, \mu, \gamma) \leq k$ such that*

$$\mathbb{E}_{\lambda,\mu}\Big[\langle P(\mathbf{A}, \mathbf{B}), \sigma\sigma^{\mathrm{T}}\rangle\Big] \geq n\delta\mathbb{E}_{\lambda,\mu}\Big[\|P(\mathbf{A}, \mathbf{B})\|_F^2\Big]^{\frac{1}{2}} \tag{12}$$

for some $\delta := \delta(\varepsilon, C, c, \lambda, \mu, \gamma) > 0$.

To prove this lemma, we will show that an entry-wise version holds:

$$\mathbb{E}_{\lambda,\mu}\Big[P_{i_1,i_2}(\mathbf{A}, \mathbf{B}) \cdot \sigma_{i_1}\sigma_{i_2}\Big] \geq \delta\mathbb{E}_{\lambda,\mu}\Big[P_{i_1,i_2}(\mathbf{A}, \mathbf{B})^2\Big]^{\frac{1}{2}}. \tag{13}$$

By construction, we have that $\mathbb{E}_{\lambda,\mu}[P_{i_1,i_2}(\mathbf{A},\mathbf{B}) \mid \sigma_{i_1}\sigma_{i_2}] = \sigma_{i_1}\sigma_{i_2}$. As a result, it is easy to check that (13) is implied by the following:

$$\mathbb{E}_{\lambda,\mu}[(\sigma_{i_1}\sigma_{i_2})^2] \cdot \sum_{\alpha,\beta\in\mathcal{L}(i_1,i_2,k,l)} \mathbb{E}_{\lambda,\mu}[p_\alpha \cdot p_\beta] \leq \frac{1}{\delta^2} \cdot \sum_{\alpha,\beta\in\mathcal{L}(i_1,i_2,k,l)} \mathbb{E}_{\lambda,\mu}[p_\alpha \cdot (\sigma_{i_1}\sigma_{i_2})]\mathbb{E}_{\lambda,\mu}[p_\beta \cdot (\sigma_{i_1}\sigma_{i_2})]. \tag{14}$$

Intuitively, this inequality is saying that the correlation between different self avoiding walks is not too large. The right hand side of the inequality is easy to control: note that $\mathbb{E}_{\lambda,\mu}[p_\alpha \cdot (\sigma_{i_1}\sigma_{i_2})] = \mathbb{E}_{\lambda,\mu}[(\sigma_{i_1}\sigma_{i_2})^2] = 1$, so the right hand side is equal to $\delta|\mathcal{L}(i_1,i_2,k,l)|^2$. We see that this is given by

$$|\mathcal{L}(i_1,i_2,k,l)|^2 = (1+o(1))\binom{k}{l}^2 n^{2(k-1)}p^{2l}. \tag{15}$$

Thus we are left to control correlation between p_α and p_β, given on the left hand side of (14).

5.2 Correlation of SAWs: Proof of (14)

We want an upper bound on the left hand side of (14), so we need to control the correlation $\mathbb{E}[p_\alpha p_\beta]$, for paths $\alpha,\beta \in \mathcal{L}(i_1,i_2,k,l)$. For this, we have to keep track of the number of intersections. Let $\tilde{a}$ be the number of $\hat{A}$-edge intersections, and $\tilde{b}$ be the number of $\hat{B}$-edge intersections. In particular, we must have $\tilde{a} \leq k-l$, and $\tilde{b} \leq 2l$.

Note that the contribution to correlation depends only on how many intersections there are, and does not depend on the other edges of α and β. Computations show that each $\hat{A}$-edge intersection contributes a factor of $O(n/\lambda^2)$, and each $\hat{B}$-edge intersection contributes a factor of $O(n/\mu)$, so the total contribution of such an intersection would simply be

$$O(1)\left(\frac{n}{\lambda^2}\right)^{\tilde{a}}\left(\frac{n}{\mu}\right)^{\tilde{b}}.$$

