Title
Testing in Network Models with Community Structure

Author
Zhang, Linfan

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Testing in Network Models with Community Structure

A dissertation submitted in partial satisfaction
of the requirements for the degree Doctor of Philosophy
in Statistics

by

Linfan Zhang

2022

ABSTRACT OF THE DISSERTATION

Testing in Network Models with Community Structure

by

Linfan Zhang

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2022

Professor Arash Ali Amini, Chair

We consider the problem of testing in network models with community structures. In the first part, we propose a goodness-of-fit test for degree-corrected stochastic block models (DCSBM). The test is based on an adjusted chi-square statistic for measuring equality of means among groups of n multinomial distributions with $d_1, \ldots, d_n$ observations. In the context of network models, the number of multinomials, n, grows much faster than the number of observations, d_i, corresponding to the degree of node i, hence the setting deviates from classical asymptotics. We show that a simple adjustment allows the statistic to converge in distribution, under null, as long as the harmonic mean of $\{d_i\}$ grows to infinity. When applied sequentially, the test can also be used to determine the number of communities. Since the test statistic does not rely on a specific alternative, its utility goes beyond sequential testing and can be used to simultaneously test against a wide range of alternatives outside the DCSBM family. We show the effectiveness of the approach by extensive numerical experiments with simulated and real data. In the second part, we provide theoretical guarantees for label consistency in generalized k-means problems, with an emphasis on the overfitted case where the number of clusters used by the algorithm is more than the ground truth. We provide conditions under which the estimated labels are close to a refinement of the true cluster labels. We consider both exact and approximate

recovery of the labels. Our results hold for any constant-factor approximation to the k-means problem. The results are also model-free and only based on bounds on the maximum or average distance of the data points to the true cluster centers. These centers themselves are loosely defined and can be taken to be any set of points for which the aforementioned distances can be controlled. We show the usefulness of the results with applications to some manifold clustering problems.

The dissertation of Linfan Zhang is approved.

Chad J. Hazlett

Ying Nian Wu

Hongquan Xu

Arash Ali Amini, Committee Chair

University of California, Los Angeles

2022

To my mother, father and

my advisor Professor Arash Ali Amini

Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

I would like to extend my deepest gratitude towards my advisor Professor Arash Amini for supporting my PhD journey along the way. His wisdom keeps enlightening me about statistical field and beyond. His encouragement helps me go through the ebbs and flows of research projects. Thanks for prompt feedback and insightful discussion, without which this dissertation would never be possible.

I would like to thank Professor Chad Hazlett for providing me the opportunity to work on the kernel scaling computation project. I gained much valuable knowledge outside my main research areas through his guidance and wisdom. I also like to thank Professor Ying Nian Wu for teaching some of the most valuable machine learning courses, deeply enriching my understanding. Last but not least, thanks Professor Hongquan Xu for being the best department chair I can hope for, always standing for students, faculty and staff in our department. I am grateful for all their endeavor serving as my committee.

I would like to thank Professor Linda Zanontian for making my TA experience so much fun. I like to also thank our department staff for their efforts in keeping the department running. Thanks Enrique and Verghese for helping me with all the IT stuff. Thanks Laurie for being a reliable source for everything despite her huge amount of work.

I would like to thank my fellow friends in the department for the companionship: Jiayu Wu, Kexin Li, Kun Zhou, Xiaofeng Gao, Yifei Xu and Jireh Huang. Special thanks to Yizhou Zhao for the love and support. Without them, life at UCLA would be black and white.

Finally, I would like to thank my mother and father for everything. You make me everything I am today.

VITA

Education

B.S., Statistics 2013 - 2017
Zhejiang University

Publications

Zhang, L. & Amini, A.A. (2021). Label consistency in overfitted generalized k-means. *Neural Information Processing Systems (NeurIPS)* , 34 (2021): 7965-7977

Zhang, L. & Amini, A.A. (2020). Adjusted chi-square test for degree-corrected block models. *arXiv*. https://arxiv.org/abs/2012.15047

Chapter 1

Introduction

Network analysis has become an increasingly prominent part of data analysis as the developments in the age of the internet and in various sciences, especially life and social sciences, have produced a substantial collection of network data. Given a network, it is of interest to understand its structure, which is often done by finding communities or clusters. Probabilistic network models such as the Stochastic Block Model (SBM) [HLL83] and its variant the Degree-Corrected Stochastic Block Model (DCSBM) [KN11] are commonly used to recover the community structure from network data. Both models use a latent variable, the node label, to categorize nodes in a network into different communities. In the SBM, the probability of an edge formation between two nodes depends on the communities they belong to. The DCSBM incorporates an additional propensity parameter to determine the edge probability, allowing heterogeneous node degrees within a community. Fitting network data to probabilistic models is a heated topic in the literature [RCY11a]. On the other hand, how well these network models fit the data, the so-called goodness-of-fit question, is studied comparatively much less. In this section, we give a brief review on the Degree-Corrected Stochastic Block Model (DCSBM) and the spectral clustering as the most popular community detection algorithm to fit network data. Then we introduce some existing goodness-of-fit methods in the literature and the relevant model selection problems in the network.

1.1 Degree-Corrected Stochastic Block Model

The fundamental characteristic of a network block model is that each node i has a latent variable $z_i \in [K]$ indicating which block it belongs to. The z_i can be regarded as independent draws from class prior $\pi \in \{x \in \mathbb{R}_+^K : \sum_{k=1}^K x_k = 1\}$ such that $\mathbb{P}(z_i = k) = \pi_k$. Given a node label vector $z = (z_i) \in [K]^n$, a DCSBM with connectivity matrix $B \in [0,1]^{K \times K}$ and connection propensity vector $\theta = (\theta_i) \in \mathbb{R}_+^n$, assumes the following structure for the mean of adjacency matrix A,

$$\mathbb{E}[A_{ij} \mid z] = \theta_i \theta_j B_{z_i z_j}, \quad \forall\, i \neq j. \tag{1.1}$$

One further assumes that A is symmetric and the entries $A_{ij}, i < j$ are drawn independently, while $A_{ii} = 0$ for all i. Common choices for the distribution of each element, A_{ij}, are Bernoulli and Poisson. In this dissertation, unless otherwise stated, we assume the Poisson distribution for derivations, following the original DCSBM paper [KN11]. The Poisson assumption simplifies the arguments and provides computational advantages. The Stochastic Block Model (SBM) is a special case of (1.1) with $\theta_i = 1$ for all i. Throughout this dissertation, we define $\mathcal{C}_k = \{i : z_i = k\}$ to be the k-th cluster and $n_k = |\mathcal{C}_k|$ to be its size. Let $B = (\nu_n/n)B_0$, where ν_n is a scaling factor and B^0 has its maximum entry at most 1. Let $\theta_max = \max_i \theta = 1$. Such setup has convenience that the expected average degree of a DCSBM is of order ν_n.

The SBM and its degree-corrected variant have been the subject of intense study in recent years and numerous methods have been developed for fitting them. Mostly people are interested in finding clusters in networks generated from SBM or DCSBM. A very incomplete list includes modularity maximization [NG04; BC09], likelihood-based approaches such as the profile likelihood [BC09; ZLZ12], the pseudo-likelihood [Ami+13] and the variational likelihood [DPR08; Bic+13; ZZ20], spectral methods based on the adjacency matrix [RCY11b; CCT12; QR13; Fis+13; YP14; LR15a; CRV15; JY16; ABH16; ZA19], the non-backtracking matrix [Krz+13] and the Bethe-Hessian matrix [SKZ14], semidefinite relaxations [ABH16; AL18; LCX18; FC19], local refinements [MNS16b;

Gao+17; Gao+18; LZ17; ZA20b], message-passing algorithms [Dec+11; ZM14; AS15; MNS16a] and Bayesian approaches [SN97; HW08; MS12; Suw+16; PV18; PAL19]. Many of these methods are based on the assumption that the number of communities K is given and most come with consistency guarantees, when the data is generated from the corresponding model with K communities. We refer to [Abb18] for a review of the theoretical limits of community detection in SBMs.

1.1.1 Spectral Clustering

We briefly recall the spectral clustering, one of the most prevalent community detection algorithm for SBM and DCSBM. There are many variants of the spectral clustering in networks, and we use the regularized spectral clustering whenever community detection is performed in this dissertation. Let $d_i = \sum_{j=1}^{n} A_{ij}$ be the degree of node i and $D = \mathrm{diag}(d_i)$ is $n \times n$ diagonal matrix with d_i on the diagonal. Algorithm 1 gives a brief outline the method. The regularization greatly improve clustering performance in sparse networks with degree heterogeneity [Ami+13]. With column normalization on eigenvectors, it also works well under DCSBM [QR13]. Let $d_i = \sum_{j=1}^{n} A_{ij}$ be the degree of node i and $D = \mathrm{diag}(d_i)$ is $n \times n$ diagonal matrix with d_i on the diagonal.

Algorithm 1: Spectral clustering with regularization

Input : Adjacency matrix A, number of clusters K, regularization parameter for the Laplacian τ

Output : A label vector $\widehat{z} \in [K]^n$

1 Compute the regularized graph Laplacian $L_\tau = D_\tau^{-1/2} A D_\tau^{-1/2}$, where $D_\tau = D + \tau I$.

2 Find the eigenvectors $X_1, \ldots, X_K \in \mathbb{R}^n$ corresponding to the K largest eigenvalues of L_τ. Putting them together column-wise to get $X = [X_1, \ldots, X_K] \in \mathbb{R}^{n \times K}$.

3 Treat each row of X as a point in $\mathbb{R}^K$, and run k-means with K clusters. Output the cluster label.

The spectral clustering, like many other community detection algorithms, requires a input of the cluster number K, which one can not know the true number unless the underlying generating model is revealed. There are two kinds of solutions to the unknown cluster number in the literature. One is to regard it as a goodness-of-fit problem and test whether a K-block model is a good fit. The other one is to make it a model selection

problem and design an evaluation criteria to select the optimal K from a candidate pool. In the following section, we review some of the methods in both ways.

1.2 Goodness-of-Fit Test

Despite the efforts in finding clusters, how well these network models fit the data, the so-called goodness-of-fit question, is studied comparatively much less. Prominent work in this area include the graphical approach of [HGH08] for general network models, and the recent work of Bickel and Sarkar [BS16] and its extension by Lei [Lei16], on a spectral goodness-of-fit test for the SBM. Developing goodness-of-fit tests specifically for the DCSBM is more challenging and to the best of our knowledge has not been considered so far, except for the work of Karwa et al. [Kar+16] on the related β-SBM. We give a brief overview of each method below and a more extensive comparison with our proposed method is deferred to later Section 2.3.3.

Graph Statistics Comparison Hunter, Goodreau, and Handcock [HGH08] discussed the goodness-of-fit in general graphs, in which nodes could have additional information as covariates aside from edge relationship. Their idea is to compare a set of observed graph statistics with the range of the same statistics obtained by simulating many graphs from the fitted models. If the observed graph statistic is far from the range, then the model fits poorly and vice versa. The challenges are to choose appropriate network statistic for comparison and determine specific rules about the goodness-of-fit.

Likelihood Ratio and BIC The likelihood ratio test assess the goodness-of-fit for two network models by the ratio of their likelihoods [Yan+14a; WB17; YFS18; MSZ18]. For a DCSBM with K clusters, the complete likelihood is

$$P(A, z \mid B, \pi, \theta) = \prod_i \pi_{z_i} \prod_{i<j} \frac{(\theta_i \theta_j B_{z_i z_j})^{A_{ij}}}{A_{ij}!} e^{-\theta_i \theta_j B_{z_i z_j}}. \tag{1.2}$$

As z is unobservable, the likelihood for A with parameter space $\Theta_K = \{B, \pi, \theta\}$ is the sum of the complete likelihood w.r.t. the label z:

$$P(A \,|\, B, \pi, \theta) = \sum_{z \in [K]^n} P(A, z \,|\, B, \theta, \pi) \tag{1.3}$$

Consider a hypothesis testing with the null hypothesis H_0: K-block model and the alternative hypothesis H_a: K'-model, the log likelihood ratio is

$$L_{K,\,K'} = \log \frac{\sup_{\{B,\pi,\theta\} \in \Theta_{K'}} P(A \,|\, B, \pi, \theta)}{\sup_{\{B,\pi,\theta\} \in \Theta_K} P(A \,|\, B, \pi, \theta)}. \tag{1.4}$$

Wang and Bickel [WB17] showed that if the true model has K blocks and the degree grows in a polylog rate $\nu_n / \log \to \infty$, then when $K' < K$, for constants C and σ, $\nu_n^{-1} n^{-1/2} L_{K,\,K'} \rightsquigarrow N(C\sqrt{n}, \sigma^2)$ and when $K' > K$, $\nu_n^{-1} n^{-1/2} L_{K,\,K'} = O_P(1)$. Note that the result holds for both SBM and DCSBM. The limitation for this likelihood ratio test is that it requires a specific alternative with a different cluster number to compare the likelihoods. It can be turned into a Bayesian Information Criterion (BIC) type of model selection method to find the optimal cluster number among a bunch of candidates. By adding a penalized term, we get

$$\beta(K') = \sup_{\{B,\pi,\theta\} \in \Theta_{K'}} \ell(B, \pi, \theta \,|\, A) - \frac{K'(K'+1)}{2} n \log n, \tag{1.5}$$

and the optimal K maximizes $\beta(K')$. It leads to a consistent estimate on the true cluster number.

In practice, computing $P(A \,|\, B, \pi, \theta)$ is prohibitive due to an exponential number of summands. Therefore, they proposed either using the EM algorithm or plugging in an estimated $\widehat{z}$ to the complete likelihood (1.2) to approximate it. The plug-in technique is also used by other likelihood-based test in network [Yan+14a; MSZ18] for its computation simplicity. To compute it, we first get the complete log-likelihood from (1.2):

$$\ell(B, \theta, \pi \mid A, z) = \sum_i \log \pi_{z_i} + \sum_{i<j} \phi(A_{ij}; \theta_i \theta_j B_{z_i z_j}), \tag{1.6}$$

where $\phi(x;\lambda) = x\log\lambda - \lambda$ for Poisson likelihood. Then given an estimated label vector $\widehat{z} \in [K]^n$ and assuming the identification constraint on θ that $\sum_{i:\widehat{z}_i=k} \theta_i = |\{i : \widehat{z}_i = k\}|$, maximizing (1.6) yields estimators for parameters

$$\hat{B}_{k\ell}(\widehat{z}) = \frac{N_{k\ell}(\hat{z})}{m_{k\ell}(\hat{z})}, \quad \hat{\theta}_i(\widehat{z}) = \frac{n_{\hat{z}_i}(\widehat{z})d_i}{\sum_{j:\hat{z}_j=\hat{z}_i} d_i}, \quad \hat{\pi}_k(\widehat{z}) = \frac{n_k(\widehat{z})}{n} \tag{1.7}$$

where $N_{k\ell}(\widehat{z})$ is the sum of the elements of A in block (k,ℓ) specified by labels $\widehat{z}$, $n_k(\widehat{z})$ is the number of nodes in community k according to $\widehat{z}$ and $m_{k\ell}(\widehat{z}) = n_k(\widehat{z})(n_\ell(\widehat{z}) - 1\{k=\ell\})$.

Spectral Test The spectral test has the largest singular value of a residual matrix as its test statistic [BS16; Lei16]. It is obtained by removing the estimated block mean effect from the observed adjacency matrix. Essentially, if A is generated by a SBM and the block mean effect is estimated appropriately, the residual matrix will approximate a generalized Wigner matrix. Given an estimated label vector $\widehat{z}$, $\hat{B} = \hat{B}(\widehat{z})$ is estimated from (1.7). Then under SBM, we have the centered and re-scaled adjacency matrix $\tilde{A}$:

$$\tilde{A}_{ij} = \frac{A_{ij} - \hat{P}_{ij}}{\sqrt{(n-1)\hat{P}_{ij}(1-\hat{P}_{ij})}} \cdot 1\{i \neq j\}, \quad \text{where } \hat{P}_{ij} = \hat{B}_{\widehat{z}_i\widehat{z}_j}. \tag{1.8}$$

Then compute the largest singular value of $\tilde{A}$ as $\sigma_1(\tilde{A})$ to get the test statistic $n^{2/3}[\sigma_1(\tilde{A}) - 2]$, which is shown to asymptotically converge to the Tracy-Widom distribution under the null Erdős-Rényi model, i.e. an SBM with $K = 1$ [BS16], and also K-SBM. Furthermore, Lei [Lei16] shows that the growth rate of $n^{2/3}\sigma_1(\tilde{A})$ is of the order $\nu_n n^{1/6}$ with a underfitted cluster number, compared to the order $n^{2/3}$ with the true cluster number. Note that this require the degree ν_n grows at least in the order of $n^{1/2}$, i.e. the dense network regime.

The natural extension of the test to DCSBM is to modify the re-scaled adjacency matrix $\tilde{A}$ as

$$\tilde{A}_{ij} = \frac{A_{ij} - \hat{P}_{ij}}{\sqrt{n\hat{P}_{ij}}} \cdot 1\{i \neq j\}, \quad \text{where } \hat{P}_{ij} = \hat{\theta}_i\hat{\theta}_j\hat{B}_{\widehat{z}_i\widehat{z}_j}. \tag{1.9}$$

However, whether the same properties of $\sigma_1(\tilde{A})$ can be extended from SBM to DCSBM is

still unknown. The major challenge is that the estimator $\hat{\theta}$ required for DCSBM from (1.7) has large variance compared to $\hat{B}$ in SBM. To see the reason intuitively, $\hat{\theta}_i$ only contains information from a single row of A with n nodes, whereas $\hat{B}_{k\ell}$ from a block of A with size of order n^2 if cluster sizes are balanced.

Exact Chi-square Test Karwa et al. [Kar+16] proposed a finite-sample test based on chi-square statistic and Monte Carlo method to evaluate the goodness-of-fit of SBM and its variants. Assuming the true cluster label z is known, and let $\mathcal{C}_k = \{i : z_i = k\}$ be the k-th cluster and $n_k = |\mathcal{C}_k|$ be its size. Let $X_{i\ell} = \sum_{j\in\mathcal{C}_\ell} A_{ij}$, $N_{k\ell} = \sum_{i\in\mathcal{C}_k} X_{i\ell}$, $m_{k\ell} = n_k(n_\ell - 1\{k = \ell\})$, $\tilde{B}_{k\ell} = N_{k\ell}/m_{k\ell}$, then the chi-square test statistic is

$$\sum_{k=1}^{K}\sum_{\ell=1}^{K}\sum_{i\in\mathcal{C}_k}\frac{\left(X_{i\ell} - n_k\tilde{B}_{k\ell}\right)^2}{n_k\tilde{B}_{k\ell}}. \tag{1.10}$$

Repeatedly construct Markov moves using Monte Carlo simulation to get new networks and compute their chi-square statistics to compare with its original counterpart (1.10) and then get the exact p-value for goodness-of-fit. In the case where z is unknown, first estimating $\widehat{z}$, then sampling new networks based on the label $\widehat{z}$ and finally applying the known-community test on each. Due to the heavy sampling, the exact test works better for small networks with only hundreds of nodes.

1.3 Model Selection

A related problem to the goodness-of-fit is model selection, that is, determining the number of communities assuming that the network is generated from some SBM (or DCSBM). It has been studied more extensively as most of the community detection algorithms require the cluster number as an input. Bayesian approaches, though computationally intensive, can estimate the structure and the number of communities simultaneously. Ideas include the use of Dirichlet process prior [PAL19] and mixture of mixture priors [NR16; Rio+17; GBP19]. Another use of model selection methods is to serve as the stopping rule in hierarchical clustering procedures [Li+20a]. Below we briefly describe some model

selection methods that later to be compared with our method.

Pseudo Likelihood Ratio Ma, Su, and Zhang [MSZ18] proposed a model selection method that combines the spectral clustering and binary segmentation to determine the number of communities in DCSBM. Let $\widehat{z} \in [K]^n$ be the label vector estimated by the spectral clustering with K clusters and $\widehat{z}^b \in [K+1]^n$ be the label vector obtained using binary segmentation technique. Then compute parameter estimators (1.6) based on $\widehat{z}$ and $\widehat{z}^b$, and furthermore get the estimated probability $\hat{P}_{ij}(\widehat{z}) = \hat{\theta}_i(\widehat{z})\hat{\theta}_j(\widehat{z})\hat{B}_{\widehat{z}_i\widehat{z}_j}(\widehat{z})$ and substituting $\widehat{z}^b$ for $\widehat{z}$, we get $\hat{P}_{ij}(\widehat{z}^b)$. The pseudo likelihood is formed as

$$PL(\widehat{z}^b, \widehat{z}) = \frac{1}{2}\sum_{i\neq j}\left(\frac{\hat{P}_{ij}(\widehat{z}^b)}{\hat{P}_{ij}(\widehat{z})} - 1\right)^2. \tag{1.11}$$

They showed that when the true cluster number $K_0 \geq 2$ and $\nu_n \gtrsim \log n$, then $PL(\widehat{z}^b, \widehat{z}) \asymp n^2$ when $K < K_0$ and $PL(\widehat{z}^b, \widehat{z}) \lesssim n/\nu_n$ when $K = K_0$. Therefore the ratio of $PL(\widehat{z}^b, \widehat{z})$ between K and $K+1$ can be used sequentially to determine the cluster number, i.e. the K that gives the smallest ratio would be the selected cluster number. A tuning threshold is needed to help determine if the true cluster number is 1. Note that they can only guarantee that the ratio between the pseudo likelihood is large when the cluster number is underfitted and small with the true cluster number, and it is not clear how it grows when the cluster number is overfitted.

Bethe-Hessian spectral method The Bethe-Hessian matrix is defined as

$$H(r) = \left(r^2 - 1\right) I - rA + D \tag{1.12}$$

where $r \in \mathbb{R}$ is a parameter. $H(r)$ is often used in community detection, and its spectral structure makes it capable of selecting the cluster number [LL22]. When $r = (\frac{1}{n}\sum_{i=1}^n d_i)^{1/2}$ or $r = \left(\frac{\sum_{i=1}^n d_i^2}{\sum_{i=1}^n d_i} - 1\right)^{1/2}$, they show that the number of negative eigenvalues of $H(r)$ equal to the community number consistently when $\nu_n \gtrsim \log n$. One of the limitations for this method, as we show in Section 2.4.1, is that it requires $\mathbb{E}[A]$ to have all eigenvalues

positive, which could be violated when B has random pattern.

Cross Validation Cross validation gains huge prevalence as a model selection method, for it is applicable to a wide range of settings. To apply this technique in network data, one needs to be extra careful in splitting nodes into groups as it deletes edges, causing information loss. In [CL18], they proposed the so-called block wise splitting. First split the whole node set into two distinct groups S_1 and S_2, then do model fitting on the matrix $A_{S_1\cdot} = (A_{ij} : \ i \in S_1, j \in [n])$, and testing on the matrix $A_{S_2S_2} = (A_{ij} : \ i \in S_2, j \in S_2)$. They show that it can be used to select the number of clusters in both SBM and DCSBM, and can further determine which model is a better fit for the data.

Li, Levina, and Zhu [LLZ20] proposed to split on all edges instead of nodes. Their method is more general and can be applied to select from any latent low rank network models. Firstly, randomly choose a subset from edge set, then given rank K, use a low-rank matrix completion algorithm on the subset to obtain $\hat{A}$ with rank K. Then fit the candidate model on $\hat{A}$ and evaluate the loss on the rest of the edges. The consistency of this method depends on how well $\hat{A}$ reflects structural properties of the underlying true mode. They showed that the error between $\hat{A}$ and the expectation of model is small when $\nu_n \gtrsim \log n$. Both of the methods are computationally costly since they require estimating communities on many random network splits.

1.4 Motivation and Contributions

Per our discussion above, the current goodness-of-fit methods are scarce and have limitations in only working with: a) specific types of alternative hypothesis; b) SBM not DCSBM; c) dense regime with degree growth at least polylog rate. Model selection methods are applicable to wider range of network models, but they can be computationally instense and have complicated procedures, and can not tell how good a model fits to the data, limited to select cluster number only. To address these, we proposed a simple goodness-of-fit test based on adjustment of the well-known chi-square statistics in Chapter 2. It works with a community detection algorithm that achieves weak recovery of true labels, i.e. the

probability of having $o(1)$ misclassification rate diminishes to 0 under the true community number. We show the exact distance of the statistic to the standard normal distribution under the null. We also show its consistency in selecting the correct community number by deriving the growth rate when the community number is underfitted. The results are applicable to the sparse regime where degrees grow at least as fast as $\log n$. The test statistic works with both DCSBM and SBM, is easy to compute and has low computation complexity. In chapter 2, we describe and analyze the test in detail. Section 2.1 is a introduction on the test in a general hypothesis tests, including proofs on the asymptotic null distribution. We extend the test to the network setting in Section 2.2, followed by numerical experiments in Section 2.4. The chapter concludes with proofs of the main result in Section 2.5.

We also notice that for some model selection methods above [WB17; MSZ18; CL18], it is much more difficult to tell an overfitting cluster number from the true cluster number than it is for an underfitting cluster name. Intuitively, this is because a K-block model is embedded in an K'-block model, where $K' > K$, as there are exponential ways of splitting labels to form more clusters. This inspires us to investigate the overfitting in a general clustering setting. We focus on the generalized k-means algorithm, which is the foundation of many clustering algorithms. The question is when fit it with a cluster number greater than the true cluster number, will the estimated clusters be (close to) a refinement of the true clusters? We explore such overfitting consistency of k-means in Chapter 3. We formalized the theory that as long as the "true" centers between different clusters are large enough, the estimated clusters are (close to) a refinement of the true clusters in the overfitting case. Here the "true" centers are not necessarily the centers generating the data points and they can be constructed mannually. This allows us to confirm the intuition that clustering achieves higher accuracy by recovering subsets of true clusters as cluster number increases. The outline of Chapter 3 is as follows. Section 3.1 gives the background of the problem. Then section 3.2 contains the main results: the center distance conditions that lead to the approximate and exact recovery of refinement of true clusters. In section 3.3, we provide numerical examples and experiments to illustrate the

application of the results. Finally, Section 3.5 has the proof of the main results.

Chapter 2

Adjusted Chi-square Test for Degree-corrected Blockmodels

In this chapter, we propose the adjusted chi-square test for measuring the goodness-of-fit of a DCSBM. The idea is as follows: Given a set of column labels, we compress the adjacency matrix by summing each row over the communities specified by the labels. Under a DCSBM, the rows of the compressed matrix will have a multinomial distribution, conditional on the node degrees d_i (i.e., the row sums). Rows in the same (row) community will have the same multinomial parameter. Thus, the problem reduces to that of testing whether groups of multinomials have equal means. The challenge is that the number of multinomials in each group is proportional to n, the total number of nodes, which grows to infinity fast, while the number of observations in each multinomial, d_i, grows much slower. We study this general multi-group testing problem in Section 2.1 and show that under mild conditions, as long as the harmonic mean $h(d_1, \ldots, d_n)$ goes to infinity, a modified version of the classical chi-square statistic, which we refer to as Adjusted Chi-square (**AC**), has standard normal distribution under the null hypothesis.

We then extend these ideas to the analysis of networks, leading to the network Adjusted Chi-square tests. It has many variants depending on the subset of the adjacency matrix it applies to and how the columns of the adjacency matrix are aggregated. We show that given a consistent set of labels, by using a subsampling scheme, the same conclusion about the null distribution holds. Using $(K+1)$-community column labels for compression, and K-community row labels when testing equality of multinomials, we

obtain a powerful test in sequential applications. We refer to this variant as **SNAC+** and discuss why it is more powerful than **SNAC** where K-community labels are used for both rows and columns. We also develop bootstrapped versions of the tests which are more robust in practice and can be applied even when the null distribution of the test statistic is difficult to compute. Moreover, we introduce a smoothing idea that can further increase the robustness of sequential model selection.

Our theoretical results are non-asymptotic, controlling the Kolomogrov distance of the distribution of the test statistic to the target, with explicit constants. The results are valid in the regime where the expected average degree of the network, λ, scales as $\gtrsim \log n$, hence applicable in the same sparsity regime where weak consistency is possible for DCSBMs. From a computational standpoint, evaluating the statistic is highly scalable, with an expected computational overhead of $O(n(\lambda + K))$ over the cost of applying the community detection algorithm. To test a sequence of DCSBMs with $K = K_1, \ldots, K_2$, the test requires an application of a community detection algorithm at most $K_2 - K_1 + 2$ times.

We show the effectiveness of these ideas with extensive experiments on simulated and real networks. The code for these experiments is available at [ZA20a]. In particular, we apply the test to the Facebook-100 dataset [Tra+11; TMP12], a collection of one hundred social networks, and find that a DCSBM (or SBM) with a small number of communities (say < 25) is far from a good fit in almost all cases. Despite the lack of fit, we show that the statistic itself can be used as an effective tool for exploring communities, due to its high sensitivity to block structure. Coupled with the smoothing idea, SNAC+ allows us to construct a *community profile* for each network, regardless of whether DCSBM is a good fit.

2.1 Adjusted Chi-square test

We start by developing a general test for the equality of the parameters among groups of multinomial observations. To set the ideas, we first consider the case of a single group and show how the classical chi-square test can be adjusted to accommodate a growing number

of multinomials. We then discuss the multi-group extension and provide quantitative bounds for the null distribution of the test statistic in this general setting.

2.1.1 Single-group Case

Let $\mathcal{P}_L$ be the probability simplex in $\mathbb{R}^L$, and consider the following problem: We have

$$X_i \sim \text{Mult}(d_i, p^{(i)}), \quad i = 1, \dots, n, \tag{2.1}$$

independently, where $X_i = (X_{i\ell}) \in \mathbb{N}^L$ and $p^{(i)} \in \mathcal{P}_L$, and we would like to test the null hypothesis

$$H_0 : \ p^{(1)} = p^{(2)} = \cdots = p^{(n)} = p. \tag{2.2}$$

Let $\psi(x, y) := (x - y)^2/y$. The chi-square statistic for testing this hypothesis is

$$\widetilde{Y}^*_{(n,d)} := \sum_{i=1}^{n} \sum_{\ell=1}^{L} \psi\big(X_{i\ell}, d_i \widetilde{p}_\ell\big), \quad \text{where} \quad \widetilde{p}_\ell = \frac{\sum_{i=1}^n X_{i\ell}}{\sum_{i=1}^n d_i}, \ \ell \in [L].$$

Here, $\widetilde{p} = (\widetilde{p}_\ell) \in \mathcal{P}_L$ is the pooled estimate of p under the null, and $d = (d_1, \dots, d_n)$. We are also using the shorthand notation $[L] := \{1, \dots, L\}$.

Standard asymptotic theory gives the following (cf. Chapter 17 in [Vaa98]): If n is fixed and $d_{\min} := \min_i d_i \to \infty$, then,

$$\widetilde{Y}^*_{(n,d)} \rightsquigarrow \chi^2_{(n-1)(L-1)}, \text{ under } H_0. \tag{2.3}$$

A heuristic for the degrees of freedom of the limiting χ^2 distribution can be given by counting parameters. In the unrestricted model, we have a total of $n(L-1)$ free parameters among $p^{(1)}, \dots, p^{(n)}$, while under the restricted null model, we only have $L-1$ free parameters. The difference gives the degrees of freedom of the limit.

The setting we are interested in, however, is the opposite of the classical setting. We would like to use the statistic when $n \to \infty$, while $d_{\min}$ is fixed or grows slowly with n.

Assuming that n is large enough so that $(n-1)(L-1) \approx n(L-1)$, (2.3) suggests that we can approximate $\widetilde{Y}^*_{(n,d)}$ in distribution by the sum of n independent χ^2_{L-1} variables, that is,

$$\widetilde{Y}^*_{(n,d)} \overset{d}{\approx} \sum_{i=1}^{n} \xi_i$$

for some i.i.d. random variables $\xi_i \sim \chi^2_{L-1}$. (The approximate inequality above is only in distribution and $\{\xi_i\}$ are not necessarily related to $\widetilde{Y}^*_{(n,d)}$.) Moreover, the central limit theorem suggests that the standardized version of $\sum_i \xi_i$ has a distribution close to a standard normal.

Based on the above heuristic argument, we propose the following adjusted test statistic:

$$\widetilde{T}^*_n = \frac{1}{\sqrt{2}}\Big(\frac{\widetilde{Y}^*_{(n,d)}}{\gamma_n} - \gamma_n\Big), \quad \text{where } \gamma_n = \sqrt{n(L-1)}. \tag{2.4}$$

Note that γ_n^2 is the expectation of $\sum_i \xi_i$ and $\sqrt{2}\gamma_n$ is its standard deviation. We refer to (2.4) as the *adjusted chi-square* (AC) statistic.

Remark 1. The name adjusted chi-square has appeared in the literature in contexts completely different from our work. For example, adjustments to the chi-square statistic to account for the dependence of individuals have been proposed by Reed [Ree04] in randomized cluster trials, and by Jung et al. [JAD01] and Ahn et al. [AJD02] in observational studies.

2.1.2 Multi-group extension

Before proceeding, let us introduce an extension of the testing problem (2.2) to groups of observations. This extension is needed for the network applications. Consider model (2.1) and assume that each observation is assigned to one of the K known groups, denoted as $[K] = \{1, \dots, K\}$. Let $g_i \in [K]$ be the group assignment of observation i and let $\mathcal{G}_k = \{i \in [n] : \ g_i = k\}$ be the kth group. We would like to test the null hypothesis that

all the observations in the same group have the same parameter vector, that is,

$$H_0 : \ p^{(i)} = p_{k*}, \ \forall i \in \mathcal{G}_k, \ k \in [K], \tag{2.5}$$

where for each $k \in [K]$, $p_{k*} = (p_{k\ell})_{\ell \in [L]} \in \mathcal{P}_L$.

In some problems, it is reasonable to assume that the groups $\mathcal{G}_k$ are known. However, in our network applications, the groups themselves are not known. In such settings, we first estimate the label vector g from data, to obtain $\hat{g}$, and then form the test statistic based on the estimated groups $\widehat{\mathcal{G}}_k = \{i : \hat{g}_i = k\}$. The resulting test is based on the extended chi-square statistic

$$\widehat{Y}_{(n,d)} = \sum_{k=1}^{K} \sum_{i \in \widehat{\mathcal{G}}_k} \sum_{\ell=1}^{L} \psi(X_{i\ell}, d_i \widehat{p}_{k\ell}) \quad \text{where} \quad \widehat{p}_{k\ell} = \frac{\sum_{i \in \widehat{\mathcal{G}}_k} X_{i\ell}}{\sum_{i \in \widehat{\mathcal{G}}_k} d_i}, \ \ell \in [L]. \tag{2.6}$$

Alternatively, we have $\widehat{Y}_{(n,d)} = \sum_{i=1}^{n} \sum_{\ell=1}^{L} \psi(X_{i\ell}, d_i \widehat{p}_{\hat{g}_i \ell})$. We also let $Y_{(n,d)}$ be the idealized version of $\widehat{Y}_{(n,d)}$ with $\widehat{p}_{k\ell}$ replaced with $p_{k\ell}$ and $\widehat{\mathcal{G}}_k$ replaced with $\mathcal{G}_k$. Let $\widehat{T}_n$ and T_n be the adjusted chi-square statistics based on $\widehat{Y}_{(n,d)}$ and $Y^{(n,d)}$, respectively, that is,

$$\widehat{T}_n = \frac{1}{\sqrt{2}}\Big(\frac{\widehat{Y}_{(n,d)}}{\gamma_n} - \gamma_n\Big), \quad T_n = \frac{1}{\sqrt{2}}\Big(\frac{Y_{(n,d)}}{\gamma_n} - \gamma_n\Big). \tag{2.7}$$

We are interested in understanding under what conditions $\widehat{T}_n$ has an approximately normal null distribution. This question is nontrivial, since we would like to allow $\{d_i\}$ as well as groups sizes $|\mathcal{G}_k|, k \in [K]$ to vary with n. Moreover, we would like to allow the groups to be estimated based on the same data we use for testing, in which case, $\hat{g}$ and $\widehat{T}_n$ are most likely statistically dependent.

We give a precise answer to the above question by quantifying the Kolomogorv distance between the distribution of $\widehat{T}_n$ and that of a standard normal variable Z, for any choice of $\{d_i\}$ and $\{|\mathcal{G}_k|\}$ that satisfy a mild set of conditions, and for consistent label estimates of a certain quality. We measure the quality of label estimation in terms of misclassification rate:

Definition 1. The misclassification rate between two label vectors $g \in [K]^n$ and $\hat{g} \in [K]^n$

is

$$\mathrm{Mis}(g, \hat{g}) = \min_{\omega} \frac{1}{n} \sum_{i=1}^{n} 1\{g_i = \omega(\hat{g}_i)\}$$

where the minimization ranges over all bijective maps $\omega : [K] \to [K]$.

Recall that for two random variables X and Y, the Kolomogrov distance between their distributions is defined as

$$d_{\mathrm{K}}(X, Y) := \sup_{t \in \mathbb{R}} \big|\mathbb{P}(X \le t) - \mathbb{P}(Y \le t)\big|. \tag{2.8}$$

For a vector $d = (d_1, \ldots, d_n)$, we write $h(d) = \big(n^{-1} \sum_{i=1}^{n} d_i^{-1}\big)^{-1}$ for the harmonic mean of its elements, and $d_{\mathrm{av}} = n^{-1} \sum_{i=1}^{n} d_i$ for the arithmetic mean. Since d has positive elements, $d_{\mathrm{av}} \ge h(d) \ge d_{\min} := \min_i d_i$. Let $\pi_k = |\mathcal{G}_k|/n$ and write $d_{\mathrm{av}}^{(k)} = \frac{1}{|\mathcal{G}_k|} \sum_{i \in \mathcal{G}_k} d_i$ for the arithmatic average of $\{d_i\}$ within group $\mathcal{G}_k$, and define

$$\omega_n := \min_k \pi_k d_{\mathrm{av}}^{(k)}, \quad d_{\max} := \max_i d_i, \quad \tau_d := \omega_n / d_{\max}. \tag{2.9}$$

The following result formalizes the heuristic argument of Section 2.1.1, by providing a quantitative finite-sample bound on the Kolomogrov distances of T_n and $\widehat{T}_n$ to a standard normal variable:

Theorem 1. *Let $X_i \sim \mathrm{Mult}(d_i, p_{k*})$, $i \in \mathcal{G}_k, k \in [K]$ be n independent L-dimensional multinomial variables, with probability vectors $p_{k*} = (p_{k\ell})$ and group labels $g = (g_i) \in [K]^n$ so that $\mathcal{G}_k = \{i : g_i = k\}$. Let $\hat{g}$ be some (estimated) group labels, potentially dependent on $\{X_i\}$ and consider $\widehat{T}_n$, based on $\hat{g}$, and T_n as in (2.7). Let $Z \sim N(0,1)$ and $\underline{p} = \min_{k,\ell} p_{k\ell}$. Assume that $\min\{h(d), L\} \ge 2$.*

(a) Then, under the null hypothesis (2.5), for all $n \ge 1$,

$$d_K(T_n, Z) \le \frac{C_{1,p}}{\sqrt{Ln}} + \frac{C_{2,p}}{h(d)} \tag{2.10}$$

where $C_{1,p} = 55/\underline{p}^4$ and $C_{2,p} = (\pi e)^{-1/2} \max\{1, \underline{p}^{-1} - L - 1\}$.

(b) Let $C_{3,p} = 2(2\underline{p}^{-1} + 1)/\tau_d$ *and pick a sequence* $\{\alpha_n\}$ *such that*

$$\alpha_n \le \min\Big\{\frac{\underline{p}}{8C_{3,p}},\ \frac{2}{C_{3,p}^2 L}\Big\}, \quad \textit{for all } n \ge 1.$$

If $d_{\max} \ge C_{3,p}L/\sqrt{2}$, $\omega_n \ge L$ *and* $\log(K\omega_n)/\omega_n \le (\underline{p}/8)^2 n$, *then under the null hypothesis* (2.5), *for all* $n \ge 1$,

$$d_K(\widehat{T}_n, Z) \le d_K(T_n, Z) + \frac{\sqrt{L}}{\underline{p}}\bigg(\sqrt{\frac{72\log(K\omega_n)}{\omega_n}} + \frac{12K\log(K\omega_n)}{\sqrt{n}} + \frac{2+K}{L}C_{3,p}\, d_{\max}\sqrt{n}\alpha_n\bigg) + 2\mathbb{P}\big(\mathrm{Mis}(\hat{g}, g) \ge \alpha_n\big). \tag{2.11}$$

Proof. Theorem 1 is composed of two parts: the first part (2.10) bounds the distance of the AC statistic with true clusters and probabilities to a standard normal, which is a direct result of Proposition 1 below; the second part (2.11) shows the distance of the AC statistic with estimated clusters and probabilities to a standard normal is close to the distance of the AC statistic with true clusters and probabilities to a standard normal, which is based on Proposition 2 below.

Proposition 1. *Let* $X_i \sim \mathrm{Mult}(d_i, p_{k*})$, $i \in \mathcal{G}_k, k \in [K]$ *be independent* L*-dimensional multinomial variables, with probability vectors* $p_{k*} = (p_{k\ell})$, *and let*

$$Y_i := \sum_{\ell=1}^{L} \psi(X_{i\ell}, d_i p_{g_i,\ell}) \quad \textit{and} \quad S_n = \frac{1}{v_n}\sum_{i=1}^{n}(Y_i - \mathbb{E}[Y_i])$$

where $v_n^2 := \sum_{i=1}^n \mathrm{var}(Y_i)$. *Moreover, let* $T_n = \frac{1}{\sqrt{2}\gamma_n}\big(\sum_{i=1}^n Y_i - \gamma_n^2\big)$ *where* $\gamma_n = \sqrt{n(L-1)}$. *Let* $\underline{p} = \min_{k,\ell} p_{k\ell}$ *and assume that* $\min\{h(d), L\} \ge 2$. *Then, with* $Z \sim N(0,1)$, *we have*

$$d_K\big(S_n, Z\big) \le \frac{55}{\underline{p}^4\sqrt{Ln}}, \tag{2.12}$$

$$d_K(T_n, Z) \le d_K(S_n, Z) + \frac{\max\{1, \underline{p}^{-1} - L - 1\}}{\sqrt{\pi e}} h(d)^{-1}. \tag{2.13}$$

Proposition 2. *Recall that* $\omega_n = \min_k \pi_k d_{av}^{(k)}$. *Under the assumptions of Theorem 1, for*

any nonnegative $u \le (\underline{p}/8)^2 n\omega_n$*, we have*

$$\begin{aligned} d_K(\widehat{T}_n, Z) \le d_K(T_n, Z) + 6KLe^{-u} + 2\mathbb{P}\big(\mathrm{Mis}(\hat{g}, g) \ge \alpha_n\big) \\ + \frac{\sqrt{L}}{\underline{p}}\Big[\sqrt{\frac{8u}{\omega_n}} + 12K\frac{u}{\sqrt{n}} + (2+K)C_{3,p}L^{-1}d_{\max}\sqrt{n}\alpha_n\Big], \end{aligned} \tag{2.14}$$

where $C_{3,p}$ *is as defined in Theorem 1.*

To obtain (2.11) in Theorem 1, we take $u = \log(K\omega_n)$. To satisfy the condition of Proposition 2, we need $\log(K\omega_n)/\omega_n \le (\underline{p}/8)^2 n$. Since $\omega_n \ge L \ge 2$ by assumption and thus $2\log(K\omega_n) \ge 1$, we have

$$6KLe^{-u} = 6L/\omega_n \le \frac{2}{\underline{p}}\sqrt{8L\log(K\omega_n)/\omega_n} = \frac{2}{\underline{p}}\sqrt{8Lu/\omega_n},$$

and the result follows. □

Note that we always have $\underline{p}^{-1} \ge L$ since the elements of p_{k*} are nonnegative and sum to one. In the proof of Theorem 1, we will show that $\mathbb{E}[Y_{(n,d)}] = \gamma_n^2$. But the standard deviation $v_n(p) := \sqrt{\mathrm{var}[Y_{(n,d)}]}$ has a more complicated form and is not equal to $\sqrt{2}\gamma_n$ in general. The proof gives an explicit expression for this variance, and we could have alternatively defined $\widehat{T}_n$ by dividing by $v_n(\widehat{p})$ instead of $\sqrt{2}\gamma_n$. Nevertheless, Theorem 1 shows that we do not lose much by using the simpler standardization by $\sqrt{2}\gamma_n$.

The condition $h(d) \ge 2$ is very mild in network application as d_i will be the degree of node i. Furthermore, the condition holds when $d_i \ge 2$ for all i and as we will discuss in later network application this can be achieved by manually filtering out those with tiny d_i. In general, for T_n to converge in distribution to the standard normal, we need $n \to \infty$ and $h(d) \to \infty$. For $\widehat{T}_n$ to converge to the normal distribution, we further need $\omega_n \to \infty$, $K\log(K\omega_n) = o(\sqrt{n})$, $\alpha_n = o(d_{\max}^{-1} n^{-1/2})$ and $\mathbb{P}\big(\mathrm{Mis}(\hat{g}, g) \ge \alpha_n\big) = o(1)$. Note that $\log(K\omega_n)/\omega_n \le (\underline{p}/8)^2 n$ and $\alpha_n \le \min\{\underline{p}/(8C_{3,p}),\ 2/(C_{3,p}^2 L)\}$ are satisfied for large n, as long as $\underline{p}$ is bounded away from zero. $d_{\max} \ge C_{3,p}L/\sqrt{2}$ also holds as $d_{\max} \ge h(d)$ and we require that $h(d) \to \infty$. When there is only one single group, the requirements are $d_{\mathrm{av}} \to \infty$, which is implied by $h(d) \to \infty$, and $\log(d_{\mathrm{av}}) = o(\sqrt{n})$.