Now we calculate the number of pairs $\alpha,\beta \in \mathcal{L}(i_1,i_2,k,l)$ that intersect on $\tilde{a}$ A edges and $\tilde{b}$ B edges. First we will calculate the number of pairs α and β that intersect on the smallest number of vertices, given the number of edge intersections. The smallest numbers of vertex intersections are $\tilde{a} + \lfloor \tilde{b}/2 \rfloor$ V_1 vertices, and $\lceil \tilde{b}/2 \rceil$ V_2 vertices. This is achieved when both paths, α and β, begin with $k-l$ consecutive A edges, and end with l B wedges, and intersect on the first $\tilde{a}$ $\hat{A}$-edges, as well as the last $\tilde{b}$ $\hat{B}$-edges. In this case, there are $2(k-1) - (\tilde{a} + \lfloor \tilde{b}/2 \rfloor)$ free V_1 vertices for the two paths to choose, and $b_1 + b_2 - \lceil \tilde{b}/2 \rceil$ free V_2 vertices for the paths to choose, which means that the total number of such pairs is

$$n^{2(k-1)-\tilde{a}-\lfloor \tilde{b}/2 \rfloor} p^{2l-\lceil \tilde{b}/2 \rceil}.$$

A similar calculation shows that number of pairs of paths that intersects on the least number of vertices contributes the leading order term. This is because if two paths intersected on r more vertices, then the total number of paths will decrease by a factor of $l^{O(r)}n^{-O(r)}$. Thus the total contribution in the correlation (in (14)) is:

$$
\begin{aligned}
&O(1)n^{2(k-l-1)-\tilde{a}-\lfloor\tilde{b}/2\rfloor}p^{2l-\lceil\tilde{b}/2\rceil}\left(\frac{n}{\lambda^2}\right)^{\tilde{a}}\left(\frac{n}{\mu}\right)^{\tilde{b}}\\
&=O(1)n^{2(k-1)}p^{2l}\left((\lambda^2)^{-\tilde{a}}\left(\frac{\mu^2}{\gamma}\right)^{-\tilde{b}/2}\gamma^{\lceil\tilde{b}/2\rceil-\tilde{b}/2}\right).
\end{aligned}
$$

Thus, summing over possible number of intersections, we have that:

$$
\begin{aligned}
&\sum_{\alpha,\beta\in\mathcal{L}(i_1,i_2,k,l)}\mathbb{E}_{\lambda,\mu}[p_\alpha\cdot p_\beta]\\
&\le O(1)n^{2(k-1)}p^{2l}\sum_{\tilde{a}\le l-m}\sum_{\tilde{b}\le 2m}\left((\lambda^2)^{-\tilde{a}}\left(\frac{\mu^2}{\gamma}\right)^{-\tilde{b}/2}\gamma^{\lceil\tilde{b}/2\rceil-\tilde{b}/2}\right)\\
&\le O(1)n^{2(k-1)}p^{2l}(\lambda^2)^{-(k-l)}\left(\frac{\mu^2}{\gamma}\right)^{-l}.
\end{aligned}
\tag{16}
$$

Thus, to achieve the bound in (14), we just need to show that (15) really is an upper bound of (16). That amounts to showing

$$
\binom{k}{l}^2\gtrsim(\lambda^2)^{-(k-l)}\left(\frac{\mu^2}{\gamma}\right)^{-l}.
$$

Let us now choose $\frac{l}{k}=\frac{\mu^2/\gamma}{\lambda^2+\mu^2/\gamma}$, and hence $\frac{k-l}{k}=\frac{\lambda^2}{\lambda^2+\mu^2/\gamma}$. Note that we are assuming $\lambda^2+\mu^2/\gamma>1$, so $\lambda^2>\frac{k-l}{k}$, and $\mu^2/\gamma>\frac{l}{k}$. As a result, we see that

$$
\begin{aligned}
&\binom{k}{l}^2=\exp\left(-2l\log\frac{l}{k}-2(k-l)\log\frac{k-l}{k}+o(1)\right)\\
&\gtrsim\exp\left(-l\log\frac{\mu^2}{\gamma}-(k-l)\log\lambda^2\right)=(\lambda^2)^{-(k-l)}\left(\frac{\mu^2}{\gamma}\right)^{-l}.
\end{aligned}
\tag{17}
$$

which is exactly what we wanted to show.