As we will see, in network applications, typically K, L and $\underline{p}$ are of constant order, $\omega_n \log \omega_n \lesssim n$ and the degrees $\{d_i\}$ are of the same order, hence $h(d) \asymp d_{\max} \asymp \omega_n$. In such settings, we obtain the rate of convergence

$$d_{\mathrm{K}}(\widehat{T}_n, Z) \lesssim \sqrt{\log \omega_n / \omega_n} + \alpha_n d_{\max} \sqrt{n} + \mathbb{P}\big(\mathrm{Mis}(\hat{g}, g) \geq \alpha_n\big).$$

2.2 Network Extension

We are now ready to apply the AC test to DCSBMs. Let $A_{n\times n}$ be the adjacency matrix of a random network on n nodes. A DCSBM with connectivity matrix $B \in [0,1]^{K\times K}$, node label vector $z = (z_i) \in [K]^n$ and connection propensity vector $\theta = (\theta_i) \in \mathbb{R}^n_+$, assumes the following structure for the mean of A,

$$\mathbb{E}[A_{ij} \mid z] = \theta_i \theta_j B_{z_i z_j}, \quad \forall\, i \neq j. \tag{2.15}$$

One further assumes that A is symmetric and the entries $A_{ij}, i < j$ are drawn independently, while $A_{ii} = 0$ for all i. Common choices for the distribution of each element, A_{ij}, are Bernoulli and Poisson. In this dissertation, unless otherwise stated, we assume the Poisson distribution for derivations, following the original DCSBM paper [KN11]. The Poisson assumption simplifies the arguments and provides computational advantages. We show in simulations that the tests so-derived work well in the Bernoulli case when the network is sparse. The SBM is a special case of (1.1) with $\theta_i = 1$ for all i.

2.2.1 NAC Family of Tests

The network AC test can be performed on a general submatrix $A_{S_2 S_1} = (A_{ij} : i \in S_2, j \in S_1)$ of the adjacency matrix, for $S_1, S_2 \subseteq [n]$. We first present this general form, though one can assume $S_1 = S_2 = [n]$ on the first reading. Consider another label vector on S_1, say $\widehat{y} = (\widehat{y}_j)_{j\in S_1} \in [L]^{S_1}$—for some L that can be different from K. Let $R = (R_{k\ell}) \in \mathbb{R}^{K\times L}_+$

be the weighted confusion matrix between z_{S_1} and $\widehat{y}$, given by

$$R_{k\ell} = \frac{1}{|S_1|} \sum_{j \in S_1} \theta_j 1\{z_j = k, \widehat{y}_j = \ell\}. \tag{2.16}$$

Now compress each row of $A_{S_2 S_1}$ by summing w.r.t. $\widehat{y}$, defined as $X = (X_{i\ell}) \in \mathbb{R}_+^{|S_2| \times L}$, with

$$X_{i\ell}(\widehat{y}) = \sum_{j \in S_1} A_{ij} 1\{\widehat{y}_j = \ell\}. \tag{2.17}$$

Assuming that $\widehat{y}$ is deterministic, we have

$$\begin{aligned}\mathbb{E}[X_{i\ell}(\widehat{y})] = \sum_{j \in S_1} B_{z_i z_j} \theta_i \theta_j 1\{\widehat{y}_j = \ell\} &= \theta_i \sum_{k=1}^{K} B_{z_i k} \sum_{j \in S_1} \theta_j 1\{z_j = k, \widehat{y}_j = \ell\} \\ &= |S_1|\, \theta_i (BR)_{z_i \ell}.\end{aligned}$$

Let $d_i = \sum_{j \in S_1} A_{ij}$ be the degree of node i in S_2. Under the Poisson model, $(A_{ij}, j \in S_1)$ is a vector of independent Poisson cooridnates. It is well-known that such a vector has a multinomial distribution conditional on the sum of its entries. That is,

$$X_{i*}(\widehat{y}) \mid d_i \sim \text{Mult}(d_i, \rho_{z_i *}), \tag{2.18}$$

where $\rho_{z_i *}$ denotes the z_ith row of $\rho = (\rho_{k\ell}) \in [0,1]^{K \times L}$, defined as

$$\rho_{k\ell} = \frac{(BR)_{k\ell}}{\sum_{\ell'} (BR)_{k\ell'}}. \tag{2.19}$$

In other words, conditioned on the degree sequence $d = (d_i, i \in S_2)$, all the rows of X corresponding to z-community k, have multinomial distributions with probability vector ρ_{k*}. This observation allows us to apply the AC test developed in Section 2.1.2, to test whether all the rows with $z_i = k$, have the same multinomial distribution.

Now, consider two estimated label vectors $\widehat{z} = (\widehat{z}_i) \in [K]^n$ and $\widehat{y} = (\widehat{y}_i) \in [L]^{S_1}$. Let $\hat{\mathcal{C}}_k = \{i \in [n] : \widehat{z}_i = k\}$, $\widehat{\mathcal{G}}_k = \hat{\mathcal{C}}_k \cap S_2$ and $\tilde{n} = |S_2|$. Consider the multi-group version of

the AC statistic based on $\widehat{z}$ and $\widehat{y}$:

$$\widehat{T}_n = \frac{1}{\sqrt{2}}\Big(\frac{1}{\gamma_{\tilde{n}}}\sum_{k=1}^{K}\sum_{i\in\widehat{\mathcal{G}}_k}\sum_{\ell=1}^{L}\psi\big(X_{i\ell}(\widehat{y}), d_i\widehat{\rho}_{k\ell}\big) - \gamma_{\tilde{n}}\Big) \tag{2.20}$$

where $\gamma_{\tilde{n}} = \sqrt{\tilde{n}(L-1)}$ and

$$\widehat{\rho}_{k\ell} = \frac{\sum_{i\in\widehat{\mathcal{G}}_k} X_{i\ell}(\widehat{y})}{\sum_{i\in\widehat{\mathcal{G}}_k} d_i}, \;\; k\in[K],\, \ell\in[L]. \tag{2.21}$$

The above construction specifies a family of test statistics, depending on the choices of label vectors $\widehat{z}$ and $\widehat{y}$, and subsets S_1 and S_2. We refer to this family, as the NAC family of tests. The acronym NAC stands for Network Adjusted Chi-square, since the test is the natural extension of the adjusted chi-square test, introduced earlier, to networks.

2.2.2 Full Version

We now single out two specifc members of the NAC family. Let $S_1 = S_2 = [n]$ and consider the following choices for $\widehat{z}$ and $\widehat{y}$:

1. FNAC: $\widehat{y} = \widehat{z}$ and $\widehat{z}$ is an estimated label vector with K communities,

2. FNAC+: $\widehat{z}$ and $\widehat{y}$ are estimated label vectors with K and $L = K+1$ communities.

The acronym FNAC stands for Full NAC, where "full" refers to the choice $S_1 = S_2 = [n]$, and we will use full NAC to mean FNAC and FNAC+ together. There are two main reasons for introducing the FNAC+ version with $L = K+1$ column communities. Firstly, FNAC only works when $K \geq 2$; when $K = L = 1$, (2.18) leads to a noninformative statistic for FNAC, becasue, then, $X_{i*} = d_i$ almost surely, conditioned on d_i. FNAC+ on the other hand still produces an informative statistic when $K = 1$. Secondly, the choice $L = K+1$ makes FNAC+ especially powerful in determining the number of communities by sequential testing from below, as we discuss extensively in Section 2.3.2.

Remark 2. The NAC family of tests are easily applicable to non-square and nonsymmetric adjacency matrices, with potentially unequal number of communities or clusters for the

rows and columns. In particular, they can be used to test directed or bipartite DCSBMs or SBMs. In addition, they can be easily applied if the cluster structure of one side is known but not the other. For example, they can be used for model selection and goodness-of-fit testing in problems involving clustering and biclustering of (Poisson) count arrays, a common task in contemporary bioinformatics [AH10]. More specifically, the biclustering problem on a Poisson count array corresponds to having an array $A_{n\times m} = (A_{ij})$, where

$$A_{ij} \sim \text{Poi}(B_{z_i y_j}),$$

independently across $i \in [n]$ and $j \in [m]$. Here $z = (z_i) \in [K]^n$ and $y = (y_j) \in [L]^m$ are the unknown clusters of rows and columns, respestively. The goal of biclustering is to recover estimates of z and y, hence simultaneously clustering rows and columns of $A = (A_{ij})$, given only an instance of A. It is clear from Section 2.2.1, that an NAC test with K and L matching the number of row and column communities, resepectively, is immediately applicable in this case. In this dissertation, we focus on the symmetric DCSBM for simplicity. All the results hold in the general nonsymmetric case as well, with suitable modifications.

2.2.3 Subsampled Version

To determine the exact asymptotic null distribution of full NAC statistics, we can extend from Theorem 1, but there are three obstacles to overcome. The first two are related to the dependence among $X_{i*}(\widehat{y})$, whereas Theorem 1 requires them being independent. Firstly, $\widehat{y}$ depends on the entire adjacency matrix A when estimated using all nodes (unless it equals to the true label z), leading to $X_{i\ell}(\widehat{y})$, the sum of entries in A based on $\widehat{y}$, delicately dependent among each other. Specifically, the aggregation of i-th row of A is related to j-th row through $\widehat{y}$. The other difficulty is the symmetry of A which makes $X_{i*}(\widehat{y})$ and $X_{j*}(\widehat{y})$ (mildly) dependent through the shared element $A_{ij} = A_{ji}$, even when $\widehat{y} = z$, the true label vector. The last obstacle is due to the sparsity of network data. Networks can contain many singletons violating the $h(d) \to \infty$ condition in Theorem 1.

We now introduce a particular subsampling scheme to circumvent the above obstacles.

It involves a sampling step so that: a) $\widehat{y}$ no longer depends on entries of A needed to be summed; b) the symmetry is broken. It also has a filtering step to leave out nodes with small degree, so that $h(d)$ is large. The detailed procedures are as follows:

1. Fit K communities to the whole network (i.e., entire A) to get labels $\widehat{z} \in [K]^n$ and clusters $\hat{\mathcal{C}}_k = \{i : \widehat{z}_i = k\}$.

2. (Sampling) Choose a subset $S_1 \subset [n]$ by including each index $i \in [n]$, independently, with probability 1/2. Let $S_2 = [n] \setminus S_1$ be the complement of S_1.

3. Fit L communities to $A_{S_1 S_1} = (A_{ij} : i, j \in S_1)$, to learn the label vector $\widehat{y}$ on S_1.

4. Form the (partial) degrees $d_i := \sum_{j \in S_1} A_{ij}$ for all $i \in S_2$.

5. (Quantile filtering) Within each $\widehat{\mathcal{G}}_k = \hat{\mathcal{C}}_k \cap S_2$, keep nodes with d_i at least the σ-th quantile of all d_i in $\widehat{\mathcal{G}}_k$ to form $\widehat{\mathcal{G}}'_k$. Let $S'_2 = \bigcup_{k=1}^K \widehat{\mathcal{G}}'_k$.

6. Perform the test on $A_{S'_2 S_1}$ using row labels $\widehat{z}_{S'_2}$ and column labels $\widehat{y}$ from Step 3.

We refer to the above as the subsampled version of FNAC as SNAC, and similarly for FNAC+. Also, we use subsampled NAC to represent SNAC and SNAC+ together. Note that step 4, the quantile filtering step, can be skipped when degrees are mostly large or we do not care about the null distribution being standard normal, e.g. using bootstrap debiasing in Section 2.2.4 to determine the critical region. In such case, we will have $S'_2 = S_2$ and perform the test on $A_{S_2 S_1}$. In Section 2.3, we show that, under the null model, the distributions of the test statistics of the subsampled NAC are close to a standard normal. Furthermore, they are large when the model is underfitted, i.e., the presumed number of communities is smaller than that of the true model. Particularly, SNAC+ is more powerful and versatile than SNAC. Such properties allow us to use SNAC+ for assessing the goodness-of-fit of DCSBM or SBM to an observed network and to determine the number of clusters in community detection.

2.2.4 Bootstrap Debiasing

Per our discussion above, without subsampling, the full version statistics do not have a standard normal null distribution. However, they are expected to have more power in the test as they utilize all the nodes in the network, so they are still great options to use in practice. The remedy is to use bootstrap simulation to determine their critical regions holistically. In addition, bootstrap helps remove the shift of SNAC and SNAC+ from a standard normal null distribution with a Bernoulli network, since the null distribution result is based on Poisson generation. Therefore, with bootstrap, both the full and subsampled version would gain more power in the test.

Now we discuss the details of bootstrap. Given adjacency matrix A, the null hypothesis that the number of communities is K, and the test statistic $\hat{T} = \hat{T}(A)$, the bootstrap debiasing is performed as follows:

1. Fit a K-community SBM to A and get label estimates $\hat{z}$ and connectivity matrix $\hat{B}$.
2. For $j = 1, \ldots, J$, sample $A^{(j)} \sim \text{SBM}(\hat{z}, \hat{B})$ and evaluate the test statistic $\hat{T}^{(j)}$ based on $A^{(j)}$.
3. Construct the debiased statistic $\hat{T}^{(\text{boot})} = (\hat{T} - \hat{\mu})/\hat{\sigma}$ where $\hat{\mu}$ and $\hat{\sigma}$ are the sample mean and the standard deviation of $\{\hat{T}^{(j)}\}_{j=1}^{J}$.

Note that we sample from SBM instead of DCSBM, since the estimator of θ could have a large variance. The test rejects for large values of $\hat{T}^{(\text{boot})}$ (or $|\hat{T}^{(\text{boot})}|$), with the threshold set, assuming that $\hat{T}^{(\text{boot})}$ has (approximately) a standard normal distribution under null. An alternative to debiasing is to use the empirical quantiles of $\{T^{(j}\}$ to set the critical threshold. We, however, found that the debiasing approach performs better in practice. A similar idea is used in [Lei16] for the spectral test as well. See Appendix A.1 for a comparison of all the bootstrap methods in simulations.

2.2.5 Model Selection

A goodness-of-fit test can also be used as a model selection method, through a process of sequential testing. In particular, we can use the full (with bootstrap debiasing) and the

subsampled NAC statistics to determine the number of communities when fitting DCSBM models.

The idea is to test the null hypothesis of K communities, starting with $K = K_{\min}$, which is usually taken to be 1, and increasing K to $K + 1$ if the null is rejected. The process is repeated until we can no longer reject the null or a preset maximum number of communities, $K_{\max}$, is reached. The value of K on which we stop is selected as the optimal number of communities. We refer to this procedure as *sequential testing from below*. There is also the possibility of starting at $K = K_{\max}$ and working backwards. Testing from below is, however, more advantageous, especially if one expects a small number of communities a priori.

The rejection thresholds for the subsampled NAC can be determined based on the standard normal distribution. For the full NAC, we need to apply the bootstrap debiasing of Section 2.2.4 before comparing the statistic with the threshold. Theorem 3 provides a theoretical guarantee for the consistency of the sequential testing from below, when the subsampled NAC is used. An empirical comparison of the model selection performance of this approach, with existing methods, is provided in Section 2.4.1.

2.3 Analysis of Subsampled NAC

We now provide a theoretical analysis of the subsampled NAC. We consider a DCSBM with K_0 true community, and the edge probability matrix $B = (\nu_n/n)B^0$ where ν_n is a scaling factor and B^0 satisfies

$$\min_{k,\ell} B^0_{k\ell} \geq \tau_B \cdot \max_{k,\ell} B^0_{k\ell}. \tag{2.22}$$

Let $\mathcal{C}_k = \{i \in [n] : z_i = k\}$ be the true community k. We assume that

$$n_k := |\mathcal{C}_k| \geq \tau_{\mathcal{C}}\, n, \quad \theta_i \geq \tau_\theta \cdot \max_i \theta_i \tag{2.23}$$

for all $k \in [K_0]$ and $i \in [n]$. Here, $\tau_B, \tau_{\mathcal{C}}$ and τ_θ are in $(0, 1]$ and measure the deviation of the corresponding parameters from being balanced. To make ν_n identifiable, we further assume without loss of generality that $\|B^0\|_\infty := \max_{k,\ell} B^0_{k\ell} = 1$ and $\|\theta\|_\infty := \max_i \theta_i = 1$.

We also require the community detection algorithm to have weak recovery of the true communities well and does not produce extremely small communities when the presumed number of clusters is close to K_0.

Assumption 1. *The community detection algorithm applied with K communities to the DCSBM described above, producing labels $\{\widehat{z}_i\}$, satisfies:*

(a) Weak consistency: when $K = K_0$, $\mathbb{P}\big(\mathrm{Mis}(\widehat{z}, z) \geq \alpha_n\big) = o(1)$ for a sequence $\{\alpha_n\}$.

(b) Stability: $|\{i : \widehat{z}_i = k\}| \geq \tau_0 n$ for all $k \in [K]$, when $K \in \{1, \ldots, K_0 - 1, K_0 + 1\}$.

The weak consistency of the algorithm allows us to focus on the event where $\widehat{z}$ is close to z under the null model, and the stability allows us to lower-bound ρ, defined in (2.19). Condition (a) in Assumption 1 is known as the almost exact recovery, and it is well-known that if $\nu_n \gtrsim \log n$, there are algorithms that can achieve it [Abb18]. Note that the growth rate of ν_n is roughly that of the expected average degree (EAD) of the network, assuming that B^0, $\{n_k/n\}_k$ and the distribution of $\{\theta_i\}$ are roughly constant. Hence, assuming $\nu_n \gtrsim \log n$ imposes a mild restriction on the network, requiring the EAD to grow at least as $\log n$. The stability condition (b) is even milder and can be guaranteed by explicitly enforcing it in the algorithm. If the size of a recovered community is too small relative to n, we merge it with another community. Whether a specific community detection algorithm satisfies this condition automatically without explicit enforcement is an interesting research question.

2.3.1 Null Distribution

To state further assumptions, we define the following constants:

$$c_1 := (1 - \sigma)\frac{\tau_{\mathcal{C}}}{5K_0}, \quad C_1 := \tau_\theta^2 \tau_{\mathcal{C}} \min_h \|B^0_{h*}\|_1, \tag{2.24}$$

$$\tau_a := \tau_\theta \tau_B \tau_{\mathcal{C}}, \quad \tau_\rho := \tau_\theta \tau_B \tau_0. \tag{2.25}$$

We make the following assumptions:

$$\frac{\nu_n}{\log n} \geq \frac{1000}{C_1}, \quad \frac{n}{\log n} \geq \frac{400}{3C_1} \vee \frac{300}{\tau_{\mathcal{C}}}, \quad \sigma \leq 2\tau_{\mathcal{C}}/3, \quad \alpha_n \leq \frac{\tau_{\mathcal{C}}(1-\sigma)}{5(1+\sigma)}. \tag{2.26}$$

Theorem 2 (Null distribution)**.** *Consider an $n \times n$ adjacency matrix A that is generated from a Poisson DCSBM with K_0 blocks, satisfying (2.22) and (2.23). Let $\widehat{z} \in [K_0]^n$ be an estimated label vector of A satisfying assumption in 1 and $\widehat{y} \in [L]^{|S_1|}$ be an estimated label vector of $A_{S_1S_1}$ satisfying the stability assumption in 1. Let $\widehat{T}_n$ be the test statistic of the subsampled NAC.*

Let $\beta_n = \log[(3/4)K_0^2\nu_n]$ and assume in addition that $L \geq 2$, $\beta_n/\nu_n \leq C_6 n$, $\nu_n \geq \max\{12L/(5c_1C_1K_0), 2\sqrt{2}C_2L/C_1\}$, and $\alpha_n \leq \min\{\frac{\tau_\rho}{8C_2}, \frac{2}{LC_2^2}\}$, where C_1 is as defined in (2.25), $C_2 = 54/(5c_1C_1\tau_\rho)$ and $C_6 = (\frac{\tau_\rho}{8})^2 \frac{5}{12} c_1C_1K_0$. Then,

$$d_K(\widehat{T}_n, Z) \leq \frac{C_3}{\sqrt{Ln}} + \frac{C_4}{C_1\nu_n} + \frac{19\sqrt{L}}{\tau_\rho}\left(\sqrt{\frac{\beta_n}{c_1C_1K_0\nu_n}} + \frac{K_0\beta_n}{\sqrt{(1-\sigma)n}}\right) + \tag{2.27}$$

$$C_5\nu_n\sqrt{n}\alpha_n + 3\mathbb{P}(\mathrm{Mis}(\widehat{z}, z) \geq \alpha_n) \tag{2.28}$$

where $C_3 = 87/[(1-\sigma)^{1/2}\tau_\rho^4] + 7$, $C_4 = 4(\pi e)^{-1/2}\max\{1, \tau_\rho^{-1} - L - 1\}$ and $C_5 = 15K_0(1 + K_0/2)C_2/[4(1-\sigma)\tau_\rho\sqrt{L}]$.

Note that the above bound applies to both SNAC and SNAC+ as only the stability assumption on $\widehat{y}$ is required in the proof. Assuming the common scaling $\log n \lesssim \nu_n \lesssim \sqrt{n}$, we obtain

$$d_{\mathrm{K}}(\widehat{T}_n, Z) \lesssim \sqrt{\frac{\log \nu_n}{\nu_n}} + \alpha_n\nu_n\sqrt{n} + \mathbb{P}(\mathrm{Mis}(\widehat{z}, z) \geq \alpha_n).$$

In order to have the null distribution close to a standard normal, we need to require $\alpha_n = o((\nu_n\sqrt{n})^{-1})$, and there are many community detection algorithms that can achieve this [QR13; Gao+17].

2.3.2 Consistency

Power against Underfitting

Next, we consider the consistency of the subsampled NAC when applied in sequential testing from below, to determine the number of communities. In particular, we analyze its power in distinguishing the null hypothesis $H_0 : K = K_0$ from the alternative $H_1 : K < K_0$. Theorem 3 provides a lower bound on the growth rate of the test statistic $\widehat{T}$ under the alternative.

Recall that $\widehat{y}$ are labels derived for nodes S_1 based on $A_{S_1S_1}$. Let parameters $\rho_{k\ell}$ be defined as in (2.19), and let

$$\omega_2 = \frac{1}{18}\tau_\theta^2\tau_a^2 c_1^2 \min_{k,h\in[K_0]:\, k\neq h} \frac{1}{L}\|\rho_{k*} - \rho_{h*}\|_2^2. \tag{2.29}$$

Note that ω_2 is a random quantity due to the randomness in $\widehat{y}$.

Theorem 3 (Power)**.** *Let A be an $n \times n$ adjacency matrix generated from a Poisson DCSBM with $K_0 \geq 2$ blocks that satisfies* (2.22) *and* (2.23)*. Let $\widehat{T}_n$ be the the subsampled NAC test statistic* (2.20) *formed as detailed in Section 2.2.3, with $K < K_0$ and $L = K + 1$ communities, estimated from a community detection algorithm satisfying stability in Assumption 1. Moreover, let $C_7 := c_1C_1/10$ and assume that $(\log n)/\nu_n \leq C_1\tau_\rho^2/64$ and consider the event*

$$\Omega_n = \left\{ \max\left(\frac{1}{C_7\nu_n},\ \frac{768}{\tau_\rho^3}\sqrt{\frac{\log n}{C_1\nu_n}} \right) \leq \omega_2 \right\}. \tag{2.30}$$

Then, with probability at least $1 - 9Ln^{-1} - \mathbb{P}(\Omega_n^c) - \mathbb{P}(\mathrm{Mis}(\widehat{z}, z) \geq \alpha_n)$,

$$\widehat{T}_n \geq C_7\,\omega_2\,\nu_n\sqrt{Ln}.$$

Quantity ω_2 that appears in Theorem 3 is random (via $\{\rho_{k\ell}\}$) and depends on the specific community detection algorithm used to form the test statistic. As discussed below, for any reasonable algorithm, under mild conditions on the connectivity matrix, we expect

ω_2 to be of constant order as $n \to \infty$, i.e., $\omega_2 \asymp 1$. In particular, we expect to have $\mathbb{P}(\omega_2 \geq c_2) \to 1$ for some constant $c_2 > 0$, as $n \to \infty$. Then, we have $\mathbb{P}(\Omega_n^c) \to 0$, as long as $(\log n)/\nu_n \leq c_2$.

Under these assumptions, Theorem 2 shows that for a given significance level $\alpha > 0$, the subsampled NAC statistic $\widehat{T}_n \asymp 1$ with probability $1 - \alpha$ when $K = K_0$, while Theorem 3 guarantees that $\widehat{T}_n \gtrsim \nu_n\sqrt{n}$, w.h.p., when $K < K_0$. This shows that the subsampled NAC with a constant threshold or one that grows slower than $\nu_n\sqrt{n}$, leads to consistent model selection when applied sequentially from below (i.e., with $K < K_0$). In short, model selection consistency of the subsampled NAC only requires $(\log n)/\nu_n = O(1)$, that is, the expected degree should grow no slower than $\log n$.

Comparison between SNAC+ and SNAC

SNAC+ is expected to be more powerful than SNAC in sequential testing from below, as its column compression $\widehat{y}$ is estimated with $L = K + 1$ communities. Let us consider the hardest case in Theorem 3, that is, testing the null hypothesis $K = K_0 - 1$ against the alternative $K = K_0$. When $L = K + 1 = K_0$, the estimated column labels $\widehat{y}$ is close to the true labels z, under the weak recovery Assumption 1. Recalling the definition of the confusion matrix from (2.16), we roughly obtain $R = \mathrm{diag}(\widetilde{\pi}_k)$, where $\widetilde{\pi}_k = \frac{1}{|S_1|}\sum_{j \in S_1} \theta_j 1\{z_j = k\}$ for all $k \in [K_0]$. Then, $\rho_{k\ell} = B^0_{k\ell}\widetilde{\pi}_\ell / (\sum_{\ell'} B^0_{k\ell'}\widetilde{\pi}_{\ell'})$. Note that both B^0 and $\{\widetilde{\pi}_k\}$ are stable as $n \to \infty$. In particular, although the entries of B vanish under the scaling $\nu_n/n \to 0$, the entries of $(\rho_{k\ell})$ do not. To guarantee that $\omega_2 > 0$, it is enough that the $K_0 \times K_0$ matrix $(B^0_{k\ell}\widetilde{\pi}_\ell)$ has no two colinear rows, a mild identifiability condition. On the other hand when $L = K_0 - 1$, the multinomial parameter matrix $\rho \in \mathbb{R}^{K_0 \times (K_0 - 1)}$ have row entries as weighted average of its counterpart when $L = K_0$. We refer to [WB17] for an example of how the weighted mixture of the rows of the connectivity matrix B emerges in the underfitted case and thus ρ is mixed in the similar way. Due to this averaging, the row distance of matrix ρ would be smaller compared to when $L = K_0$ and thus ω_2 smaller. Moreover, consider the extreme case, SBM with planted partition B and equal community sizes. If the community detection algorithm can recover a superset of

the true communities when underfiting, as shown, for example, for the spectral clustering in [MSZ18], ρ would have identical rows and thus ω_2 is close 0, making SNAC powerless.

Now let us look at an simple example which supports our argument above. Consider an SBM with $K_0 = 3$, equal-sized communities and a planted-partition B with p on the diagonal and q on the off-diagonal

$$B = \begin{pmatrix} p & q & q \\ q & p & q \\ q & q & p \end{pmatrix}.$$

Now we test the null hypothesis $K = 2$. When $L = 2$, consider a column label $\widehat{y}$ that has cluster 2 and 3 combined into one, then the confusion matrix R defined in (2.16) and BR are

$$R = \begin{pmatrix} 1/3 & 0 \\ 0 & 1/3 \\ 0 & 1/3 \end{pmatrix}, \quad BR = \begin{pmatrix} p/3 & 2q/3 \\ q/3 & (p+q)/3 \\ q/3 & (p+q)/3 \end{pmatrix}$$

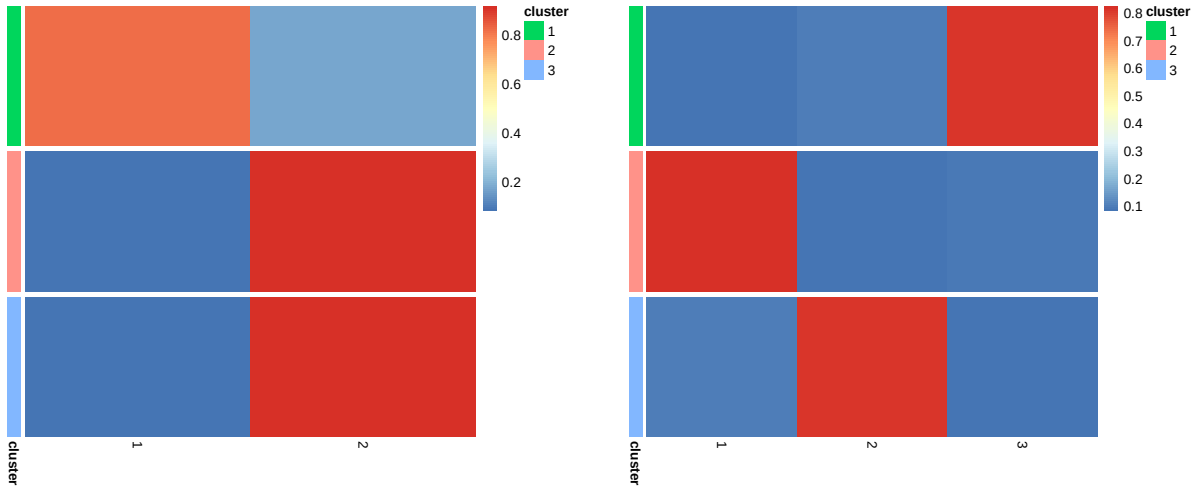
Recall that the multinomial probability ρ (2.19) is determined by standardizing rows in BR. Therefore $\rho_{2*} = \rho_{3*}$, and if the row label equals to $\widehat{y}$, then the test fails. On the other hand, when $L = 3$ and $\widehat{y} = z$, we have

$$R = \begin{pmatrix} 1/3 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/3 \end{pmatrix}, \quad BR = \begin{pmatrix} p/3 & q/3 & q/3 \\ q/3 & p/3 & q/3 \\ q/3 & q/3 & p/3 \end{pmatrix}$$

Therefore, the multinomial probability ρ is proportional to B. Then the ω_2 in (2.29) is positive and by Theorem 3, we have the consistency of SNAC+.

We can observe the same phenomenon in practice. Consider an SBM with $K = 3$ equal-sized clusters, planted partitioned B with out-in-ratio $\beta = 0.1$, average degree as 10 and node number $n = 300$. The null hypothesis is $K = 2$. For simplicity, we focus on the full version FNAC and FNAC+ and compare the expectation of (X_{i*}), i.e. ρ_{z_i*},

which is determined by the column label $\widehat{y}$. In FNAC, $\widehat{y}$ is obtained by performing the spectral clustering on the entire adjacency matrix with $L = 2$ clusters and it has true cluster 2 and 3 merged as a single cluster. Similarly, $\widehat{y}$ in FNAC+ has $L = 3$ clusters from the spectral clustering, and it is very close to the true cluster label with only one node mis-classified. Figure 2.1 shows the heatmap of (ρ_{z_i*}), $i = 1, \ldots, n$ for the above FNAC (left side) and FNAC+ (right side). Because of the merging in $\widehat{y}$ with $L = 2$, the left heatmap shows that cluster 2 and 3 has the same multinomial probabilities. Whereas $\widehat{y}$ with $L = 3$ is close to the true label vector, so the right heatmap shows the 3 clusters have distinct multinomial probabilities. This corroborates the discussion above. Lastly, when computing the statistics, both FNAC and FNAC+ have row labels from the spectral clustering with $K = 2$, which has cluster 2 and 3 merged. Since these two clusters have the same multinomial probabilities when $L = 2$, FNAC would be small and thus fail to reject the null. However, they have different probabilities when $L = 3$, hence FNAC+ would be large as we compute the chi-square statistic on a mixed cluster and it would reject the null.



(a)FNAC with $L = 2$ (b) FNAC+ with $L = 3$

Figure 2.1: Heatmaps of multinomial probability matrix, where the i-th row equal to ρ_{z_i*}. The left side shows the case for FNAC with $L = 2$ and the right for FNAC+ with $L = 3$. The column labels are both obtained through spectral clustering with $L = 2$ and $L = 3$ respectively. The row labels are true cluster labels and are indicated using the color bar on the left.

The above argument also shows the advantage of using labels estimated from community detection algorithm to serve as $\widehat{y}$ rather than using some random labels. When

$\widehat{y}$ contains randomly assigned K_0 labels, the confusion matrix R has non-zero weights as entries and thus ρ is weighted average of B_0 based on R. Whereas when $\widehat{y}$ is close to true labels, R is diagonal. Therefore, the row distances of ρ with random $\widehat{y}$ are smaller compared to with esimtated $\widehat{y}$.

Power against DCLVM

In this section we show that subsampled NAC is consistent for distinguishing DCSBM from a general class of degree-corrected latent variable models (DCLVM). We consider a K-community DCLVM, with degree parameter θ, label vector $z^* \in [K^*]^n$, mixture components $\{\mathbb{Q}_k^*\}_{k=1}^K$ and latent variables $\{x_i\}_{i=1}^n \subset \mathcal{X}$, a network model specified as follows: Given $\{x_i\}$, each (i,j) is drawn independently (of other edges) from a Poisson distribution with mean

$$p_{ij} := \mathbb{E}[\, A_{ij} \,|\, x_i,\ x_j] = \frac{\nu_n}{n}\theta_i\theta_j g(x_i, x_j)$$

and $x_i \sim \mathbb{Q}_{z_i^*}^*$ independently across i. The mixture components $\{\mathbb{Q}_k^*\}$ are distributions on the space $\mathcal{X}$, and when they are different they impose some latent community structure. An example, with specific forms for $g(\cdot,\cdot)$ and $\{\mathbb{Q}_k\}$ is given in Section 2.4.1. Here, we consider the general case, with minimal assumptions on $g(\cdot,\cdot)$ and $\{\mathbb{Q}_k\}$. We use similar assumptions on θ as in the DCSBM, namely,

$$\max_i \theta_i = 1, \quad \theta_i \geq \tau_\theta,\ \forall i \in [n]$$

Without loss of generality, we assume that g has range $[0,1]$, by rescaling ν_n if need be.

Without strong assumptions on $\{\mathbb{Q}_k\}$, the distribution of x_i is a nonparametric mixture model which, in general, is not identifiable. One can shift mass from one of $\{\mathbb{Q}_k^*\}$ to the other ones or create a new component, and redefine the label vector to get the same distribution. For example, suppose that we start with a two-community model with components $\mathbb{Q}_1^*$ and $\mathbb{Q}_2^*$. We relabel each x_i by assigning it the new label $z_i \in [K]$ (rather than z_i^*). The same model for x_i can be stated as $x_i \sim \mathbb{Q}_{z_i}$ for new mixture components

$\mathbb{Q}_k = \pi_{k1}\mathbb{Q}_1^* + \pi_{k2}\mathbb{Q}_2^*$ which are convex combinations of the original ones. We refer to $\{\mathbb{Q}_k\}$ as the mixture components induced by z. The result that we present here applies to any of these parameterizations.

Assume that we perform the subsampled NAC with K row communities and L column communities. Let $\widehat{z} \in [K]^n$ be the estimated label vector based on the entire adjacency matrix A and $\widehat{y} \in [L]^{|S_1|}$ that based on $A_{S_1 S_1}$. We assume that there are deterministic vectors $z \in [K]^n$ and $y \in [L]^n$, and sequences $\{\alpha_n\}$ and $\{\kappa_n\}$ such that the following event

$$\mathcal{M}_n := \{\mathrm{Mis}(\widehat{z}, z) \le \alpha_n \text{ and } \mathrm{Mis}(\widehat{y}, y_{S_1}) \le \kappa_n\}, \tag{2.31}$$

has probability converging to 1, as $n \to \infty$. Note that we do not require z (or y) to be the original z^*. Letting $n_k = |\{i : z_i = k\}|$, we assume

$$n_k \ge \tau_{\mathcal{C}} n, \quad \forall k \in [K]. \tag{2.32}$$

Let $\{\mathbb{Q}_k, k \in [K]\}$ be the mixture components induced by label vector z that appears in (2.31). Define

$$h_k(x) := \mathbb{E}[g(x, \xi)], \quad \xi \sim \mathbb{Q}_k, \ k \in [K].$$

We assume that there is an almost sure event Γ with the following property: There exists a constant $\tau_h > 0$ and $r_1, r_2 \ldots, r_K \in [K]$ such that on Γ, we have

$$\forall k \in [K], \ \forall i \in \mathcal{C}_k, \ h_{r_k}(x_i) \ge \tau_h. \tag{2.33}$$

Note that (2.33) can be equivalently stated as $h_{r_{z_i}}(x_i) \ge \tau_h$ for all i. Condition (2.33) is mild and is satisfied for example if for any $k \in [K]$, one of $h_r(\cdot), r \in [K]$ is uniformly bounded below over the support of $\mathbb{Q}_k$. We also define

$$H_\ell(x) := \frac{\sum_k h_k(x) R_{k\ell}}{\sum_{\ell'} \sum_k h_k(x) R_{k\ell'}}, \quad R_{k\ell} := \frac{1}{2} \sum_{j=1}^n \theta_j 1\{z_j = k, \, y_j = \ell\}. \tag{2.34}$$

There exists a sequence $\{\ell_k\}_{k=1}^K$ such that

$$R_{k\ell_k} \geq \frac{1}{L}\sum_{\ell=1}^{L} R_{k\ell}, \quad \forall k \in [K]. \tag{2.35}$$

Fix one such sequence and consider the following quantities

$$\begin{aligned} \vartheta_{k\ell} &:= \mathrm{var}(H_\ell(x)), \quad \text{where } x \sim \mathbb{Q}_k \\ \underline{\vartheta} &:= \min_k \vartheta_{k\ell_k}. \end{aligned} \tag{2.36}$$

Let $c_2 = \tau_{\mathcal{C}}\tau_h\tau_\theta^2/100$, $\tau_\rho = \tau_{\mathcal{C}}\tau_h\tau_\theta/(2L)$ and $\zeta_n = \max\{1, L\sqrt{\nu_n/n}, L/\sqrt{\nu_n \log n}\}$. We need the following assumptions:

$$\sqrt{\frac{\log n}{n}} \leq \frac{2}{9}\frac{\tau_\rho^2}{K} \quad n \geq 2, \tag{2.37}$$

$$\alpha_n \leq \sqrt{\frac{\log n}{\nu_n}} \leq \frac{21\tau_{\mathcal{C}}^2\tau_h c_2^2}{L^2}, \quad \frac{n\kappa_n}{\nu_n} \leq 4c_2\tau_\rho, \tag{2.38}$$

$$\underline{\vartheta} \geq \frac{L^3}{c_2^3\tau_\rho^3}\max\left\{\frac{2\zeta_n}{\tau_\rho\tau_{\mathcal{C}}}\sqrt{\frac{\log n}{\nu_n}}, \ \frac{1}{5c_2}\frac{n\kappa_n}{\nu_n}\right\}. \tag{2.39}$$

Theorem 4. *Let A be an $n \times n$ adjacency matrix generated from a Poisson DCLVM with K blocks that satisfies* (2.22) *and* (2.23). *Let $\widehat{T}_n$ be the subsampled NAC test statistic* (2.20) *formed as detailed in Section 2.2.3, with K and L communities, with estimated label vector $\widehat{z}$. Moreover, assume* (2.33) *and* (2.37)–(2.39). *Then, with probability at least* $1 - 12KLn^{-1} - Kn^{-c} - \mathbb{P}(\mathcal{M}_n^c)$,

$$\widehat{T}_n \geq \frac{49c_2^3}{\sqrt{L}}\,\underline{\vartheta}\sqrt{n}\nu_n.$$

The theorem roughly states the following: As long as the community detection algorithm produces row and columns labels that converge to some deterministic labels z and y and the rates $\alpha_n \sim \sqrt{(\log n)/\nu_n}$ and $\kappa_n \sim \nu_n/n$ respectively, and the resulting induced mixture components $\{\mathbb{Q}_k\}$ lead to a positive minimum variance $\underline{\vartheta}$, as defined in 2.36, then SNAC(+) are consistent in rejecting the underlying DCLBM model, with $\widehat{T}_n \gtrsim \sqrt{n}\nu_n \to \infty$. Note that $\underline{\vartheta} > 0$, unless there exists a sequence of constants $a_1, \ldots, a_K$

such that $\sum_r a_r h_r(x) = 0$ for $\mathbb{Q}_k$-almost all x. That is, unless $\{h_r\}_{r=1}^K$ satisfy a non-trivial linear constraint under $\mathbb{Q}_k$, the condition $\underline{\vartheta} > 0$ is guaranteed. An example where the condition $\underline{\vartheta} > 0$ is violated is when all $h_r(\cdot)$ are constant functions, as is the case for a DCSBM, consistent with the fact that we should not be able to reject a DCSBM.

Remark 3. The constant $\frac{1}{2}$ in the defintion of $R_{k\ell}$ in (2.34) is for the convenience in the proof. It can be changed to any other prefactor (including $\frac{1}{n}$) since $H_\ell(x)$ is invariant to a rescaling of $R_{k\ell}$.

2.3.3 Comparison with the Existing Literature

The closest work in the literature to ours is the spectral goodness-of-fit test for SBMs [BS16; Lei16]. Roughly speaking, Lei [Lei16] shows that, under a K-SBM, $n^{2/3}(\sigma_1(\tilde{A}) - 2)$ has a type-1 Tracy-Widom distribution asymptotically, where $\sigma_1(\cdot)$ denotes the largest singular value, and $\tilde{A}$ is a standardized version of the adjacency matrix, calculated based on fitting a K-SBM (see Section 1.2). This result requires the entries of the connectivity matrix B to be bounded away from zero which excludes the sparse regime $\nu_n/n \to 0$ we consider here. Moreover, Lei's Theorem 3.3 provides an asymptotic power guarantee. Translating the results to our notation, assuming that the true model has more communities that the fitted model, the result shows that $n^{2/3}\sigma_1(\tilde{A}) \gtrsim \nu_n n^{1/6}$ w.h.p. Since under the true model $n^{2/3}\sigma_1(\tilde{A}) \approx 2n^{2/3}$, one obtains a consistent test as long as $\nu_n n^{1/6} \gg n^{2/3}$, that is, $\nu_n \gg n^{1/2}$. This required scaling is in fact better than what is stated in [Lei16]. Nevertheless, $\nu_n \gg n^{1/2}$ is far from the sparse regime $\nu_n \asymp \log n$ that our results allow. More importantly, it is not clear how to extend the spectral test to the degree-corrected setting. In Section 1.2, we discuss the natural extension of the spectral test to the DCSBM and study its performance empirically. Due to the difficulty of estimating the θ parameter of DCSBM, theoretical guarantees for this (naive) extension are not easy to obtain. Our SNAC+ test avoids explicitly estimating θ, by conditioning on the degrees which leads to the cancellation of individual θ_i in the resulting multinomial distributions. In practice, convergence to the Tracy-Widom distributions is known to be slow, whereas convergence to the normal distribution for SNAC+ happens quite fast (at a rate at most $\approx \nu_n^{-1/2}$ as

we showed).

Another work with connections to ours is that of Karwa et al. [Kar+16] where chi-square statistics for the goodness-of-fit testing of SBM and β-SBM are proposed. They introduce a block-corrected chi-square statistic for the SBM that uses the idea of block compression and has resemblance to our FNAC statistics. The similarity is, however, superficial, since we work conditional on the degrees, hence the parameters we consider are not the connectivity parameters B but their normalized versions ρ (compare equation (5) in [Kar+16] with our equation (2.20)). The ρ parameters have many desirable features; for example, they do not vanish in the sparse regime ($\nu_n/n \to 0$) while the connectivity parameters B do, making the corresponding chi-square statistic numerically very unstable (due to the division by these vanishing parameters). The cancellation of the degree-propensity parameters θ_i in ρ is another key advantage, allowing us to use the same statistic in the degree-corrected case. In contrast, Karwa et al. [Kar+16] devise another test (with no compression) for the β-SBM (a close cousin of DCSBM, in the sparse regime) which requires $O(n^2)$ operations to compute. Another novelty of our approach relative to [Kar+16] is the idea of block compression with $K+1$ communities instead of K which leads to a dramatic increase in the power of the test.

Another major difference with [Kar+16] is their interest in computing exact p-values which requires enumerating all graphs with a given sufficient statistic as the observed one. (For example, for an SBM with known community structure, this translates to enumerating all graphs that have the exact same number of edges between communities as that of the observed network). Although, Karwa et al. develop clever sampling schemes to traverse this space, to get an accurate p-value, one has to sample a prohibitively large number of graphs in general, rendering the approach infeasible beyond small networks. In addition, their main arguments are for block models with a given community structure, and to get around the unknown nature of the communities in practice, they propose sampling the community labels and applying the known-community test on each. The space of all labels is again exponentially large (of size K^n), and one requires a a very large sample to get any reasonable estimate, making the approach infeasible for large

networks. The authors acknowledge this difficulty and suggest using labels obtained by spectral clustering in practice. One then has to worry about the dependence of these labels on the same data the test is computed from, a point where we carefully address in this dissertation. The asymptotic distributions we obtain for the adjusted statistics are very good approximations for large networks and allow us to apply the tests with minimal computational overhead to even networks of millions of nodes.

Compared with likelihood ratio (LR) tests [Yan+14b; WB17], our approach is more general since LR tests require a specific alternative model to compare with (often another SBM or DCSBM), while in goodness-of-fit testing, only the null has to be specified. In addition, rigorous results on LR tests, such as [WB17], often work with a computationally intractable version of the test where the label parameter z is marginalized by summing over K^n possibilities. In practice, these tests are often implemented by approximating the sum via variational inference or plug-in MLE estimators for which the theoretical results of goodness-of-fit do not extend, though the consistency of selecting the correct community number is retained.

A pseudo-LR approach with rigorous guarantees is developed in [MSZ18]. As in [WB17], the focus there, too, is on model selection and comparing DCSBM models, specifying both the null and alternative models, in contrast to FNAC tests. Our approach is comparable to that of [MSZ18] when applied sequentially for model selection, but FNAC family of tests are computationally more efficient: (1) Computing the test statistic of [MSZ18] has $O(n^2)$ computational complexity, whereas due to the column compression, we require only $O(M)$ where M is the number of edges. (2) [MSZ18] creates new labels by binary segmentation, but we save time by reusing the labels estimated by the community detection algorithm. In addition, their consistency results are based on the assumption that the community detection algorithm merges the true communities when it underfits and splits them when it overfits. However, our test only imposes the mild assumption that connectivity parameters are distinguishable among communities, allowing it to be compatible with many community detection algorithms.

As for the degree requirement, our method only requires $\nu_n \gtrsim \log n$, similar to model

selection approaches in [MSZ18; LL22; CL18], and slightly better than those of [LLZ20; WB17] that require $\nu_n/\log n \to \infty$. In contrast, the spectral goodness-of-fit test [BS16; Lei16] has a much more severe requirement ($\nu_n \gg n^{1/2}$) as discussed earlier.

In addition, our work is also related to post-selection inference in clustering [Kim+17; GBW20]. In general clustering setting, it is of interest to test if the estimated clusters are statistically significant. It is commonly done by testing the difference in mean between clusters, but it could suffer inflated type I error if just using its usual null distribution [GBW20]. So more often bootstrap sampling is used [Kim+17]. Here we provide a different solution to use the adjusted chi-square statistic to have a valid null distribution.

2.4 Numerical Experiments

We now illustrate the performance of FNAC+ and SNAC+ on simulated and real networks. We use regularized spectral clustering [Ami+13] as the community detection algorithm, since it is widely used, computationally efficient and conjectured to satisfy Assumption 1 [Abb+20]. Given the number of communities K, the spectral clustering estimates the community labels by applying k-means clustering to the rows of the matrix formed by the K leading eigenvectors of the normalized Laplacian. Regularization is attained by adding $\tau d_{\text{av}}/n$ (where d_{av} is the network average degree) to every entry of the adjacency matrix before forming the Laplacian. This regularization is known to improve the performance in the sparse regime ($d_{\text{av}} \ll n$) [LLV17; ZR18].

Along with the FNAC tests, we consider the following approaches for comparison: Likelihood ratio test (LR) [WB17], Bayesian information criteria (BIC) [WB17], adjusted spectral test (AS) [Lei16], Bethe-Hessian spectral approach (BH) [LL22], network cross-validation (NCV) [CL18] and edge cross-validation (ECV) [LLZ20]. Details for each method can be found in Section 1.2 and 1.3. Note that for LR, we use the plug-in estimators based on labels obtained from the regularized spectral clustering.

For AS, we use the Poisson version of the re-scaled adjacency matrix defined in (1.9). As mentioned previously, though this version has not been proved to have the Tracy-Widom null distribution yet, it is the most natural extension to DCSBM. Moreover,

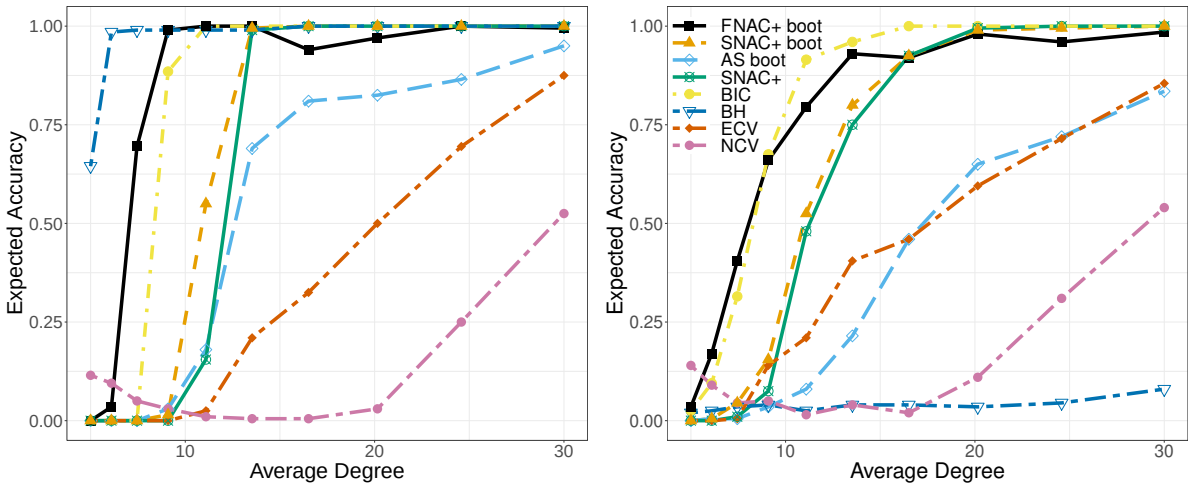


Figure 2.2: Expected accuracy of selecting the true number of communities versus expected average degree of the network. The data follows a DCSBM with $n = 5000$, $K = 4$, $\theta_i \sim \text{Pareto}(3/4, 4)$ and balanced community sizes. The connectivity matrices are B_1 (left) and B_2 (right), as defined in the text.

this significantly improves the computational performance for sparse matrices, since then $\tilde{A}$ can be written as the sum of a sparse matrix and a term involving the product of diagonal and low-rank matrices. This allows fast computation of $\tilde{A}x$ for any vector x, hence allows the singular value computation to scale to very large networks. In some simulations, we also consider AS-SBM, where we use the SBM estimate for $\hat{P}_{ij}$, which is obtained by setting $\hat{\theta}_i = 1$ in (1.9).

2.4.1 Simulations

As discussed earlier, a goodness-of-fit test can be used in a sequential manner to perform model selection. We now provide simulations showing that, when applied sequentially, the family of FNAC+ are consistent, and competitive with other model selection approaches. Here, we report results for samples from Bernoulli DCSBMs. Since we work with sparse networks, the Bernoulli model will be very close to its Poisson version. This was empirically confirmed, as we did not see a significant drop in performance for the FNAC+ tests in our simulations, under a Bernoulli model relative to the Poisson.

Model Selection Performance

Let Pareto(x_0, α) denote a Pareto distribution with scale parameter x_0 and shape parameter α, so that its mean is $\alpha x_0/(\alpha - 1)$. We simulate data from a K-block DCSBM with connection propensity $\theta_i \sim$ Pareto$(3/4, 4)$, and a connectivity matrix which is one of the following:

1. $B_1 \propto (1-\beta)I_K + \beta \mathbf{1}\mathbf{1}^T$, that is, a simple planted partition model with out-in-ratio β,

2. $B_2 \propto \gamma R + (1-\gamma)Q$, where $\gamma \in (0,1)$, R is a random symmetric permutation matrix, and Q a symmetric matrix with i.i.d. Unif$(0,1)$ entries on and above diagonal.

Here, $\mathbf{1}$ is the all-ones vector. In both cases, the matrices are normalized to have a given expected average degree λ. The simple planted partition model B_1 generates a very homogeneous assortative network. Model B_2 creates a more general model by employing the permutation, allowing a mix of assortative and dissortative communities. Model B_2 is in general harder to fit.

Figure 2.2 illustrates the model selection accuracy of various methods for the following setup: $n = 5000$, true $K = 4$ with balanced community sizes, $\beta = 0.2$ and $\gamma = 0.3$. For the goodness-of-fit tests FNAC+, SNAC+ and AS, we use sequential testing from below to estimate K. In each case, the rejection threshold is set to have a significance level of 10^{-6} (under null). For tests with bootstrap debiasing, the number of bootstrap simulations is 10. Figure 2.2 shows the expected model selection accuracy versus the expected average degree λ for each method. The accuracy is obtained by averaging over 200 replications. As λ increases, the problem gets easier and we expect the performance of consistent methods to improve.

For both models B_1 and B_2, the performance of FNAC+ and BIC are close and they outperform other approaches, except for the BH in the case of the B_1 model. Note, however, that BH performs extremely poorly under B_2, showing that associativity is necessary for its consistency. In fact, as pointed out in [LL22], BH requires all the eigenvalues of $\mathbb{E}[A]$ to be positive, which is violated with positive probability under the B_2 model. The

two versions of SNAC+ perform very close to each other and ranked after the FNAC+ and BIC pair. That the performance of the bootstrap SNAC+ is very close to that of SNAC+ with the theoretical threshold, corroborates the accuracy of the null distribution in Theorem 2. The spectral test (AS) performs reasonably well for model B_1, albeit ranked after SNAC+, but relatively poorly under B_2. The cross-validation approaches generally underperform other approaches for model selection, with ECV significantly outperforming NCV.

We include more variants of out-in-ratio and class prior settings in Appendix A.2, which also has examples where FNAC+ significantly outperforms BIC.

ROC Curves

Another way to measure the performance of a test statistic is by means of its Receiver Operating Characteristic (ROC) curve, that is, the power of the test as a function of Type I error; equivalently, the true positive rate (TPR) as a function of the false positive rate (FPR). The ROC curve reveals the best possible performance of a statistic for a given testing problem (one achieved by setting the optimal threshold). Here, we compare the ROC curves of the FNAC+ tests to the likelihood ratio (LR) and spectral (AS) test, for the problem of testing the null hypothesis of $K = 4$ versus the alternative of $K + 1 = 5$ communities. This is an example of "testing from below" which is encountered in sequential model selection.

For the null hypothesis, we consider a simple DCSBM with $K = 4$ communities, having a connectivity matrix of type B_1, introduced in Section 2.4.1, with $\beta = 0.1$. For the alternative, we consider two cases: (a) a DCSBM with $K + 1 = 5$ and otherwise similar parameters to the null DCSBM, and (b) a degree-corrected latent variable model (DCLVM) with $K + 1 = 5$ communities generated as follows: Given a set of latent node variables $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ with $d = K + 1$, the adjacency matrix $A = (A_{ij})$ is generated as a symmetric matrix, with independent Bernoulli entries above the diagonal, with

$$\mathbb{E}[\, A_{ij} \mid x, \theta \,] \;\propto\; \theta_i \theta_j e^{-\|x_i - x_j\|^2} \quad \text{and} \quad x_i = 2e_{z_i} + w_i \tag{2.40}$$

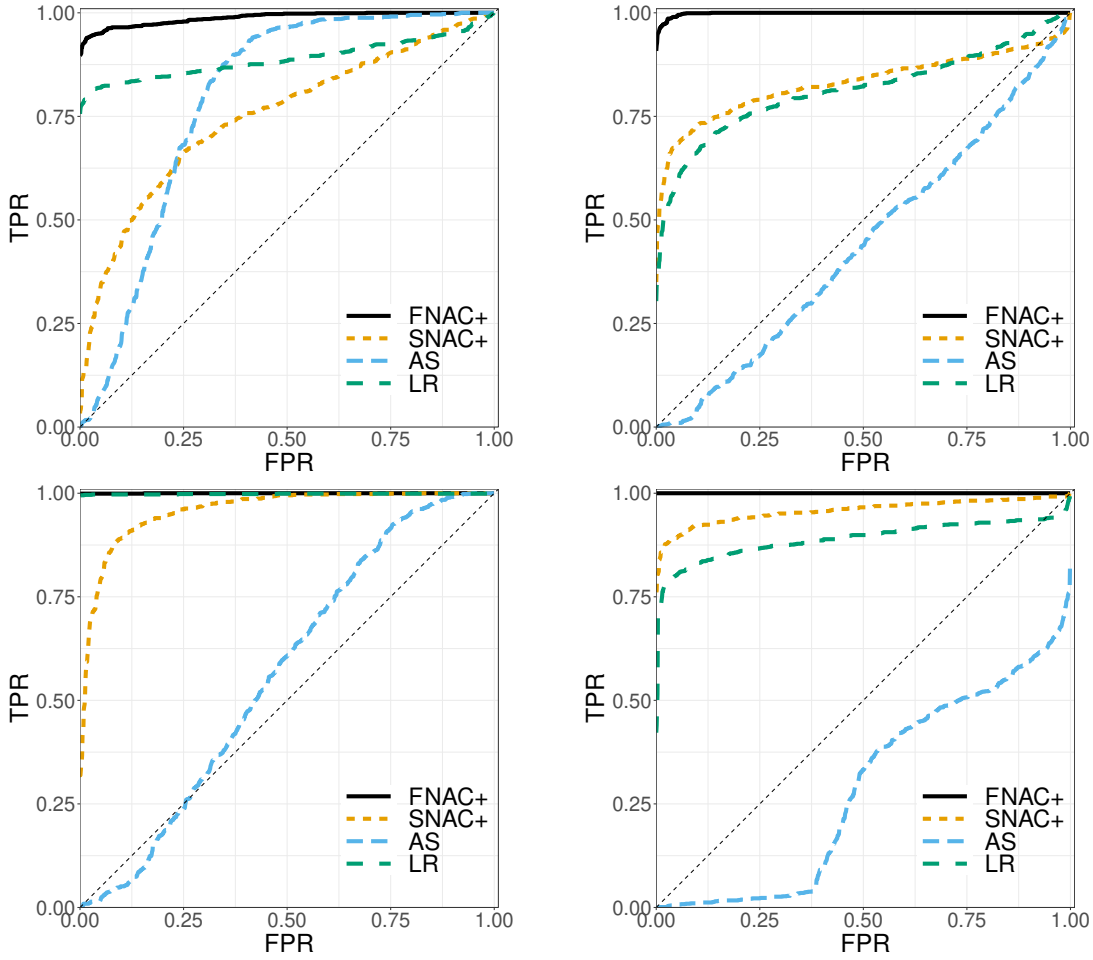


Figure 2.3: ROC plots for testing 4 versus 5 community models. Top and bottom rows correspond to $n = 2000$ and $n = 10000$, respectively. Left and right columns correspond to the DCSBM and DCLVM alternatives, respectively.

where e_k is the kth basis vector of $\mathbb{R}^d$, $w_i \sim N(0, I_d)$ and $\{z_i\} \subset [K+1]^n$ are multinomial labels (similar to the DCSBM labels). In other words, the latent positions $\{x_i\}$ are drawn from a Gaussian mixture model with $K+1 = 5$ components, living in $\mathbb{R}^{K+1}$. The proportionality constant in (2.40) is chosen such that the overall network has expected average degree λ. For all the models, including the null and the two alternatives, the underlying prior on the labels is taken to be proportional to an arithmetic progression: $\mathbb{P}(z_i = k) \propto k$ to produce unequal community sizes, and we let $\theta_i \sim \text{Pareto}(3/4, 4)$. For DCSBM, we use average degree $\lambda = 15$ and DCLVM $\lambda = 8$.

Figure 2.3 illustrates the resulting ROC curves. As expected, increasing n generally improves the performance (except for AS). Both FNAC+ and LR are almost perfect tests for differentiating the two DCSBMs at $n = 10^4$. In all cases, the FNAC+ is more powerful

Table 2.1: Statistics on the FB-100 dataset. Qu. is a short-hand for quartile.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
n	769	4444	9950	12083	17033	41554
Mean deg.	39	65	77	77	88	116
3rd Qu. deg.	54	91	110	108	124	166
Max. deg.	248	673	1202	1787	2123	8246

than the sub-sampled version, SNAC+. This is expected since SNAC+ relies on half the data. Note that as n increases, SNAC+ greatly improves which can be attributed to the label estimation procedure achieving almost exact recovery, even at half the size of the original network. Note that AS generally is much less competitive compared to LR or FNAC+. This is expected since the spectral test relies on a general statistic that is not tailored to the blocked nature of the adjacency matrix of a DCSBM.

FNAC+ is almost perfect test for DCLVM even at $n = 2000$, whereas LR test underperfroms under the DCLVM. This is also expected, since the LR test incorporates the likelihood of a DCSBM for the alternative, which is mismatched to the actual alternative model. This experiment shows the power of FNAC+ family in rejecting against models outside the family of DCSBM. It highlights the advantage of goodness-of-fit over likelihood-ratio testing where one does not have to specify explicitly alternatives, hence can test against many alternatives simultaneously. Companion results for the problem of testing 4 versus 3 communities are reported in Figure A.3 and testing DCSBM and DCLVM with 4 communities are in Figure A.4 of Appendix A.3.

2.4.2 Goodness-of-fit Testing

The main utility of a goodness-of-fit test is to assess how well real data fits the model. Let us investigate how well a DCSBM fits the real networks from the Facebook-100 dataset [Tra+11; TMP12], hereafter referred to as FB-100. This dataset is a collection of 100 social networks, each the entire Facebook network within one university from a date in 2005. The networks vary considerably in size and degree characteristics; some statistics are provided in Table 2.1.

Figure 2.4 shows the violin plots of the SNAC+ statistic with filtering $\sigma = 0.2$ versus

the number of communities for the entire FB-100 data. The variation at each K is due to the variability of SNAC+ over the 100 networks in the dataset. Each FB network has a corresponding synthesized 3-cluster DCSBM network that resembles its degree distribution. Violin plots are also shown for those synthesized DCSBM networks for comparison. For model parameters, each synthesized DCSBM has its own θ parameter proportional to the corresponding FB network degree vector, but they all share the same connectivity matrix B, which is the corresponding MLE based on all the FB networks. To get the shared B, we first apply spectral clustering with $K = 3$ to each FB network $A^{(s)}$, $s = 1, \ldots, 100$ and get estimated label z. Then get the block sum matrix $N^{(s)}$ and block size matrix $M^{(s)}$ respectively with elements $N_{k\ell}(\widehat{z})$ and $m_{k\ell}(\widehat{z})$ defined below the equation (1.7). Finally get $B = \sum_s N^{(s)} / \sum_s M^{(s)}$, where the summation and division matrix operations are element wise. The community sizes are balanced. Kolmogorov–Smirnov test was performed between each FB and synthesized DCSBM network's degrees, and 84 out of such 100 pairs have p-value greater than 0.05, indicating the their closeness. The plots show a marked deviation of FB-100 networks from a DCSBM model as measured by SNAC+ goodness-of-fit test. If the networks were generated from a DCSBM, one would expect the distribution of SNAC+ to drop to within a narrow band around zero once K surpasses the true number of communities. Only at $K = 25$ a small fraction of FB-100 networks have SNAC+ values within, say, the interval $[-5, 5]$, showing that a DCSBM with $K < 25$ is not a good model for any of these networks. Even at $K = 25$, the majority of FB-100 networks are still ill-fitted.

On the other hand, we observe that SNAC+ with filtering at $\sigma = 0.2$ is nearly normal distributed for $K = 3$, while remaining large for $K = 1$ and $K = 2$. This corroborates the results of both Theorem 2 and Theorem 3 that predict exactly this behavior. Note that this conclusion holds despite the variation in the sizes and average degrees of the simulated networks, showing the insensitivity of the null distribution of SNAC+ to those parameters, as predicted by the theory.

Examining the FB-100 data further, one observes that most networks show some very high degree nodes that seem to skew the result of community detection as well as

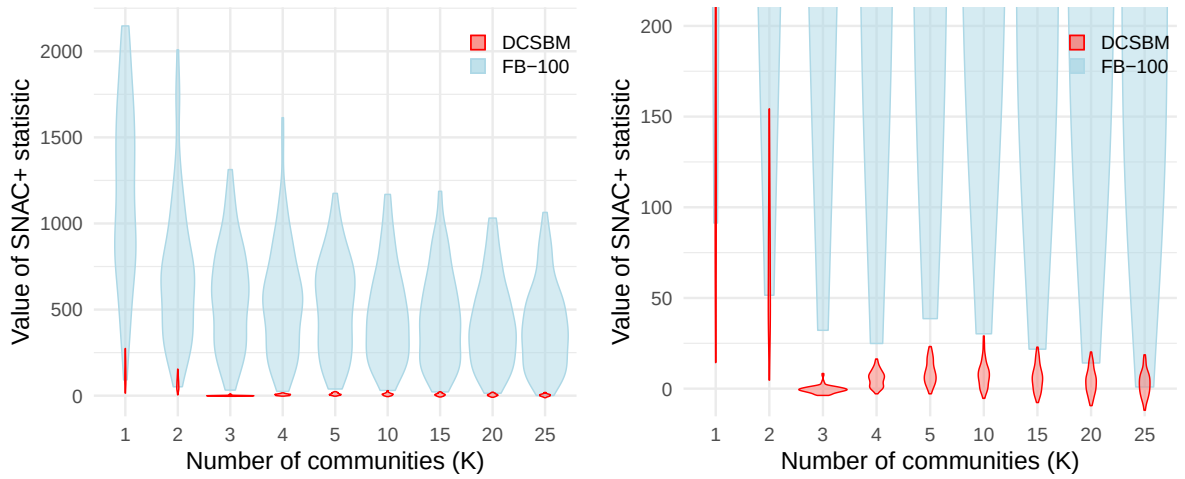


Figure 2.4: Comparing the goodness-of-fit of DCSBM to the FB-100 dataset versus a dataset simulated from a DCSBM with $K = 3$ communities, with the same sizes and average degree characteristics as those of FB-100. The left plot is the zoomed-in version of the right plot.

graph drawing algorithms. This can also be inferred from the significant divide between the third quartile and the maximum degree in Table 2.1. Le et al. [LLV17] have also shown that abnormally high degrees can obstruct community detection. Treating these high-degree nodes as outliers, one could ask what happens if we remove them and refit the model? Figure 2.5 shows the result of performing the same experiment, but applied to the *reduced* FB-100 networks, obtained by restricting to the (induced) subnetwork formed by nodes having degrees below the 3rd quartile (i.e., the 75 percentile). Table 2.2 shows the statistics on these reduced networks, revealing less skewed degree distributions compared to the original data. Figure 2.5 shows that the reduction leads to an overall improvement in the fit: More networks have SNAC+ values that drop to near zero and this happens for lower values of K. This shows the effectiveness of goodness-of-fit testing, in the sense that it allows us to test the hypothesis that removing the high-degree nodes causes a better DCSBM fit. Nevertheless, Figure 2.5 shows that the majority of the reduced networks are still far from a DCSBM with few number of communities.

Table 2.2: Statistics on the reduced FB-100 dataset.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
n	544	3293	7356	8930	12601	30590
Mean deg.	11	20	24	24	28	36
3rd Qu. deg.	16	29	34	34	40	52
Max. deg.	38	74	89	87	101	149

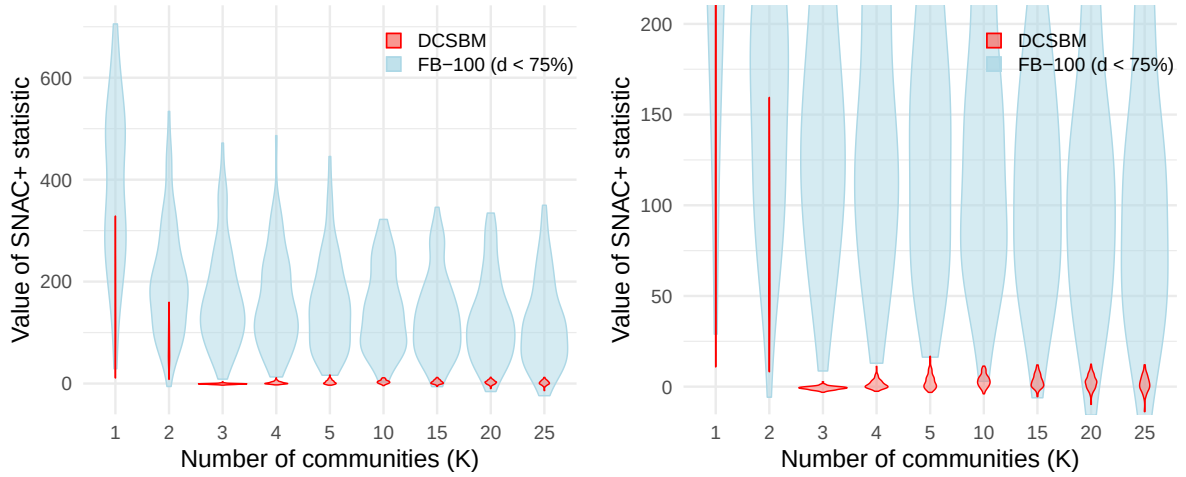


Figure 2.5: Similar to Figure 2.4 but with Facebook networks reduced by restricting to nodes with degrees below the 75 percentile.

2.4.3 Exploring Community Structure

As demonstrated in Section 2.4.2, a DCSBM (with small K) is not a good fit for most of the networks in FB-100. Even in such cases, SNAC+ has utility beyond testing and can be used to reveal community structure in networks. We demonstrate this by using the reduced FB-100 networks constructed in Section 2.4.2. Recall that SNAC+ quantifies the similarity within each estimated community, and a smaller value means that the rows in an estimated community share a similar connection pattern to other communities. Therefore, sharp drops in the value of SNAC+, as K varies can signal the existence of community structure. For a sequence of SNAC+ statistics with increasing K, there could be an *elbow* where continuing to increase K does not bring a significant decrease in the statistic, or a *dip* where SNAC+ starts to increase. These two types of points signal that it is not worthwhile to continue increasing K. Furthermore, these transitions are often much more dramatic for FNAC+ family of tests than the competing methods and can be easily identified by eyeballing the plots.

Figure 2.6 shows the normalized statistic plots for two networks from FB-100. The plots show the normalized value of SNAC+, FNAC+, AS, AS-SBM and negative BIC statistics for $K = 1, \ldots, 13$. The statistics are normalized to fall in the range $[-1, 1]$ by dividing by their largest absolute value, for each test, respectively. This allows us to compare the trend of each statistic as K increases among different methods.

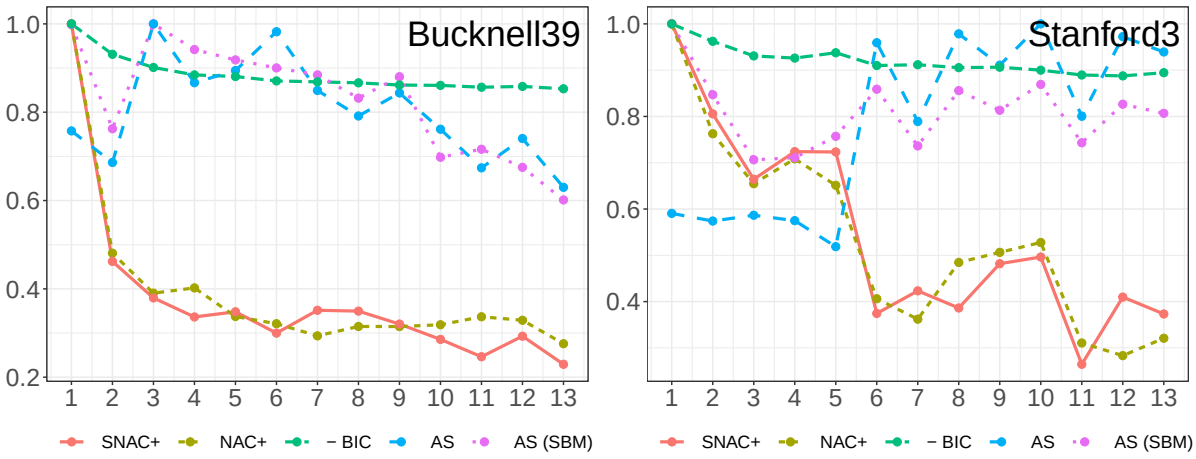


Figure 2.6: Normalized statistics versus the candidate number of communities (K).

For many of the FB-100 networks, SNAC+ and FNAC+ share a similar pattern, with rapid drops followed by the flattening or increase of the statistic, signaling strong community structures. In contrast, AS and AS-SBM generally do not show strong trends, while negative BIC barely fluctuates at all when K increases. For example, for the Bucknell network (Figure 2.6), there is one sharp elbow at $K = 2$ for the FNAC+ tests. The Stanford network shows an elbow/dip at $K = 3$ and a similar elbow/dip at $K = 6$ in FNAC+ tests. This suggests that the network has two levels of community structure (cf. Figure 2.8), an interesting phenomenon not captured by other statistics. Note that AS-SBM captures the community structure at $K = 3$ for the Stanford network (with a dip at $K = 3$) while missing the $K = 6$ possibility. The AS version (employing degree-correction) behaves contrary to expectation in this case and misses both structures.

Community Profiles

We now consider a more quantitative approach to constructing a community profile based on the value of SNAC+. We take advantage of the randomness in SNAC+ due to subsampling, as a natural measure of the uncertainty of the community structure. For each K, we calculate SNAC+ several times (each time using a random split of the nodes) and and then fit a smooth function to the resulting points, treating the problem as a nonparametric regression. Here, we consider smoothing splines but other approaches such as Gaussian kernel ridge regression can be equally useful. The estimated smooth function

provides what we refer to as *a community profile* for the network. This profile can be used for comparing and classifying networks as well as determining possible good choices of the number of communities. The subsampling and smoothing provide a degree of robustness to these profiles as illustrated below.

Instead of eyeballing a plot for its elbows and dips, we can rely on the derivatives of the community profile to guide us. We quantify the elbow as the point where the second derivative has the largest value and the dip as where the first derivative turns positive for the first time. Alternatively, one can use the point with the largest curvature as the elbow point [HO93]. However, we have found, empirically, that the second derivative, as a proxy for the curvature, is much more accurate in capturing the elbow as determined by a human observer.

Figure 2.7 provides instances of three most common patterns of community profiles for the FB-100 networks. For each plot, we show community profiles using two smoothness levels: (1) the dashed red line corresponding to smoothness level set by generalized cross-validation (GCV) [GHW79], and (2) the solid line providing a smoother fit, corresponding to `spar` $= 0.3$, where `spar` is the smoothness parameter in base `R`'s implementation of smoothing splines. The GCV version is usually rougher and captures subtle changes, whereas the solid black fit is smoother and more robust. For each of the two fitted curves, the values of K corresponding to the elbow and dip, as estimated by the derivatives, are given on each plot (with the elbow point recorded first). For example, the Harvard network shows an elbow at $K = 6$ and a dip at $K \approx 3.2$ according to the smoother profile. Compared with normalized plots (Figure 2.6), community profiles show less randomness and the quantified elbows and dips are consistent with those identified by a human observer. It is worth noting that our maximum second derivative criterion for identifying the elbows, surprisingly, almost always returned an integer in these experiments (i.e., no rounding is performed in reporting the elbow points).

The first row in Figure 2.7 shows a single-elbow pattern, and the second row a single first dip (possibly followed by minor smaller dips later on). The third row illustrates a pattern with more than one significant drop, corresponding to multiple elbow/dips.

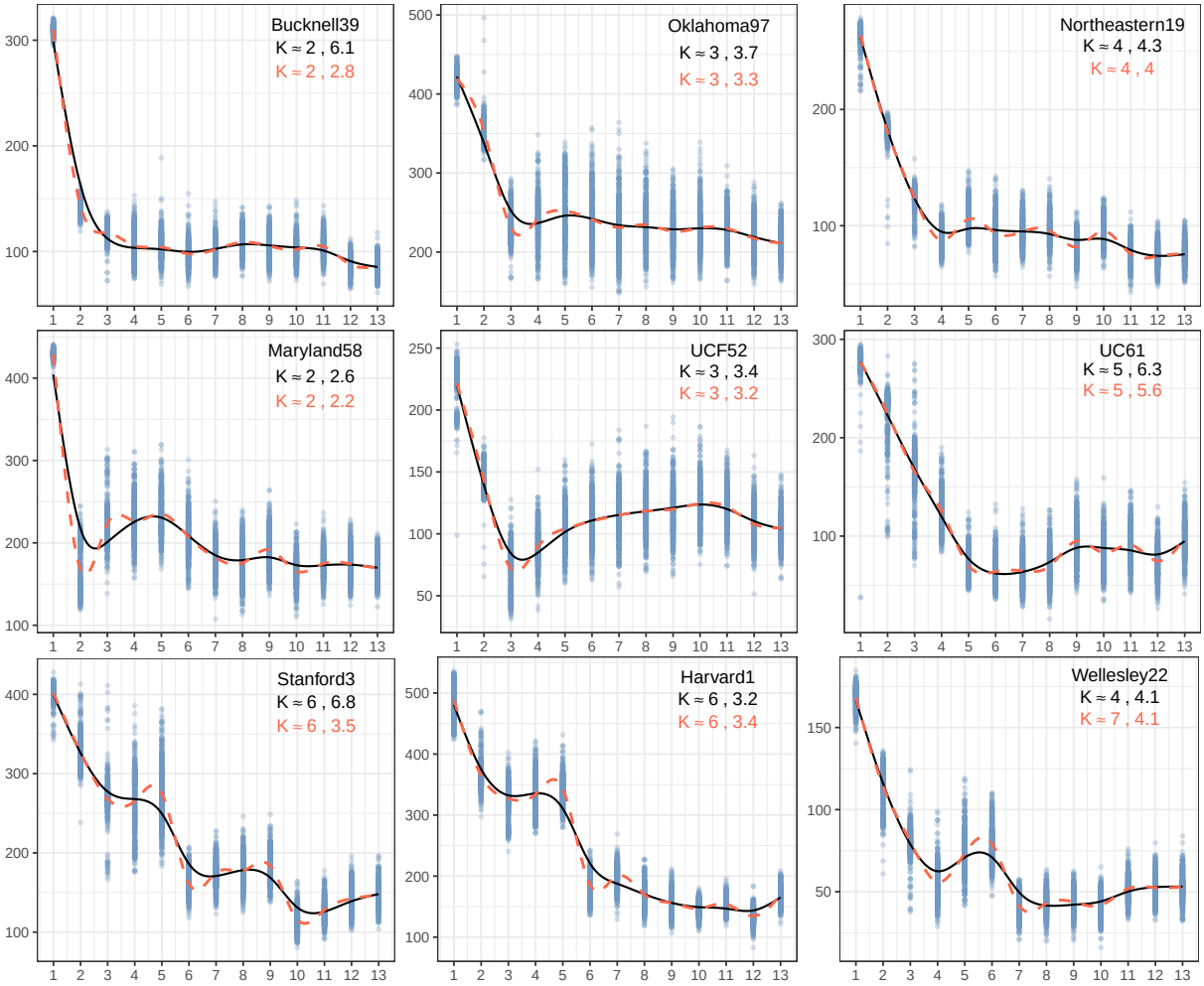


Figure 2.7: Community profile plots. The solid and dashed lines show the smoothed SNAC+ statistic versus the candidate number of communities (K). The dots each represent the SNAC+ value for a random split of the network. The difference between the solid and dashed lines is the smoothness level of the fitted smoothing spline.

This interesting multi-stage behavior is exhibited by a few of the FB-100 networks, and suggests the possibility of breaking the networks into communities in multiple (potentially hierarchical) ways. As mentioned earlier, these multi-stage structures are only captured by SNAC+ among the competing methods. This case illustrates the subtlety of community detection in real networks, showing that insisting on fitting the networks with a single K could lead to missing interesting substructures. We also point out that having an elbow/dip at $K = 2$ is very common for the FB-100 networks; we refer to the additional profile plots provided in Appendix A.4.

Note that in addition to revealing community structure, the absolute value of the profile curves in Figure 2.7 is also informative and measures the distance of the network

form a DCSBM. Since SNAC+ is guaranteed to be centered around zero under a DCSBM, the networks with a larger absolute value of SNAC+ are further away from a DCSBM. For example, Figure 2.7 shows that Wellesley with $K = 4$ communities, having an average SNAC+ value ≈ 60 is a much better fit to DCSBM than Maryland with $K = 2$ communities, showing an average SNAC+ value ≈ 200.

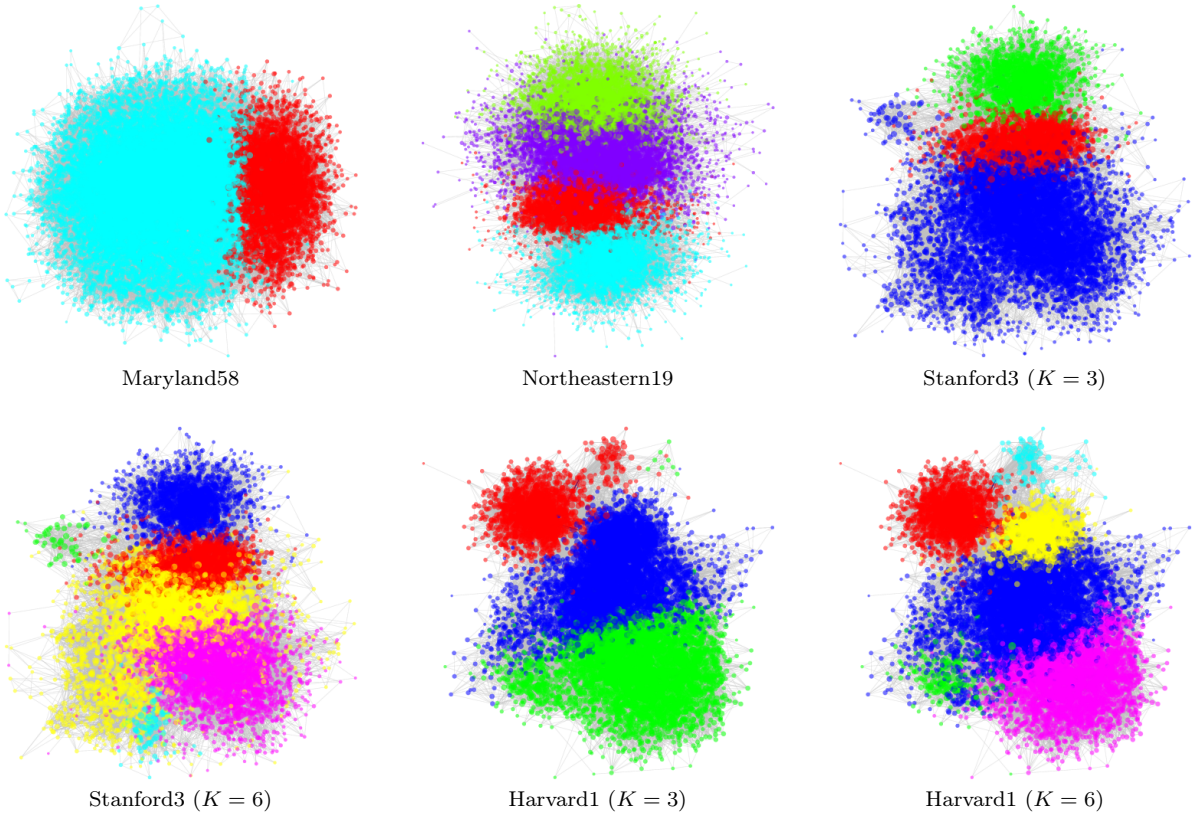


Figure 2.8: FB-100 network plots. The colors specify the estimated communities.

Figure 2.8 shows community structure of some of the FB-100 networks with nodes colored according to their estimated community label. The Stanford and Harvard networks are shown both with $K = 3$ and $K = 6$ estimated communities, as suggested by the two stages of their community profiles. We note that for both of these networks either of these two divisions into communities is visually sensible, with $K = 6$ apparently capturing more refined substructures within the $K = 3$ division. (It is interesting to note that the $K = 6$ partition in each case is not a strict refinement of the $K = 3$ partition, but rather close to being a refinement.) The community structures shown for Maryland and Northeastern are based on the optimal K predicted by their profile plots, and they too make sense visually.

In the Appendix A.4, we also provide normalized and profile plots (Figure A.7) for the political blog network [AG05] which is widely used as a benchmark for community detection. The profile plot shows an elbow at $K = 2$, as identified by the second derivative, matching the expected ground truth of two communities corresponding to the Democratic and Republican parties.

2.5 Proofs of Main Results

2.5.1 Additional Proofs of Theorem 1

We prove Propositions 1 and 2, which are used in the proof of Theorem 1.

Proof of Proposition 1

We first introduce two lemmas used in the proof of Proposition 1. Lemma 1 is on the mean and variance of the chi-square statistic. Lemma 2 is a general result on the growth rate of the third central moment of the empirical variance of a sum of independent variables. Applying this result to a chi-square statistic, we can bound its third central moment by some constant. Plugging them in to Essen's bound, we show the sum of such chi-square statistics normalized by its mean and standard deviation, is close to the standard normal distribution. Next we show that replacing the exact standard deviation in the normalized sum with a simpler form (to get T_n) pays a small price in the distance to the standard normal distribution (Lemma 24).

Recall that $\psi(x, y) := (x - y)^2/y$.

Lemma 1 (Variance of the chi-square statistic)**.** *Let $X = (X_1, \ldots, X_L) \sim \text{Mult}(d, p)$, where $p = (p_1, \ldots, p_L)$ is a probability vector and let $Y := \sum_{\ell=1}^{L} \psi(X_\ell, dp_\ell)$. Then, for $L \geq 2$,*

$$\begin{aligned} \mathbb{E}[Y] &= L - 1, \\ \text{var}(Y) &= \Big(1 - \frac{1}{d}\Big) 2(L-1) + \frac{1}{d}\Big(\frac{L}{h(p)} - L^2\Big). \end{aligned}$$

In particular, $\mathrm{var}(Y) \geq (1 - 1/d)2(L-1)$

Note that we always have $L/h(p) \geq L^2$ since $\sum_\ell p_\ell = 1$. Hence, the variance of Y is a convex combination of two nonnegative terms. Furthermore, if $d \geq 2$, $\mathrm{var}(Y) \geq L - 1$.

Lemma 2 (Central moment growth)**.** *Let* $\{W_1, \ldots, W_n\}$ *be a sequence of i.i.d. zero mean random variables with finite moments of order 6, and let* $X_n = \sum_{i=1}^n W_i$*. Then, the third central moment of* X_n^2 *is* $O(n^3)$*:*

$$\mathbb{E}\big|X_n^2 - \mathbb{E}X_n^2\big|^3 \leq C_{W_1} n^3,$$

where C_{W_1} *is a constant that only depends on the first 6 moments of* W_1*. For the case where* $W_1 = \alpha(Z - p)$ *with* $Z \sim \mathrm{Ber}(p)$ *and* $\alpha \in \mathbb{R}$*, one can take* $C_{W_1} = 34.5\,\alpha^6 p(1-p)$*.*

Proof Proposition 1. By Esseen's bound for non-identically distributed summands [She10],

$$d_{\mathrm{K}}(S_n, Z) \leq \frac{C_0}{(v_n^2)^{3/2}} \sum_{i=1}^n \mathbb{E}|Y_i - \mathbb{E}[Y_i]|^3 \tag{2.41}$$

for some constant $C_0 \in [0.41, 0.56]$. By Lemma 1, $\mathrm{var}(Y_i) \geq (1 - d_i^{-1})2(L-1)$. Then, using assumption $h(d) \geq 2$,

$$v_n^2 = \sum_{i=1}^n \mathrm{var}(Y_i) \geq n\big(1 - h(d)^{-1}\big)2(L-1) \geq n(L-1). \tag{2.42}$$

Next, we bound the third central moment of Y_i. Let $Z_{i\ell} = (X_{i\ell} - d_i p_{g_i\ell})/p_{g_i\ell}$. We have $Y_i = \sum_\ell p_{g_i\ell} Z_{i\ell}^2/d_i$. We can write $Z_{i\ell} = \sum_{j=1}^{d_i}(W_j - p_{g_i\ell})/p_{g_i\ell}$, where $W_j \overset{i.i.d.}{\sim} \mathrm{Ber}(p_{g_i\ell})$. By Lemma 2, $\mathbb{E}|Z_{i\ell}^2 - \mathbb{E}Z_{i\ell}^2|^3 \leq C_{p_{g_i\ell}} d_i^3$, for some constant $C_{p_{g_i\ell}}$ that only depends on $p_{g_i\ell}$. Then,

$$\begin{aligned}
\mathbb{E}|Y_i - \mathbb{E}[Y_i]|^3 &= \mathbb{E}\Big|\sum_{\ell=1}^L p_{g_i\ell}(Z_{i\ell}^2 - \mathbb{E}Z_{i\ell}^2)/d_i\Big|^3 \\
&\leq \sum_{\ell=1}^L \frac{p_{g_i\ell}}{d_i^3}\mathbb{E}\big|Z_{i\ell}^2 - \mathbb{E}Z_{i\ell}^2\big|^3 \leq \sum_{\ell=1}^L p_{g_i\ell}\Big(\frac{34.5}{p_{g_i\ell}^6} p_{g_i\ell}(1 - p_{g_i\ell})\Big)
\end{aligned} \tag{2.43}$$

where the first inequality is the discrete Jensen's inequality applied to convex function

$x \mapsto |x|^3$, that is, $|\sum_\ell q_\ell x_\ell|^3 \le \sum_\ell q_\ell |x_\ell|^3$ for any $\{x_\ell\}$ and probability vector $q = (q_\ell)$. Combining (2.41), (2.42) and (2.43) gives

$$\sqrt{n}\, d_{\mathrm{K}}(S_n, Z) \le \frac{34.5 C_0}{(L-1)^{3/2}} \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{L} \frac{1}{p_{g_i \ell}^4} \le 34.5 C_0 2^{3/2} \frac{1}{L^{1/2} \underline{p}^4} \le \frac{55}{L^{1/2} \underline{p}^4}$$

using $p_{g_i \ell} \ge \underline{p}$ for all i and ℓ, $L - 1 \ge L/2$ and $C_0 \le 0.56$.

To prove (2.13), let $\beta_n = v_n/(\sqrt{2}\gamma_n)$, so that $T_n = \beta_n S_n$. By Lemma 24,

$$d_{\mathrm{K}}\big(T_n, Z\big) \le d_{\mathrm{K}}\big(S_n, Z\big) + \frac{\zeta_n}{\sqrt{2\pi e}}, \quad \zeta_n := \frac{|\beta_n - 1|}{\min\{\beta_n, 1\}}.$$

It remains to bound ζ_n. Let $d_{\mathcal{G}_k} = (d_i, i \in \mathcal{G}_k)$ and $n_k = |\mathcal{G}_k|$. By Lemma 1,

$$\begin{aligned} v_n^2 &= \sum_{i=1}^{n} \big(1 - d_i^{-1}\big) 2(L-1) + d_i^{-1} \big(L h(p_{g_i *})^{-1} - L^2\big) \\ &= \sum_k n_k \Big[\big(1 - h(d_{\mathcal{G}_k})^{-1}\big) 2(L-1) + h(d_{\mathcal{G}_k})^{-1} \big(L h(p_{k*}) - L^2\big) \Big] \end{aligned}$$

where the second line follows by breaking the sum as $\sum_{i=1}^n (\cdots) = \sum_{k=1}^K \sum_{i \in \mathcal{G}_k} (\cdots)$ and using $n_k h(d_{\mathcal{G}_k})^{-1} = \sum_{i \in \mathcal{G}_k} d_i^{-1}$. To simplify, let $\alpha_k = h(d_{\mathcal{G}_k})^{-1}$. Then,

$$\beta_n = \frac{v_n}{\sqrt{2}\gamma_n} = \Big(\sum_k \pi_k (1 + \alpha_k b_k) \Big)^{1/2}, \quad b_k := \frac{L h(p_{k*})^{-1} - L^2}{2(L-1)} - 1.$$

where $\pi_k = n_k/n$. Since $L \le h(p_{k*})^{-1} \le \underline{p}^{-1}$ and $L/2 \le L - 1$, we have

$$0 \le b_k + 1 \le \frac{L(\underline{p}^{-1} - L)}{2(L-1)} \le \underline{p}^{-1} - L. \tag{2.44}$$

Let $u = \sum_k \pi_k \alpha_k b_k$ and note that $\beta_n = \sqrt{1+u}$. We have $0 < \sum_k \pi_k \alpha_k = h(d)^{-1} \le 1/2$, by assumption. Moreover $b_k \ge -1$ for all k from (2.44). It follows that $u \ge -1/2$.