Note that we do not need to know λ^2 and $\frac{\mu^2}{\gamma}$ exactly to construct the estimator, because (17) can still be satisfied as long as we chooce k and l such that $1-\lambda<\frac{l}{k}<\frac{\mu}{\sqrt{\gamma}}$.Thus our estimator allows for some room of misspecification.

6. Numerical experiments

In this section, we supplement our analytic results with some numerical experiments. In particular, we explore the validity of our results in regimes beyond ones studied so far in this paper.

Finite sample recovery performance using Belief Propagation Note that our recovery algorithm is based on counting SAWs of a given length on the graph. While this can be accomplished in quasi-polynomial time, it is still not practical for networks with a few thousand nodes. An attractive alternative is provided by iterative algorithms such as Belief propagation (Mezard and Montanari (2009)) or spectral algorithms (e.g. PCA). Indeed, in the setting of community detection (i.e. $\mu = 0$), the original conjectures for the weak recovery threshold were based on the analysis of a linearized Belief propagation algorithm, which directly yields a spectral algorithm based on the non-backtracking walk (Krzakala et al. (2013)). The validity of this approach has been established rigorously in follow up work (Bordenave et al. (2015)). On the other extreme, for $\lambda = 0$, natural spectral algorithms for community recovery are derived based on the sample covariance matrix (Baik et al. (2005)). Here, we numerically study the finite sample performance of a linearized Belief Propagation algorithm proposed originally in (Deshpande et al., 2018, Section 6). Original experiments in Deshpande et al. (2018) suggest that this algorithm attains the threshold for weak recovery—establishing this rigorously is an important direction for future work. Here we study the finite sample performance of this algorithm in depth (see Fig 2 for details).

The linearized Belief Propagation algorithm studied here is indeed different from the SAW based algorithm discussed in the previous sections—yet they are conceptually intimately related. For the sparse Stochastic Block Model ($\mu = 0$), it is well known that the naive spectral algorithm fails due to the presence of high-degree vertices. Computing the largest eigenvalue is roughly equivalent to counting closed walks of a long length (of order $\log n$) and these walks are dominated by those passing through the high-degree vertices; consequently, the largest eigenvalue depends solely on the highest degree vertices, and does not capture the latent community structure. The SAW based algorithm (originally introduced in Massoulié (2014)) and the non-backtracking walk address the same challenge in slightly different ways. Both these algorithms restrict the class of closed walks to ameliorate the harmful effects of high-degree vertices. We believe the linearized Belief Propagation algorithm (analyzed in this section) and the SAW based algorithm discussed in the previous sections are similarly related; the linearized Belief Propagation based algorithm has the added benefit that it is computationally cheap and easily implementable. As mentioned above, a formal proof of its statistical efficacy remains a direction for future enquiry.

Distributional assumptions on the covariates Recall our assumption $B_i = \sqrt{\frac{\mu}{n}}\sigma_i u + Z_i$, where $Z_i \sim \mathcal{N}(0, I_p)$. It is natural to wonder about the necessity of this distributional assumption. We explore this in two ways—

(i) We assume that the Z_i are iid gaussian, but correlated. Specifically, we assume that $\mathrm{cov}(Z_1(k), Z_1(l)) = \rho \in (0,1)$, for $\rho \in (0,1)$, $1 \leq k \neq l \leq p$. In this case, $B_i \sim \mathcal{N}(0, \Sigma + \frac{\mu}{n}\sigma\sigma^\top)$, where $\Sigma_{kk} = 1$, $\Sigma_{kl} = \rho$ for $1 \leq k \neq l \leq p$. We run the linearized Belief propagation algorithm described above, and test the null hypothesis. The empirical performance of this test is shown in Fig 3. This suggests that for $\rho > 0$, the problem is statistically easier than the $\rho = 0$ case.

(ii) We test the robustness of the gaussianity assumption of the covariates. For this experiment, we sample the coordinates of Z_i as independent random signs from $\mathrm{Unif}\{\pm 1\}$, instead of Gaussians. Thus the first two moments of the covariates matches the one for the Gaussian distribution. Fig 4 plots the performance of the linearized Belief Propagation algorithm. The empirical results suggest that the performance of this algorithm does not depend strongly on the gaussianity of the covariates. We expect similar results for our SAW based algorithms. We emphasize that our experiments do

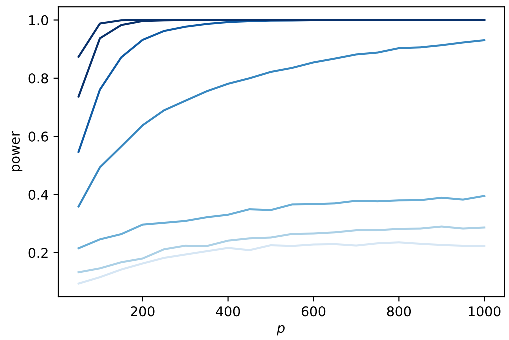

Figure 2: We fix $\gamma = 4/5$, and vary p from 50 to 1000. Each line corresponds to the choice of $\mu = \frac{\sqrt{3\gamma}}{2}$, $\lambda = \frac{1}{2} + h$, where $h = -0.06, -0.04, -0.02, 0, 0.02, 0.04, 0.06$, going from bottom to top. Note that $h = 0$ lies on the boundary of the detection curve. We plot the empirical power of the linearized BP test introduced in Deshpande et al. (2018). We observe that as $n, p \to \infty$, the curves exhibit a thresholding behavior—for $h > 0$, the empirical power converges to 1, while for $h < 0$ it is bounded away from 1. This shows that the finite sample performance is in line with the theoretical predictions for moderate problem sizes.

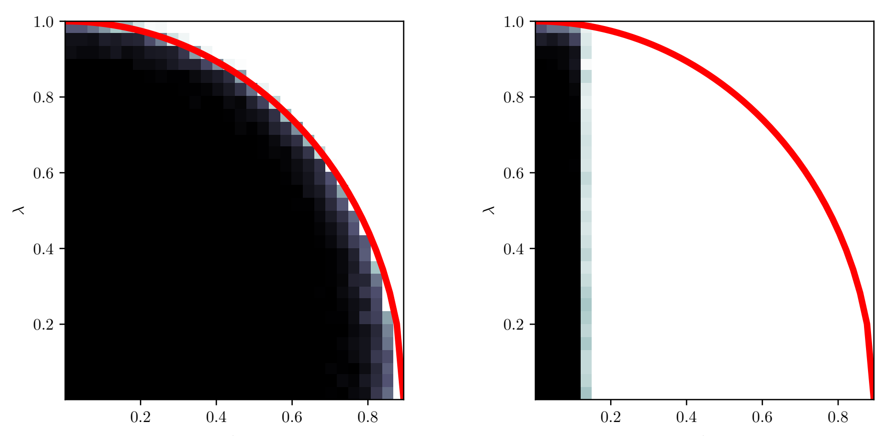

Figure 3: We set $n = 800$, $p = 1000$. The node covariates are drawn iid from an equi-correlated gaussian model with correlation parameter ρ. We plot the empirical power of the linearized BP test. The darker and lighter regions correspond to the power being closer to zero and one respectively. The left figure corresponds to $\rho = 0.001$, while the right figure corresponds to $\rho = 0.01$. Note that the $\rho = 0.001$ case is qualitatively similar to the $\rho = 0$ case studied in this paper. On the other hand $\rho = 0.01$ exhibits markedly different behavior, and suggests that detection is substantially easier in this setting.

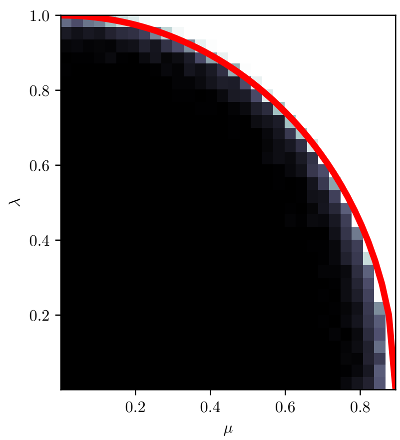

Figure 4: Set $n = 800$, $p = 1000$. The node coordinates are iid Rademacher. We again plot the empirical power of the linearized BP test. We see that the performance of the test is exactly the same as that in the gaussian covariates setting.

not suggest that the statistical threshold for detection remains unchanged—in fact, there might exist better tests which explicitly exploits the distributional knowledge of the Z_i vectors. However, our results point to the robustness of specific procedures, which we believe is of practical significance, as real data might often not be truly gaussian. Results on the universality of Approximate Message Passing algorithms (Bayati et al. (2015); Chen and Lam (2021)) also suggest that the performance of these algorithms should remain the same as long as the noise has suitably light tails. We believe this could also be an interesting direction for future research.