If $u \ge 0$, then $\beta_n \ge 1$ and $\zeta_n = \beta_n - 1 \le \frac{1}{2}u$, using the inequality $\sqrt{1+x} \le 1 + x/2$ which holds for all $x \ge -1$. If $u < 0$, then $\beta_n \in (0, 1)$, and

$$\zeta_n = \frac{1}{\beta_n} - 1 = \frac{1}{\sqrt{1 - |u|}} - 1 \le \sqrt{2}|u|,$$

using $|u| \le 1/2$ and the inequality $(1-x)^{-1/2} \le 1 + \sqrt{2}x$ which holds for $0 \le x \le 0.77$. We have $|b_k| \le \max\{1, \underline{p}^{-1} - L - 1\}$, hence $\zeta_n \le \sqrt{2}\max\{1, \underline{p}^{-1} - L - 1\}h(d)^{-1}$. The proof is complete. □

Proof of Proposition 2

Our strategy for proving Proposition 2 is to show that $\widehat{T}_n$ is close to T_n via a chain of intermediate counterparts—namely $\widetilde{T}_n$ and $\widetilde{T}_n^*$—defined by replacing estimated clusters and probabilities with their true versions; see (2.45) and the subsequent paragraph. The fact that the chi-square statistic does not change very much when the probabilities are slightly perturbed (Lemma 3) helps us show that $\widetilde{T}_n$ is close to $\widetilde{T}_n^*$ and $\widetilde{T}_n^*$ is close to T_n.

It remains to show that $\widehat{T}_n$ is close to $\widetilde{T}_n$. Here, the probabilities defining the underlying chi-square statistics are the same (both estimated), but the clusters are different (estimated versus true). For this step, we use a uniform bound to avoid the dependence of the estimated clusters on the same data used to form the statistic. This is where we need $d_{\max}\alpha_n\sqrt{n} = o(1)$.

Once we show that $\widehat{T}_n$ is close to T_n with high probability, we use the fact that for two random variables close to each other, their Kolmogorv distances to the standard normal distribution are also close (Lemma 4).

Throughout the proof, there will be a parameter u and a derived parameter δ based on u. We set u in the end to balance all the terms; see the discussion after the statement of Proposition 2. But in reading the proof, it could help to consider the case where all d_i are of the same order say $d_i \asymp d$. Then u will scale like $\log d$ and hence δ defined in (2.48) sclaes as $\delta = O(\sqrt{\log d/(nd)})$.

We are now ready to give the detailed proof. First, we state the auxiliary lemmas.

Lemma 3. *Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^d$ and $y, y+v \in \mathbb{R} \setminus \{0\}$, and consider the function $G(v) = \sum_{i=1}^n d_i \psi(x_i, y+v)$ where $\{d_i\}$ are nonnegative and $\psi(s,t) = (s-t)^2/t$. Let $R = \sum_i d_i x_i - d_+ y$ where $d_+ = \sum_{i=1}^n d_i$, and assume further that $|v| \le |y|/2$. Then,*

$$|G(v) - G(0)| \le \frac{2|v|}{|y|}\big[G(0) + 2|R| + |v|d_+\big].$$

Lemma 4. *Let $\delta \in [0, 1/2]$ and $\varepsilon > 0$. Then, for any two random variables $\widehat{T}_n$ and T_n, and $Z \sim N(0,1)$*

$$d_K(\widehat{T}_n, Z) \le d_K(T_n, Z) + \frac{1}{2}(\delta + \varepsilon) + \mathbb{P}\big(|\widehat{T}_n - T_n| \ge \delta T_n + \varepsilon\big).$$

Next, we introduce the intermediaries between $\widehat{T}_n$ and T_n. Consider

$$Y(\{\mathcal{G}_k\}, \{p_{k\ell}\}) := \sum_{k=1}^{K} \sum_{i \in \mathcal{G}_k} \sum_{\ell=1}^{L} \psi(X_{i\ell}, d_i p_{k\ell})$$

and let

$$\begin{aligned} \widehat{Y} &= Y(\{\widehat{\mathcal{G}}_k\}, \{\widehat{p}_{k\ell}\}), \\ \widetilde{Y} &= Y(\{\mathcal{G}_k\}, \{\widehat{p}_{k\ell}\}), \\ \widetilde{Y}^* &= Y(\{\mathcal{G}_k\}, \{\widetilde{p}_{k\ell}\}), \\ Y &= Y(\{\mathcal{G}_k\}, \{p_{k\ell}\}) \end{aligned} \tag{2.45}$$

where, for $k \in [K]$ and $\ell \in [L]$,

$$\widehat{p}_{k\ell} = \frac{\sum_{i \in \widehat{\mathcal{G}}_k} X_{i\ell}}{\sum_{i \in \widehat{\mathcal{G}}_k} d_i}, \quad \widetilde{p}_{k\ell} = \frac{\sum_{i \in \mathcal{G}_k} X_{i\ell}}{\sum_{i \in \mathcal{G}_k} d_i}. \tag{2.46}$$

We define the corresponding T-statistics based on Y-statistics, via the relation $Y = \sqrt{2}\gamma_n T + \gamma_n^2$. For example,

$$\widehat{Y} = \sqrt{2}\gamma_n \widehat{T}_n + \gamma_n^2$$

and similarly for $\widetilde{T}_n$, $\widetilde{T}_n^*$ and T_n. The rest of proof is devoted to showing that $\widetilde{T}_n^*$ is close to T_n, $\widetilde{T}_n$ is close to $\widetilde{T}_n^*$ and $\widehat{T}_n$ is close to $\widetilde{T}_n$.

Controlling probability estimates We first show that the probabilities in (2.46) are close to their true counterparts, $p_{k\ell}$. Let

$$X_{+\ell}^{(k)} = \sum_{i \in \mathcal{G}_k} X_{i\ell}, \quad d_+^{(k)} = \sum_{i \in \mathcal{G}_k} d_i,$$

and

$$\omega_{n,1/2} = \Big(\sum_k (\pi_k d_{\text{av}}^{(k)})^{1/2} \Big)^2, \quad \omega_{n,1} = \sum_k \pi_k d_{\text{av}}^{(k)}, \tag{2.47}$$

and $\widehat{\Delta}_{k\ell} = \widehat{p}_{k\ell} - \widetilde{p}_{k\ell}$ and $\widetilde{\Delta}_{k\ell} = \widetilde{p}_{k\ell} - p_{k\ell}$.

First, we control $\widetilde{\Delta}_{k\ell}$. Let

$$\delta_k := 2(u/d_+^{(k)})^{1/2}, \quad \delta := \max_k \delta_k, \tag{2.48}$$

for $u \geq 0$ in the statement of the proposition, and consider the event

$$\mathcal{B} := \Big\{ \max_\ell |\widetilde{\Delta}_{k\ell}| \leq \delta_k, \ \forall k \in [K] \Big\}. \tag{2.49}$$

Lemma 5. $\mathbb{P}(\mathcal{B}^c) \leq 2KLe^{-u}$ *whenever* $u \leq \min_k d_+^{(k)}$.

Recalling the defintion of ω_n in (2.9), we note that $\min_k d_+^{(k)} = n\omega_n$. Then, $u \leq (\underline{p}/8)^2 n\omega_n \leq \min_k d_+^{(k)}$ where the first inequality is by assumption. Hence, the condition of Lemma 16 holds and $\mathcal{B}$ is a high probability event. For the rest of the proof, we work on $\mathcal{B}$. Moreoever, $u \leq (\underline{p}/8)^2 n\omega_n \leq (\underline{p}/8)^2 d_+^{(k)}$ for all k, from which it follows that $\delta \leq \underline{p}/4$. Since on $\mathcal{B}$, we have $\max_{k,\ell} |\widetilde{\Delta}_{k\ell}| \leq \delta$, then for all k, ℓ,

$$\widetilde{p}_{k\ell} \geq p_{k\ell} - \delta \geq \underline{p}/2. \tag{2.50}$$

Next, we control $\widehat{\Delta}_{k\ell}$. Recall that $\tau_d = \omega_n / d_{\max}$ as defined in (2.9). Let

$$\mathcal{M}_n := \{\text{Mis}(\widehat{z}, z) \leq \alpha_n\}.$$

Lemma 6. *Assume that* $\alpha_n \leq \tau_d \, \underline{p}/2$ *and* $\delta \leq \underline{p}/2$ *and let*

$$\hat{\delta} := \frac{6}{\underline{p} \, \tau_d} \alpha_n. \tag{2.51}$$

Then, on $\mathcal{B} \cap \mathcal{M}_n$, *we have* $|\widehat{\Delta}_{k\ell}| \leq \hat{\delta} \cdot \widetilde{p}_{k\ell}$ *for all* k *and* ℓ.

Since by assumption in Theorem 1, $\alpha_n \le \underline{p}/(8C_{3,p})$, we have $\hat{\delta} \le \underline{p}/8$, hence applying Lemma 6, we obtain

$$\widehat{p}_{k\ell} \ge \widetilde{p}_{k\ell} - \hat{\delta} \ge \underline{p}/2 - \underline{p}/8 \ge \underline{p}/4.$$

Controlling $\widetilde{T}_n^*$ in terms of T_n Apply Lemma 3 with $x_i = X_{i\ell}/d_i$, $y = p_{k\ell}$ and $v = \widetilde{p}_{k\ell} - p_{k\ell} = \widetilde{\Delta}_{k\ell}$. The condition $|v| \le |y|/2$ of the lemma is satisfied on $\mathcal{B}$, as long as $\delta \le \underline{p}/2$, which is the case as established earlier. Let

$$G_{k\ell}(\widetilde{\Delta}_{k\ell}) = \sum_{i \in \mathcal{G}_k} d_i \psi(X_{i\ell}/d_i, p_{k\ell} + \widetilde{\Delta}_{k\ell}).$$

We have $\widetilde{Y}^* = \sum_{k,\ell} G_{k\ell}(\widetilde{\Delta}_{k\ell})$ and $Y = \sum_{k,\ell} G_{k\ell}(0)$, hence

$$\begin{aligned}
|\widetilde{Y}^* - Y| &\le \sum_{k,\ell} |G_{k\ell}(\widetilde{\Delta}_{k\ell}) - G_{k\ell}(0)| \\
&\le 2 \sum_{k,\ell} \frac{|\widetilde{\Delta}_{k\ell}|}{p_{k\ell}} \Big[G_{k\ell}(0) + 2|X_{+\ell}^{(k)} - d_+^{(k)} p_{k\ell}| + |\widetilde{\Delta}_{k\ell}| d_+^{(k)} \Big] \\
&= 2 \sum_{k,\ell} \frac{|\widetilde{\Delta}_{k\ell}|}{p_{k\ell}} \Big[G_{k\ell}(0) + 3|\widetilde{\Delta}_{k\ell}| d_+^{(k)} \Big]
\end{aligned}$$

where we have used $X_{+\ell}^{(k)} - d_+^{(k)} p_{k\ell} = d_+^{(k)} \widetilde{\Delta}_{k\ell}$ since $\widetilde{p}_{k\ell} = X_{+\ell}^{(k)}/d_+^{(k)}$. By assumption $p_{k\ell} \ge \underline{p}$ for all k and ℓ. Hence,

$$\begin{aligned}
\sqrt{2}\gamma_n |\widetilde{T}_n^* - T_n| = |\widetilde{Y}^* - Y| &\le \frac{2}{\underline{p}} \Big[\delta \sum_{k,\ell} G_{k\ell}(0) + 3L \sum_k \delta_k^2 d_+^{(k)} \Big] \\
&= \frac{2}{\underline{p}} \Big[\delta(\sqrt{2}\gamma_n T_n + \gamma_n^2) + 12LKu \Big].
\end{aligned}$$

Then, on $\mathcal{B}$, we have

$$|\widetilde{T}_n^* - T_n| \le \frac{2}{\underline{p}} \Big[\delta(T_n + \sqrt{nL/2}) + 12Ku\sqrt{L/n} \Big]$$

using $\sqrt{nL/2} \le \gamma_n \le \sqrt{nL}$ which holds for $L \ge 2$. Since $2\delta/\underline{p} \le 1/2$, we can apply Lemma 4 to get

$$d_{\mathrm{K}}(\widetilde{T}_n^*, Z) \le d_K(T_n, Z) + \frac{1}{\underline{p}}\Big[\delta(1+\sqrt{nL/2}) + 12Ku\sqrt{L/n}\Big] + \mathbb{P}(\mathcal{B}^c). \tag{2.52}$$

Controlling $\widetilde{T}_n$ in terms of $\widetilde{T}_n^*$ We consider the event $\mathcal{B} \cap Mc_n$ from now on. We apply Lemma 3 with $x_i = X_{i\ell}/d_i$, $y = \widetilde{p}_{k\ell}$ and $v = \widehat{p}_{k\ell} - \widetilde{p}_{k\ell} = \widehat{\Delta}_{k\ell}$. Condition $|v| \le |y|/2$ of the lemma is satisfied, as long as $\hat{\delta} \le \underline{p}/2$, which is the case as established earlier. Letting

$$F_{k\ell}(\Delta) := \sum_{i \in \mathcal{G}_k} d_i \psi(X_{i\ell}/d_i, \widetilde{p}_{k\ell} + \Delta),$$

Lemma 3 implies

$$\begin{aligned}
|F_{k\ell}(\widehat{\Delta}_{k\ell}) - F_{k\ell}(0)| &\le \frac{2|\widehat{\Delta}_{k\ell}|}{\widetilde{p}_{k\ell}}\Big(F_{k\ell}(0) + 2|X_{+\ell}^{(k)} - d_+^{(k)}\widetilde{p}_{k\ell}| + |\widehat{\Delta}_{k\ell}|d_+^{(k)}\Big) \\
&\le \frac{4}{\underline{p}}|\widehat{\Delta}_{k\ell}|\Big(F_{k\ell}(0) + \big[2|\widetilde{\Delta}_{k\ell}| + |\widehat{\Delta}_{k\ell}|\big]d_+^{(k)}\Big),
\end{aligned}$$

where we have used $d_+^{(k)}\widetilde{\Delta}_{k\ell} = X_{+\ell}^{(k)} - d_+^{(k)}\widetilde{p}_{k\ell}$ and $\widetilde{p}_{k\ell} \ge \underline{p}/2$ on event $\mathcal{B}$; see (2.50). We have $\widetilde{Y} = \sum_{k,\ell} F_{k\ell}(\widehat{\Delta}_{k\ell})$ and $\widetilde{Y}_* = \sum_{k,\ell} F_{k\ell}(0)$. It follows that

$$\begin{aligned}
\sqrt{2}\gamma_n|\widetilde{T}_n - \widetilde{T}_n^*| = |\widetilde{Y} - \widetilde{Y}^*| &\le \sum_{k,\ell}|F_{k\ell}(\widehat{\Delta}_{k\ell}) - F_{k\ell}(0)| \\
&\le \frac{4}{\underline{p}}\hat{\delta}\Big(\sum_{k,\ell}F_{k\ell}(0) + L\sum_k\big[2\delta_k + \hat{\delta}\big]d_+^{(k)}\Big) \\
&= \frac{4}{\underline{p}}\hat{\delta}\Big(\widetilde{Y}_* + 2L\sum_k \delta_k d_+^{(k)} + L\hat{\delta}d_+\Big).
\end{aligned}$$

Using $d_+^{(k)} = d_{\mathrm{av}}^{(k)}\pi_k n$, and the defintions (2.47) and (2.48), we obtain

$$\sum_k \delta_k d_+^{(k)} = 2\sum_k (u d_+^{(k)})^{1/2} = 2\sqrt{nu\omega_{n,1/2}}.$$

Noting that $d_+ = n\omega_{n,1}$, we have

$$\sqrt{2}\gamma_n|\widetilde{T}_n - \widetilde{T}_n^*| \le \frac{4}{\underline{p}}\hat{\delta}\Big(\sqrt{2}\gamma_n\widetilde{T}_n^* + \gamma_n^2 + 4L\sqrt{nu\omega_{n,1/2}} + L\omega_{n,1}\hat{\delta}n\Big).$$

Using $\sqrt{nL/2} \le \gamma_n \le \sqrt{Ln}$, we obtain, on $\mathcal{B}\cap\mathcal{M}_n$,

$$|\widetilde{T}_n - \widetilde{T}_n^*| \le \frac{4\hat{\delta}}{\underline{p}}\Big(\widetilde{T}_n^* + \sqrt{nL/2} + 4\sqrt{Lu\omega_{n,1/2}} + \omega_{n,1}\hat{\delta}\sqrt{Ln}\Big).$$

Recalling that $\hat{\delta}/\underline{p} \le 1/8$, Lemma 4 gives

$$\begin{aligned} d_{\mathrm{K}}(\widetilde{T}_n, Z) \le d_{\mathrm{K}}(\widetilde{T}_n^*, Z) \ &+ \mathbb{P}(\mathcal{B}^c \cup \mathcal{M}_n^c) \\ &\frac{2\hat{\delta}}{\underline{p}}\Big(1 + \sqrt{nL/2} + 4\sqrt{Lu\omega_{n,1/2}} + \omega_{n,1}\hat{\delta}\sqrt{Ln}\Big). \end{aligned} \tag{2.53}$$

Controlling $\widehat{T}_n$ in terms of $\widetilde{T}_n$ Working $\mathcal{B}\cap\mathcal{M}_n$ and recalling $\widehat{p}_{k\ell} \ge \underline{p}/4$,

$$\sum_\ell \psi(X_{i\ell}, d_i\widehat{p}_{k\ell}) = \frac{d_i}{\widehat{p}_{k\ell}}\sum_\ell (X_{i\ell}/d_i - \widehat{p}_{k\ell})^2 \le 8d_i/\underline{p},$$

where we have used the following result:

Lemma 7. $\max_{x,y\in\mathcal{P}_L}\|x-y\|^2 = 2$, *where $\mathcal{P}_L$ is the probability simplex in $\mathbb{R}^L$.*

Letting $\mathcal{H}_k = \mathcal{G}_k \Delta \widehat{\mathcal{G}}_k := (\mathcal{G}_k \setminus \widehat{\mathcal{G}}_k) \cup (\widehat{\mathcal{G}}_k \setminus \mathcal{G}_k)$,

$$|\widehat{Y} - \widetilde{Y}| \le \sum_{k,\ell}\sum_{i\in\mathcal{H}_k}\psi(X_{i\ell}, d_i\widehat{p}_{k\ell}) \le \frac{8}{\underline{p}}\sum_k\sum_{i\in\mathcal{H}_k} d_i \le \frac{8d_{\max}}{\underline{p}}\alpha_n n$$

using $\sum_k |\mathcal{H}_k| \le \alpha_n n$. Hence, on $\mathcal{B}\cap\mathcal{M}_n$, we have

$$|\widehat{T}_n - \widetilde{T}| \le \frac{1}{\sqrt{2}\gamma_n}|\widehat{Y} - \widetilde{Y}| \le \frac{8d_{\max}\alpha_n\sqrt{n}}{\underline{p}\sqrt{L}}$$

using $\sqrt{2}\gamma_n \ge \sqrt{nL}$. Applying Lemma 4,

$$d_{\mathrm{K}}(\widehat{T}_n, Z) \le d_{\mathrm{K}}(\widetilde{T}_n, Z) + \frac{4d_{\max}\alpha_n\sqrt{n}}{\underline{p}\sqrt{L}} + \mathbb{P}(\mathcal{B}^c \cup \mathcal{M}_n^c). \tag{2.54}$$

Putting the pieces together Combining (2.52), (2.53) and (2.54), we have

$$
\begin{aligned}
d_{\mathrm{K}}(\widehat{T}_n, Z) \leq d_K(T_n, Z) + \\
&\frac{1}{\underline{p}}\Big[\delta(1+\sqrt{nL/2}) + 12Ku\sqrt{L/n}\Big] + \mathbb{P}(\mathcal{B}^c) + \\
&\frac{2\hat{\delta}}{\underline{p}}\Big(1+\sqrt{nL/2} + 4\sqrt{Lu\omega_{n,1/2}} + \omega_{n,1}\hat{\delta}\sqrt{Ln}\Big) + \mathbb{P}(\mathcal{B}^c \cup \mathcal{M}_n^c) + \\
&\frac{4d_{\max}\alpha_n\sqrt{n}}{\underline{p}\sqrt{L}} + \mathbb{P}(\mathcal{B}^c \cup \mathcal{M}_n^c).
\end{aligned}
$$

Using $1+\sqrt{nL/2} \leq 2\sqrt{nL/2}$ and the union bound,

$$
\begin{aligned}
d_{\mathrm{K}}(\widehat{T}_n, Z) - d_K(T_n, Z) \leq \\
&\frac{\sqrt{L}}{\underline{p}} \cdot \Big[2\hat{\delta}\Big(\sqrt{2n} + 4\sqrt{u\omega_{n,1/2}} + \omega_{n,1}\hat{\delta}\sqrt{n}\Big) + \frac{4}{L}d_{\max}\alpha_n\sqrt{n} + \\
&\delta\sqrt{2n} + 12K\frac{u}{\sqrt{n}}\Big] + 3\mathbb{P}(\mathcal{B}^c) + 2\mathbb{P}(\mathcal{M}_n^c).
\end{aligned}
$$

Substituting $\hat{\delta} = C_{3,p}\alpha_n$, where $C_{3,p} := 6/(\underline{p}\tau_d)$, and $\delta\sqrt{n} = 2\sqrt{u/\omega_n}$, and the upper bound on $\mathbb{P}(\mathcal{B}^c)$ from Lemma 16, we obtain after some rearranging,

$$
\begin{aligned}
d_{\mathrm{K}}(\widehat{T}_n, Z) - d_K(T_n, Z) \leq \\
&\frac{\sqrt{L}}{\underline{p}} \cdot \Big[\sqrt{\frac{8u}{\omega_n}} + 12K\frac{u}{\sqrt{n}} + 2\alpha_n\zeta_n\Big] + 6KLe^{-u} + 2\mathbb{P}\big(\mathrm{Mis}(\hat{g}, g) \geq \alpha_n\big)
\end{aligned}
$$

where

$$
\zeta_n = \big(\sqrt{2}C_{3,p} + 2L^{-1}d_{\max} + \omega_{n,1}C_{3,p}^2\alpha_n\big)\sqrt{n} + 4C_{3,p}\sqrt{u\omega_{n,1/2}}. \tag{2.55}
$$

As $d_{\max} \geq C_{3,p}L/\sqrt{2}$, $\alpha_n \leq 2/(C_{3,p}^2 L)$ and $u \leq (\underline{p}/8)^2 n\omega_n$ by assumption, and since $\omega_{n,1/2} \leq K^2 d_{\max}$ and $\omega_n \leq d_{\max}$, we obtain

$$
\begin{aligned}
\zeta_n &\leq 6d_{\max}\sqrt{n}/L + \underline{p}KC_{3,p}d_{\max}\sqrt{n}/2 \\
&\leq (1+K/2)C_{3,p}L^{-1}d_{\max}\sqrt{n}
\end{aligned}
$$

where the last inequality is due to $C_{3,p} \geq 6$ and $\underline{p} \leq L^{-1}$. The result follows.

2.5.2 Proofs of Theorems 2 and 3

We start by setting up notation and deriving some preliminary results that are common to both proofs. Throughout, K_0 denotes the true number of communities. Recall that $S_1 \subset [n]$ is determined by including any element of $[n]$ with probability $1/2$, and $S_2 = [n] \setminus S_1$. Then define $d_i := \sum_{j \in S_1} A_{ij} = \sum_{j=1}^n A_{ij} U_j$ where $U_j = 1\{j \in S_1\}, j \in [n]$ is an independent Ber$(1/2)$ sequence. We often work conditioned on S_1, $A_{S_1 S_1}$ and $(d_i, i \in S_2)$. Let $\mathcal{F}$ be the σ-field generated by these variables:

$$\mathcal{F} = \sigma\big(S_1, A_{S_1,S_1}, (d_i, i \in S_2)\big), \tag{2.56}$$

and let $\mathbb{E}^{\mathcal{F}}$ and $\mathbb{P}^{\mathcal{F}}$ denote the expectation and probability operators, conditioned on $\mathcal{F}$. We assume without loss of generality that the community detection algorithm is nonrandomized, so that conditioned on $\mathcal{F}$, $\widehat{y}$ is fixed. (Otherwise, we add the independent source of randomness used by the algorithm to $\mathcal{F}$.) The idea in both proofs is to first condition on $\mathcal{F}$ and derive bounds assuming parameters are all fixed. Then we can leave out $\mathcal{F}$ thanks to the fact that parameters are all bounded with high probability as we show below.

Controlling $\rho_{k\ell}$, d_i and $|\mathcal{G}_k|$ Recall the definition of $X_{i\ell}(\widehat{y})$ in (2.17), then conditioned on $\mathcal{F}$, for $i \in S_2$, $X_{i*}(\widehat{y}) \sim \text{Mult}(d_i, \rho_{z_i *})$ independently and thus $\mathbb{E}^{\mathcal{F}}[X_{i\ell}(\widehat{y})] = d_i \rho_{z_i \ell}$, $z_i \in [K_0]$ where

$$\rho_{k\ell} = \frac{\sum_{h=1}^{K_0} B^0_{kh} R_{h\ell}}{\sum_{\ell'=1}^{L} \sum_{h=1}^{K_0} B^0_{kh} R_{h\ell'}} \geq \tau_B \tau_\theta \frac{\sum_{j \in S_1} 1\{\widehat{y}_j = \ell\}}{|S_1|} \geq \tau_B \tau_\theta \tau_0 := \tau_\rho, \tag{2.57}$$

where the last inequality is due to the stability assumption 1. It follows that $\underline{\rho} := \min_{k,\ell} \rho_{k\ell} \geq \tau_\rho$.

Letting $d_i^* = \mathbb{E}[d_i]$, we have

$$d_i^* = \frac{1}{2}\sum_{j=1}^{n}\mathbb{E}[A_{ij}] = \frac{1}{2}\theta_i a_{z_i}, \quad \text{where} \quad a_h := \sum_{k=1}^{K_0} B_{hk}\sum_{j\in\mathcal{C}_k}\theta_j \tag{2.58}$$

for all $h \in [K_0]$. Let us derive some bounds on d_i^*. Recalling that $\theta_{\max} = 1$,

$$a_h \geq \tau_\theta \theta_{\max} n_k \sum_k B_{hk} = \tau_\theta n_k \frac{\nu_n}{n}\|B^0_{h*}\|_1 \geq \tau_\theta \nu_n \tau_{\mathcal{C}} \min_{h'}\|B^0_{h'*}\|_1. \tag{2.59}$$

Using the definition of C_1 in (2.25), we obtain

$$d_i^* \geq \frac{1}{2}C_1\nu_n, \quad \forall i. \tag{2.60}$$

Let $a_{\max} = \max_h a_h$. Since $|\mathcal{C}_k| \leq n$, we have $a_{\max} \leq \nu_n \cdot \max_h \|B^0_{h*}\|_1$. Combining with assumption (2.59), we obtain

$$a_h \geq \tau_a a_{\max}, \quad \tau_a := \tau_\theta \tau_B \tau_{\mathcal{C}}. \tag{2.61}$$

Since $\|B^0\|_\infty = 1$ by assumption, we have

$$d_i^* \leq \frac{1}{2}K_0\,\nu_n, \quad \forall i. \tag{2.62}$$

Recall that $\mathcal{C}_k = \{i \in [n] :\ z_i = k\}$ is the true community k, and we let $n_k = |\mathcal{C}_k|$. We also let $\mathcal{G}_k := \{i \in S_2 :\ z_i = k\} = \mathcal{C}_k \cap S_2$. We often work on the following event:

$$\mathcal{A} = \Big\{|\mathcal{G}_k| \in [0.4n_k, 0.6n_k],\ \forall k \in [K_0]\Big\} \cap \Big\{d_i \in \Big[\frac{d_i^*}{2}, \frac{3d_i^*}{2}\Big],\ \forall i \in [n]\Big\}. \tag{2.63}$$

Note that $\mathcal{A}$ is deterministic conditioned on $\mathcal{F}$. The next lemma guarantees that this event holds with high probability:

Lemma 8. *Under the scaling assumption* (2.26), $\mathbb{P}(\mathcal{A}^c) \leq 7n^{-1}$.

From S_2 to S_2' Recall that in the subsampled NAC, we first use random sampling to get S_2 and then use quantile filtering with threshold σ-th quantile in each estimated cluster to get S_2'. We spend most part of proofs below to work with S_2 for simplicity and their close connections. Recall that in each $\widehat{\mathcal{G}}_k = \{i \in S_2 : \widehat{z}_i = k\}$ we keep nodes with degrees at least its σ-th quantile to get $\widehat{\mathcal{G}}_k' = \{i \in \widehat{\mathcal{G}}_k : d_i \geq d^k_{(\lfloor \sigma |\mathcal{G}_k| \rfloor)}\}$. It follows that $|\widehat{\mathcal{G}}_k'| \geq (1-\sigma)|\widehat{\mathcal{G}}_k|$ and $|S_2'| \geq (1-\sigma)|S_2|$, where $S_2' = \cup_{k=1}^K \widehat{\mathcal{G}}_k'$.

Now let us get a lower bound for the size of $\mathcal{G}_k' := \mathcal{G}_k \cap S_2'$, which is used in the proof of Theorem 2 and 3. Given an estimated label vector $\widehat{z}$ and a true label vector z, recall the definition of event

$$\mathcal{M}_n := \{\text{Mis}(\widehat{z}, z) \leq \alpha_n\}.$$

Then on $\mathcal{M}_n$, $|\mathcal{G}_k| - \alpha_n n \leq |\widehat{\mathcal{G}}_k| \leq |\mathcal{G}_k| - \alpha_n n$. Recall that $|\mathcal{G}_k| \geq 0.4\tau_{\mathcal{C}} n$ on $\mathcal{A}$. Therefore, on $\mathcal{M}_n \cap \mathcal{A}$,

$$\begin{aligned}
|\mathcal{G}_k'| &\geq |\mathcal{G}_k \cap \widehat{\mathcal{G}}_k'| \\
&\geq |\mathcal{G}_k \cap \widehat{\mathcal{G}}_k| - |\widehat{\mathcal{G}}_k \setminus \widehat{\mathcal{G}}_k'| \\
&\geq |\mathcal{G}_k| - \alpha_n n - \sigma|\widehat{\mathcal{G}}_k| \\
&\geq |\mathcal{G}_k| - \alpha_n n - \sigma(|\mathcal{G}_k| + \alpha_n n) \\
&\geq (1-\sigma)0.4\tau_{\mathcal{C}} n - (1+\sigma)\alpha_n n \\
&\geq 0.2(1-\sigma)\tau_{\mathcal{C}} n := c_1 K_0 n
\end{aligned}$$

where the last inequality is by assumption that $\alpha_n \leq \frac{\tau_{\mathcal{C}}(1-\sigma)}{5(1+\sigma)}$.

Proof of Theorem 2

The proof has two parts. The first part is to get an upper bound of the Kolmogorov distance conditioned on subsampling random field, resulting from combining Lemma 9 and Theorem 1. The second part is to show that on event $\mathcal{A} \cap \mathcal{M}_n$, the random quantities are bounded by constants.

For a random variable Y, let $\mathcal{L}(Y \mid \mathcal{F})$ be the law of Y conditioned on $\mathcal{F}$ and let

$$d_{\mathrm{K}}\big(\mathcal{L}(Y \mid \mathcal{F}), Z\big) = \sup_{t\in\mathbb{R}} |\mathbb{P}(Y \le t \mid \mathcal{F}) - \mathbb{P}(Z \le t)|. \tag{2.64}$$

Lemma 9. *For any random variables, Y and Z, any event $\mathcal{B}$ and any σ-field $\mathcal{F}$, we have*

$$\begin{aligned} d_K(Y,Z) &\le \mathbb{E}\big[d_K\big(\mathcal{L}(Y \mid \mathcal{F}), Z\big)\big], \\ |d_K(Y,Z) - d_K(Y1_{\mathcal{B}}, Z)| &\le \mathbb{P}(\mathcal{B}^c). \end{aligned}$$

Applying Lemma 9 and since the Kolmogorov distance is bounded above by 1, we obtain

$$\begin{aligned} d_{\mathrm{K}}(\widehat{T}_n, Z) &\le \mathbb{E}^{\mathcal{F}}\big[d_{\mathrm{K}}\big(\mathcal{L}(\widehat{T}_n \mid \mathcal{F}), Z\big)\big] \\ &\le \mathbb{E}^{\mathcal{F}}[d_{\mathrm{K}}\big(\mathcal{L}(\widehat{T}_n \mid \mathcal{F}), Z\big) \cdot 1_{\mathcal{A}\cap\mathcal{M}_n}] + \mathbb{P}(\mathcal{A}^c) + \mathbb{P}(\mathcal{M}_n^c). \end{aligned} \tag{2.65}$$

Recall that with subsampling, we first get subset S_2. For each estimated cluster $\hat{\mathcal{C}}_k$, we have $\widehat{\mathcal{G}}_k = \hat{\mathcal{C}}_k \cap S_2$. With quantile filtering in each $\widehat{\mathcal{G}}_k$, we are left with $\widehat{\mathcal{G}}'_k$. And the test is performed on $A_{S'_2 S_1}$, where $S'_2 = \cup_{k=1}^K \widehat{\mathcal{G}}_k$. Conditioned on $\mathcal{F}$, $X_{i\ell}(\widehat{y})$ for $i \in S_2$ is distributed as $\mathrm{Mult}(d_i, \rho_{z_i *})$ as discussed in (2.18). Let $\mathcal{G}'_k = \mathcal{G}_k \cap S'_2$. Then we can apply Theorem 1 (assuming its conditions hold) with $\hat{g} = \widehat{z} \cap S'_2$, $g = z \cap S'_2$, $\mathrm{Mis}(\hat{g}, g) \ge \frac{\alpha_n}{0.4(1-\sigma)}$, $h(d)$ as the harmonic mean of $(d_i, i \in S'_2)$, $d_{\mathrm{av}}^{(k)}$ as the arithmetic average of $(d_i, i \in \mathcal{G}'_k)$, $\pi_k = |\mathcal{G}'_k|/|S'_2|$, $\omega_n = \min_k \pi_k d_{\mathrm{av}}^{(k)}$, $d_{\max} = \max_{i\in S'_2} d_i$, and $\tau_d = \omega_n / d_{\max}$ to the conditional law of $\widehat{T}_n$ given $\mathcal{F}$ to obtain

$$\begin{aligned} d_{\mathrm{K}}\big(\mathcal{L}(\widehat{T}_n \mid \mathcal{F}), Z\big) \le{}& \frac{C_{1,\rho}}{\sqrt{L|S'_2|}} + \frac{C_{2,\rho}}{h(d)} \\ &+ \frac{\sqrt{L}}{\underline{\rho}} \left(\sqrt{\frac{72\log(K_0\omega_n)}{\omega_n}} + \frac{12K_0\log(K_0\omega_n)}{\sqrt{|S'_2|}} \right) \\ &+ \frac{(2+K_0)C_{3,\rho} d_{\max}\sqrt{n}\alpha_n}{0.4(1-\sigma)\underline{\rho}\sqrt{L}} + 2\mathbb{P}\Big(\mathrm{Mis}(\hat{g}, g) \ge \frac{\alpha_n}{0.4(1-\sigma)} \,|\, \mathcal{F} \Big) \end{aligned} \tag{2.66}$$

The constants $C_{1,\rho}$, $C_{2,\rho}$ and $C_{3,\rho}$ depend on the ρ matrix defined in (2.19). Note that

$h(d)$ and ω_n, although in general random, are deterministic conditioned on $\mathcal{F}$.

Now we bound (2.66) on $\mathcal{A}$, and without further specification the following results are all based on $\mathcal{A}$. Recall that $|\mathcal{G}_k'| \geq c_1 K_0 n$ on event $\mathcal{M}_n \cap \mathcal{A}$. Then, we obtain

$$1 \geq \pi_k = \frac{|\mathcal{G}_k'|}{|S_2'|} \geq \frac{c_1 K_0 n}{0.6n} = \frac{5c_1 K_0}{3}.$$

Since $\frac{1}{4}C_1\nu_n \leq d_{\text{av}}^{(k)} \leq \frac{3}{4}K_0\,\nu_n$, for all $k \in [K_0]$ then

$$\frac{5}{12}c_1 C_1 K_0 \nu_n \leq \omega_n \leq \frac{3}{4}K_0\,\nu_n, \quad \tau_d \geq \frac{5}{9}c_1 C_1$$

Recall that $\underline{\rho} \geq \tau_\rho$ from (2.57), then

$$C_{3,\rho} = \frac{6}{\underline{\rho}\,\tau_d} \leq \frac{54}{5c_1 C_1 \tau_\rho} := C_2.$$

Furthermore,

$$\frac{1}{h(d)} = \frac{1}{|S_2'|}\sum_{i\in|S_2'|} d_i^{-1} \leq \frac{2}{|S_2'|}\sum_{i\in|S_2'|}(d_i^*)^{-1} \leq \frac{4}{C_1\nu_n}.$$

To bound the probability of the missclassification rate, we first note that

$$\text{Mis}(\hat{g}, g) \leq \frac{\sum_{i\in S_2'} 1\{g_i \neq \hat{g}_i\}}{|S_2'|} \leq \frac{\sum_{i=1}^n 1\{z_i \neq \widehat{z}_i\}}{|S_2'|} = (n/|S_2'|)\,\text{Mis}(\widehat{z}, z).$$

Furthermore, $|S_2'|/n \geq 0.4(1-\sigma)$ on the event $\mathcal{A}$. Then

$$\begin{aligned}
\mathbb{P}\Big(\,\text{Mis}(\hat{g}, g) \geq \frac{\alpha_n}{0.4(1-\sigma)}\,|\,\mathcal{F}\Big)1_{\mathcal{A}\cap\mathcal{M}_n} &\leq \mathbb{P}\Big(\frac{n}{|S_2'|}\,\text{Mis}(\widehat{z}, z) \geq \frac{\alpha_n}{0.4(1-\sigma)}\,|\,\mathcal{F}\Big)1_{\mathcal{A}} \\
&= \mathbb{P}\Big(\Big\{\frac{n}{|S_2'|}\,\text{Mis}(\widehat{z}, z) \geq \frac{\alpha_n}{0.4(1-\sigma)}\Big\} \cap \mathcal{A}\,|\,\mathcal{F}\Big) \\
&\leq \mathbb{P}(\text{Mis}(\widehat{z}, z) \geq \alpha_n\,|\,\mathcal{F}).
\end{aligned}$$

Note that $\log(K_0\omega_n) \le \log\big((3/4)K_0^2\nu_n\big) := \beta_n$. Therefore,

$$\mathbb{E}^{\mathcal{F}}\Big[d_{\mathrm{K}}\big(\mathcal{L}(\widehat{T}_n \mid \mathcal{F}), Z\big)\cdot 1_{\mathcal{A}\cap\mathcal{M}_n}\Big] \le \frac{C_3-21}{\sqrt{Ln}} + \frac{C_4}{C_1\nu_n} + \frac{19\sqrt{L}}{\tau_\rho}\left(\sqrt{\frac{\beta_n}{c_1C_1K_0\nu_n}} + \frac{K_0\beta_n}{\sqrt{(1-\sigma)n}}\right) + C_5\nu_n\sqrt{n}\alpha_n + 2\mathbb{P}(\mathrm{Mis}(\widehat{z},z)\ge\alpha_n)$$

where $C_3 = 87/[(1-\sigma)^{1/2}\tau_\rho^4] + 7$, $C_4 = 4(\pi e)^{-1/2}\max\{1, \tau_\rho^{-1} - L - 1\}$ and $C_5 = 15K_0(1 + K_0/2)C_2/[4(1-\sigma)\tau_\rho\sqrt{L}]$ are as defined in the statement of the theorem. The conditions of Theorem 1 hold on $\mathcal{A}$, if $L \ge 2$, $C_1\nu_n/4 \ge 2$, $\frac{5}{12}c_1C_1K_0\nu_n \ge L$, $C_1\nu_n/4 \ge C_2L/\sqrt{2}$, $\beta_n \le (\frac{\tau_\rho}{8})^2\frac{5}{12}c_1C_1K_0\nu_n n := C_6\nu_n n$, and $\alpha_n \le \frac{\tau_\rho}{8C_2} \wedge \frac{2}{LC_2^2}$. Then the result follows by using (2.65) and $(C_3-7)/\sqrt{nL} + 7n^{-1} \le C_3/\sqrt{nL}$ as $\sqrt{nL} \le n$.

Proof of Theorem 3

The proof has two main components. The first part is to show that when $K < K_0$, there exists a mixed cluster that contains two large pieces of true communities. Note that this holds for any underfitted labels and is not a assumption bounded by a specific community detection algorithm. The second part is to show the chi-square statistic is large in that mixed cluster. This is done by first showing statistics are close to its parameters with high probability as in Lemma 10. Then we further show that chi-square statistic is close to its counterpart when substitute the observed statics with their corresponding parameters in Lemma 11. Finally, by calculation, the chi-square statistic over two chunks with different probabilities is large. For simplicity, the following proof is based on S_2 (only sampling is performed). At the end we will show the case for S_2' (both sampling and quantile filtering are performed).

Recall that $\hat{\mathcal{C}}_k = \{i : \widehat{z}_i = k\}$ and $\hat{n}_k = |\hat{\mathcal{C}}_k|$. For $r \in [K_0]$, a true community $\mathcal{C}_r$ is partitioned into $\hat{\mathcal{C}}_{k,r} = \{i : \widehat{z}_i = k, z_i = r\}$, $k \in [K]$. With such partition, we have for each $r \in [K_0]$, there exists $k_r \in [K]$ such that $|\hat{\mathcal{C}}_{k_r,r} \cap S_2| \ge |\mathcal{C}_r \cap S_2|/K$. Since $K < K_0$, there are $r_1, r_2 \in [K_0]$ such that $r_1 \ne r_2$ and $k_{r_1} = k_{r_2} =: \hat{k}$. Note that $\hat{k}$ is random and potentially dependent on A. Without loss of generosity, let $r_1 = 1$ and $r_2 = 2$. Therefore,

$\hat{\mathcal{C}}_{\hat{k}}$ contains "large" pieces $\hat{\mathcal{C}}_{\hat{k},1}$ and $\hat{\mathcal{C}}_{\hat{k},2}$ of two different true communities 1 and 2, and we will show below that this furthermore makes $\widehat{\mathcal{G}}_{\hat{k}} = \hat{\mathcal{C}}_{\hat{k}} \cap S_2$ having substantial size. First recalling that $\mathcal{G}_r = \mathcal{C}_r \cap S_2$, on event $\mathcal{A}$, we have

$$|\hat{\mathcal{C}}_{\hat{k},1} \cap S_2| \geq |\mathcal{C}_1 \cap S_2|/K \geq |\mathcal{G}_1|/K_0 \geq (0.4\tau_{\mathcal{C}}/K_0)n \geq c_1 n, \tag{2.67}$$

where $c_1 = (1-\sigma)\frac{\tau_{\mathcal{C}}}{5K_0}$ as in (2.24) and the same goes for $|\hat{\mathcal{C}}_{\hat{k},2} \cap S_2|$. Therefore, we have $|\widehat{\mathcal{G}}_{\hat{k}}| \geq |\hat{\mathcal{C}}_{\hat{k},1} \cap S_2| + |\hat{\mathcal{C}}_{\hat{k},2} \cap S_2| \geq 2c_1 n$. Then we will focus on $\widehat{\mathcal{G}}_{\hat{k}}$ in the following argument. Let $\bigcup_{r=1}^{K_0} \hat{\mathcal{T}}_r$ be a disjoint partition of $\widehat{\mathcal{G}}_{\hat{k}}$ into the true communities, where $\hat{\mathcal{T}}_r = \{i \in S_2 : \widehat{z}_i = k, z_i = r\} = \hat{\mathcal{C}}_{k,r} \cap S_2$. Some $\hat{\mathcal{T}}_r$ might be empty, but we can safely ignore those empty sets and focus on the two big chunk $\hat{\mathcal{T}}_1$ and $\hat{\mathcal{T}}_2$ as we discussed above. Since $\hat{\mathcal{T}}_r \in \mathcal{C}_r \cap S_2$, then for any $i \in \hat{\mathcal{T}}_r$, we have $\mathbb{E}^{\mathcal{F}}[\xi_{i\ell}] = \rho_{r\ell}$, where $\rho_{r\ell}$ is defined based on (2.18) and (2.19). Let us define

$$\hat{\alpha}_r = \sum_{i \in \hat{\mathcal{T}}_r} d_i, \quad \hat{\beta}_r := \frac{\hat{\alpha}_r}{\hat{\alpha}_+}, \quad \bar{\rho}_\ell := \sum_{r=1}^{K_0} \hat{\beta}_r \rho_{r\ell},$$

where $\hat{\alpha}_+ = \sum_r \hat{\alpha}_r = \sum_{i \in \widehat{\mathcal{G}}_{\hat{k}}} d_i$. Consider the event

$$\mathcal{E} := \Big\{ \max_{r,\ell} \max_{i \in \hat{\mathcal{T}}_r} |\xi_{i\ell} - \rho_{r\ell}| \leq \varepsilon_n := 4\sqrt{\frac{\log n}{C_1 \nu_n}}. \Big\} \tag{2.68}$$

The following lemma shows that $\mathcal{E}$ holds with high probability and we work on $\mathcal{E}$ for the rest of the proof. Note that its assumption is satisfied since we have a stronger assumption $\frac{\log n}{\nu_n} \leq \frac{C_1 \tau_\rho^2}{64}$.