Acknowledgments

SS thanks Yash Deshpande, Andrea Montanari and Elchanan Mossel for helpful conversations. SS was partially supported by a Harvard Dean's Competitive Fund Award.

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *arXiv preprint arXiv:1703.10146*, 2017.

Christopher Aicher, Abigail Z Jacobs, and Aaron Clauset. Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221–248, 2014.

Ahmed El Alaoui and Michael I Jordan. Detection limits in the high-dimensional spiked rectangular model. *arXiv preprint arXiv:1802.07309*, 2018.

Greg W Anderson and Ofer Zeitouni. A clt for a band matrix model. *Probability Theory and Related Fields*, 134(2):283–338, 2006.

Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Annals of Probability*, pages 1643–1697, 2005.

Ramnath Balasubramanyan and William W Cohen. Block-lda: Jointly modeling entity-annotated text and entity-entity links. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 450–461. SIAM, 2011.

Debapratim Banerjee. Contiguity and non-reconstruction results for planted partition models: the dense case. *Electronic Journal of Probability*, 23, 2018.

Debapratim Banerjee and Zongming Ma. Asymptotic normality and analysis of variance of log-likelihood ratios in spiked random matrix models. *arXiv preprint arXiv:1804.00567*, 2018.

Mohsen Bayati, Marc Lelarge, and Andrea Montanari. Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2): 753–822, 2015.

Norbert Binkiewicz, Joshua T Vogelstein, and Karl Rohe. Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377, 2017.

Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1347–1357. IEEE, 2015.

Cécile Bothorel, Juan David Cruz, Matteo Magnani, and Barbora Micenkova. Clustering attributed graphs: models, measures and methods. *Network Science*, 3(3):408–444, 2015.

Jonathan Chang and David M Blei. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, pages 124–150, 2010.

Wei-Kuo Chen and Wai-Kit Lam. Universality of approximate message passing algorithms. *Electronic Journal of Probability*, 26:1–44, 2021.

Hong Cheng, Yang Zhou, and Jeffrey Xu Yu. Clustering large attributed graphs: A balance between structural and attribute similarities. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(2):12, 2011.

TA Dang and Emmanuel Viennet. Community detection based on structural and attribute similarities. In *International conference on digital society (icds)*, pages 7–12, 2012.

Yash Deshpande, Subhabrata Sen, Andrea Montanari, and Elchanan Mossel. Contextual stochastic block models. In *Advances in Neural Information Processing Systems*, pages 8581–8593, 2018.

Ahmed El Alaoui, Florent Krzakala, and Michael Jordan. Fundamental limits of detection in the spiked wigner model. *Annals of Statistics*, 48(2):863–885, 2020.

Jaume Gibert, Ernest Valveny, and Horst Bunke. Graph embedding in vector spaces by node attribute statistics. *Pattern Recognition*, 45(9):3072–3083, 2012.

Stephan Gunnemann, Ines Farber, Sebastian Raubach, and Thomas Seidl. Spectral subspace clustering for graphs with feature vectors. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 231–240. IEEE, 2013.

Tuan-Anh Hoang and Ee-Peng Lim. On joint modeling of topical communities and personal interest in microblogs. In *International Conference on Social Informatics*, pages 1–16. Springer, 2014.

Peter D Hoff. *Random effects models for network data.* na, 2003.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Samuel B Hopkins and David Steurer. Efficient bayesian estimation from few samples: community detection and related problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 379–390. IEEE, 2017.

Svante Janson. Random regular graphs: Asymptotic distributions and contiguity. *Combinatorics, Probability and Computing*, 4(4):369–405, 1995.