Lemma 10. *Assume* $\frac{\log n}{\nu_n} \leq \frac{C_1}{4}$, *we have* $\mathbb{P}(\mathcal{E}^c \cap \mathcal{A}) \leq 2Ln^{-1}$.

We next show that $\widehat{\rho}_{\hat{k}\ell}$ is close to $\bar{\rho}_\ell$ for all $\ell \in [L]$. We have

$$\begin{aligned}\Big| \sum_{i \in \widehat{\mathcal{G}}_{\hat{k}}} X_{i\ell}(\widehat{y}) - \sum_r \hat{\alpha}_r \rho_{r\ell} \Big| &= \Big| \sum_r \sum_{i \in \hat{\mathcal{T}}_r} d_i \xi_{i\ell} - \sum_r \sum_{i \in \hat{\mathcal{T}}_r} d_i \rho_{r\ell} \Big| \\ &\leq \sum_r \sum_{i \in \hat{\mathcal{T}}_r} d_i \big| \xi_{i\ell} - \rho_{r\ell} \big| \leq \varepsilon_n \sum_r \hat{\alpha}_r = \varepsilon_n \hat{\alpha}_+.\end{aligned}$$

Dividing by $\hat{\alpha}_+$ and recalling the definition of $\widehat{\rho}_{k\ell}$ in (2.21), we get

$$\big|\widehat{\rho}_{\hat{k}\ell} - \bar{\rho}_\ell\big| = \Big|\frac{\sum_{i\in\widehat{\mathcal{G}}_{\hat{k}}} X_{i\ell}(\widehat{y})}{\sum_{i\in\widehat{\mathcal{G}}_{\hat{k}}} d_i} - \frac{\sum_r \hat{\alpha}_r\rho_{r\ell}}{\hat{\alpha}_+}\Big| \le \varepsilon_n, \quad \forall \ell \in [L]. \tag{2.69}$$

We now apply the following lemma:

Lemma 11. *Let $\psi(x,y) = (x-y)^2/y$. Consider (x,y) and (x',y') in $[0,1]\times[1/c_1,1]$, where $c_1 > 1$, such that $\max\{|x-x'|,|y-y'|\} \le \varepsilon \le 1$. Then,*

$$\big|\psi(x',y') - \psi(x,y)\big| \le 12c_1^3\,\varepsilon. \tag{2.70}$$

Note that $\bar{\rho}_\ell$ is a weighted sum of $(\rho_{r\ell})$, hence $\bar{\rho}_\ell \ge \underline{\rho} \ge \tau_\rho$ using (2.57). Furthermore, by assumption $\varepsilon_n \le \tau_\rho/2$, then combined with (2.69), we have $\min\{\widehat{\rho}_{\hat{k}\ell}, \bar{\rho}_\ell\} \ge \tau_\rho/2$ for all $\ell \in [L]$. Therefore, we can apply Lemma 11 with $c_1 = 2/\tau_\rho$ to obtain

$$\big|\psi(\xi_{i\ell}, \widehat{\rho}_{\hat{k}\ell}) - \psi(q_{r\ell}, \bar{q}_\ell)\big| \le 96\tau_\rho^{-3}\varepsilon_n, \quad \forall i \in \hat{\mathcal{T}}_r,\ \forall \ell \in [L].$$

For two vectors $x, y \in \mathbb{R}^L$, let us write $\Psi(x,y) = \sum_\ell \psi(x_\ell, y_\ell)$. Let $\xi_i = (\xi_{i\ell})$, and $\widehat{\rho}_{u*} = (\widehat{\rho}_{u\ell})$, and set $\widehat{Y}_+^{(u)} = \sum_{i\in\widehat{\mathcal{G}}_u} d_i\Psi(\xi_i, \widehat{\rho}_{u*})$ for any $u \in [K]$. By the triangle inequality,

$$\widehat{Y}_+^{(\hat{k})} \ge \sum_r \sum_{i\in\hat{\mathcal{T}}_r} d_i\big(\Psi(\rho_{r*}, \bar{\rho}_*) - 96\tau_\rho^{-3}\varepsilon_n L\big) = \sum_r \hat{\alpha}_r\big(\Psi(\rho_{r*}, \bar{\rho}_*) - 96\tau_\rho^{-3}\varepsilon_n L\big)$$

where $\rho_{r*} = (\bar{\rho}_{r\ell})$ and $\bar{\rho}_* = (\bar{\rho}_\ell)$. Dividing by $\hat{\alpha}_+$, we have

$$\frac{1}{\hat{\alpha}_+}\widehat{Y}_+^{(\hat{k})} \ge \sum_r \hat{\beta}_r\Psi(\rho_{r*}, \bar{\rho}_*) - 96\tau_\rho^{-3}\varepsilon_n L. \tag{2.71}$$

Let us define $\omega_1 := \sum_r \hat{\beta}_r(\rho_{r*}, \bar{\rho}_*)$.

Controlling ω_1 Recall $|\hat{\mathcal{T}}_1|, |\hat{\mathcal{T}}_2| \geq c_1 n$ and the definition of a_h in (2.58). On the event $\mathcal{A}$, for $u = 1, 2$, we have

$$\hat{\beta}_u := \frac{\sum_{i\in\hat{\mathcal{T}}_u} d_i}{\sum_r \sum_{i\in\hat{\mathcal{T}}_r} d_i} \geq \frac{1}{3}\frac{\sum_{i\in\hat{\mathcal{T}}_u} d_i^*}{\sum_r \sum_{i\in\hat{\mathcal{T}}_r} d_i^*} = \frac{1}{3}\frac{\sum_{i\in\hat{\mathcal{T}}_u} \theta_i a_u}{\sum_r \sum_{i\in\hat{\mathcal{T}}_r} \theta_i a_r}$$
$$\geq \frac{1}{3}\frac{\tau_\theta \theta_{\max} a_u |\hat{\mathcal{T}}_u|}{\theta_{\max} a_{\max} |\widehat{\mathcal{G}}_{\hat{k}}|} \geq \frac{1}{3}\tau_\theta \tau_a c_1$$

using $a_h \geq \tau_a a_{\max}$ from (2.61), $\theta_i \geq \tau_\theta \theta_{\max}$ and $|\widehat{\mathcal{G}}_{\hat{k}}| \leq n$. We have

$$\omega_1 = \sum_r \hat{\beta}_r \sum_\ell \frac{(\rho_{r\ell} - \bar{\rho}_\ell)^2}{\bar{\rho}_\ell} = \sum_\ell \frac{1}{\bar{\rho}_\ell} \sum_r \hat{\beta}_r (\rho_{r\ell} - \bar{\rho}_\ell)^2.$$

The inner summation is the variance of a random variable taking values $\{\rho_{r\ell}\}$ with probabilities $\{\hat{\beta}_r\}$. Applying Lemma 29 in the Supplement (Section B.4) and recalling the definition of ω_2 from (2.29), we have

$$\omega_1 \geq \frac{1}{\max_\ell \bar{\rho}_\ell} \frac{1}{2} \hat{\beta}_1 \hat{\beta}_2 \sum_\ell (\rho_{1\ell} - \rho_{2\ell})^2 \geq \frac{1}{18} \tau_\theta^2 \tau_a^2 c_1^2 \|\rho_{1*} - \rho_{2*}\|^2 \geq L\omega_2 \tag{2.72}$$

since $\max_\ell \bar{\rho}_\ell \leq \max_{k,\ell} \rho_{k\ell} \leq 1$.

Putting the pieces together On Ω_n, we have $96\tau_\rho^{-3}\varepsilon_n \leq \frac{1}{2}\omega_2$, which combined with (2.72) and (2.71), gives $\frac{1}{\hat{\alpha}_+}\widehat{Y}_+^{(\hat{k})} \geq \frac{1}{2}L\omega_2$. Recall that, on $\mathcal{A}$, $|\widehat{\mathcal{G}}_{\hat{k}}| \geq 2c_1 n$ and $d_i \geq C_1\nu_n/4$. Then, on $\Omega_n \cap \mathcal{E} \cap \mathcal{A}$,

$$\widehat{Y}_+^{(\hat{k})} \geq \frac{1}{4} c_1 C_1 L \omega_2 n \nu_n.$$

Furthermore, $\tilde{n} = |S_2| \leq 0.6n$ on $\mathcal{A}$, hence $\gamma_{\tilde{n}} = \sqrt{\tilde{n}(L-1)} \leq \sqrt{0.6nL}$ and we have

$$\widehat{T}_n = \frac{1}{\sqrt{2}}\Big(\frac{1}{\gamma_{\tilde{n}}}\sum_{u=1}^K \widehat{Y}_+^{(u)} - \gamma_{\tilde{n}}\Big) \geq \frac{1}{\sqrt{2}}\Big(\frac{1}{\gamma_{\tilde{n}}}\widehat{Y}_+^{(\hat{k})} - \gamma_{\tilde{n}}\Big)$$
$$\geq \sqrt{\frac{n}{2}}\left(\frac{c_1 C_1 L \omega_2 \nu_n/4}{\sqrt{0.6L}} - \sqrt{0.6L}\right).$$

On Ω_n, we have $\frac{1}{2}(c_1 C_1 L\omega_2\nu_n/4) \geq 0.6L$, hence on $\Omega_n \cap \mathcal{E} \cap \mathcal{A}$, we obtain

$$\widehat{T}_n \geq \sqrt{\frac{n}{2}} \left(\frac{c_1 C_1 L\omega_2\nu_n/8}{\sqrt{0.6L}} \right) \geq \frac{c_1 C_1}{10} \omega_2 \nu_n \sqrt{Ln}. \tag{2.73}$$

Furthermore, we note that

$$\mathbb{P}((\Omega_n \cap \mathcal{E} \cap \mathcal{A})^c) \leq \mathbb{P}(\Omega_n^c) + 2Ln^{-1} + 2(7n^{-1}) \leq \mathbb{P}(\Omega_n^c) + 9Ln^{-1}$$

using Lemmas 8 and 10 and $L \geq 2$.

Lastly, we extend the proof to subsampling with quantile filtering when the test is performed on $A_{S_2' S_1}$. The inequality (2.67) needs to be updated with respect to S_2':

$$\begin{aligned} |\hat{\mathcal{C}}_{\hat{k},1} \cap S_2'| \geq |\mathcal{C}_1 \cap S_2'|/K &\geq |\mathcal{G}_1 \cap S_2'|/K_0 \\ &\geq c_1 n \end{aligned}$$

on event $\mathcal{A} \cap \mathcal{M}_n$. Then (2.73) is true under event $\Omega_n \cap \mathcal{E} \cap \mathcal{A} \cap \mathcal{M}_n$ with probability at least $1 - \mathbb{P}(\Omega_n^c) - \mathbb{P}(\mathcal{M}_n^c) - 9Ln^{-1}$. The proof of Theorem 3 is complete.

2.5.3 Proof of Theorem 4

Let $\mathcal{C}_k = \{i \in [n] : \ z_i = k\}$ be the community k defined by label vector z, and $n_k = |\mathcal{C}_k|$. We also let $\mathcal{G}_k := \{i \in S_2 : \ z_i = k\} = \mathcal{C}_k \cap S_2$. Consider event

$$\mathcal{A}_1 = \big\{ |\mathcal{G}_k| \in [0.4n_k, 0.6n_k], \ \forall k \in [K] \big\}. \tag{2.74}$$

Since $|\mathcal{C}_k \cap S_1| = n_k - |\mathcal{G}_k|$, on $\mathcal{A}_1$, we also have

$$|\mathcal{C}_k \cap S_1| \in [0.4n_k, 0.6n_k], \quad \forall k \in [K]. \tag{2.75}$$

Under the assumption $0.4\tau_{\mathcal{C}} n \geq 2$, we have,

$$|\mathcal{G}_k| \geq 2, \ \forall k \in [K], \quad \text{on } \mathcal{A}_1. \tag{2.76}$$

From the proof of Lemma 8, we obtain:

Lemma 12. $\mathbb{P}(\mathcal{A}_1^c) \le n^{-1}$ *if* $\frac{\log n}{n} \le \tau_c/300$.

For two σ-fields $\mathcal{F}$ and $\mathcal{H}$, we write $\mathcal{F} \vee \mathcal{H} = \sigma(\mathcal{F} \cup \mathcal{H})$ for the σ-field generated by their union. Recall that with subsampling, the set $S_1 \subset [n]$ is determined by including any element of $[n]$, indepenently with probability $1/2$, and $S_2 = [n] \setminus S_1$. Let $d_i = \sum_{j \in S_1} A_{ij}$ and consider the following σ-fields

$$\begin{aligned} \mathcal{F}_0 &= \sigma\big(S_1\big), \\ \mathcal{F}_1 &= \mathcal{F}_0 \vee \sigma(x_{S_2}), \\ \mathcal{F}_2 &= \mathcal{F}_1 \vee \sigma(x_{S_1}) = \mathcal{F}_0 \vee \sigma(x_{[n]}), \\ \mathcal{F} &= \mathcal{F}_2 \vee \sigma\big((d_i, i \in S_2)\big), \end{aligned} \tag{2.77}$$

where $x_{S_2} = (x_i, i \in S_2)$, and similarly for x_{S_1}, and $x_{[n]} = (x_1, \ldots, x_n)$. Note that conditioned on $\mathcal{F}_0$, $\widehat{y}$ is fixed, and conditioned on $\mathcal{F}_2$, (p_{ij}) is fixed.

We first consider the case where $\widehat{y}_{S_1} = y_{S_1}$. In this case, we drop the dependence of $X_{i\ell}(\widehat{y})$ (defined in (2.17)) on $\widehat{y}$, and write

$$X_{i\ell} := \sum_{j \in S_1} A_{ij} 1\{y_j = \ell\}. \tag{2.78}$$

Since conditioned on $\mathcal{F}_2$, $x_{[n]}$ are fixed, it follows that

$$X_{i\ell} \mid \mathcal{F}_2 \sim \mathrm{Poi}(q_{i\ell}), \quad \text{where} \quad q_{i\ell} := \sum_{j \in S_1} p_{ij} 1\{y_j = \ell\}, \tag{2.79}$$

independently across ℓ. Since for $i \in S_2$, the sum of $X_{i\ell}$ over ℓ is d_i, and when we condition on $\mathcal{F}$, we are also conditioning on $d_i, i \in S_2$, we obtain

$$(X_{i\ell})_{\ell=1}^L \mid \mathcal{F} \sim \mathrm{Mult}\big(d_i, (\rho_{i\ell})_{\ell=1}^L\big) \quad \text{where} \quad \rho_{i\ell} := \frac{q_{i\ell}}{\sum_{\ell'} q_{i\ell'}}. \tag{2.80}$$

independently across $i \in S_2$.

Controlling conditional probabilities Let us set

$$\tau_\rho := \frac{C_8}{2L\tau_\theta} = \frac{\tau_{\mathcal{C}}\tau_h\tau_\theta}{2L}. \tag{2.81}$$

As the first step in the proof, we show that $\rho_{i\ell}$ is close to $H_\ell(x_i)$. More specifically, the following event

$$\mathcal{R} := \Big\{ |\rho_{i\ell} - H_\ell(x_i)| \le \frac{4K}{\tau_\rho}\sqrt{\frac{\log n}{n}},\ \forall i \in S_2,\ \forall \ell \in [L] \Big\} \tag{2.82}$$

holds with high probability:

Lemma 13. *There is an event $\mathcal{W}$ such that*

$$\mathcal{R} \supseteq \Gamma \cap \mathcal{W}, \quad \textit{and} \quad \mathbb{P}(\mathcal{W}^c) \le 4LKn^{-1}$$

whenever $\frac{\log n}{n} < (\frac{\tau_\rho}{4K})^2$.

Controlling the degrees Consider the event

$$\mathcal{A}_2 := \big\{ d_i \in [0.16C_8\nu_n, 0.96\nu_n],\ \forall i \in S_2 \big\} \tag{2.83}$$

The next lemma guarantees that $\mathcal{A}_2$ holds with high probability:

Lemma 14. *There is an event $\mathcal{D}$ such that*

$$\mathcal{A}_2 \supseteq \mathcal{A}_1 \cap \Gamma \cap \mathcal{D} \quad \textit{and} \quad \mathbb{P}(\mathcal{D}^c \cap \mathcal{A}_1) \le 2.2n^{-1}$$

wherever $\frac{\log n}{n} \le 0.04C_8^2$ *and* $\frac{\log n}{\nu_n} \le 0.001C_8$.

From now on, let $\mathcal{A} := \mathcal{A}_1 \cap \mathcal{A}_2$. Let $d_+^k = \sum_{i\in\mathcal{G}_k} d_i$ and $\omega_n = \min_k d_+^k/|S_2|$. On $\mathcal{A}$,

we have

$$\min_k d_+^k \geq (0.16C_8\nu_n)(0.4n_k) \geq 0.064\tau_{\mathcal{C}}C_8\, n\nu_n, \tag{2.84}$$

$$\omega_n \geq (8/75)\tau_{\mathcal{C}}C_8\nu_n \tag{2.85}$$

using $0.4n \leq |S_2| \leq 0.6n$.

Controlling probability estimates Recall $\widehat{\mathcal{G}}_k = \{i \in S_2 : \widehat{z}_i = k\}$, and let

$$\widehat{\rho}_{k\ell} = \frac{\sum_{i\in\widehat{\mathcal{G}}_k} X_{i\ell}}{\sum_{i\in\widehat{\mathcal{G}}_k} d_i}, \quad \widetilde{\rho}_{k\ell} = \frac{1}{d_+^k}\sum_{i\in\mathcal{G}_k} X_{i\ell}, \quad \bar{\rho}_{k\ell} = \frac{1}{d_+^k}\sum_{i\in\mathcal{G}_k} d_i\rho_{i\ell}, \tag{2.86}$$

$\widehat{\Delta}_{k\ell} = \widehat{\rho}_{k\ell} - \widetilde{\rho}_{k\ell}$ and $\widetilde{\Delta}_{k\ell} = \widetilde{\rho}_{k\ell} - \bar{\rho}_{k\ell}$. To control these deviations, we first show that $\bar{\rho}_{k\ell_k}$ is lower-bounded, where ℓ_k is as in (2.33):

Lemma 15. *Assume* $\frac{\log n}{n} \leq (\frac{\tau_\rho^2}{4K})^2$. *Then, with* $\{\ell_k\}_{k=1}^K$ *as defined in* (2.35), *on* $\Gamma \cap \mathcal{R}$, *we have*

$$\rho_{i\ell_{z_i}} \geq \tau_\rho \quad \forall i \in S_2.$$

Combined with definition of $\bar{\rho}_{k\ell}$ in (2.86), Lemma 15 immediately implies that under the same condition, on $\Gamma \cap \mathcal{R}$,

$$\bar{\rho}_{k\ell_k} \geq \tau_\rho, \quad \forall k \in [K]. \tag{2.87}$$

Note that $2\tau_\rho = \tau_{\mathcal{C}}\tau_h\tau_\theta/L$ hence $2\tau_\rho \leq 1$.

Next, we show that $\widetilde{\Delta}_{k\ell}$ is small by considering the following event

$$\mathcal{B} := \Big\{ \max_\ell |\widetilde{\Delta}_{k\ell}| \leq \frac{8}{\sqrt{\tau_{\mathcal{C}}C_8}}\sqrt{\frac{\log n}{n\nu_n}} =: \delta,\ \forall k \in [K] \Big\}. \tag{2.88}$$

Lemma 16. $\mathbb{P}(\mathcal{B}^c \cap \mathcal{A}) \leq 2Ln^{-1}$ *whenever* $\frac{\log n}{n\nu_n} \leq 0.064\tau_{\mathcal{C}}C_8$.

To simplify the notation, let us define

$$\mathcal{N}_n := \mathcal{A} \cap \mathcal{B} \cap \mathcal{M}_n \cap \mathcal{R} \cap \Gamma. \tag{2.89}$$

The next step is to control $\widehat{\Delta}_{k\ell_k}$:

Lemma 17. *Assume that $\alpha_n \le \tau_{\mathcal{C}}\tau_\rho C_8/18$. Then, on $\mathcal{N}_n$,*

$$|\widehat{\Delta}_{k\ell_k}| \le \frac{54}{\tau_\rho \tau_{\mathcal{C}} C_8}\, \alpha_n\, \widetilde{\rho}_{k\ell_k}, \quad \forall k \in [K]. \tag{2.90}$$

Combining (2.88) and (2.90), on $\mathcal{N}_n$, we have

$$\begin{aligned} |\widehat{\rho}_{k\ell_k} - \bar{\rho}_{k\ell_k}| &\le \frac{54}{\tau_\rho \tau_{\mathcal{C}} C_8}\alpha_n + \frac{8}{\sqrt{\tau_{\mathcal{C}} C_8}}\sqrt{\frac{\log n}{n\nu_n}} \\ &\le \frac{58}{\tau_\rho \tau_{\mathcal{C}} C_8}\sqrt{\frac{\log n}{\nu_n}}. \end{aligned} \tag{2.91}$$

The second inequality uses $\tau_{\mathcal{C}} C_8 \le 1$ and $2\tau_\rho \le 1$ to replace the prefactor of the second term with $4/(\tau_\rho \tau_{\mathcal{C}} C_8)$ and then uses the assumption $\alpha_n \le \sqrt{(\log n)/\nu_n}$ to combine the two terms. We note that the fast rate $\sqrt{(\log n)/(n\nu_n)}$ of the second term is not helpful since it will be dominated later in the argument by the slow rate $\sqrt{(\log n)/\nu_n}$ needed to control (2.92).

Let $\xi_{i\ell} = X_{i\ell}/d_i$ for $i \in [S_2]$. Then $\mathbb{E}^{\mathcal{F}}[\xi_{i\ell}] = \rho_{i\ell}$. Consider the event

$$\mathcal{E} := \Big\{ \max_{i \in S_2,\ \ell \in [L]} |\xi_{i\ell} - \rho_{i\ell}| \le 5\sqrt{\frac{\log n}{C_8 \nu_n}} \Big\}. \tag{2.92}$$

Then, we have:

Lemma 18. $\mathbb{P}(\mathcal{E}^c \cap \mathcal{A}) \le 2Ln^{-1}$ *whenever* $\frac{\log n}{\nu_n} \le 0.16 C_8$.

Controlling chi-square statistics Now, let us define

$$\varepsilon_n := \frac{58}{\tau_\rho \tau_{\mathcal{C}} C_8}\sqrt{\frac{\log n}{\nu_n}}. \tag{2.93}$$

Then, on $\mathcal{N}_n \cap \mathcal{E}$, we have

$$|\widehat{\rho}_{k\ell_k} - \bar{\rho}_{k\ell_k}| \le \varepsilon_n, \quad |\xi_{i\ell} - \rho_{i\ell}| \le \varepsilon_n \tag{2.94}$$

for all k, ℓ and $i \in S_2$. This follows by recalling that $\tau_{\mathcal{C}}, \tau_\rho, C_8 \le 1$. Combining (2.87), (2.94) and the assumption $\varepsilon_n \le \tau_\rho/2$, we obtain

$$\min\{\widehat{\rho}_{k\ell_k}, \bar{\rho}_{k\ell_k}\} \ge \tau_\rho/2, \quad \text{on } \mathcal{N}_n. \tag{2.95}$$

Hence, we can apply Lemma 11 with $c_1 = 2/\tau_\rho$, using (2.94) to obtain that, on $\mathcal{N}_n \cap \mathcal{E}$,

$$\left|\psi(\xi_{i\ell_k}, \widehat{\rho}_{k\ell_k}) - \psi(\rho_{i\ell_k}, \bar{\rho}_{k\ell_k})\right| \le 96\tau_\rho^{-3}\varepsilon_n, \quad \forall i \in \mathcal{G}_k,\ \forall k \in [K_0].$$

Define $\widetilde{Y} := \sum_{k=1}^K \sum_{i\in\mathcal{G}_k} d_i \psi(\xi_{i\ell_k}, \widehat{\rho}_{k\ell_k})$. Then, on $\mathcal{N}_n \cap \mathcal{E}$,

$$\begin{aligned}
\widetilde{Y} &\ge \sum_{k=1}^K \sum_{i\in\mathcal{G}_k} d_i \big[(\rho_{i\ell_k} - \bar{\rho}_{k\ell_k})^2 - 96\tau_\rho^{-3}\varepsilon_n\big] \\
&= \sum_{k=1}^K d_+^k \Big[\sum_{i\in\mathcal{G}_k} \frac{d_i}{d_+^k}(\rho_{i\ell_k} - \bar{\rho}_{k\ell_k})^2 - 96\tau_\rho^{-3}\varepsilon_n\Big]
\end{aligned}$$

where the first inequality also uses $\psi(x,y) \ge (x-y)^2$ for $x, y \in [0,1]$. Let

$$\varpi_k = \sum_{i\in\mathcal{G}_k} \frac{d_i}{d_+^k}(\rho_{i\ell_k} - \bar{\rho}_{k\ell_k})^2, \quad k \in [K]. \tag{2.96}$$

Note that ϖ_k is the variance of a random variable taking value $\rho_{i\ell_k}$ with probability d_i/d_+^k for $i \in \mathcal{G}_k$. Recalling that $\vartheta_{k\ell} := \mathrm{var}(H_\ell(x))$ when $x \sim \mathbb{Q}_k$, we have the following:

Lemma 19. *Assume* $\frac{\log n}{n} \le \min\{\frac{\tau_\rho^2}{4}, \tau_{\mathcal{C}}\}$. *Then, there is an event* $\mathcal{H}$ *on which*

$$\varpi_k \ge \frac{C_8^2}{144}\vartheta_{k\ell_k} - \frac{C_8}{8}\tau_\theta L\sqrt{\frac{\log n}{n}}, \quad k \in [K] \tag{2.97}$$

and we have $\mathbb{P}(\mathcal{H}^c \cap \mathcal{A} \cap \mathcal{R}) \le Kn^{-c}$.

Let $\mu_n := \max\{1, L\sqrt{\nu_n/n}\}$ and

$$\tilde{\varepsilon}_n = \frac{58}{\tau_\rho \tau_C C_8} \mu_n \sqrt{\frac{\log n}{\nu_n}}. \tag{2.98}$$

so that $\varepsilon_n \le \tilde{\varepsilon}_n$. We have

$$\frac{C_8}{8} \tau_\theta L \sqrt{\frac{\log n}{n}} \le \mu_n \sqrt{\frac{\log n}{\nu_n}} \le \tilde{\varepsilon}_n.$$

It follows that on $\mathcal{N}_n \cap \mathcal{E} \cap \mathcal{H}$, we have

$$\widetilde{Y} \ge \sum_{k=1}^{K} d_+^k (\varpi_k - 96\tau_\rho^{-3}\varepsilon_n) \ge \sum_{k=1}^{K} d_+^k \Big(\frac{C_8^2}{144} \vartheta_{k\ell_k} - 97\tau_\rho^{-3}\tilde{\varepsilon}_n \Big)$$

Recalling $\underline{\vartheta} := \min_k \vartheta_{k\ell_k}$ and by assumption $\frac{C_8^2}{144}\underline{\vartheta} \ge 2 \cdot 97\tau_\rho^{-3}\tilde{\varepsilon}_n$, we get

$$\widetilde{Y} \ge \frac{C_8^2}{288}\underline{\vartheta} \sum_{k=1}^{K} d_+^k \ge \frac{C_8^3}{1800}\underline{\vartheta}\, n\nu_n \tag{2.99}$$

using $d_i \ge 0.16 C_8 \nu_n$ on $\mathcal{A}_2$; see (2.83).

Let $\widehat{Y} = \sum_{k=1}^{K} \sum_{i \in \widehat{\mathcal{G}}_k} d_i \psi(\xi_{i\ell_k}, \widehat{\rho}_{k\ell_k})$ and $\mathcal{H}_k = \mathcal{G}_k \Delta \widehat{\mathcal{G}}_k := (\mathcal{G}_k \setminus \widehat{\mathcal{G}}_k) \cup (\widehat{\mathcal{G}}_k \setminus \mathcal{G}_k)$. Note that $\sum_k |\mathcal{H}_k| \le \alpha_n n$ on event $\mathcal{M}_n$. Therefore, on $\mathcal{N}_n \cap \mathcal{E}$

$$\begin{aligned} |\widehat{Y} - \widetilde{Y}| &= \sum_{k=1}^{K} \sum_{i \in \mathcal{H}_k} d_i \psi(\xi_{i\ell}, \widehat{\rho}_{k\ell_k}) \\ &\le \frac{2}{\tau_\rho} \sum_{k=1}^{K} \sum_{i \in \mathcal{H}_k} d_i \le \frac{1.92}{\tau_\rho} \alpha_n n \nu_n. \end{aligned} \tag{2.100}$$

The second inequality follows from (2.95) and noting that $(\xi_{i\ell} - \widehat{\rho}_{k\ell_k})^2 \le 1$. The third inequality is by (2.83).

The assumption $\frac{C_8^3}{1800}\underline{\vartheta} \ge 2 \cdot \frac{1.92}{\tau_\rho}\alpha_n$ combined with (2.99) and (2.100) gives

$$\widehat{Y} \ge \frac{C_8^3}{3600}\underline{\vartheta}\, n\nu_n. \tag{2.101}$$

Estimated column labels Finally, we consider the case where the column labels $\widehat{y}$ are estimated using the community detection algorithm. Let $X'_{i\ell} = \sum_{j\in S_1} A_{ij} 1\{\widehat{y}_j = \ell\}$ and $\xi'_{i\ell} = X'_{i\ell}/d_i$,

$$\widehat{Y}' = \sum_{k=1}^{K} \sum_{i\in\widehat{\mathcal{G}}_k} d_i \psi(\xi'_{i\ell}, \widehat{\rho}'_{k\ell_k}), \quad \text{where } \widehat{\rho}'_{k\ell} = \frac{\sum_{i\in\widehat{\mathcal{G}}_k} X'_{i\ell}}{\sum_{i\in\widehat{\mathcal{G}}_k} d_i}.$$

On $\mathcal{M}_n$, we have $|X_{i\ell} - X'_{i\ell}| \le n\kappa_n$. Letting

$$\varepsilon'_n := \frac{n\kappa_n}{0.16 C_8 \nu_n},$$

it follows that on $\mathcal{A}\cap\mathcal{M}_n$,

$$|\xi_{i\ell} - \xi'_{i\ell}| \le n\kappa_n/d_i \le \varepsilon'_n.$$

Assuming $\alpha_n \le 0.2\tau_{\mathcal{C}}$, we have $|\widehat{\mathcal{G}}_k| \ge (0.4\tau_{\mathcal{C}} - \alpha_n)n \ge 0.2\tau_{\mathcal{C}} n$ on $\mathcal{A}$. Then, on $\mathcal{A}\cap\mathcal{M}_n$,

$$|\widehat{\rho}_{k\ell_k} - \widehat{\rho}'_{k\ell_k}| \le \frac{n\kappa_n}{0.16 C_8 \nu_n \cdot 0.2\tau_{\mathcal{C}} n} = \frac{\varepsilon'_n}{0.2\tau_{\mathcal{C}} n} \le \varepsilon'_n$$

where we have used the assumption $n \ge 5/\tau_{\mathcal{C}}$.

Recall that $\widehat{\rho}_{k\ell_k} \ge \tau_\rho/2$ on event $\mathcal{N}_n$. Then, by the assumption that $\varepsilon'_n \le \tau_\rho/4$, we have $\min\{\widehat{\rho}_{k\ell_k}, \widehat{\rho}'_{k\ell_k}\} \ge \tau_\rho/4$. We can, then, apply Lemma 11 with $c_1 = 4/\tau_\rho$ to obtain

$$\big|\psi(\xi_{i\ell_k}, \widehat{\rho}_{k\ell_k}) - \psi(\xi'_{i\ell_k}, \widehat{\rho}'_{k\ell_k})\big| \le 768\tau_\rho^{-3}\varepsilon'_n, \quad \forall i \in \mathcal{G}_k,\ \forall k \in [K].$$

Furthermore, on $\mathcal{N}_n \cap \mathcal{A}$,

$$|\widehat{Y}' - \widehat{Y}| \le 768\tau_\rho^{-3}\varepsilon'_n \cdot (0.96\nu_n) \cdot (0.6n) \le \frac{2765}{C_8\tau_\rho^3}\kappa_n n^2 \tag{2.102}$$

Combining (2.101) and (2.102), on event $\mathcal{N}_n \cap \mathcal{E} \cap \mathcal{H}$

$$\widehat{Y}' \ge \Big(\frac{C_8^3}{3600}\underline{\vartheta} - \frac{2765}{C_8\tau_\rho^3}\kappa_n n/\nu_n\Big) n\nu_n \ge \frac{C_8^3}{7200}\underline{\vartheta} n\nu_n,$$

assuming $\underline{\vartheta} C_8^3/7200 \geq \frac{2765}{C_8 \tau_\rho^3} \kappa_n n/\nu_n$.

Furthermore, $\tilde{n} = |S_2| \leq n$, hence $\gamma_{\tilde{n}} = \sqrt{\tilde{n}(L-1)} \leq \sqrt{nL}$ and we have

$$\widehat{T}_n \geq \frac{1}{\sqrt{2}}\Big(\frac{\widehat{Y}'}{\gamma_{\tilde{n}}} - \gamma_{\tilde{n}}\Big) \geq \sqrt{\frac{Ln}{2}}\left(\frac{C_8^3}{7200L}\underline{\vartheta}\,\nu_n - 1\right) \geq \frac{C_8^3}{14400\sqrt{2L}}\underline{\vartheta}\sqrt{n}\nu_n.$$

where the last inequality is by assumption $\underline{\vartheta} \geq 14400L/(C_8^3\nu_n)$.

Finally, we put together the probabilities. From Lemma 12 and 14,

$$\mathbb{P}(\mathcal{A}) \geq \mathbb{P}(\mathcal{A}_1 \cap \mathcal{D}) = \mathbb{P}(\mathcal{A}_1) - \mathbb{P}(\mathcal{A}_1 \cap \mathcal{D}^c) \geq 1 - 3.2n^{-1}.$$

Furthermore, with Lemma 13, 16, 18 and 19,

$$\begin{aligned}
\mathbb{P}(\mathcal{N}_n \cap \mathcal{E} \cap \mathcal{H}) &= \mathbb{P}\Big(\mathcal{A} \cap \mathcal{B} \cap \mathcal{M}_n \cap \mathcal{R} \cap \mathcal{E} \cap \mathcal{H}\Big) \\
&= \mathbb{P}(\mathcal{A} \cap \mathcal{R} \cap \mathcal{M}_n) - \mathbb{P}\Big(\mathcal{A} \cap \mathcal{R} \cap \mathcal{M}_n \cap (\mathcal{B} \cap \mathcal{E} \cap \mathcal{H})^c\Big) \\
&\geq \mathbb{P}(\mathcal{A} \cap \mathcal{R} \cap \mathcal{M}_n) - \mathbb{P}(\mathcal{A} \cap \mathcal{B}^c) - \mathbb{P}(\mathcal{A} \cap \mathcal{E}^c) - \mathbb{P}(\mathcal{A} \cap \mathcal{R} \cap \mathcal{H}^c) \\
&\geq 1 - 3.2n^{-1} - 4LKn^{-1} - \mathbb{P}(\mathcal{M}_n^c) - 2Ln^{-1} - 2Ln^{-1} - Kn^{-c} \\
&\geq 1 - 12KLn^{-1} - Kn^{-c} - \mathbb{P}(\mathcal{M}_n^c).
\end{aligned}$$

Simplifying the assumptions The following is a list of all the assumptions we used in the proof:

$$
\begin{array}{lll}
n \geq 5/\tau_{\mathcal{C}} & \frac{\log n}{n} \leq \tau_c/300 & \frac{\log n}{n} < (\frac{\tau_\rho}{4K})^2 \\
\frac{\log n}{n} \leq 0.04 C_8^2 & \frac{\log n}{\nu_n} \leq 0.001 C_8 & \frac{\log n}{n} \leq (\frac{\tau_\rho^2}{4K})^2 \\
\frac{\log n}{\nu_n} \leq 0.064 \tau_{\mathcal{C}} C_8 n & \alpha_n \leq \tau_{\mathcal{C}} \tau_\rho C_8/18 & \alpha_n \leq \sqrt{(\log n)/\nu_n} \\
\frac{\log n}{\nu_n} \leq 0.16 C_8 & \varepsilon_n = \frac{58}{\tau_\rho \tau_{\mathcal{C}} C_8}\sqrt{\frac{\log n}{\nu_n}} \leq \tau_\rho/2 & \frac{\log n}{n} \leq \min\{\frac{\tau_\rho^2}{4}, \tau_{\mathcal{C}}\} \\
\frac{C_8^2}{144}\underline{\vartheta} \geq 2 \cdot 97 \tau_\rho^{-3} \mu_n \varepsilon_n & \frac{C_8^3}{1800}\underline{\vartheta} \geq 2 \cdot \frac{1.92}{\tau_\rho}\alpha_n & \varepsilon_n' = n\kappa_n/(0.16 C_8 \nu_n) \leq \tau_\rho/4 \\
\frac{C_8^3}{7200}\underline{\vartheta} \geq \frac{2765}{C_8 \tau_\rho^3}\kappa_n n/\nu_n & \underline{\vartheta} \geq 14400 L/(C_8^3 \nu_n) & \alpha_n \leq 0.2\tau_{\mathcal{C}}
\end{array}
$$

We recall that

$$
c_2 = \frac{C_8}{100} = \frac{\tau_{\mathcal{C}} \tau_h \tau_\theta^2}{100}, \quad \tau_\rho = \frac{C_8}{2\tau_\theta L} = \frac{50 c_2}{\tau_\theta L} = \frac{\tau_{\mathcal{C}} \tau_h \tau_\theta}{2L}.
$$

The conditions on $\frac{\log n}{n}$ can summarized as follows:

$$
\sqrt{\frac{\log n}{n}} \leq \min\left\{\sqrt{\frac{\tau_{\mathcal{C}}}{300}},\ \frac{\tau_\rho}{2},\ 20c_2,\ \frac{\tau_\rho^2}{4K}\right\}. \tag{2.103}
$$

We also note that if $n \geq 2$, then $\frac{\log n}{n} \leq \tau_{\mathcal{C}}/300$ implies $n \geq 5/\tau_{\mathcal{C}}$. Since $\tau_\rho^2 \leq \tau_\rho$, we can drop $\frac{\tau_\rho}{2}$ from (2.103). Similarly,

$$
\frac{\tau_\rho^2}{4K}/(20c_2) = \frac{\tau_\rho^2}{4K}\frac{50}{20 L \tau_\rho \tau_\theta} = \frac{5}{8}\frac{\tau_\rho}{KL\tau_\theta} = \frac{5}{16}\frac{\tau_{\mathcal{C}}\tau_h}{KL^2} \leq 1,
$$

hence we can also drop $20c_2$ from (2.103). Since $\tau_\rho^2/(4K) \geq \tau_{\mathcal{C}}^2 \tau_h^2 \tau_\theta^2/(18KL^2) = \frac{2}{9}\frac{\tau_\rho^2}{K}$ and $\sqrt{\tau_{\mathcal{C}}/300} \geq \tau_{\mathcal{C}}^2/18$, condition (2.103) holds under assumption (2.37).

The condition $\varepsilon_n \leq \tau_\rho/2$ is

$$
\frac{0.58}{\tau_{\mathcal{C}} c_2}\sqrt{\frac{\log n}{\nu_n}} \leq \frac{\tau_\rho^2}{2} = \frac{1250 c_2^2}{\tau_\theta^2 L^2}
$$

which is equivalent to

$$\sqrt{\frac{\log n}{\nu_n}} \leq \frac{1250}{0.58}\frac{\tau_{\mathcal{C}} c_2^3}{\tau_\theta^2 L^2} = \frac{1250}{0.58}\frac{\tau_{\mathcal{C}}^2 \tau_h c_2^3}{\tau_{\mathcal{C}}\tau_h \tau_\theta^2 L^2} = \frac{1250}{58}\frac{\tau_{\mathcal{C}}^2 \tau_h c_2^3}{c_2 L^2}.$$

The condition is satisfied if

$$\sqrt{\frac{\log n}{\nu_n}} \leq 21\frac{\tau_{\mathcal{C}}^2 \tau_h c_2^2}{L^2} \tag{2.104}$$

which is what is assumed in (2.38).

The three upper bounds on $\frac{\log n}{\nu_n}$ can be combined into

$$\frac{\log n}{\nu_n} \leq c_2 \min\{0.1,\, 6.4\tau_{\mathcal{C}} n\}$$

Since $n \geq 5/\tau_{\mathcal{C}}$, we have $6.4\tau_{\mathcal{C}} n \geq 0.1$, hence it is enough that $\frac{\log n}{\nu_n} \leq 0.1 c_2$. Next, since $c_2 \leq 0.01$, we have

$$21\frac{\tau_{\mathcal{C}}^2 \tau_h c_2^2}{L^2} \leq 21\tau_{\mathcal{C}}^2 \tau_h c_2^2 \leq 21 c_2^2 \leq 0.21 c_2 \leq \sqrt{0.1 c_2}$$

showing that (2.104) is already enough to guarantee this condition.

The three assumptions on α_n can be combined into

$$\alpha_n \leq \min\left\{\frac{100}{36}\tau_{\mathcal{C}}^2 \tau_h \tau_\theta c_2,\ \sqrt{\frac{\log n}{\nu_n}}\right\}$$

Since $c_2 \leq 0.01\tau_\theta$, we have

$$21\tau_{\mathcal{C}}^2 \tau_h c_2^2 \leq 0.21\tau_{\mathcal{C}}^2 \tau_h \tau_\theta c_2$$

showing that $\alpha_n \leq \sqrt{\frac{\log n}{\nu_n}}$ together with (2.104) is enough to guarantee both upper bounds on α_n.

The conditions involving $\underline{\vartheta}$ are implied by

$$\underline{\vartheta} \geq \max\left\{ \frac{3L^3\mu_n\varepsilon_n}{c_2^2\tau_\rho^3},\ \frac{0.007L}{\tau_\rho c_2^3}\alpha_n,\ \frac{L^3}{5\tau_\rho^3 c_2^4}\frac{\kappa_n n}{\nu_n},\ \frac{0.0144L}{c_2^3\nu_n} \right\} \tag{2.105}$$

where we recall

$$\mu_n := \max\{1, L\sqrt{\nu_n/n}\}, \quad \varepsilon_n = \frac{0.58}{\tau_\rho\tau_{\mathcal{C}}c_2}\sqrt{\frac{\log n}{\nu_n}}.$$

We have

$$\frac{3L^3\mu_n\varepsilon_n}{c_2^2\tau_\rho^3} \leq \frac{2L^3}{\tau_\rho^4\tau_{\mathcal{C}}c_2^3}\zeta_n\sqrt{\frac{\log n}{\nu_n}}$$

Similarly, using $\tau_\rho, \tau_{\mathcal{C}} \leq 1$,

$$\frac{0.0144L}{c_2^3\nu_n} \leq \frac{0.0144}{c_2^3}\zeta_n\sqrt{\frac{\log n}{\nu_n}} \leq \frac{2L^3}{\tau_\rho^4\tau_{\mathcal{C}}c_2^3}\zeta_n\sqrt{\frac{\log n}{\nu_n}}$$

and using assumption $\alpha_n \leq \sqrt{\frac{\log n}{\nu_n}}$

$$\frac{0.007}{\tau_\rho c_2^3}\alpha_n \leq \frac{0.007}{\tau_\rho c_2^3}\sqrt{\frac{\log n}{\nu_n}} \leq \frac{2L^3}{\tau_\rho^4\tau_{\mathcal{C}}c_2^3}\zeta_n\sqrt{\frac{\log n}{\nu_n}}.$$

It follows that the assumption (2.39) in the statement of theorem is enough to guarantee (2.105).

Finally, condition $n\kappa_n/(16c_2\nu_n) \leq \tau_\rho/4$ is equivalent to what is stated in (2.38). The proof is complete.

Chapter 3

Label consistency in overfitted generalized k-means

3.1 Introduction

Consider the problem of clustering data points sampled according to some probability measure μ from a normed space $\mathcal{X}$ with norm $\|\cdot\|_{\mathcal{X}}$. In the ideal setting, the generalized k-means clustering minimizes the population cost function

$$Q(\xi;\mu) := \Big(\int \min_{1\le\ell\le L} \|x-\xi_\ell\|_{\mathcal{X}}^p d\mu(x)\Big)^{1/p} \tag{3.1}$$

where $\xi = (\xi_1,\dots,\xi_L) \in \mathcal{X}^L$ is a set of L vectors in $\mathcal{X}$, for some fixed integer L. In practical data analysis, we are given a sample $\{x_1,\dots,x_n\}$ drawn from μ and solve an empirical version of (3.1), namely,

$$\widehat{Q}(\xi) = Q(\xi;\mathbb{P}_n) := \Big(\frac{1}{n}\sum_{i=1}^n \min_{1\le\ell\le L} \|x_i-\xi_\ell\|_{\mathcal{X}}^p\Big)^{1/p}. \tag{3.2}$$

Here, $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is the empirical measure associated with the sample and δ_x is the point mass measure at x. The minimizer of $\widehat{Q}(\cdot)$ over $\mathcal{X}^L$ is denoted as $\widehat{\xi} = (\widehat{\xi}_1,\dots,\widehat{\xi}_L)$ and each point x_i is assigned a cluster label $\widehat{z}_i := \operatorname{argmin}_\ell \|x_i - \widehat{\xi}_\ell\|_{\mathcal{X}}$.

Meanwhile, we assume that each data point x_i also has a true cluster label $z_i \in [K] := \{1,\dots,K\}$ which is determined solely by an unknown data-generating process.

These true labels are not necessarily related to the optimal solutions of (3.1) or (3.2). To distinguish the two, we refer to the clustering induced by (z_i) as the **true clustering**, while a clustering that minimizes the generalized k-means cost function (3.2), i.e., the clustering induced by $(\widehat{z}_i)$, is referred to as an **optimal k-means clustering**. In this chapter, we consider the *label consistency* problem, that is, how close the labels $(\widehat{z}_i)$ estimated by k-means clustering are to the true labels (z_i). Note that we allow the number of k-means clusters L to be different from the true number of clusters K.

In the above formulation, the case where $p = 2$, $\mathcal{X} = \mathbb{R}^d$ and $\|\cdot\|_{\mathcal{X}}$ is the Euclidean norm leads to the classical and widely used k-means problem. Much of the theoretical analysis of k-means has been performed in this case. Early work has focused on how close the optimization problems based on the empirical and ideal cost functions (3.2) and (3.1) are to each other, where the closeness is measured in terms of the recovered centers (i.e., $\widehat{\xi}$ and ξ) or the optimal value of the objective function.

Such consistency results are proved, for the global minimizers of (3.2), in the early work of [Mac67; Pol81] and also in [Pol82; Lin02] from the vector quantization perspective. These classical results do not directly apply to the performance of the k-means in practice, mainly because solving (3.2) is NP-hard and approximation methods are usually applied to deal with it. Also, considerations of the label consistency problem are absent from this line of work since no true clustering, external to the k-means problem, is assumed to exist.

More recently, there has been more interest in the consistency of practical k-means algorithms [Kło20; LZ16] as well as the label consistency problem. Lu and Zhou [LZ16] obtain sharp bounds on the label consistency of the Lloyd's algorithm [Llo82] under a sub-Gaussian mixture model. Semidefinite programming (SDP) relaxation is another popular technique for deriving polynomial-time approximations to the k-means problem [PW07]. Its label consistency has been studied when data is generated from the stochastic ball model [Awa+15; Igu+17], sub-Gaussian mixtures [MVW16; FC18; GV19], the Stochastic Block Model (SBM) [GV19] and general models [Li+20b]. Convex clustering is another relaxation method whose label consistency has been discussed in [Zhu+14; Pan+17; JVZ20; STY21]. The literature on community detection in SBM, a network clustering problem, is

also mainly focused on label consistency and inspires our work here; see [Abb17; ZA19] for a review of those results. For label consistency in kernel spectral clustering, see [AR21].

In this chapter, we study the label consistency of approximate solutions of the generalized k-means problem (3.2) when $L \geq K$. Our focus will be on the overfitted case where $L > K$. This is often relevant in practice since the data-generating process may have a natural number of clusters K that is unknown a priori. An example is the sub-Gaussian mixture with K components. More interesting examples are given in Section 3.3. All the aforementioned works on label consistency exclusively consider the correctly-fitted case $L = K$. We show that when the data-generating process admits a set of centers that satisfy certain separation conditions, estimated labels with $L \geq K$ clusters, are close to a *refinement* of the true labels. These bounds reduce to the label consistency criteria for $L = K$, but have no counterpart in the literature for $L > K$.

Overfitting in k-means is considered in [Wei16; MRS20] where it is shown to improve the approximation factor (see Assumption 2(b)) of certain polynomial-time k-means algorithms. Analysis of the approximation factor is concerned with how close one can get to the optimal value of the k-means objective function. In contrast, we are concerned with the label recovering problem and not directly concerned with how well the objective function is approximated. Our work is also aligned with the recent trend of *beyond worst case* analysis of the NP-hard problems [CAS17], where the performance of the algorithms are considered assuming that there are some meaningful structures in the data (e.g., true clusters). We refer to Section 3.2.2 for a more detailed comparison with this literature.

Our results are algorithm-free in the sense that they apply to any algorithm that achieves a constant-factor approximation to the optimal objective. They are also model-free in the sense that we do not make any explicit assumption on the data-generating process. This is important in practice, since many common data models, such as sub-Gaussian mixtures, are often too simplified to capture real clustering problems. We provide examples of more complicated data models in Section 3.3 and show how our general results can provide new insights for these models. Since k-means clustering often appears as a building block in many sophisticated clustering algorithms, we believe our

results will be of broad interest in understanding the performance of clustering algorithms in overfitted settings.

Notation. $Q(\xi; \mu)$ is only dependent on the set of values among the coordinates of ξ. Although we view ξ as a vector (for which the order of elements matter), with some abuse of notation, we view $Q(\cdot; \mu)$ as a set function (mapping $2^{\mathcal{X}}$ to $\mathbb{R}$) that is only sensitive to the set of values represented by ξ. This justifies using the the same symbol for the function irrespective of the number of coordinates of ξ, i.e., the number of clusters. The reason to keep ξ as an (ordered) vector is to make the cluster labels well-defined. For simplicity, let $\|\cdot\| = \|\cdot\|_{\mathcal{X}}$. For the case where $\mathcal{X} \subset \mathbb{R}^d$, one often takes $\|\cdot\|$ to be the Euclidean norm, but our results are valid for any norm on $\mathbb{R}^d$, and more broadly any normed space $\mathcal{X}$.

3.2 Main Results

We first state assumptions about the k-means clustering algorithm.

Assumption 2. *Consider an algorithm for the generalized k-means problem* (3.2)*, referred to as ALG(p) hereafter, and let $\widehat{\xi}^{(L)} \in \mathcal{X}^L$ and $\widehat{\xi}^{(K)} \in \mathcal{X}^K$ be its estimated centers when applied with L and K clusters, respectively. Let $L \geq K$. Assume that ALG(p) has the following properties, for all input sequences (x_i):*

(a) Efficiency: The Voronoi cell of every estimated center $\widehat{\xi}_\ell^{(L)}$ contains at least one of (x_i).

(b) κ-approximation: $\widehat{Q}(\widehat{\xi}^{(K)}) \leq \kappa \cdot \min_{\xi \in \mathcal{X}^K} \widehat{Q}(\xi)$, and similarly with K replaced by L.

Efficiency can be achieved by substituting centers whose Voronoi cells have an empty intersection with $\{x_i\}$, by those having the opposite property. For κ-approximation, the factor κ can depend on the number of clusters K (or L). For example, the k-means++ algorithm has $\kappa = O(\log K)$, with high probability over the initialization [AV06]. However, there are also constant-factor approximation algorithms for k-means where $\kappa = O(1)$ independent of K (or L) [Mat00; Kan+04; KSS04]. For example, with local search, k-means++ can achieve a constant-factor approximation [LS19]. In addition,

κ-approximation is not required for all inputs. That is, we are not concerned with the worst-case approximation factor. The κ in Assumption 2(b) is the approximation factor of the algorithm on the specific data under consideration. It is enough for an algorithm to achieve good approximation only on the data of interest.

For some of the results, Assumption 2(b) can be replaced with the following modified version: (b') κ-approximation only for K clusters plus a mononoticity assumption: $\widehat{Q}(\widehat{\xi}^{(L)}) \leq \widehat{Q}(\widehat{\xi}^{(K)})$. Mononoticity is also a reasonable requirement and obviously true for the exact k-means solutions.

Next, we extend the definition of the misclassification rate to the overfitted case.

Definition 2. The (generalized) misclassification rate between two label vectors $z \in [K]^n$ and $\widehat{z} \in [L]^n$, with $K \leq L$, is

$$\mathrm{Mis}(z, \widehat{z}) = \min_{\omega} \frac{1}{n} \sum_{i=1}^{n} 1\{z_i \neq \omega(\widehat{z}_i)\},$$

where the minimization ranges over all surjective maps $\omega : [L] \to [K]$.

When $L = K$, a surjective map ω is necessarily a bijection and the above becomes the usual definition of misclassification rate when the number of clusters is correctly identified. In this case, $\mathrm{Mis}(z, \widehat{z}) = 0$ means that there is a one-to-one correspondence between the estimated and true clusters. The generalized definition above allows us to extend this notion of exact recovery to the case $L > K$. In particular, $\mathrm{Mis}(z, \widehat{z}) = 0$ when $L > K$, if and only if $\widehat{z}$ is a *refinement* of z. To see this, note that $\mathrm{Mis}(z, \widehat{z}) = 0$ implies the existence of a map $\omega : [L] \to [K]$ such that $\omega(\widehat{z}_i) = z_i$ for all i. This in turn is equivalent to: $\widehat{z}_i = \widehat{z}_{i'} \implies z_i = z_{i'}$, which is the refinement relation for the associated clusters. In general, $\mathrm{Mis}(z, \widehat{z})$ is small if $\widehat{z}$ is close to a refinement of z.

We also use the (optimal) matching distances between elements of two vectors viewed as sets.

Definition 3. For $\xi \in \mathcal{X}^L$ and $\xi^* \in \mathcal{X}^K$, define the ℓ_∞ and ℓ_p optimal matching distances

as

$$d_\infty(\xi, \xi^*) = \min_\sigma \max_{1\le k\le K} \|\xi_{\sigma(k)} - \xi_k^*\|, \quad d_p(\xi, \xi^*) = \min_\sigma \Big(\sum_{k=1}^K \|\xi_{\sigma(k)} - \xi_k^*\|^p \Big)^{1/p},$$

where $\sigma : [K] \to [L]$ ranges over all injective maps.

For $K = L$, d_∞ is an upper bound on the Hausdorff distance between the two sets. Obviously, we have $d_\infty \le d_p$ for any $p \ge 1$.

3.2.1 Distance to True Centers

Let $z = (z_i)_{i=1}^n \in [K]^n$ be a given set of true labels for the data points $(x_i)_{i=1}^n$. In addition, our results are stated in terms of a set of vectors $\xi^* = (\xi_k^*)_{k=1}^K$ which we refer to as the "true centers". Throughout, ξ^* will be only vaguely specified. The only requirement on ξ^* is that together with the observed data points (x_i) and the true labels (z_i), they satisfy the deviation bounds in each theorem, e.g., $\max_{1\le i\le n} \|x_i - \xi_{z_i}^*\| \le \eta$ in Theorem 5, etc. Let $\pi_k = \sum_{i=1}^n 1\{z_i = k\}/n$ be the proportion of observed data points in true cluster k and let $\pi_{\min} = \min_k \pi_k$.

We let $\widehat{\xi}$ be a solution of the k-means algorithm with $L \ge K$ centers and let $\widehat{z}_i \in \operatorname{argmin}_\ell \|x_i - \widehat{\xi}_\ell\|$ be the corresponding estimated labels. Our first result provides guarantees for exact label recovery, in the extended sense of recovering a refinement of the true partition when $L > K$ and the exact partition when $L = K$.

Theorem 5 (Exact recovery)**.** *Consider a vector of (true) centers $\xi^* \in \mathcal{X}^K$ and labels $(z_i)_{i=1}^n \in [K]^n$. Pick $\eta, \delta > 0$ such that $\max_{1\le i\le n} \|x_i - \xi_{z_i}^*\| \le \eta$, and*

$$\min_{(k,k'):\, k\ne k'} \|\xi_k^* - \xi_{k'}^*\| \ge \delta. \tag{3.3}$$

Consider an algorithm ALG(p) for problem (3.2)*, satisfying Assumption 2, and let $(\widehat{z}_i)_{i=1}^n \in [L]^n$ and $\widehat{\xi} \in \mathcal{X}^L$ be the estimated labels and centers of ALG(p) applied with the $L \ge K$.*

Then,

$$\frac{\delta}{\eta} > 2\frac{(1+\kappa)}{\pi_{\min}^{1/p}} + 4 \quad \Longrightarrow \quad \mathrm{Mis}(z,\widehat{z}) = 0, \quad d_p(\widehat{\xi},\xi^*) \le \frac{(1+\kappa)\eta}{\pi_{\min}^{1/p}}. \tag{3.4}$$

When $L = K$, the assertion Mis $= 0$ means that there is a permutation σ on $[K]$ such that $\sigma(\widehat{z}_i) = z_i$ for all i, that is, we have the exact recovery of labels (z_i) in the classical sense. When $L > K$, Theorem 5 guarantees the exact recovery of a refinement of the true labels (z_i).

Example 1 (Stochastic Ball Model)**.** Assume that data are generated from the stochastic ball model considered in [NW15], where $x_i = \xi^*_{z_i} + r_i$ with r_i sampled independently from a distribution supported on the unit ball in $\mathbb{R}^d$. Here, $\{\xi^*_k\}_{k=1}^K \subset \mathbb{R}^d$ are a fixed set of centers. Clearly, we can take $\eta = 1$ in Theorem 5. Then, any κ-approximate k-means algorithm achieves exact recovery when $\delta > 2 + 2(1+\kappa)/\sqrt{\pi_{\min}}$ for $L = K$. In the overfitted case, when $\delta > 4 + 2(1+\kappa)/\sqrt{\pi_{\min}}$, the estimated label vector is an exact refinement of the true labels (z_i). □

In the above example, although it is intuitively clear that for a sufficiently large δ, the solution of the k-means problem should achieve exact label recovery (in the extended sense), Theorem 5 allows us to provide a provable guarantee for any constant-factor approximation, with an explicit bound on the separation parameter δ.

We now turn to approximate recovery where the misclassification rate is small.

Theorem 6 (Approximate Recovery)**.** *Consider a vector of (true) centers $\xi^* \in \mathcal{X}^K$ and labels $(z_i)_{i=1}^n \in [K]^n$. Pick $\varepsilon, \delta > 0$ such that $(\frac{1}{n}\sum_{i=1}^n \|x_i - \xi^*_{z_i}\|^p)^{1/p} \le \varepsilon$, and (3.3) holds. Consider an algorithm ALG(p) for problem (3.2), satisfying Assumption 2, and let $(\widehat{z}_i)_{i=1}^n \in [L]^n$ and $\widehat{\xi} \in \mathcal{X}^L$ be the estimated labels and centers of ALG applied with the $L \ge K$. Then, for any $c > 2$,*

$$\frac{\delta}{\varepsilon} > \frac{(1+\kappa)c}{\pi_{\min}^{1/p}} \quad \Longrightarrow \quad \mathrm{Mis}(z,\widehat{z}) < K(1+\kappa)^p c^p \Big(\frac{\varepsilon}{\delta}\Big)^p, \quad d_p(\widehat{\xi},\xi^*) \le \frac{(1+\kappa)\varepsilon}{\pi_{\min}^{1/p}}. \tag{3.5}$$

The key difference between Theorems 5 and 6 is the bounds assumed on the deviations

$D_i := \|x_i - \xi^*_{z_i}\|, i \in [n]$. Theorem 5 assumes a bound on the maximum distance to true centers, $\max_i D_i$, while Theorem 6 assumes a bound on an average distance, $(\frac{1}{n}\sum_i D_i^p)^{1/p}$, leading to a more relaxed condition.

Example 2 (Sub-Gaussian mixtures)**.** Let us assume that the data is generated from a K-component sub-Gaussian mixture model $x_i = \xi^*_{z_i} + d^{-1/2} w_i$, where $w_i = (w_{i1}, \ldots, w_{id}) \in \mathbb{R}^d$ is a zero mean sub-Gaussian noise vector with sub-Gaussian parameter σ_i, and $z_i \in [K]$ is the latent label of the ith observation. Here we define the sub-gaussian vector as: A random vector $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ is sub-gaussian if the one-dimensional marginals $u^T X$ are sub-gaussian random variables for all $u \in \mathbb{R}^d$ [Ver18, Definition 3.4.1]. This is an extension of the sub-Gaussian mixture model considered in [EK10]. Determining whether $(\xi^*_k)_{k=1}^K$ is actually the solution of the population problem (3.1) is, itself, challenging and the answer may depend on the exact distribution of $\{w_i\}$. Nevertheless, our results allow us to treat (ξ^*_k) as the true centers. Below we sketch how Theorem 6 applies in this case. First we have the following lemma

Lemma 20. *Let $w_i = (w_{i1}, \ldots, w_{id}) \in \mathbb{R}^d$ be a zero mean sub-Gaussian noise vector with sub-Gaussian parameter σ_i for $i \in [n]$. Let $\sigma_{\max} = \max_i \sigma_i$ and set $\alpha_i^2 := \mathbb{E}\|d^{-1/2} w_i\|_2^2$ and $\bar{\alpha}_n^2 := \frac{1}{n}\sum_{i=1}^n \alpha_i^2$. Assume that there is a numerical constant $C > 0$ such that $\bar{\alpha}_n^2 \le C\sigma_{\max}^2$. Then, we have*

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \frac{\|w_i\|^2}{d} > 2\bar{\alpha}_n^2\Big) \le \exp\Big(-c_1 n \frac{\bar{\alpha}_n^2}{\sigma_{\max}^4}\Big).$$

for some numerical constant $c_1 > 0$.

Let $p_n = e^{-c_1 n \bar{\alpha}_n^4/\sigma_{\max}^4}$. Note that $\|x_i - \xi^*_{z_i}\|^2 = \|w_i\|^2/d$, then taking $\varepsilon^2 = 2\bar{\alpha}_n^2$ and $p = 2$ in Theorem 6, we have that with probability at least $1 - p_n$,

$$\frac{\delta^2}{2\bar{\alpha}_n^2} > \frac{(1+\kappa)^2 c^2}{\pi_{\min}} \quad \Longrightarrow \quad \mathrm{Mis}(z, \widehat{z}) \le 2K(1+\kappa)^2 c^2 \Big(\frac{\bar{\alpha}_n}{\delta}\Big)^2,$$

where δ is as in (3.3) and $c > 2$. In a general problem, $\bar{\alpha}_n$, $\sigma_{\max}$ and δ all can vary with n. In order to have label consistency for an ALG(2) algorithm, it is enough to have $\bar{\alpha}_n/\delta = o(1)$ and $n\bar{\alpha}_n^4/\sigma_{\max}^4 \to \infty$. The consistency here is based on the extended Definition 2 and

includes the overfitted case in which a refinement of the true labels is consistently recovered. We note that the model in this example includes a very general Gaussian mixture model as a special case, namely the case $w_i \sim N(0, \Sigma_i)$ where the covariance matrices $\Sigma_i \in \mathbb{R}^{d\times d}$ are allowed to vary with each data point. In this case, one can take $\sigma_{\max} = \max_{1\le i\le n} \|\Sigma_i\|$ where $\|\cdot\|$ denotes the operator norm, and $\bar{\alpha}_n^2 := \frac{1}{n}\sum_{i=1}^n \operatorname{tr}(\Sigma_i)/d$. □

Remark 4. Theorem 6 provides an upper bound on the misclassification rate when a certain separation condition is satisfied. To simplify, consider the case $K = \kappa = p = 2$ and take $c = 2.1$. Then, Theorem 6 implies the following: For every $\beta > 0$, there exists a constant $c_1(\beta, \pi_{\min}) > 0$ such that if

$$\delta/\varepsilon \ge c_1(\beta, \pi_{\min}), \tag{3.6}$$

then any 2-factor k-means algorithm will have Miss $\le \beta$ to the target labels. The next proposition shows that condition (3.6) is sharp up to constants.

Proposition 3. *There exists a family of datasets $\{(x_i, z_i)\}_{i=1}^n$, with $K = 2$ balanced true clusters (i.e., $\pi_{\min} = 1/2$) and parameterized by true center separation δ and $\varepsilon = (\frac{1}{n}\sum_{i=1}^n \|x_i - \xi^*_{z_i}\|^2)^{1/2}$ with the following property: Given any constant $\beta \in (0, 1/2)$, there exists a constant $c_2(\beta) > 0$, such that if $\delta/\varepsilon < c_2(\beta)$, then any 2-factor k-means approximation algorithm with $L = 2$ clusters has misclassification rate satisfying $\frac{1}{2} - \beta \le$ Miss $\le \frac{1}{2}$. Moreover, any 2-factor k-means approximation algorithm with $L = 4$ clusters will recover a perfect refinement of the original clusters in the above setting.*

This proposition shows that if the separation condition (3.6) is reversed, one can force the performance of any k-means algorithm to be arbitrarily close to that of random guessing. The true centers in Proposition 3 are the natural centers implied by the k-means cost function for the true labels, that is, $\xi^*_k = \frac{1}{n}\sum_i x_i 1\{z_i = k\}$ for $k = 1, 2$. One can take $c_1(\beta, \pi_{\min}) = 6.3\max(1/\pi_{\min}, 2/\beta)^{1/2}$ and $c_2(\beta) = \sin(\tan^{-1}(\sqrt{\beta/45}))$ for the constants in (3.6) and Proposition 3.

3.2.2 Connection to Distribution Stability

The separation condition (3.6) is related to the *distribution stability* introduced in [ABS10]. Roughly speaking distribution stability plus the following property implies our condition:

(D1) For every pair of distinct clusters C_k and C_ℓ with centers ξ_k^* and ξ_ℓ^*, there is a point $x \in C_\ell$ such that $\|x - \xi_k^*\| \le \|\xi_\ell^* - \xi_k^*\|$.

That is, every cluster C_ℓ has points which are closer than ξ_ℓ^* to the centers of other clusters. This property is quite mild and one expects it to hold with high probability if the distribution of the points have positive density w.r.t. to the (full-dimensional) Lebesgue measure in a ball around the center. The above seems to suggest that distribution stability is slightly weaker than our condition (3.6). However, in the presence of (D1), we can also significantly relax distribution stability to arrive at our condition

First recall that the distribution stability for the K-means assumes the following [ABS10]:

$$\|x - \xi_k^*\|^2 \ge \beta \cdot \frac{\mathrm{OPT}_K}{n_k}, \quad \text{for all } x \notin C_k,$$

where $\mathrm{OPT}_K = \sum_{i=1}^n \|x_i - \xi_{z_i}^*\|^2$ for the K-means optimal cluster labels $\{z_i\} \subset [K]^n$ and optimal centers $\{\xi_k^*\}$. Here, $C_k = \{i : z_i = k\}$ and $n_k = |C_k|$.

In our setting, we do not necessarily need to work with the optimal K-means clustering. So let us generalize the notion as follows: The data $\{x_i\}$ is β-distributed with respect to cluster labels $\{z_i\}$ and centers $\{\xi_k^*\}$ if

$$\|x - \xi_k^*\|^2 \ge \beta \cdot \sum_{i=1}^n \|x_i - \xi_{z_i}^*\|^2 / n_k, \quad \text{for all } x \notin C_k,$$

where $C_k = \{i : z_i = k\}$ and $n_k = |C_k|$. Setting $\frac{1}{n}\sum_{i=1}^n \|x_i - \xi_{z_i}^*\|^2 = \varepsilon^2$ and recalling $\pi_k = n_k/n$, the condition is equivalent to

$$\|x - \xi_k^*\| \ge \frac{\sqrt{\beta} \cdot \varepsilon}{\sqrt{\pi_k}}, \quad \text{for all } x \notin C_k. \tag{3.7}$$

Let us strengthen the condition slightly and consider the following notion instead

$$\|x - \xi_k^*\| \geq \frac{\sqrt{\beta} \cdot \varepsilon}{\sqrt{\pi_{\min}}}, \quad \text{for all } x \notin C_k, \tag{3.8}$$

where $\pi_{\min} = \min_k \pi_k$. This is without loss of generality: We could have stated our results with separate center separation parameters for each cluster, i.e., $\delta_k = \min_{\ell \neq k} \|\xi_k^* - \xi_\ell^*\|$, in which case we could directly compare with the original version (3.7). We opted for the simpler global center separation for simplicity.

Now assume that the data is β-distributed and in addition:

(D1) For all distinct pairs (k, ℓ), there is $x \in C_\ell$ such that $\|x - \xi_k^*\| \leq \|\xi_\ell^* - \xi_k^*\|$.

That is, every cluster C_ℓ has points which are closer than ξ_ℓ^* to the centers of other clusters. Then, it follows that

$$\frac{\delta}{\varepsilon} \geq \frac{\sqrt{\beta}}{\sqrt{\pi_{\min}}} \tag{3.9}$$

which is our separation condition. (Recall that $\delta = \min_{k \neq \ell} \|\xi_k^* - \xi_\ell^*\|$). In fact, in the presence of (D1), we can relax β-distribution stability as follows: Assume (D1) and for the x in (D1) assume that the inequality in (3.8) holds. Then, our separation condition (3.9) follows. Note that (D1) is quite mild and one expects it to hold almost always if there is some full-dimensional randomness in the distribution of the points in a cluster.

Alternatively, our separation condition can be written equivalently as

$$\|x - \xi_k^*\| \geq \frac{\sqrt{\beta} \cdot \varepsilon}{\sqrt{\pi_{\min}}}, \quad \text{for all } x \in \{\xi_\ell^*\}_{\ell \neq k} \tag{3.10}$$

Comparing (3.10) and (3.8), the conditions are somewhat close, but different. Neither condition directly follow from the other one in general. Note also that although in the discussion above, we refer to ξ_k^* as the center of C_k, in our general setting ξ_k^* need not be the optimal center $\frac{1}{n_k} \sum_{i \in C_k} x_i$.

3.2.3 Distance to Fake Centers

We now extend Theorem 6, to allow for "fake" centers $\{\widetilde{\xi}_\ell\}_{\ell=1}^L$ and the corresponding labels $\{\widetilde{z}_i\}$. These can be constructed to reduce the required distance to the data points (x_i).

Theorem 7 (Approximate Recovery, II)**.** *For a fixed $L \geq K$, consider a vector of constructed centers $\widetilde{\xi} \in \mathcal{X}^L$, constructed labels $\widetilde{z} = (\widetilde{z}_i)_{i=1}^n \in [L]^n$ and true labels $z = (z_i)_{i=1}^n \in [K]^n$. Assume that $\widetilde{z}$ is a refinement of z, i.e. there is $\widetilde{\omega} : [L] \to [K]$ such that $\widetilde{\omega}(\widetilde{z}_i) = z_i$ for all $i \in [n]$. Pick $\varepsilon, \delta > 0$ such that*

$$\Big(\frac{1}{n}\sum_{i=1}^n \|x_i - \widetilde{\xi}_{\widetilde{z}_i}\|^p\Big)^{1/p} \leq \varepsilon, \qquad \min_{\ell_1 \neq \ell_2,\ \widetilde{\omega}(\ell_1) \neq \widetilde{\omega}(\ell_2)} \|\widetilde{\xi}_{\ell_1} - \widetilde{\xi}_{\ell_2}\| \geq \delta \tag{3.11}$$

Consider an algorithm ALG(p) for problem (3.2)*, satisfying Assumption 2, and let $(\widehat{z}_i)_{i=1}^n \in [L]^n$ be the estimated label vector of ALG(p) applied with L clusters. Then, for any $c > 2$,*

$$\frac{\delta}{\varepsilon} > \frac{(1+\kappa)c}{\pi_{\min}^{1/p}} \quad \Longrightarrow \quad \mathrm{Mis}(z, \widehat{z}) < K(1+\kappa)^p c^p \Big(\frac{\varepsilon}{\delta}\Big)^p. \tag{3.12}$$

The advantage of Theorem 7 is that when the desired number of clusters L increases, the bound on the misclassification rate can go down: In some applications, by carefully constructing the fake centers $\widetilde{\xi}$, we can make ε smaller as L increases, while roughly maintaining the minimum separation among fake centers associated with the true clusters. If successful, this implies that a refinement of the true clustering can be achieved even when it is hard to recover the true clustering itself. In the following section, we show how this strategy can be applied to some manifold clustering problems.

3.3 Overfitting Cases

We now illustrate applications of Theorem 7 in settings where it is hard to recover true clusters, based on the ideal K, but it is possible to obtain accurate refinements by overfitting. The idea is to consider clusters that look like submanifolds of $\mathbb{R}^d$.

3.3.1 Mixture of Curves

We say that a random variable x has a (ρ, σ) *sub-Gaussian curve distribution* if $x = \gamma(t)$ where $t \in \mathbb{R}$ has a sub-Gaussian distribution with parameter σ and $\gamma : \mathbb{R} \to \mathbb{R}^d$ is a locally ρ-Lipschitz map. i.e., $\|\gamma(t) - \gamma(s)\| \leq \rho|t - s|$ for all $t, s \in \mathbb{R}$ such that $|t - s| \leq \frac{1}{\rho}$.

Proposition 4. *Assume that $(x_i)_{i=1}^n$ are independent draws from a K-component mixture of (ρ, σ) sub-Gaussian curve distributions. That is, $x_i = \gamma_{z_i}(t_i)$ where $z_i \in [K]$, $t_i \sim \mathbb{Q}_{z_i}$ independently across i, each $\mathbb{Q}_k$ is a sub-Gaussian distribution on $\mathbb{R}$ with parameter σ, and each γ_k is locally ρ-Lipschitz. Let $\mathcal{C}_k$ be the support of the distribution of $\gamma_k(t)$ where $t \sim \mathbb{Q}_k$. Assume that*

$$\min_{x \in \mathcal{C}_k,\, y \in \mathcal{C}_{k'}} \|x - y\| \geq \delta > 0, \quad \textit{for all } k \neq k'.$$

Then, there exist a constant $C = C(K, \delta, \rho, \sigma, \kappa)$ such that any ALG(2) satisfying Assumption 2 applied with $L_n \leq C\sqrt{n \log n}$ clusters recovers a perfect refinement of z with probability $\geq 1 - n^{-1}$.

The significance of this result is that one recovers a perfect refinement with the number of partitions $L_n = o(n)$. It is trivial to obtain a perfect refinement with $L_n = n$, but not so with $L_n/n \to 0$. This is especially the case since one can achieve quite complex cluster configurations with the model in Proposition 4, for some of which applying k-means with K clusters will have a misclassification rate bounded away from zero. Section 3.4 provides some such examples where the true cluster centers coincide, causing any k-means algorithm applied with the true K to incur a substantial error. See also Section 3.3.3 for a discussion of whether $L_n = O(\sqrt{n \log n})$ can be improved.

Various extensions of Proposition 4 are possible. We have the following extension to the noisy setting.

Corollary 1. *Assume that the data is given by $y_i = x_i + \frac{1}{\sqrt{d}} w_i$ for $i \in [n]$ where (x_i) are as given in Proposition 4 and w_i are sub-Gaussian noise vectors as in Example 2. Then, under the same assumptions as in Proposition 4, ALG(2) applied with $L_n \leq C\sqrt{n \log n}$*

achieves a misclassification rate $\lesssim K(\bar{\alpha}_n/\delta)^2 + \frac{1}{n}$ *with probability* $\geq 1 - p_n - n^{-1}$ *where* $\bar{\alpha}_n$ *and* p_n *are defined in Example 2.*

Proof. We first construct fake centers $(\widetilde{\xi}_\ell)$ for (x_i) as in the proof of Proposition 4 and treat them as the fake centers for y_i. By the triangle inequality,

$$\Big(\frac{1}{n}\sum_{i=1}^n \|y_i - \widetilde{\xi}_{\widetilde{z}_i}\|^2\Big)^{1/2} \leq \Big(\frac{1}{n}\sum_{i=1}^n \|x_i - \widetilde{\xi}_{\widetilde{z}_i}\|^2\Big)^{1/2} + \Big(\frac{1}{n}\sum_{i=1}^n \|w_i/\sqrt{d}\|^2\Big)^{1/2} \leq \varepsilon + \sqrt{2}\bar{\alpha}_n$$

holds with probability at least $1 - p_n - n^{-1}$. The result follows by applying Theorem 7. □

Corollary 1 shows that one can achieve consistent clustering (in the generalized sense) with $L_n = o(n)$ clusters assuming that the noise-to-signal ratio $\bar{\alpha}_n/\delta \to 0$ and $n\bar{\alpha}_n^4/\sigma_{\max}^4 \to \infty$; the same conditions needed in the sub-Gaussian mixture example. Again, this result is significant since even in the noiseless case ($\bar{\alpha}_n = 0$), consistent recovery is not possible with $L = K$ for some mixtures of curve models.

3.3.2 Mixture of Higher-order Submanifolds

A version of Proposition 4 can be stated for a higher-dimensional version of the mixture-of-curves model, if we consider generalized k-means problems with $p > 2$. We say that a random variable x has a (ρ, σ, r) *sub-Gaussian manifold distribution* if $x = \gamma(t)$ where $t \in \mathbb{R}^r$ has a sub-Gaussian distribution with parameter σ and $\gamma : \mathbb{R}^r \to \mathbb{R}^d$ is a locally ρ-Lipschitz map. i.e., $\|\gamma(t) - \gamma(s)\| \leq \rho\|t - s\|$ for all $t, s \in \mathbb{R}^r$ such that $\|t - s\| \leq \frac{1}{\rho}$.

Proposition 5. *Assume that* $(x_i)_{i=1}^n$ *are independent draws from a* K*-component mixture of sub-Gaussian manifold distributions, with parameters* (ρ, σ, r_k) *for* $k \in [K]$*, and let* $r = \max_{r\in[K]} r_k$*. Let* $\mathcal{C}_k$ *be the support of the distribution of the* k*th component. Assume that*

$$\min_{x\in\mathcal{C}_k,\, y\in\mathcal{C}_{k'}} \|x - y\| \geq \delta > 0, \quad \textit{for all } k \neq k'.$$

Then, there exist a constant $C = C(K, \delta, \rho, \sigma, r, \kappa)$ *such that any ALG(p) satisfying Assumption 2, applied with* $L_n \leq C(n^{1/p}\sqrt{\log n})^r$ *clusters recovers a perfect refinement of*

z with probability $\geq 1 - n^{-1}$. *In particular, for* $p > r$, *we have perfect refinement recovery with* $L_n = o(n)$ *clusters, with high probability.*

3.3.3 Discussion

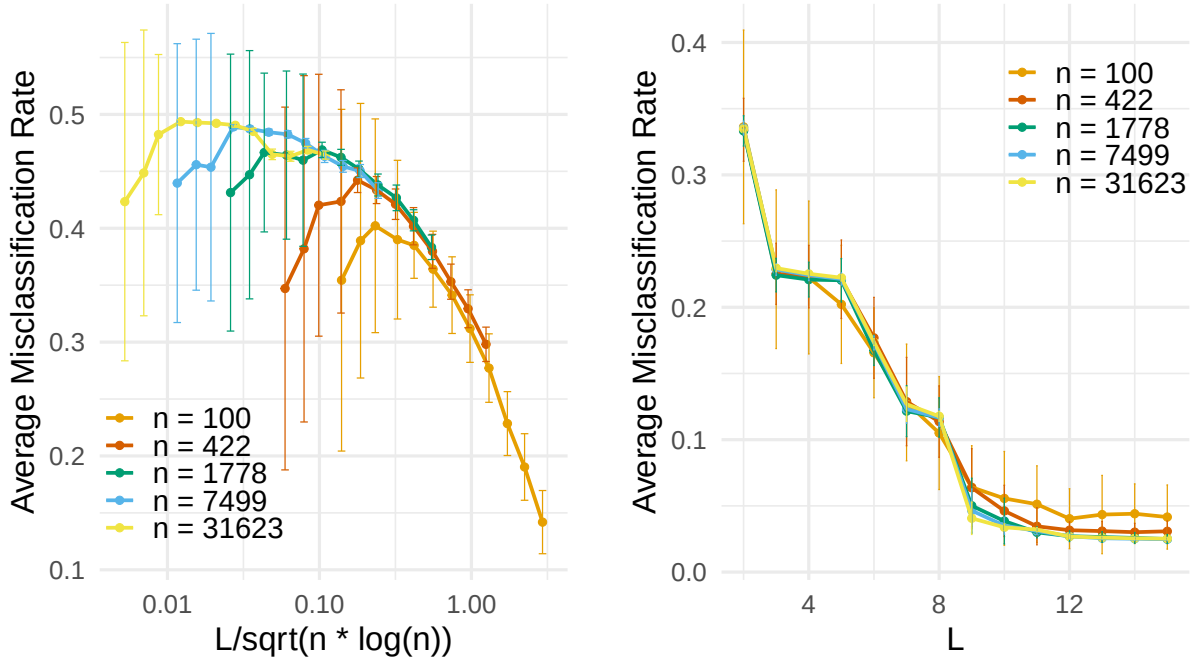
Proposition 4 and 5 show that perfect refinement for sub-gaussian mixture-of-curves model can be achieved when the number of clusters grows as $L_n = O(\sqrt{n \log n})$. To the best of our knowledge, this is the first such result in the literature, that is, an upper bound on the minimum number of clusters needed to achieve a perfect refinement of the true clusters. What remains for future investigations to determine is how tight this bound is. Empirically, we have found examples of the mixture-of-curves model for which $L_n \asymp 1$ seems to enough, but also an example where $L_n \asymp \sqrt{n \log n}$ seems to be the required scaling. Figure 3.1(a) shows a noisy circle-torus model (cf. Section 3.4.2) with $R = 10, r = 2$ and $\sigma = 1$ that demonstrates the scaling $L_n \asymp \sqrt{n \log n}$. Here, we plot the average misclassification rate over 32 repetitions vs $L_n/\sqrt{n \log n}$ for various n. The fact that these plots coincide with each other for different n suggests that there is sharp threshold $\tau_n = C_1\sqrt{n \log n}$ such that with $L_n > \tau_n$, perfect refinement recovery is possible and with $L_n < \tau_n$, impossible. Figure 3.1(b) shows an example that exhibits $L_n \asymp 1$ threshold: A line-circle model (cf. Section 3.4.1) with parameters $\delta = 4$, $\sigma = 1$ and line standard deviation $= 7$.

The fact that, empirically, there are examples for which L_n has to grow as fast as $\sqrt{n \log n}$ for a perfect refinement recovery, suggests that the result of Proposition 4 may be sharp up to constants, over the class of mixture-of-curves distributions considered.

3.4 Numerical Experiments

3.4.1 Line-circle Model

We first consider the (noiseless) line-circle model in $\mathbb{R}^3$, an example of mixture-of-curves. This model has two clusters: (1) The uniform distribution on the circumference of a circle in the xz-plane, centred at the origin, and (2) the standard Gaussian distribution on the y axis. The minimum separation δ between the two clusters is the radius of the circle.



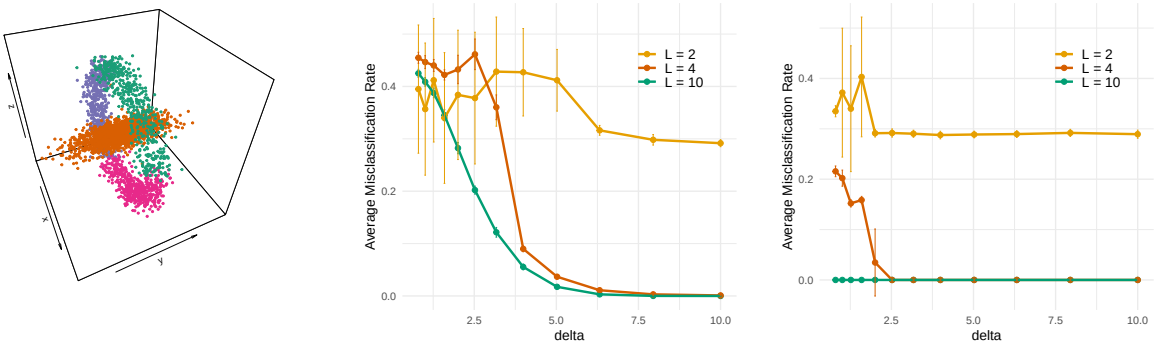
(a) circle-torus model

(b) line-circle model

Figure 3.1: Examples of mixture-of-curve models that exhibit (a) $L_n \asymp \sqrt{n \log n}$ and (b) $L_n = O(1)$ refinement recovery threshold.

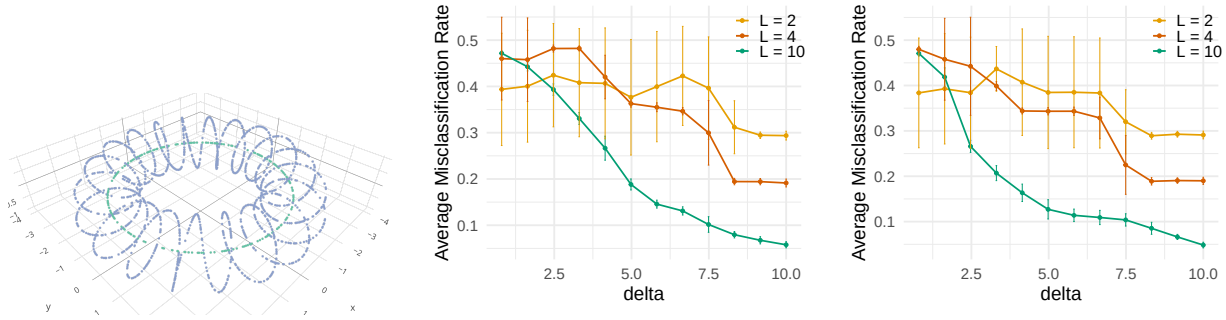
We also consider the noisy version of this model where we add $N(0, \sigma^2 I_3)$. We sample data points with equal probability from the two clusters. It is nearly impossible for the k-means to correctly label these two clusters when $L = 2$, since the centers of the two clusters coincide. Figure 3.2 shows the scatter plot of the data simulated from the noisy line-circle model, with noise level $\sigma = 0.1$, $n = 3000$ and $\delta = 3$. Here, the noise level is set low for better illustration. Different colors are used to label data points based on the output of k-means clustering with $L = 4$, and this demonstrates that each estimated cluster is a subset of a true cluster.

The result aligns with Theorem 7. Although, the true centers coincide (with the origin) when $L = 2$, by increasing L, we can create fake centers on the line and the circle to have separation close to δ and thus get a small missclassification rate. The other two panels in Figure 3.2 show the average missclassification rate over 32 repetitions versus δ, for both the noiseless and noisy ($\sigma = 1$) line-circle model. Both show that the misclassification rate is negatively associated with δ and L when $L > 2$.



(a) Noisy ($L = 4, \delta = 3$) (b) Noisy (c) Noiseless

Figure 3.2: Line-circle model: (a) Scatter plot for the noisy version. The colors show the $L = 4$ estimated clusters by k-means. (b) and (c) show the (generalized) misclassification rate versus δ, the radius of the circle, in the noisy and noiseless versions of the model.



(a) True clusters (noiseless) (b) Noisy (c) Noiseless

Figure 3.3: Circle-torus model: (a) Scatter plot for the noiseless version. Colors are used to separate two true clusters. (b) and (c) show the (generalized) misclassification rate versus δ, the radius of the circle, in the noisy and noiseless versions of the model.

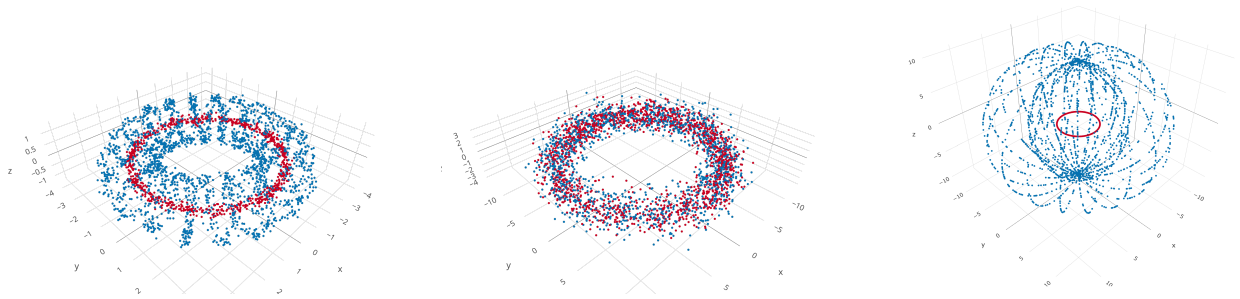
3.4.2 Circle-torus Model

The circle-torus model is a mixture of two parts: (1) The uniform distribution on the circumference of a circle in the xy-plane, at the origin, and (2) a torus built around this circle. Parametrically, these two clusters can be defined via the following equations,

$$\begin{aligned} x_1 &= R\cos(t) & & & x_2 &= (R + r\cos(mt))\cos(t) \\ y_1 &= R\sin(t) & & \text{and} & y_2 &= (R + r\cos(mt))\sin(t) \\ z_1 &= 0 & & & z_2 &= r\sin(mt). \end{aligned} \tag{3.13}$$

Here R is the radius of the circle on the plane and also the distance from the center of the tube to the center of the torus. r is the radius of the tube and it is also the minimal

distance between two clusters. We also created a noisy version by adding $N(0, \sigma^2 I_3)$ to the model. Figure 3.3 shows the geometry of the two clusters in the case $R = 3, r = 1$ and $\sigma = 0$. The other two panels in Figure 3.3 show the average missclassification rate over 32 repetitions versus $\delta := r$, for both the noiseless and noisy ($\sigma = 1$) circle-torus model. In both cases, we let $R = 3$ and very r (i.e., δ), from 0.1 to 10. In Figure 3.4, we include additional scatter plots of the circle-torus model for various settings of the parameters (R, r, σ). Figure 3.4(a) is the noisy version of Figure 3.3(a) with noise level $\sigma = 0.1$. Figure 3.4(b) shows that for sufficiently small r and high noise, the two clusters are nearly indistinguishable. Figure 3.4(c) shows the scatter plot for $R = 3$ and $r = 10$; it is an example of how the model looks like when $R < r$.

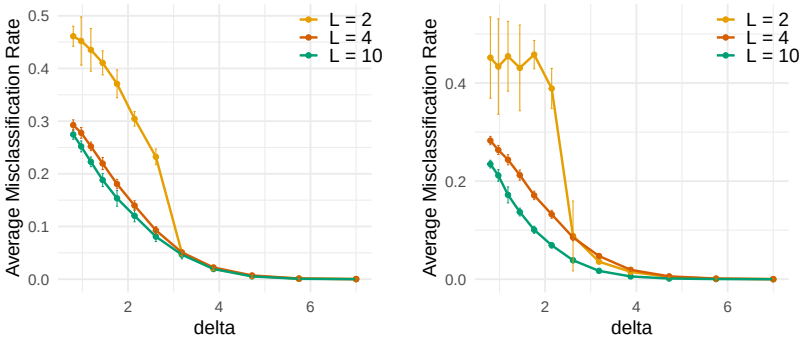


(a) $R = 3$, $r = 1$ and $\sigma = 0.1$. (b) $R = 10$, $r = 1$ and $\sigma = 1$. (c) $R = 3$, $r = 10$ and $\sigma = 0$.