Svante Janson, Tomasz Luczak, and Andrzej Rucinski. *Random graphs*, volume 45. John Wiley & Sons, 2011.

Iain M Johnstone and Alexei Onatski. Testing in high-dimensional spiked models. *Annals of Statistics*, 48(3):1231–1254, 2020.

Varun Kanade, Elchanan Mossel, and Tselil Schramm. Global and local information in clustering labeled block models. *IEEE Transactions on Information Theory*, 62(10): 5906–5917, 2016.

Myunghwan Kim and Jure Leskovec. Latent multi-group membership graph model. *arXiv preprint arXiv:1205.4546*, 2012.

Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

Marc Lelarge, Laurent Massoulié, and Jiaming Xu. Reconstruction in the labelled stochastic block model. *IEEE Transactions on Network Science and Engineering*, 2(4):152–163, 2015.

Jure Leskovec and Julian J Mcauley. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547, 2012.

Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703. ACM, 2014.

Marc Mezard and Andrea Montanari. *Information, physics, and computation.* Oxford University Press, 2009.

Elchanan Mossel and Jiaming Xu. Local algorithms for block models with side information. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 71–80. ACM, 2016.

Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, pages 1–44, 2013.

Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015.

Jennifer Neville, Micah Adler, and David Jensen. Clustering relational data using attribute and link information. In *Proceedings of the text mining and link analysis workshop, 18th international joint conference on artificial intelligence*, pages 9–15. San Francisco, CA: Morgan Kaufmann Publishers, 2003.

Mark EJ Newman and Aaron Clauset. Structure and inference in annotated networks. *Nature Communications*, 7:11863, 2016.

Alexei Onatski, Marcelo J Moreira, and Marc Hallin. Asymptotic power of sphericity tests for high-dimensional data. *The Annals of Statistics*, 41(3):1204–1231, 2013.

Alexei Onatski, Marcelo J Moreira, and Marc Hallin. Signal detection in high dimension: The multispiked case. *The Annals of Statistics*, 42(1):225–254, 2014.

Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of pca i: Spiked random matrix models. *The Annals of Statistics*, 46(5): 2416–2451, 2018.

Robert W. Robinson and Nicholas C. Wormald. Almost all cubic graphs are hamiltonian. *Random Structures & Algorithms*, 3(2):117–125, 1992.

Robert W. Robinson and Nicholas C. Wormald. Almost all regular graphs are hamiltonian. *Random Structures & Algorithms*, 5(2):363–374, 1994.

Arlei Silva, Wagner Meira Jr, and Mohammed J Zaki. Mining attribute-structure correlated patterns in large attributed graphs. *Proceedings of the VLDB Endowment*, 5(5):466–477, 2012.

Laura M Smith, Linhong Zhu, Kristina Lerman, and Allon G Percus. Partitioning networks with node attributes by compressing information flow. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 11(2):15, 2016.

Gian-Carlo Wick. The evaluation of the collision matrix. *Physical review*, 80(2):268, 1950.

Zhiqiang Xu, Yiping Ke, Yi Wang, Hong Cheng, and James Cheng. A model-based approach to attributed graph clustering. In *Proceedings of the 2012 ACM SIGMOD international conference on management of data*, pages 505–516. ACM, 2012.

Bowei Yan and Purnamrita Sarkar. Covariate regularized community detection in sparse graphs. *Journal of the American Statistical Association*, pages 1–12, 2020.

Jaewon Yang, Julian McAuley, and Jure Leskovec. Community detection in networks with node attributes. In *Data Mining (ICDM), 2013 IEEE 13th international conference on*, pages 1151–1156. IEEE, 2013.

Tianbao Yang, Rong Jin, Yun Chi, and Shenghuo Zhu. Combining link and content for community detection: a discriminative approach. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 927–936. ACM, 2009.

Hugo Zanghi, Stevenn Volant, and Christophe Ambroise. Clustering based on random graph model embedding vertex features. *Pattern Recognition Letters*, 31(9):830–836, 2010.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Community detection in networks with node features. *Electronic Journal of Statistics*, 10(2):3153–3178, 2016.

Yang Zhou, Hong Cheng, and Jeffrey Xu Yu. Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment*, 2(1):718–729, 2009.