Figure 3.4: Scatter plots for the circle-torus model. True clusters are distinguished by their color.

3.4.3 Line-Gaussian Model

Figure 3.5 shows the results for a line-Gaussian mixture model: $x_i = \xi^*_{z_i} + \Sigma^{1/2}_{z_i} w_i \in \mathbb{R}^2$ where $\xi^*_1 = (0, \delta)$ and $\xi^*_2 = (0, 0)$, $w_i \sim N(0, I_2)$, $\Sigma_1 = I_2$ and $\Sigma_2 = \text{diag}(\sigma^2, 0)$. Here, we have set $\sigma = 5$ and sampled $n = 3000$ data points with equal probability from the two clusters. We also consider its noisy version by setting all the zero elements in Σ_2 to 0.7, which makes the model a general Gaussian mixture. Figure 3.5 shows the average missclassification rate over 32 repetitions for different L. The results are consistent with Theorem 7 showing that as δ increases, the misclassification rate decreases.



(a) Noisy (b) Noiseless

Figure 3.5: Line-Gaussian model: The (generalized) misclassification rate versus δ, the distance of the Gaussian center to the line, in the (a) noisy and (b) noiseless versions of the model.

3.5 Proofs of Main Results

Let us first recall a fact from functional analysis. Consider the space of functions $f : [n] \to \mathcal{X}$ and let us equip $[n]$ with the uniform probability measure ν_n. Then, from the theory of Lebesgue-Bochner spaces, $\|f\|_p := (\int \|f(\omega)\|_{\mathcal{X}}^p \, d\nu_n(\omega))^{1/p}$ defines a proper norm on this function space, turning it into a Banach space $L^p(\nu_n; \mathcal{X})$. In particular, the triangle inequality holds for this norm. Note that $\|f\|_p = (\frac{1}{n}\sum_{i=1}^n \|f(i)\|_{\mathcal{X}}^p)^{1/p}$. We will frequently invoke the triangle inequality in $L^p(\nu_n, \mathcal{X})$.

Let $\mu^* := \sum_k \pi_k \delta_{\xi_k^*} = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_{z_i}^*}$ be the empirical measure associated with $\{\xi_{z_i}^*\}$. Recalling definition (3.1) of the population cost function, we have, for any $\xi \in \mathcal{X}^L$,

$$Q(\xi; \mu^*)^p = \sum_{k=1}^K \pi_k \min_{1\le\ell\le L} \|\xi_k^* - \xi_\ell\|^p = \frac{1}{n}\sum_{i=1}^n \min_{1\le\ell\le L} \|\xi_{z_i}^* - \xi_\ell\|^p. \tag{3.14}$$

We start with three lemmas that are proved in Appendix C.2:

Lemma 21. *Let ALG(p) be a k-means algorithm satisfying Assumption 2(b') and let $\widehat{\xi}$ be its output for L clusters. Furthermore, assume $(\frac{1}{n}\sum_{i=1}^n \|x_i - \xi_{z_i}^*\|^p)^{1/p} \le \varepsilon$. Then $Q(\widehat{\xi}; \mu^*) \le (1+\kappa)\varepsilon$.*

Lemma 22 (Curvature)**.** *For every $\xi \subset \mathcal{X}^L$ and $\xi^* \in \mathcal{X}^K$, with $L \ge K$,*

$$Q(\xi; \mu^*) \ge \pi_{\min}^{1/p}\Big(d_p(\xi, \xi^*) \wedge \frac{\delta}{2}\Big).$$

Lemma 23. *Assume that* $\max_{1\le i\le n}\|x_i-\xi^*_{z_i}\|\le\eta$ *and* $d_\infty(\widehat{\xi},\xi^*)\le\gamma$. *When* $L=K$, *if* $\delta>2\gamma+2\eta$, *there exists a bijective function* $\omega:[K]\to[K]$ *satisfying* $\omega(\widehat{z}_i)=z_i,\ \forall i\in[n]$. *When* $L>K$, *if* $\delta>2\gamma+4\eta$, *there exists a surjective function* $\omega:[L]\to[K]$ *satisfying* $\omega(\widehat{z}_i)=z_i,\ \forall i\in[n]$.

Proof of Theorem 5. As $(\frac{1}{n}\sum_{i=1}^n\|x_i-\xi^*_{z_i}\|^p)^{1/p}\le\max_{1\le i\le n}\|x_i-\xi^*_{z_i}\|\le\eta$, combining Lemma 21 and 22, we have

$$\Big(d_p(\widehat{\xi},\xi^*)\wedge\frac{\delta}{2}\Big)\le\frac{Q(\widehat{\xi},\mu^*)}{\pi_{\min}^{1/p}}\le\frac{(1+\kappa)\eta}{\pi_{\min}^{1/p}}.$$

By the condition on δ in (3.4), we have $\delta/2>(1+\kappa)\eta/\pi_{\min}^{1/p}$. Then, $d_\infty(\widehat{\xi},\xi^*)\le d_p(\widehat{\xi},\xi^*)\le\gamma:=(1+\kappa)\eta/\pi_{\min}^{1/p}$, which also makes the assumption in Lemma 23 that $\delta>2\gamma+4\eta$ valid. Finally, the result follows from Lemma 23. □

Proof of Theorem 6. The argument is similar to one that has appeared in recent literature [LR15b; Jin15; ZA19]. From the proof of Lemma 21 in Appendix C.2, we have

$$Q(\widehat{\xi};\mu^*)\le\Big(\frac{1}{n}\sum_{i=1}^n\|\xi^*_{z_i}-\widehat{\xi}_{\widehat{z}_i}\|^p\Big)^{1/p}\le(1+\kappa)\varepsilon.$$

By Lemma 22

$$\Big(d_p(\widehat{\xi},\xi^*)\wedge\frac{\delta}{2}\Big)\le\frac{Q(\widehat{\xi},\mu^*)}{\pi_{\min}^{1/p}}\le\frac{(1+\kappa)\varepsilon}{\pi_{\min}^{1/p}}.$$

By the separation assumption in (3.5), $\delta/2>(1+\kappa)\varepsilon/\pi_{\min}^{1/p}$. Hence $d_p(\widehat{\xi},\xi^*)\le(1+\kappa)\varepsilon/\pi_{\min}^{1/p}$. Let $\mathcal{C}_k=\{i:z_i=k\}$, $|\mathcal{C}_k|=n_k$, and set $T_k:=\{i\in\mathcal{C}_k:\|\xi^*_{z_i}-\widehat{\xi}_{\widehat{z}_i}\|\le\delta/c\}$. Letting $S_k=\mathcal{C}_k\setminus T_k$, we obtain

$$|S_k|\delta^p/c^p<\sum_{i\in S_k}\|\xi^*_{z_i}-\widehat{\xi}_{\widehat{z}_i}\|^p\le\sum_{i=1}^n\|\xi^*_{z_i}-\widehat{\xi}_{\widehat{z}_i}\|^p\le n(1+\kappa)^p\varepsilon^p.$$

Therefore,

$$\frac{|S_k|}{n_k}<\frac{n(1+\kappa)^pc^p\varepsilon^p}{n_k\delta^p}\le1.$$

The last inequality is by assumption $\delta > (1+\kappa)c\varepsilon/\pi_{\min}^{1/p}$. Hence, T_k is not empty. Furthermore, we argue that if $i \in T_k$ and $j \in T_\ell$ for $k \neq \ell$, i.e. $z_i \neq z_j$, then $\widehat{z}_i \neq \widehat{z}_j$. Assume otherwise, that is, $\widehat{z}_i = \widehat{z}_j$. Then

$$\|\xi_k^* - \xi_\ell^*\| \;\leq\; \|\xi_k^* - \widehat{\xi}_{\widehat{z}_i}\| + \|\xi_\ell^* - \widehat{\xi}_{\widehat{z}_j}\| \;\leq\; 2\delta/c \;<\; \delta$$

causing a contradiction.

Let $\mathcal{L}_k := \{\widehat{z}_i : i \in T_k\}$ and $\mathcal{L} = \bigcup_{k=1}^K \mathcal{L}_k$. Define a function $\omega : \mathcal{L} \to [K]$ by setting $\omega(\ell) = k$ for all $\ell \in \mathcal{L}_k$ and $k \in [K]$. By the property of $\{T_k\}$ shown above, $\mathcal{L}_k, k \in [K]$ are disjoint and nonempty sets. This implies that ω is well-defined and surjective. Extend ω to a surjective $\omega : [L] \to [K]$ by arbitrarily defining it for $[L] \setminus \mathcal{L}$. Note that $\widehat{z}_i \in \mathcal{L}_k$ implies $z_i = k$. It follows that $\omega(\widehat{z}_i) = z_i$ for all $\widehat{z}_i \in \mathcal{L}$, and

$$\frac{1}{n}\sum_{i=1}^n 1\{z_i \neq \omega(\widehat{z}_i)\} \leq \frac{n - |\mathcal{L}|}{n} = \sum_{k=1}^K \frac{|S_k|}{n} < \frac{K(1+\kappa)^p c^p \varepsilon^p}{\delta^p}.$$

The result follows. □

Proof of Theorem 7. By assumption, κ-approximation holds for both K and L clusters. Then,

$$\widehat{Q}(\widehat{\xi}) \leq \kappa\, \widehat{Q}_{\min}^{(L)}, \quad \text{where} \quad \widehat{Q}_{\min}^{(L)} := \min_{\xi \in \mathcal{X}^L} \widehat{Q}(\xi).$$

Since $\widehat{Q}_{\min}^{(L)} \leq (\frac{1}{n}\sum_{i=1}^n \|x_i - \widetilde{\xi}_{\widetilde{z}_i}\|^p)^{1/p} \leq \varepsilon$, by the triangle inequality in $L^p(\nu_n, \mathcal{X})$,

$$\Big(\frac{1}{n}\sum_{i=1}^n \|\widetilde{\xi}_{\widetilde{z}_i} - \widehat{\xi}_{\widehat{z}_i}\|^p\Big)^{1/p} \leq \Big(\frac{1}{n}\sum_{i=1}^n \|x_i - \widetilde{\xi}_{\widetilde{z}_i}\|^p\Big)^{1/p} + \Big(\frac{1}{n}\sum_{i=1}^n \|x_i - \widehat{\xi}_{\widehat{z}_i}\|^p\Big)^{1/p} \leq (1+\kappa)\varepsilon.$$

Let $T_k := \{i \in \mathcal{C}_k : \|\widetilde{\xi}_{\widetilde{z}_i} - \widehat{\xi}_{\widehat{z}_i}\| \leq \delta/c\}$ and $S_k = \mathcal{C}_k \setminus T_k$. Then,

$$|S_k|\delta^p/c^p < \sum_{i \in S_k} \|\widetilde{\xi}_{\widetilde{z}_i} - \widehat{\xi}_{\widehat{z}_i}\|^p \leq \sum_{i=1}^n \|\widetilde{\xi}_{\widetilde{z}_i} - \widehat{\xi}_{\widehat{z}_i}\|^p \leq n(1+\kappa)^p \varepsilon^p.$$

Therefore,

$$\frac{|S_k|}{n_k} < \frac{n(1+\kappa)^p c^p \varepsilon^p}{n_k \delta^p} \leq 1$$

The last inequality is by assumption $\delta \geq (1+\kappa)c\varepsilon/\pi_{\min}^{1/p}$. Hence T_k is not empty. Next we argue that if $i \in T_k$, $j \in T_\ell$ for $k \neq \ell$, i.e. $z_i \neq z_j$, then $\widehat{z}_i \neq \widehat{z}_j$. Assume otherwise, that is $\widehat{z}_i = \widehat{z}_j$. Since $\widetilde{z}$ is a refinement of z, $z_i \neq z_j$ implies $\widetilde{z}_i \neq \widetilde{z}_j$ and $\widetilde{\omega}(\widetilde{z}_i) \neq \widetilde{\omega}(\widetilde{z}_j)$. By the triangle inequality,

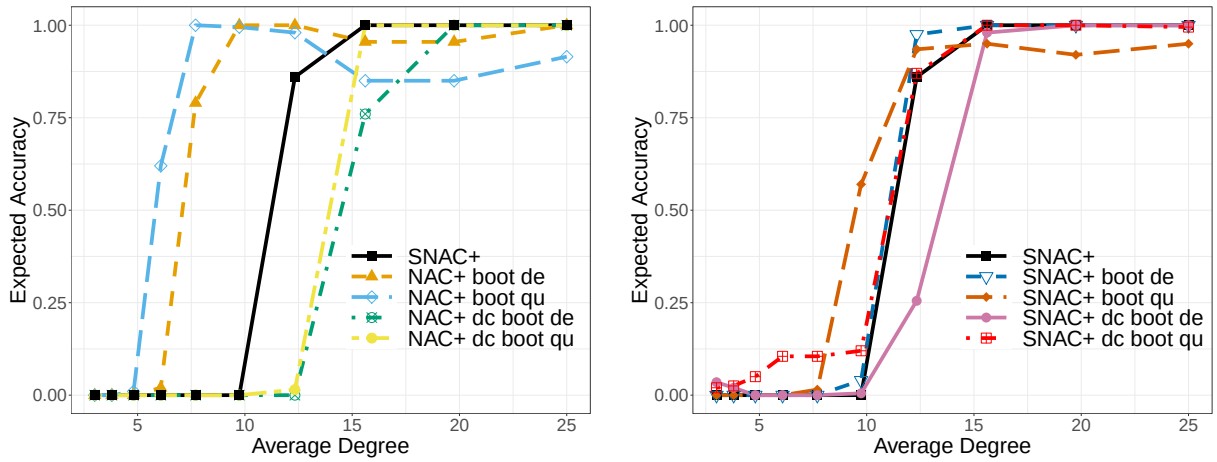
$$\|\widetilde{\xi}_{\widetilde{z}_i} - \widetilde{\xi}_{\widetilde{z}_j}\| \leq \|\widetilde{\xi}_{\widetilde{z}_i} - \widehat{\xi}_{\widehat{z}_i}\| + \|\widetilde{\xi}_{\widetilde{z}_j} - \widehat{\xi}_{\widehat{z}_j}\| \leq 2\delta/c < \delta$$

causing a contradiction. The rest of the proof follows that of Theorem 6. □

Chapter A

Extra Simulations in Chapter 2

A.1 Bootstrap Comparison



(a) Variants of NAC+ (b) Variants of SNAC+.

Figure A.1: Comparing different bootstrap approaches using expected accuracy of selecting the true number of communities versus expected average degree of the network. SNAC+ is shown in both plots as a benchmark. Details of each method in the legend is explained in the text.

In Figure A.1, we compare different approaches for bootstrapping and rejecting in SNAC+ and FNAC+. As we discussed in Section 2.2.4, for a significance level α, there are four versions of bootstrapping and rejecting, and we use different suffixes below with SNAC+ and FNAC+ to represent them in Figure A.1.

1. "boot de": bootstrapping using SBM samples and obtaining their mean and standard deviation to standardize the original statistic and rejecting the null hypothesis with α critical threshold from the standard normal;

2. "boot qu": bootstrapping with SBM samples and using their α-quantile as the rejection threshold;

3. "dc boot de": same as "boot de" except bootstrapping with DCSBM instead;

4. "dc boot qu": same as "boot qu" except bootstrapping with DCSBM instead.

All four versions are applied to FNAC+ (left plot) and SNAC+ (right plot). Both include the plain SNAC+ as the comparison baseline. The simulation data follows a DCSBM with $n = 5000$, $K = 4$, $\theta_i \sim \text{Pareto}(3/4, 4)$, connectivity matrix as B_1 defined in Section 2.4.1 and balanced community sizes. In both SNAC+ and FNAC+, the "boot de" approach has the most stable performance and that is why we use it in simulations of Section 2.4.1.

A.2 Model Selection

Figure A.2 shows model selection accuracy with four variants of DCSBM parameters. All plots has DCSBM with parameters $n = 5000$, $\theta_i \sim Pareto(3/4, 4)$. The top row is generated with a generalized version of B_1 as the connectivity matrix, given by

$$B_3 \propto (1 - \beta)\,\text{diag}(w) + \beta \mathbf{1}\mathbf{1}^T.$$

It is evident that B_1 is a special case of B_3 where w is the an all-ones vector. Here, we set $w = (1, 2, 3, 1)$ under $K = 4$ and the top left plot shows the case where the DCSBM has unbalanced community sizes proportional to $(1, 1, 2, 3)$ and the right plot shows balanced community sizes. The bottom row is generated based on the planted partition model, but with different community sizes and out-in-ratio than that in Figure 2.2. The bottom left side has unbalanced community sizes proportional to $(1, 2, 3, 4)$ and out-in-ratio $\beta = 0.2$ and the right side has balanced community sizes and out-in-ratio $\beta = 0.3$. All methods have lower accuracy in the unbalanced setting except for the AS. BH is affected the most while FNAC+ the least. The robustness of FNAC+ could be because its performance mainly relies on the full version of ρ and unbalanced sizes retain its rows' distinction. However, the SNAC+ is still affected by the unbalanced community sizes because of the

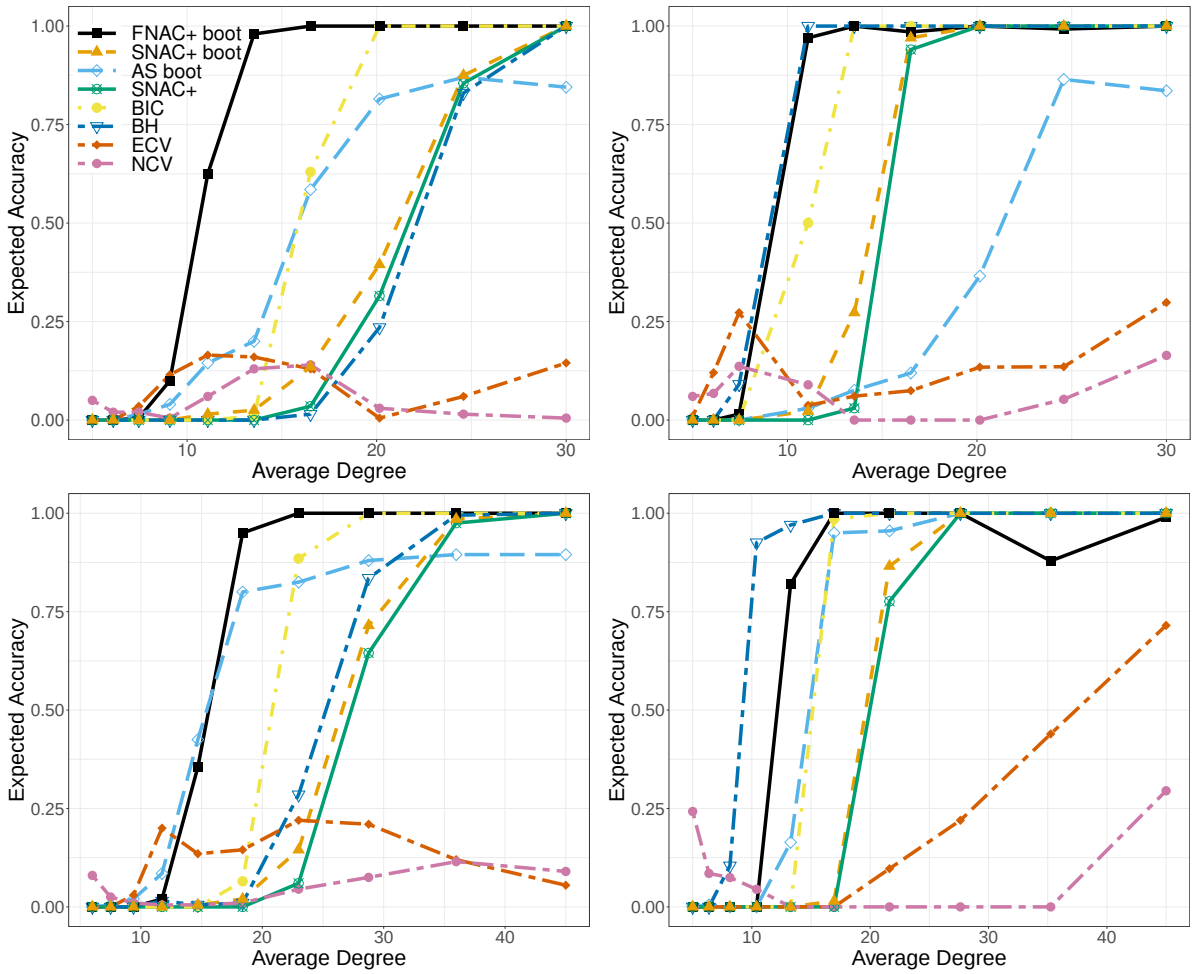


Figure A.2: Expected accuracy of selecting the true number of communities versus expected average degree of the network. The data follows a DCSBM with $n = 5000$, $\theta_i \sim \text{Pareto}(3/4, 4)$. The first row is generated with $\beta = 0.2$, connectivity matrix B_3. The top left plot has unbalanced community sizes proportional to $(1, 1, 2, 3)$ and the right plot has balanced community sizes. The second row is generated with connectivity matrix B_1. The bottom left plot has community sizes proportional to $(1, 2, 3, 4)$ and out-in-ratio $\beta = 0.2$. The bottom right plot has balanced community sizes and out-in-ratio $\beta = 0.3$.

increased difficulty in recovering the correct labels and the increased variance in ρ due to subsampling.

A.3 ROC Curves

We consider additional testing with $H_0 : K = 4$ vs. $H_a : K = 3$. Other DCSBM simulating parameters are the same as in Section 2.4.1. Figure A.3 shows ROC curves for the null being DCSBM with $K = 4$ and two alternatives: a DCSBM with $K = 3$ (left) and a DCLVM with $K = 3$ (right). In addition, we also have $n = 2000$ for the upper row and

$n = 10000$ for the lower. Similar to Figure 2.3, the performance of the tests get better as n increases. FNAC and AS tests are nearly perfect (achieve 100% recovery for very small type I error) when the alternative is DCLVM. The LR test is almost perfect in distinguishing two DCSBMs but has very poor power when the alternative is DCLVM.

We also include the test $H_0 : K = 4$, DCSBM vs. $H_a : K = 4$, DCLVM with similar parameters in Figure A.4. It shows that FNAC tests are still able to reject when the true model is a DCLVM with the same number of communties as the DCSBM. Note that we have excluded the LR test in this case, since it is the likelihood ratio of two fitted DCSBMs with different number of communities, but here we have models with the same number of communities.

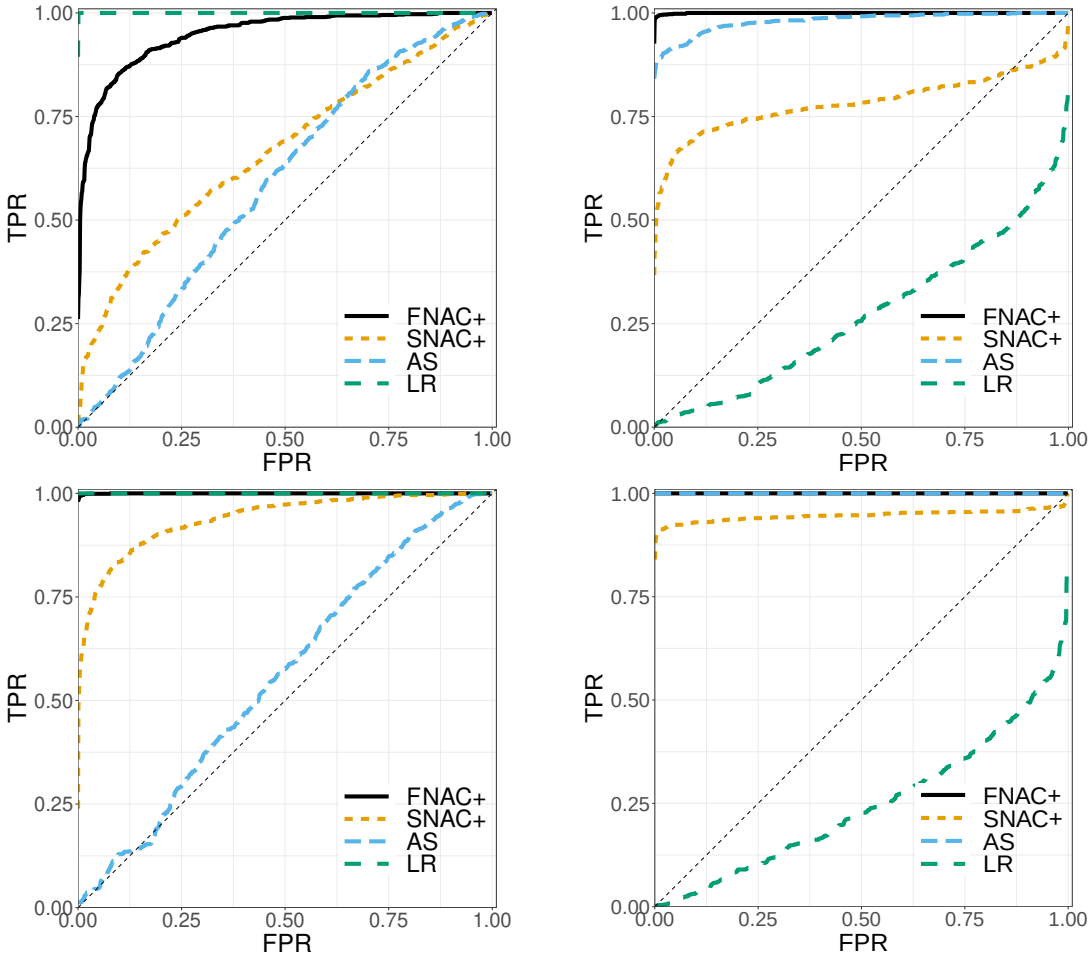


Figure A.3: ROC plots for testing 4- versus 3-community models. Top and bottom rows correspond to $n = 2000$ and $n = 10000$, respectively. Left and right columns correspond to the DCSBM and DCLVM alternatives, respectively.

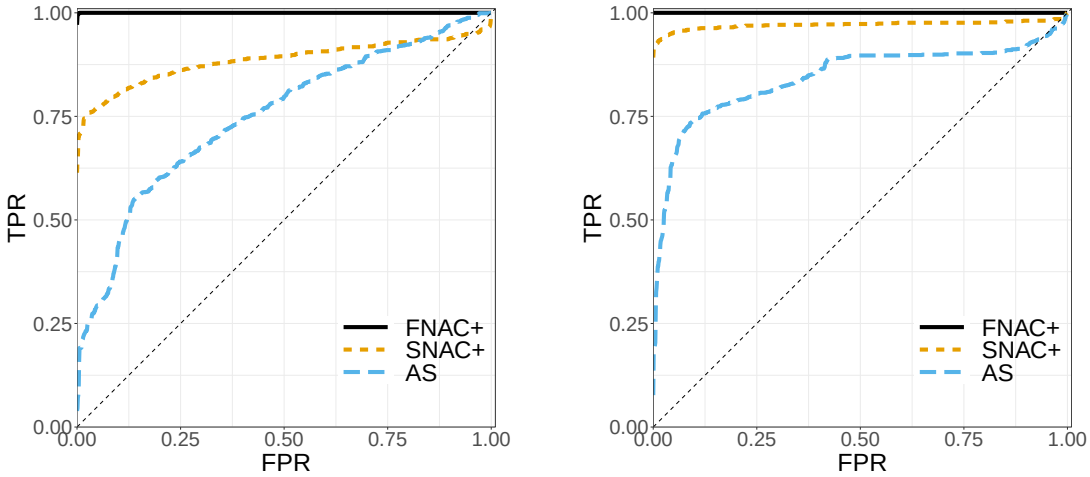


Figure A.4: ROC plots for testing $H_0 : K = 4$ DCSBM vs. $H_a : K = 4$ DCLVM. Left has $n = 2000$ and right $n = 10000$.

A.4 Extra Real Network Examples

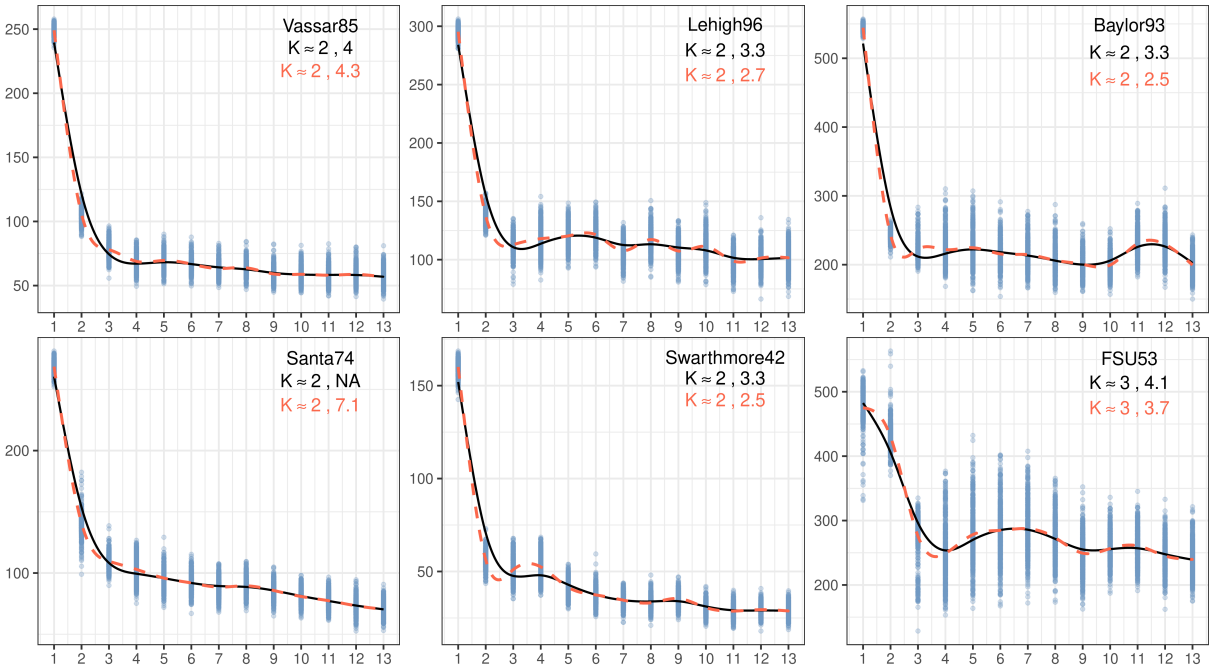


Figure A.5: More examples on community profile plots from FB-100. They show a single elbow/dip pattern.

Figures A.5 and A.6 provide more profile plots for the networks in the FB-100 dataset. The former collection shows profile plots with one-elbow pattern and the latter shows higher variability of SNAC+ statistics with multi-stage elbows/dips. We also point out that the Caltech network in Figure A.6 is the only FB-100 network for which SNAC+ drops to nearly zero (at $K = 10$) within the range of candidate K. However, the statistic

continues to decrease afterwards and does not show any dips/elbows like others. This suggests that although we cannot reject the null hypothesis of a DCSBM (with $K = 10$) in this case, a DCSBM still might not be a good model for the network. That we cannot reject the null is most likely due to the small community sizes we get with $K = 10$, leading to an insufficient signal.

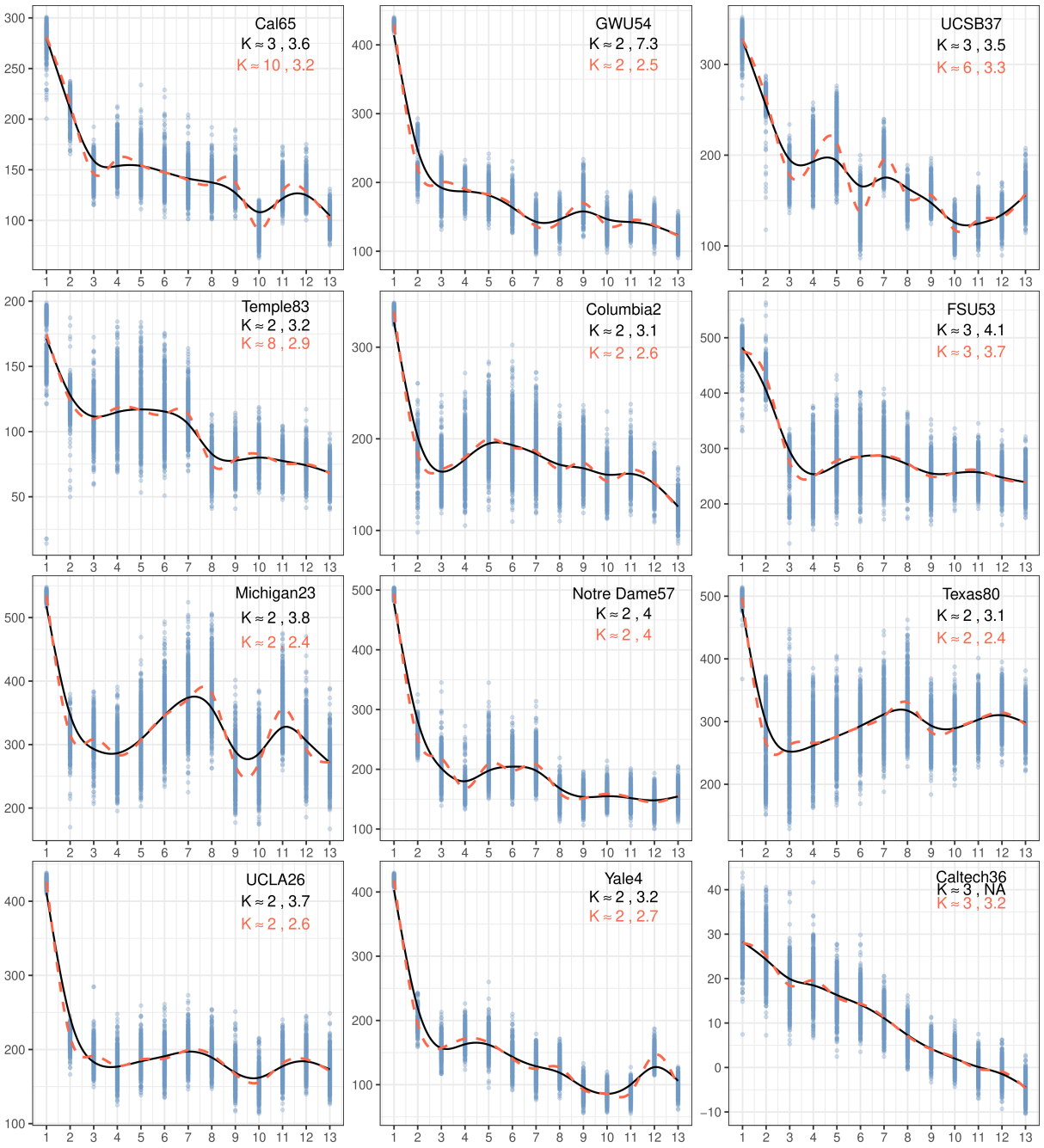


Figure A.6: More examples on community profile plots from FB-100. They show a multiple elbows/dips pattern.

Figure A.7 shows the profile plot for the political blog network and its community

structure. In the profile plot, the elbow point identified by the largest second derivative is at $K = 2$, matching the presumed ground truth number of communities in this case. The colored community structure also shows that the fitted two-community model gives a reasonable split of the nodes.

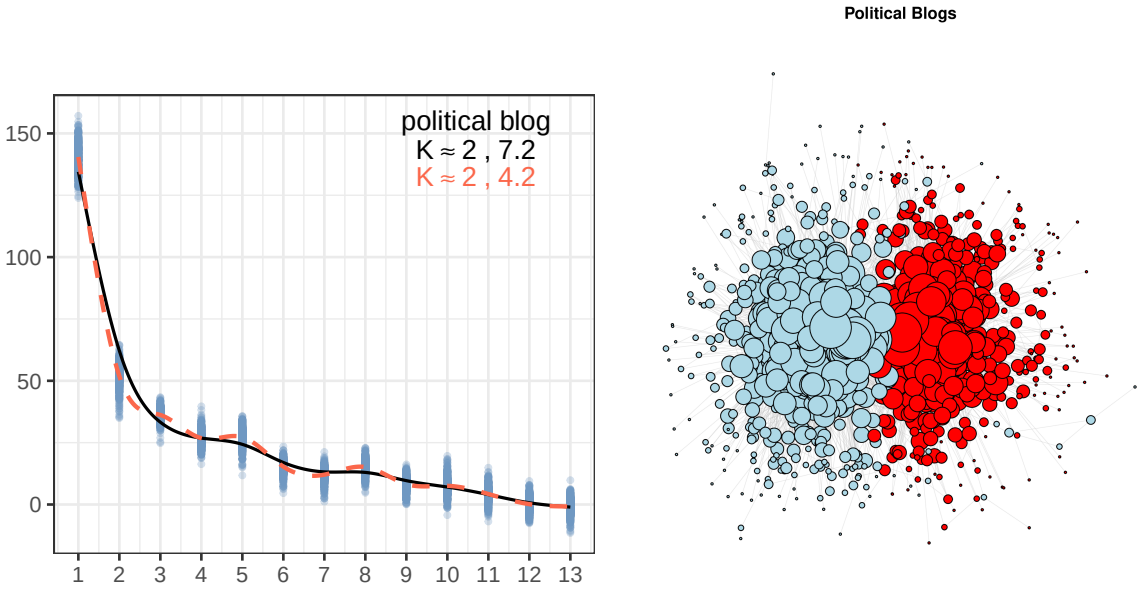


Figure A.7: Political blog network: profile plot (left) and community structure (right)

Chapter B

Remaining Proofs in Chapter 2

B.1 Lemmas in the Proof of Theorem 1

B.1.1 Lemmas in the Proof of Propostion 1

We first derive some useful relations between the moments and cumulants of a random variable that are used in the proofs of Lemma 1 and 2. In particular, for Lemma 2, we use the following observation: The central moments of sums of i.i.d. random variables grow "slowly". To develop an intuition for this observation, recall that

$$\mu_4(X) = \kappa_4(X) + 3\kappa_2^2(X) \tag{B.1}$$

where X is any random variable, $\mu_r(X)$ is its rth order central moment, and $\kappa_r(X)$ is the corresponding rth order cumulant. Assume that X can be written as a sum of i.i.d. variables $\{Y_1, \ldots, Y_n\}$, that is, $X = \sum_{i=1}^n Y_i$. Cumulants are additive over independent sums, hence $\kappa_r(X) = \sum_{i=1}^n \kappa_r(Y_i) = n\kappa_r(Y_1)$. It follows that

$$\mu_4(X) = n\kappa_4(Y_1) + 3n^2\kappa_2^2(Y_1) = O(n^2) \tag{B.2}$$

assuming $\kappa_r(Y_1) = O(1)$. In other words, $\mu_4(X)$ scales at half the rate of the worst-case scaling of the 4th power of a sum of n deterministic terms (i.e., $O(n^2)$ instead of $O(n^4)$). By using $\kappa_4(Y_1) = \mu_4(Y_1) - 3\kappa_2^2(Y_1)$ and $\kappa_2(Y_1) = \mu_2(Y_1)$, we can express the constants

in (B.2) in terms of the central moments of Y_1,

$$\mu_4(X) = n\mu_4(Y_1) + 3n(n-1)\mu_2^2(Y_1) \sim 3\mu_2^2(Y_1)n^2. \tag{B.3}$$

A similar idea holds for higher-order central moments, an example of which is Lemma 2.

Proof of Lemma 1. For the expectation, we note that $\mathbb{E}(X_\ell - dp_\ell)^2 = p_\ell(1-p_\ell)$, hence $\mathbb{E}\psi(X_i, dp_\ell) = 1 - p_\ell$ and the result follows since $\sum_\ell (1-p_\ell) = L-1$. We now turn to the variance. Let $\widetilde{X} = X - dp = \sum_{i=1}^d \widetilde{U}_i$, where $\widetilde{U}_i = U_i - p$ and $U_i \sim \text{Mult}(1, p)$, independently. We have

$$d^2\mathbb{E}Y^2 = \sum_{\ell=1}^L \frac{\mathbb{E}\widetilde{X}_\ell^4}{p_\ell^2} + \sum_{\ell \neq \ell'}^L \frac{\mathbb{E}\widetilde{X}_\ell^2 \widetilde{X}_{\ell'}^2}{p_\ell p_{\ell'}}.$$

Noting that $\widetilde{X}_\ell = \sum_i \widetilde{U}_{i\ell}$, we obtain

$$\mathbb{E}(\widetilde{X}_\ell^2 \widetilde{X}_{\ell'}^2) = \mathbb{E}\Big(\sum_{i_1,i_2} \widetilde{U}_{i_1\ell}\widetilde{U}_{i_2\ell'}\Big)^2 = \sum_{i_1,\,i_2,\,i_3,\,i_4} \mathbb{E}[\widetilde{U}_{i_1\ell}\widetilde{U}_{i_2\ell}\widetilde{U}_{i_3\ell'}\widetilde{U}_{i_4\ell'}],$$

where all four indices running from 1 to d. We can categorize the general term $\mathbb{E}[\widetilde{U}_{i_1\ell}\widetilde{U}_{i_2\ell}\widetilde{U}_{i_3\ell'}\widetilde{U}_{i_4\ell'}]$ based on how many different values i_1, i_2, i_3 and i_4 take. If i_1, i_2, i_3 and i_4 take 3 or 4 different values, the term is zero by independence. The remaining three cases are summarized below:

$$\mathbb{E}[\widetilde{U}_{i_1\ell}\widetilde{U}_{i_2\ell}\widetilde{U}_{i_3\ell'}\widetilde{U}_{i_4\ell'}] = \begin{cases} \mathbb{E}[\widetilde{U}_{1\ell}^2] \cdot \mathbb{E}[\widetilde{U}_{1\ell'}^2], & i_1 = i_2 \neq i_3 = i_4, \\ \big(\mathbb{E}[\widetilde{U}_{1\ell}\widetilde{U}_{1\ell'}]\big)^2, & i_1 = i_3 \neq i_2 = i_4 \\ & \quad \text{or } \ i_1 = i_4 \neq i_2 = i_3, \\ \mathbb{E}[\widetilde{U}_{1\ell}^2\widetilde{U}_{1\ell'}^2], & i_1 = i_3 = i_2 = i_4, \end{cases}$$

which simplifies to

$$\mathbb{E}[\widetilde{U}_{i_1\ell}\widetilde{U}_{i_2\ell}\widetilde{U}_{i_3\ell'}\widetilde{U}_{i_4\ell'}] = \begin{cases} p_\ell(1-p_\ell)p_{\ell'}(1-p_{\ell'}), & i_1 = i_2 \neq i_3 = i_4, \\ p_\ell^2 p_{\ell'}^2, & i_1 = i_3 \neq i_2 = i_4 \\ & \quad \text{or } \ i_1 = i_4 \neq i_2 = i_3, \\ p_\ell p_{\ell'}(p_\ell + p_{\ell'} - 3p_\ell p_{\ell'}), & i_1 = i_3 = i_2 = i_4. \end{cases}$$

The first two cases follow easily from independence. $\mathbb{E}[\widetilde{U}_{1\ell}^2] = \mathrm{var}(U_{1\ell}) = p_\ell(1-p_\ell)$ and $\mathbb{E}[\widetilde{U}_{1\ell}\widetilde{U}_{1\ell'}] = \mathrm{cov}(U_{1\ell}, U_{1,\ell'}) = -p_\ell p_{\ell'}$. The third case follows, after some algebra, from the following observation:

$$(\widetilde{U}_{1\ell}, \widetilde{U}_{1\ell'}) = \begin{cases} (-p_\ell, -p_{\ell'}) & \text{w.p. } 1-(p_\ell + p_{\ell'}) \\ (-p_\ell, 1-p_{\ell'}) & \text{w.p. } p_{\ell'} \\ (1-p_\ell, -p_{\ell'}) & \text{w.p. } p_\ell \end{cases}.$$

To sum up, for $\ell \neq \ell'$, we have

$$\begin{aligned} \mathbb{E}[\widetilde{X}_\ell^2 \widetilde{X}_{\ell'}^2] &= (d^2-d)p_\ell p_{\ell'}[(1-p_\ell)(1-p_{\ell'}) + 2p_\ell p_{\ell'}] + dp_\ell p_{\ell'}(p_\ell + p_{\ell'} - 3p_\ell p_{\ell'}) \\ &= dp_\ell p_{\ell'}\big[(d-1) + (2-d)(p_\ell + p_{\ell'}) + (3d-6)p_\ell p_{\ell'})\big]. \end{aligned}$$

Let $\alpha := \sum_\ell p_\ell^2$. Using $\sum_{\ell \neq \ell'} p_\ell p_{\ell'} = 1-\alpha$ and $\sum_{\ell \neq \ell'} p_\ell = \sum_{\ell \neq \ell'} p_{\ell'} = L-1$, we have

$$\frac{1}{d}\sum_{\ell \neq \ell'} \frac{\mathbb{E}[\widetilde{X}_\ell^2 \widetilde{X}_{\ell'}^2]}{p_\ell p_{\ell'}} = (d-1)(L^2-L) + 2(2-d)(L-1) + (3d-6)(1-\alpha),$$

for $\ell \neq \ell'$. Next, we consider the case $\ell = \ell'$. Let κ_n and μ_n denote nth order cumulants

and central moments of $\widetilde{U}_{1\ell}$. By (B.3),

$$\begin{aligned}\mathbb{E}[\widetilde{X}_\ell^4] &= d\mu_4 + 3d(d-1)\mu_2^2 \\ &= d[p_\ell(1-p_\ell)^4 + p_\ell^4(1-p_\ell)] + 3d(d-1)p_\ell^2(1-p_\ell)^2 \\ &= dp_\ell^2[1/p_\ell + (3d-7) + (12-6d)p_\ell + (3d-6)p_\ell^2].\end{aligned}$$

We obtain

$$\frac{1}{d}\sum_{\ell=1}^{L}\frac{\mathbb{E}\widetilde{X}_\ell^4}{p_\ell^2} = \frac{L}{h(p)} + L(3d-7) + (12-6d) + (3d-6)\alpha.$$

Putting the pieces together, we have

$$d\,\mathbb{E}Y^2 = d(L^2-1) + \frac{L}{h(p)} - L(L+2) + 2.$$

Combining with $\mathrm{var}(Y) = \mathbb{E}Y^2 - (L-1)^2$ and some algebra finishes the proof. □

Proof of Lemma 2. Let $\{W_i'\}$ be an independent copy of $\{W_i\}$, and let $X_n' = \sum_{i=1}^n W_i'$. The function $x \mapsto |x|^3$ is convex on $\mathbb{R}$. Applying Jensen's inequality with respect to X' and the Cauchy-Schwartz inequality in probability,

$$\mathbb{E}\big|X_n^2 - \mathbb{E}X_n^2\big|^3 \le \mathbb{E}\big|X_n^2 - (X_n')^2\big|^3 \le \big[\mathbb{E}\big|X_n + X_n'\big|^6\big]^{1/2}\big[\mathbb{E}\big|X_n - X_n'\big|^6\big]^{1/2}.$$

For a random variable U, write $\kappa_i(U)$ for its ith cumulant Then,

$$\begin{aligned}\kappa_i(X_n + X_n') &= \kappa_i(X_n) + \kappa_i(X_n') = 2n\kappa_i(W_1), \\ \kappa_i(X_n - X_n') &= \kappa_i(X_n) + (-1)^i\kappa_i(X_n') = 2n\kappa_i(W_1)\cdot 1\{i \text{ is even}\}.\end{aligned}$$

Recall that the 6th central moment μ_6 of any random variable can be written in terms of its cumulants $\{\kappa_i\}$ as follows: $\mu_6 = \kappa_6 + 15\kappa_4\kappa_2 + 10\kappa_3^2 + 15\kappa_2^3$. Writing $\widetilde{\kappa}_i = \kappa_i(W_1)$,

and applying this relation to $X_n + X_n'$ and $X_n - X_n'$, we have

$$\mathbb{E}\big|X_n + X_n'\big|^6 = \mu_6(X_n + X_n') = 2n\widetilde{\kappa}_6 + 60n^2\widetilde{\kappa}_4\widetilde{\kappa}_2 + 40n^2\widetilde{\kappa}_3^2 + 120n^3\widetilde{\kappa}_2^3,$$
$$\mathbb{E}\big|X_n - X_n'\big|^6 = \mu_6(X_n - X_n') = 2n\widetilde{\kappa}_6 + 60n^2\widetilde{\kappa}_4\widetilde{\kappa}_2 + 120n^3\widetilde{\kappa}_2^3.$$

Let $C_{W_1} = 2|\widetilde{\kappa}_6| + 60|\widetilde{\kappa}_4|\widetilde{\kappa}_2 + 40\widetilde{\kappa}_3^2 + 120\widetilde{\kappa}_2^3$. Then, $\mathbb{E}\big|X_n \pm X_n'\big|^6 \le C_{W_1} n^3$ and the result follows.

For the case of where $W_1 = \alpha(Z - p)$ where $Z \sim \mathrm{Ber}(p)$, let $\kappa_i = \kappa_i(Z)$ and note that $\widetilde{\kappa}_i = \alpha^i \kappa_i$. It follows that

$$C_{W_1} = \alpha^6\big(2|\kappa_6| + 60|\kappa_4|\kappa_2 + 40\kappa_3^2 + 120\kappa_2^3\big).$$

Next, we have $\kappa_2 = p(1-p)$, $\kappa_3 = \kappa_2(1-2p)$, $\kappa_4 = \kappa_2(1-6\kappa_2)$, $\kappa_6 = \kappa_2\big(1-30\kappa_2(1-4\kappa_2)\big)$. We have $\kappa_2 \in [0, 1/4]$, hence $\kappa_3/\kappa_2 \in [-1, 1]$, $\kappa_4/\kappa_2 \in [-\frac{1}{2}, 1]$ and $\kappa_6/\kappa_2 \in [-\frac{7}{8}, 1]$. It follows that $|\kappa_r| \le \kappa_2 \le 1/4$ for all $r = 3, 4, 6$. Then,

$$C_{W_1}/\alpha^6 \le 2\kappa_2 + 15\kappa_2 + 10\kappa_2 + 7.5\kappa_2 = 34.5\kappa_2$$

and the proof is complete. □

B.1.2 Lemmas in the Proof of Proposition 2

Proof of Lemma 3. We have $\sum_i d_i(x_i - y - v)^2 = \sum_i d_i(x_i - y)^2 - 2vR + d_+v^2$. Hence,

$$\sum_i d_i\psi(x_i, y + v) = \frac{\sum_i d_i(x_i - y)^2}{y + v} - \frac{2v}{y + v}R + \frac{v^2}{y + v}d_+.$$

It follows, after some algebra, that

$$\sum_i d_i[\psi(x_i, y + v) - \psi(x_i, y)] = -\frac{v}{y + v}\Big[\sum_i d_i\psi(x_i, y) + 2R - vd_+\Big].$$

We obtain

$$|G(v) - G(0)| \leq \frac{|v|}{|y+v|}\big[G(0) + 2|R| + |v|d_+\big].$$

Applying the inequality $|a|/|1+a| \leq 2|a|$ which holds for any $|a| \leq 1/2$, with $a = v/y$ finishes the proof. □

To prove Lemma 4, we first need an auxiliary lemma.

Lemma 24. *Let $T = \beta S + \alpha$ where S is random variable and $\beta, \alpha \in \mathbb{R}$ are constants, and let $Z \sim N(0,1)$. Then,*

$$d_K\big(T, Z\big) \leq d_K\big(S, Z\big) + \frac{|\beta - 1|}{\sqrt{2\pi e}\min\{|\beta|, 1\}} + \frac{|\alpha|}{\sqrt{2\pi}}.$$

Proof of Lemma 24. We have

$$\begin{aligned} d_K\big(T, Z\big)\big) &= \sup_{t\in\mathbb{R}} |\mathbb{P}(\beta S + \alpha \leq t) - \Phi(t)| \\ &= \sup_{t\in\mathbb{R}} |\mathbb{P}(S \leq t) - \Phi(\beta t + \alpha)| \\ &\leq \sup_{t\in\mathbb{R}} \big(|\mathbb{P}(S \leq t) - \Phi(t)| + |\Phi(t) - \Phi(\beta t + \alpha)|\big) \\ &= d_K\big(S, Z\big) + \sup_{t\in\mathbb{R}} |\Phi(t) - \Phi(\beta t + \alpha)|. \end{aligned}$$

Then,

$$\begin{aligned} \big|\Phi(t) - \Phi(\beta t)\big| &= \Big|\int_{\beta t}^{t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx\Big| \\ &\leq |\beta t - t| \frac{1}{\sqrt{2\pi}} e^{-\min(t, \beta t)^2/2} = \frac{1}{\sqrt{2\pi}} |\beta - 1| \cdot |t| e^{-at^2/2}, \end{aligned}$$

where $a = \min(\beta^2, 1)$. Note that $t \mapsto te^{-at^2/2}$ achieves its maximum of $1/\sqrt{ae}$ over $[0, \infty)$ at $t = 1/\sqrt{a}$. We also have $\big|\Phi(s) - \Phi(s + \alpha)\big| \leq |\alpha| \sup_{\tilde{s}} \Phi'(\tilde{s}) = \frac{1}{\sqrt{2\pi}}|\alpha|$. Putting the pieces together finishes the proof. □

Proof of Lemma 4. Let $\mathcal{A} = \{|\widehat{T}_n - T_n| \geq \delta T_n + \varepsilon\}$ and $q = \mathbb{P}(\mathcal{A})$. For any $t \in \mathbb{R}$, we have

$$\begin{aligned}\mathbb{P}(\widehat{T}_n \leq t) &\leq \mathbb{P}(\{\widehat{T}_n \leq t\} \cap \mathcal{A}^c) + \mathbb{P}(\mathcal{A}) \\ &\leq \mathbb{P}((1-\delta)T_n - \varepsilon \leq t) + q.\end{aligned}$$

Subtracting $\Phi(t) = \mathbb{P}(Z \leq t)$ from both sides, we get

$$\begin{aligned}\mathbb{P}(\widehat{T}_n \leq t) - \Phi(t) &\leq d_{\mathrm{K}}\big((1-\delta)T_n - \varepsilon, Z\big) + q \\ &\leq d_K(T_n, Z) + \frac{2\delta}{\sqrt{2\pi e}} + \frac{\varepsilon}{\sqrt{2\pi}} + q \\ &\leq d_K(T_n, Z) + \frac{1}{2}(\delta + \varepsilon) + q,\end{aligned}$$

by Lemma 24 and noting that $\min\{|1-\delta|, 1\} \geq 1/2$ by assumption. Similarly, for any $s \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}(T_n \leq s) &\leq \mathbb{P}(\{T_n \leq s\} \cap \mathcal{A}^c) + \mathbb{P}(\mathcal{A}) \\ &\leq \mathbb{P}(\widehat{T}_n \leq (1+\delta)s + \varepsilon) + q.\end{aligned}$$

Applying the change of variable $t = (1+\delta)s + \varepsilon$, adding Φ and rearranging, we obtain

$$\Phi(t) - \mathbb{P}(\widehat{T}_n \leq t) \leq \Phi(t) - \mathbb{P}((1+\delta)T_n + \varepsilon \leq t) + q,$$

and the rest of the argument follows as in the previous case. Putting the pieces together finishes the proof. □

Proof of Lemma 16. Note that $d_+^{(k)}\widehat{\Delta}_{k\ell}$ is a centered $\mathrm{Bin}(d_+^{(k)}, p_{k\ell})$ variable. Applying Proposition 8 (Section B.4), we have

$$\mathbb{P}\Big(|\widehat{\Delta}_{k\ell}| \geq \sqrt{\frac{2u}{d_+^{(k)}}} + \frac{u}{3d_+^{(k)}}\Big) \leq 2e^{-u}.$$

Then the result follows by using union bound when $u \leq \min_k d_+^{(k)}$. □

Proof of Lemma 6. Fix k and ℓ and consider $i \in \mathcal{G}_k$. Define

$$a := \frac{\sum_{i\in\widehat{\mathcal{G}}_k} X_{i\ell}}{\sum_{i\in\mathcal{G}_k} X_{i\ell}} - 1 = \frac{\sum_{i\in\widehat{\mathcal{G}}_k\setminus\mathcal{G}_k} X_{i\ell} - \sum_{i\in\mathcal{G}_k\setminus\widehat{\mathcal{G}}_k} X_{i\ell}}{\sum_{i\in\mathcal{G}_k} X_{i\ell}}.$$

On event $\mathcal{M}_n$, we have

$$\Big|\sum_{i\in\widehat{\mathcal{G}}_k\setminus\mathcal{G}_k} X_{i\ell} - \sum_{i\in\mathcal{G}_k\setminus\widehat{\mathcal{G}}_k} X_{i\ell}\Big| \le d_{\max}(|\widehat{\mathcal{G}}_k \setminus \mathcal{G}_k| + |\mathcal{G}_k \setminus \widehat{\mathcal{G}}_k|) \le d_{\max}(\alpha_n n).$$

Recall that we have $|X^{(k)}_{+\ell} - d^{(k)}_+ p_{k\ell}| \le \delta d^{(k)}_+$ on event $\mathcal{B}$. Furthermore, by assumption $\delta \le \underline{p}/2$, we obtain

$$X^{(k)}_{+\ell} \ge d^{(k)}_+\big(p_{k\ell} - \delta\big) \ge d^{(k)}_+ \underline{p}/2.$$

It follows that

$$|a| \le \frac{2(\alpha_n n) d_{\max}}{d^{(k)}_+ \underline{p}} \le \frac{2d_{\max}}{\omega_n \underline{p}}\alpha_n = \frac{2\alpha_n}{\tau_d\,\underline{p}}.$$

Similarly, letting $b := (\sum_{i\in\widehat{\mathcal{G}}_k} d_i)/(\sum_{i\in\mathcal{G}_k} d_i) - 1$, we have

$$|b| \le \frac{d_{\max}(\alpha_n n)}{d^{(k)}_+} \le \frac{d_{\max}}{\omega_n}\alpha_n = \frac{\alpha_n}{\tau_d}.$$

Then

$$\widehat{p}_{k\ell} = \frac{\sum_{i\in\widehat{\mathcal{G}}_k} X_{i\ell}}{\sum_{i\in\widehat{\mathcal{G}}_k} d_i} = \frac{(1+a)\sum_{i\in\mathcal{G}_k} X_{i\ell}}{(1+b)\sum_{i\in\mathcal{G}_k} d_i} = \frac{1+a}{1+b}\cdot\widetilde{p}_{k\ell}.$$

By assumption $\alpha_n \le \tau_d\underline{p}/2$, we have $|a| \le 1$ and $b \le 1/2$. Hence,

$$|\widehat{p}_{k\ell} - \widetilde{p}_{k\ell}| = \frac{|a-b|}{|1+b|}\cdot\widetilde{p}_{k\ell} \le \frac{|a|+|b|}{1-|b|}\cdot\widetilde{p}_{k\ell} \le 2(|a|+|b|)\cdot\widetilde{p}_{k\ell}.$$

Note that $|a| + |b| = (2\underline{p}^{-1} + 1)\frac{\alpha_n}{\tau_d} \le (3\alpha_n)/(\tau_d\underline{p})$. Then the result follows. □

Proof of Lemma 7. Let $E = \{e_\ell, \ell \in [L]\}$ be the standard basis of $\mathbb{R}^L$. Then, E is the set

of extreme points of $\mathcal{P}_L$ and $\mathcal{P}_L$ is the (closed) convex hull of E. The function $x \mapsto \|x-y\|$ is a continuous convex function, hence achieves its maximum over $\mathcal{P}_L$ at the set of extreme points. Then,

$$\max_{y\in\mathcal{P}_L}\max_{x\in\mathcal{P}_L}\|x-y\| = \max_{y\in\mathcal{P}_L}\max_{x\in E}\|x-y\| = \max_{y\in E}\max_{x\in E}\|x-y\|$$

where the last equality applies the same idea to the function $y \mapsto \|x-y\|$. The result follows since $\|e_\ell - e_k\| = \sqrt{2}$ for any $k \neq \ell$. □

B.2 Lemmas in the Proofs of Theorems 2 and 3

The following proposition, controlling the tail probability of a randomly-selected Poisson sum, is used in the proof of Lemma 8:

Proposition 6. *Let $A_j \sim \mathrm{Poi}(\lambda_j)$ and $U_j \sim \mathrm{Ber}(1/2)$ for $j = 1,\dots,n$, and assume that $\{A_j, U_j, j = 1,\dots,n\}$ are independent. Let $d = \sum_{j=1}^n A_j U_j$ and $d^* = \mathbb{E}[d]$. Then,*

$$\mathbb{P}\big(|d-d^*| \geq d^*/2\big) \leq 2e^{-0.008\,d^*} + 4e^{-0.03\,d^*/\lambda_{\max}}$$

where $\lambda_{\max} = \max_j \lambda_j$.

Proof of Propoisition 6. Let $\widetilde{d} = \sum_j \lambda_j U_j$ and $d^* = \frac{1}{2}\sum_j \lambda_j$, so that $d^* = \mathbb{E}[\widetilde{d}]$. Conditioned on $U = (U_1,\dots,U_n)$, d is a Poisson variable with mean $\widetilde{d}$. If $X \sim \mathrm{Poi}(\lambda)$, then for any $t \in (0,1]$, we have $\mathbb{P}(|X-\lambda| \geq t\lambda) \leq 2\exp(-\lambda t^2/4)$; see Lemma 30 (Section B.4). Then,

$$\mathbb{P}\big(|d-\widetilde{d}| \geq 0.2\widetilde{d} \mid U\big) \leq 2\exp(-0.01\widetilde{d}).$$

Next, we apply Proposition 8 (Section B.4) to $\widetilde{d} - d^* = \sum_j \lambda_j(U_j - 1/2)$. Since $|\lambda_j(U_j - 1/2)| \leq \lambda_{\max}$ and $\mathrm{var}(\widetilde{d}-d^*) = \sum_j \lambda_j^2/4 \leq \lambda_{\max}(d^*/2)$, we have

$$\mathbb{P}\big(\,|\widetilde{d}-d^*| \geq \sqrt{\lambda_{\max} d^* u} + \lambda_{\max} u/3\,\big) \leq 2e^{-u}.$$

Taking $u = 0.03d^*/\lambda_{\max}$, we obtain $\mathbb{P}(|\widetilde{d} - d^*| \geq 0.2d^*) \leq 2\exp(-0.03d^*/\lambda_{\max})$.

Let $\mathcal{A} = \{|d - \widetilde{d}| \geq 0.2\widetilde{d}\}$ and $\mathcal{B} = \{|\widetilde{d} - d^*| \geq 0.2d^*\}$. Note that $\mathcal{B}$ is completely determined by U. On $\mathcal{A}^c \cap \mathcal{B}^c$, we have $(0.8)^2 d^* < d < (1.2)^2 d^*$, implying $|d - d^*| < d^*/2$. It follows that

$$\mathbb{P}\big(|d - d^*| \geq d^*/2\big) \leq \mathbb{P}(\mathcal{A} \cup \mathcal{B}) \leq \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B}).$$

We have $\mathbb{P}(\mathcal{A}) = \mathbb{E}[\mathbb{P}(\mathcal{A} \mid U)1_{\mathcal{B}^c} + \mathbb{P}(\mathcal{A} \mid U)1_{\mathcal{B}}]$, hence

$$\begin{aligned}
\mathbb{P}(\mathcal{A}) &\leq \mathbb{E}[\mathbb{P}(\mathcal{A} \mid U)1_{\mathcal{B}^c} + 1_{\mathcal{B}}] \\
&\leq 2\mathbb{E}[e^{-0.01\widetilde{d}}1_{\mathcal{B}^c}] + \mathbb{P}(\mathcal{B}) \\
&\leq 2e^{-0.008d^*}\mathbb{E}[1_{\mathcal{B}^c}] + \mathbb{P}(\mathcal{B})
\end{aligned}$$

using $\widetilde{d} \geq 0.8d^*$ on $\mathcal{B}^c$. We further bound $\mathbb{E}[1_{\mathcal{B}^c}] \leq 1$. Putting the pieces together finishes the proof. □

Proof of Lemma 8. Recall that $d_i = \sum_{j=1}^n A_{ij}U_j$ where $\{U_j = 1\{j \in S_1\}\}$ is an independent Ber(1/2) sequence, and $d_i^* = \mathbb{E}[d_i]$. We also recall from (2.60) that $d_i^* \geq \frac{1}{2}C_1\nu_n$ for all $i \in [n]$. Fix $i \in [n]$. We apply Proposition 6 to d_i with $\lambda_j = \mathbb{E}[A_{ij}] = (\nu_n/n)\theta_i\theta_j B^0_{z_i z_j}$. Since $\|B^0\|_\infty = 1$ and $\theta_{\max} = 1$, we have $\max_j \lambda_j \leq \nu_n/n$, and thus

$$\frac{d_i^*}{\max_j \lambda_j} \geq \frac{C_1}{2}n \geq \frac{200}{3}\log n,$$

where the first inequality is by (2.60) and the second by assumption (2.26). Proposition 6 gives

$$\mathbb{P}\big(|d_i - d_i^*| \geq d_i^*/2\big) \leq 2e^{-0.004C_1\nu_n/2} + 4e^{-2\log n} \leq 6n^{-2}$$

since $0.004C_1\nu_n/2 \geq 2\log n$ by assumption (2.26). By union bound,

$$\mathbb{P}\big(d_i \notin [\tfrac{1}{2}d_i^*, \tfrac{3}{2}d_i^*] \text{ for some } i \in [n]\big) \leq 6n^{-1}. \tag{B.4}$$

Furthermore, $\tilde{n}_k := |\mathcal{G}_k| = n_k - |\mathcal{C}_k \cap S_1| = n_k - \sum_{i\in\mathcal{C}_k} U_i$ for all k. Applying Proposition 8 (Section B.4) with $u = 0.01n_k$, we obtain

$$\Big|\frac{\tilde{n}_k}{n_k} - \frac{1}{2}\Big| = \Big|\frac{1}{n_k}\sum_{i\in\mathcal{C}_k} U_i - \frac{1}{2}\Big| \geq \sqrt{0.01} + \frac{0.01}{3} \geq 0.1$$

with probability $\leq 2e^{-0.01n_k}$. By union bound

$$\mathbb{P}\big(\tilde{n}_k \notin [0.4n_k, 0.6n_k] \text{ for some } k \in [K_0]\big) \leq 2\sum_{k=1}^{K_0} e^{-0.01n_k} \tag{B.5}$$

$$\leq 2K_0 e^{-0.01\tau_{\mathcal{C}} n} \leq n^{-1}. \tag{B.6}$$

The last inequality is since assumption (2.26) implies $0.01\tau_{\mathcal{C}} n \geq \log(n^3) \geq \log(2K_0 n)$. The result follows by combining (B.4) and (B.6). □

B.2.1 Lemmas in the Proof of Theorem 2

Proof of Lemma 9. Let $U_t := \mathbb{P}(Y \leq t \mid \mathcal{F})$ and set $U = (U_t, t \in \mathbb{R})$ and $b_t = \mathbb{P}(Z \leq t)$. The function $f(U) = \sup_{t\in\mathbb{R}} |U_t - b_t|$ is convex, hence by Jensen's inequality

$$d_{\mathrm{K}}(Y, Z) = f(\mathbb{E}U) \leq \mathbb{E}f(U) = \mathbb{E}\big[d_{\mathrm{K}}\big(\mathcal{L}(Y \mid \mathcal{F}), Z\big)\big].$$

Next, letting $Y' := Y1_{\mathcal{B}}$, we have

$$\begin{aligned}\mathbb{P}(Y' \leq t) &\leq \mathbb{P}(\{Y' \leq t\} \cap \mathcal{B}) + \mathbb{P}(\mathcal{B}^c) \\ &= \mathbb{P}(\{Y \leq t\} \cap \mathcal{B}) + \mathbb{P}(\mathcal{B}^c) \leq \mathbb{P}(Y \leq t) + \mathbb{P}(\mathcal{B}^c)\end{aligned}$$

and

$$\mathbb{P}(Y' \leq t) \geq \mathbb{P}(\{Y' \leq t\} \cap \mathcal{B}) = \mathbb{P}(\{Y \leq t\} \cap \mathcal{B}) \geq \mathbb{P}(Y \leq t) - \mathbb{P}(\mathcal{B}^c).$$

It follows that $|\mathbb{P}(Y' \leq t) - \mathbb{P}(Y \leq t)| \leq \mathbb{P}(\mathcal{B}^c)$ for all $t \in \mathbb{R}$. An application of the triangle inequality gives $|d_{\mathrm{K}}(Y, Z) - d_{\mathrm{K}}(Y', Z)| \leq \mathbb{P}(\mathcal{B}^c)$ finishing the proof. □

B.2.2 Lemmas in the Proof of Theorem 3

Proof of Lemma 10. Recall that $\hat{\mathcal{T}}_r \subset \mathcal{C}_r \cap S_2 = \mathcal{G}_r$ and for any $i \in \mathcal{G}_r$, we have $d_i\xi_{i\ell} \sim \mathrm{Bin}(d_i, q_{r\ell})$, conditioned on $\mathcal{F}$. Thus, we can write $d_i(\xi_{i\ell} - q_{r\ell}) = \sum_{j=1}^{d_i} Z_j$ where Z_j are centered Bernoulli variables with parameter $q_{r\ell}$. Applying Proposition 8 (Appendix B.4), we have

$$\mathbb{P}^{\mathcal{F}}\Big(|\sum_j Z_j| \geq \sqrt{2vu} + \frac{u}{3}\Big) \leq 2e^{-u}, \quad u \geq 0,$$

where $v = \sum_j \mathrm{var}(Z_j)$. Since, $v = d_i q_{r\ell}(1 - q_{r\ell}) \leq d_i/4$, taking $u = 2\log n$, we have

$$\mathbb{P}^{\mathcal{F}}\Big(|\xi_{i\ell} - q_{r\ell}| \geq \sqrt{\frac{\log n}{d_i}} + \frac{2\log n}{3d_i}\Big) \leq 2n^{-2}.$$

On event $\mathcal{A}$, we have $d_i \geq d_i^*/2 \geq C_1\nu_n/4$ for all i, by (2.60). By assumption, $4\log n \leq C_1\nu_n$, hence on $\mathcal{A}$,

$$\sqrt{\frac{\log n}{d_i}} + \frac{2\log n}{3d_i} \leq 4\sqrt{\frac{\log n}{C_1\nu_n}} = \varepsilon_n.$$

We have

$$\mathcal{E}^c = \Big\{\max_{r,\ell}\max_{i\in\hat{\mathcal{T}}_r} |\xi_{i\ell} - q_{r\ell}| \geq \varepsilon_n\Big\} \subset \Big\{\max_{r,\ell}\max_{i\in\mathcal{G}_r} |\xi_{i\ell} - q_{r\ell}| \geq \varepsilon_n\Big\}.$$

Using $|\bigcup_r \mathcal{C}_{h_r}| = n$ and the union bound, we obtain $\mathbb{P}^{\mathcal{F}}(\mathcal{E}^c \cap \mathcal{A}) \leq 2(nL)\cdot n^{-2} = 2Ln^{-1}$. The lemma follows by taking the expectation of both sides and using the smoothing property of conditional expectation. □

Proof of Lemma 11. Lemma 11 follows from the following more refined result:

Lemma 25. *Let $\psi(x,y) = (x-y)^2/y$. For all (x,y) and (x',y') in $[0,1]\times[1/c_1,1]$, where $c_1 > 1$, we have*

$$\big|\psi(x',y') - \psi(x,y)\big| \leq c_2|x-y|\cdot\|\delta\| + c_3\|\delta\|^2 \tag{B.7}$$

where $\delta = (x - x', y - y')$, $c_2 = c_1\sqrt{4 + (1 + c_1)^2}$ *and* $c_3 = 4c_1^3$.

Assuming that $|x - x'| \le \varepsilon$ and $|y - y'| \le \varepsilon$, so that $\|\delta\| \le \sqrt{2}\varepsilon$, and using $|x - y| \le 1$,

$$\big|\psi(x', y') - \psi(x, y)\big| \le \sqrt{2}c_2\varepsilon + 2c_3\varepsilon^2 \le c_4 \max(\varepsilon, \varepsilon^2) \tag{B.8}$$

where $c_4 = \sqrt{2}c_2 + 2c_3$. Since $c_2 \le \sqrt{8}c_1^2$, we have $c_4 \le 12c_1^3$ and Lemma 11 follows. □

Proof of Lemma 25. The function ψ is continuously differentiable of all orders, on $\mathbb{R} \times \mathbb{R}_{++}$, with the gradient and Hessian given by

$$\nabla\psi(x, y) = \big(x/y - 1\big) \begin{bmatrix} 2 \\ -(1 + x/y) \end{bmatrix}, \quad \nabla^2\psi(x, y) = (2/y) \begin{bmatrix} 1 & -x/y \\ -x/y & x^2/y^2 \end{bmatrix}.$$

The Hessian has eigenvalues 0 and $2(x^2 + y^2)/y^3$. By Taylor expansion,

$$\psi(x', y') - \psi(x, y) = \langle \nabla\psi(x, y), \delta\rangle + \frac{1}{2}\langle \delta, \nabla^2\psi(\tilde{x}, \tilde{y}), \delta\rangle$$

where $(\tilde{x}, \tilde{y})$ is a point between (x, y) and (x', y'). Since $0 \preceq \nabla^2\psi(\tilde{x}, \tilde{y}) \preceq 2(\tilde{x}^2 + \tilde{y}^2)/\tilde{y}^3 I_2$ and $\tilde{y} \ge \min\{y, y'\} \ge 1/c_1$, we obtain

$$\big|\langle \delta, \nabla^2\psi(\tilde{x}, \tilde{y}), \delta\rangle\big| \le \frac{2(\tilde{x}^2 + \tilde{y}^2)}{\tilde{y}^3}\|\delta\|^2 \le 4c_1^3\|\delta\|^2.$$

We also have

$$\big|\langle \nabla\psi(x, y), \delta\rangle\big| \le |x/y - 1|\sqrt{4 + (1 + x/y)^2}\|\delta\| \le c_2\|\delta\|$$

using the assumption on the ranges of x and y. The result follows. □

B.3 Lemmas in the Proof of Theorem 4

Proof of Lemma 13 . For any $x \in \mathbb{R}^d$, let

$$W_\ell(x) := \sum_{j \in S_1} \theta_j g(x, x_j) 1\{y_j = \ell\}.$$

From the definition of $q_{i\ell}$ in (2.79), we have

$$q_{i\ell} = \frac{\nu_n}{n} \theta_i W_\ell(x_i).$$

To control $q_{i\ell}$, it is enough to control $W_\ell(x_i)$. Recall that $\mathcal{F}_1 = \mathcal{F}_0 \vee \sigma(x_{S_2}) = \sigma(S_1, x_{S_2})$. Note that on $\mathcal{F}_1$, both S_1 and S_2 are fixed. Then, for $i \in S_2$ and $j \in S_1$, we have

$$\begin{aligned} \mathbb{E}[g(x_i, x_j) \,|\, \mathcal{F}_1] &= \mathbb{E}[g(x_i, x_j) \,|\, x_{S_2}, S_2] \\ &= \mathbb{E}[g(x_i, x_j) \,|\, x_i] \\ &= h_{z_j}(x_i) \end{aligned}$$

where we have used the independence of x_i and x_j. It follows that

$$\begin{aligned} \mathbb{E}[W_\ell(x_i) \,|\, \mathcal{F}_1] &= \sum_{j \in S_1} \theta_j h_{z_j}(x_i) 1\{y_j = \ell\} \\ &= \sum_{k=1}^{K} h_k(x_i) \widetilde{R}_{k\ell} \end{aligned} \tag{B.9}$$

where $\widetilde{R}_{k\ell} := \sum_{j \in S_1} \theta_j 1\{z_j = k,\, y_j = \ell\}$. Furthermore, let $\widetilde{m}_\ell := \sum_{k=1}^{K} \widetilde{R}_{k\ell} = \sum_{j \in S_1} \theta_j 1\{y_j = \ell\}$. The next lemma shows that $W_\ell(x_i)$ concentrates near its conditional mean.

Lemma 26. *Assume that* $\max_{j \in S_1} \theta_j \le 1$ *and* $g(\cdot, \cdot)$ *is bounded above by* 1. *Then, for any fixed* $x \in \mathbb{R}^d$, *with* $\mathcal{F}_1$*-probability at least* $1 - 2e^{-t}$,

$$|W_\ell(x) - \mathbb{E}[W_\ell(x) \,|\, \mathcal{F}_1]| \le \sqrt{\widetilde{m}_\ell t / 2} \tag{B.10}$$

Proof of Lemma 26. Conditional on $\mathcal{F}_1$, S_1 is fixed. We note that $W_\ell(x) = F(x_{S_1})$ where $F(\cdot)$ is a function with the bounded difference property, that is, if x_{S_1} and x'_{S_1} differ only in their jth coordinate, then $|F(x_{S_1}) - F(x'_{S_1})| \le \theta_j 1\{y_j = \ell\}$ since the range of g is in $[0,1]$. By the McDiarmid's inequality, with $\mathcal{F}_1$-probability at least $1 - 2e^{-2u^2/L^2}$, we have $|W_\ell(x) - \mathbb{E}[W_\ell(x) \,|\, \mathcal{F}_1]| \le u$, where $L^2 := \sum_{j \in S_1} \theta_j^2 1\{y_j = \ell\} \le \widetilde{m}_\ell$. Taking $u^2 = tL^2/2 \le t\widetilde{m}_\ell/2$ finishes the proof. □

Applying the union bound over $(i, \ell) \in S_2 \times [L]$, we have with $\mathcal{F}_1$-conditional probability at least $1 - 2nLe^{-t}$,

$$\Big| W_\ell(x_i) - \sum_{k=1}^K h_k(x_i) \widetilde{R}_{k\ell} \Big| \le \sqrt{\widetilde{m}_\ell t}/2, \quad \forall i \in S_2,\ \ell \in [L].$$

where we have used x_{S_2} being fixed given $\mathcal{F}_1$. Taking $t = 2 \log n$ and noting that $\widetilde{m}_\ell \le n$, we can integrate out the conditional probability to get

$$\mathbb{P}\Big(\Big| W_\ell(x_i) - \sum_{k=1}^K h_k(x_i) \widetilde{R}_{k\ell} \Big| \le \sqrt{n \log n}, \quad \forall i \in S_2,\ \ell \in [L] \Big) \ge 1 - 2Ln^{-1} \tag{B.11}$$

We can write $\widetilde{R}_{k\ell} = \sum_{j=1}^n \theta_j U_j 1\{z_j = k,\, y_j = \ell\}$, for some i.i.d. Ber(1/2) sequence $\{U_j\}_{j=1}^n$. Recalling the definition of $R_{k\ell}$ from (2.34), we have

$$\mathbb{E}[\widetilde{R}_{k\ell}] = \frac{1}{2} \sum_{j=1}^n \theta_i 1\{z_j = k,\, y_j = \ell\} = R_{k\ell}.$$

Applying Proposition 8 with $v = n/4 \ge \mathrm{var}(\widetilde{R}_{k\ell})$ and $u = \log n$, we have

$$\mathbb{P}\Big(|\widetilde{R}_{k\ell} - R_{k\ell}| \ge \sqrt{\frac{n \log n}{2}} + \frac{\log n}{3} \Big) \le 2n^{-1}.$$

By union bound, with probability at least $1 - 2LKn^{-1}$,

$$|\widetilde{R}_{k\ell} - R_{k\ell}| \le \sqrt{n \log n}, \quad \forall k \in [K],\ \forall \ell \in [L]. \tag{B.12}$$

Let $\Delta_{i\ell} = W_\ell(x_i) - \sum_{k=1}^K h_k(x_i) R_{k\ell}$, and consider the event,

$$\mathcal{W} = \Big\{ |\Delta_{i\ell}| \le 2K\sqrt{n \log n}, \quad \forall i \in S_2,\ \ell \in [L] \Big\}.$$

Combining (B.11) and (B.12), using $h_k(x) \le 1$, the triangle inequality, and $K + 1 \le 2K$, we have $\mathbb{P}(\mathcal{W}^c) \le 4KLn^{-1}$.

Next we note that

$$\rho_{i\ell} = \frac{W_\ell(x_i)}{\sum_\ell' W_{\ell'}(x_i)} = \frac{\sum_k h_k(x_i) R_{k\ell} + \Delta_{i\ell}}{\sum_{\ell'} \big(\sum_k h_k(x_i) R_{k\ell'} + \Delta_{i\ell'} \big)}.$$

Furthermore, on Γ,

$$\begin{aligned} \sum_{\ell'} \sum_k h_k(x_i) R_{k\ell'} &= \frac{1}{2} \sum_k h_k(x_i) \sum_{j \in S_1} \theta_j 1\{z_j = k\} \\ &\ge \frac{1}{2} \tau_\theta h_{r_{z_i}}(x_i) n_{r_{z_i}} \\ &\ge \frac{1}{2} \tau_\theta \tau_{\mathcal{C}} \tau_h n = \tau_\rho L n \end{aligned}$$

where we have used (2.33) and the definition of τ_ρ in (2.81). By the assumption that $\tau_\rho L n > 4KL\sqrt{n \log n}$, on event $\Gamma \cap \mathcal{W}$, applying Lemma 27 below, we have for all $i \in S_2$ and $\ell \in [L]$,

$$\Big| \rho_{i\ell} - \frac{\sum_k h_k(x_i) R_{k\ell}}{\sum_{\ell'} \sum_k h_k(x_i) R_{k\ell'}} \Big| \le \frac{4K\sqrt{n \log n}}{\tau_\rho n} = \frac{4K}{\tau_\rho} \sqrt{\frac{\log n}{n}}$$

which is the event $\mathcal{R}$. That is, we have shown $\mathcal{R} \supseteq \Gamma \cap \mathcal{W}$, and the claim follows.

Lemma 27. *For $a = (a_\ell) \in \mathbb{R}_+^L \setminus \{0\}$, let $a_+ = \sum_{\ell=1}^L a_\ell$ and consider the function $U(a) = a_1/a_+$. Let $\delta \in \mathbb{R}^L$ and $\|\delta\|_\infty = \max_\ell |\delta_\ell|$. If $a_+ > L\|\delta\|_\infty$, then*

$$|U(a + \delta) - U(a)| \le \frac{(L-1)\|\delta\|_\infty .}{a_+ - L\|\delta\|_\infty}$$

In particular, $|U(a + \delta) - U(a)| \le (2L/a_+)\|\delta\|_\infty$ if $a_+ > 2L\|\delta\|_\infty$.

Proof of Lemma 27. The gradient of U at $c \in \mathbb{R}^L_{++}$ is given by

$$\nabla U(c) = \frac{1}{c_+^2}(c_+ - c_1, -c_1, \ldots, -c_1).$$

For $a, a+\delta \in \mathbb{R}^L_+$, there exist c in the line-segment connecting a and $a+\delta$ such that $U(a+\delta) - U(a) = \langle \nabla U(c), \delta \rangle$. From Hölder's inequality, we have

$$|U(a+\delta) - U(a)| \le \|\nabla U(c)\|_1 \|\delta\|_\infty$$

where

$$\|\nabla U(c)\|_1 = \frac{1}{c_+^2}\big(c_+ - c_1 + (L-1)c_1\big) \le \frac{L-1}{c_+}.$$

Noting that $c_+ \ge a_+ - L\|\delta\|_\infty$ finishes the proof. □

□

Proof of Lemma 14. For $x \in \mathbb{R}^d$, let $V(x) := \sum_{j \in S_1} \theta_j g(x, x_j)$. Recall that $d_i = \sum_{j \in S_1} A_{ij}$ and for $i \in S_2$, consider

$$\widetilde{d}_i := \mathbb{E}\big[d_i \,|\, \mathcal{F}_2\big] = \sum_{j \in S_1} p_{ij} = \theta_i \frac{\nu_n}{n} V(x_i).$$

We refer to (2.77) for the defintion of $\mathcal{F}_1, \mathcal{F}_2$, etc. Note that $\mathcal{F}_1 \subseteq \mathcal{F}_2$. Let $m = \sum_{j \in S_1} \theta_j$. Applying the same idea as in Lemma 26, we have with $\mathcal{F}_1$-conditional probability at least $1 - 2e^{-t}$,

$$\big|V(x) - \mathbb{E}[V(x) \,|\, \mathcal{F}_1]\big| \le \sqrt{tm/2} \le \sqrt{tn/2}$$

where the second inequality uses $\theta_i \le 1$ and $|S_1| \le n$. Since conditional on $\mathcal{F}_1$, $x_i, i \in S_2$ are fixed, it follows that $\mathcal{F}_1$-conditional probability at least $1 - 2ne^{-t}$,

$$\big|V(x_i) - \mathbb{E}\big[V(x_i) \,|\, \mathcal{F}_1\big]\big| \le \sqrt{tn/2}, \quad \forall i \in S_2,$$

from which we get, multiplying both sides by $\theta_i \nu_n / n$ and using $\theta_i \le 1$,

$$\big|\widetilde{d}_i - \mathbb{E}[\widetilde{d}_i \,|\, \mathcal{F}_1]\big| \le \nu_n \sqrt{t/(2n)}, \quad \forall i \in S_2. \tag{B.13}$$

Consider the event

$$\mathcal{D}_1 = \Big\{ \big|\widetilde{d}_i - \mathbb{E}\big[\widetilde{d}_i \,|\, \mathcal{F}_1\big]\big| \le \nu_n \sqrt{\frac{\log n}{n}},\ \forall\, i \in S_2 \Big\}. \tag{B.14}$$

Taking $t = 2 \log n$ in (B.14), we obtain $\mathbb{P}(\mathcal{D}_1^c \,|\, \mathcal{F}_1) \le n^{-1}$, hence $\mathbb{P}(\mathcal{D}_1^c) \le n^{-1}$ by taking the expecation of both sides.

Now let us control $\mathbb{E}\big[\widetilde{d}_i \,|\, \mathcal{F}_1\big]$. For $i \in S_2$, we have

$$\mathbb{E}[V(x_i) \,|\, \mathcal{F}_1] = \mathbb{E}[V(x_i) \,|\, x_i] = \sum_{j \in S_1} \theta_j h_{z_j}(x_i) = \sum_r \sum_{j \in S_1 \cap \mathcal{C}_r} \theta_j h_r(x_i)$$

On Γ, by (2.33), we have $h_{r_{z_i}}(x_i) \ge \tau_h$ for all $i \in [n]$. This gives,

$$\tau_\theta \tau_h |S_1 \cap \mathcal{C}_{r_{z_i}}| \ \le\ \mathbb{E}[V(x_i) \,|\, \mathcal{F}_1] \ \le\ |S_1|$$

where we have also used $0 \le h_k(\cdot) \le 1$ and $\tau_\theta \le \theta_j \le 1$. On $\mathcal{A}_1$, we have $|S_1 \cap \mathcal{C}_{r_{z_i}}| \ge 0.4 n_{r_{z_i}} \ge 0.4 \tau_{\mathcal{C}} n$ and $|S_1| \le 0.6n$. It follows that on $\Gamma \cap \mathcal{A}_1$,

$$0.4 \tau_\theta^2 \tau_h \tau_{\mathcal{C}} \nu_n \ \le\ \mathbb{E}[\widetilde{d}_i \,|\, \mathcal{F}_1] \ \le\ 0.6 \nu_n$$

for all $i \in S_2$. Recall that $C_8 = \tau_\theta^2 \tau_h \tau_{\mathcal{C}}$. Since by assumption $\sqrt{(\log n)/n} \le 0.2 C_8 \le 0.2$, it follows that on $\Gamma \cap \mathcal{A}_1 \cap \mathcal{D}_1$, we have

$$\widetilde{d}_i / \nu_n \in [0.2 C_8, 0.8], \quad \forall i \in S_2. \tag{B.15}$$

Next we show that d_i has the same growth rate as $\widetilde{d}_i$. We have $d_i \,|\, \mathcal{F}_2 \sim \mathrm{Poi}(\widetilde{d}_i)$ for all

$i \in S_2$. Consider the event

$$\mathcal{D}_2 := \{|d_i - \widetilde{d}_i| \le 0.2\widetilde{d}_i,\ \forall i \in S_2\}. \tag{B.16}$$

Applying Lemma 30, we have $\mathbb{P}(\mathcal{D}_2^c \,|\, \mathcal{F}_2) \le 2\sum_{i\in S_2} \exp(-0.01\widetilde{d}_i)$, hence

$$\begin{aligned}
\mathbb{P}(\mathcal{D}_2^c \cap \mathcal{A}_1 \cap \mathcal{D}_1) &= \mathbb{E}\big[\mathbb{P}(\mathcal{D}_2^c \cap \mathcal{A}_1 \cap \mathcal{D}_1 \,|\, \mathcal{F}_2)\big] \\
&= \mathbb{E}[\mathbb{P}(\mathcal{D}_2^c \,|\, \mathcal{F}_2) 1_{\mathcal{A}_1 \cap \mathcal{D}_1}] \\
&= \mathbb{E}[\mathbb{P}(\mathcal{D}_2^c \,|\, \mathcal{F}_2) 1_{\mathcal{A}_1 \cap \mathcal{D}_1 \cap \Gamma}] \\
&\le 2\mathbb{E}\Big[\sum_{i\in S_2} e^{-0.01\widetilde{d}_i} 1_{\mathcal{A}_1 \cap \mathcal{D}_1 \cap \Gamma}\Big] \le 1.2ne^{-0.002C_8\nu_n}.
\end{aligned}$$

where the second equality is since $\mathcal{A}_1 \cap \mathcal{D}_1$ is deterministic given $\mathcal{F}_2$, the third equality is by $\mathbb{P}(\Gamma) = 1$, and the final inequality uses (B.15) and that $|S_1| \le 0.6n$ on $\mathcal{A}_1$; see (2.74). The LHS above is also equal to $\mathbb{P}(\mathcal{D}_2^c \cap \mathcal{A}_1 \cap \mathcal{D}_1 \cap \Gamma)$. Hence,

$$\mathbb{P}(\mathcal{D}_2^c \cap \mathcal{A}_1 \cap \mathcal{D}_1 \cap \Gamma) \le 1.2n^{-1}$$

using the assumption $(\log n)/\nu_n \le C_8/1000$. We note that on $\Gamma \cap \mathcal{A}_1 \cap \mathcal{D}_1 \cap \mathcal{D}_2$, we have (B.15) and $d_i/\widetilde{d}_i \in [0.8, 1.2]$, which imply $d_i/\nu_n \in [0.16C_8, 0.96]$, that is, $\mathcal{A}_2$ hold. Let $\mathcal{D} = \mathcal{D}_1 \cap \mathcal{D}_2$. We have $\mathcal{D}^c = \mathcal{D}_1^c \uplus (\mathcal{D}_2^c \cap \mathcal{D}_1)$ where $\uplus$ denotes the disjoint union. Then,

$$\begin{aligned}
\mathbb{P}(\mathcal{D}^c \cap \mathcal{A}_1) &= \mathbb{P}(\mathcal{D}_1^c \cap \mathcal{A}_1) + \mathbb{P}(\mathcal{D}_2^c \cap \mathcal{D}_1 \cap \mathcal{A}_1) \\
&\le \mathbb{P}(\mathcal{D}_1^c) + \mathbb{P}(\mathcal{D}_2^c \cap \mathcal{A}_1 \cap \mathcal{D}_1 \cap \Gamma) \le 2.2n^{-1}
\end{aligned}$$

and the result follows. □

Proof of Lemma 15. We first develop a lower bound for $H_{\ell_{z_i}}(x_i)$. Using the ℓ_k defined in (2.35),

$$R_{k\ell_k} \ge \frac{1}{L}\sum_{\ell} R_{k\ell} \ge \frac{\tau_\theta n_k}{2L} = \frac{\tau_\theta \tau_{\mathcal{C}}}{2L} n$$

Recall that on event Γ, $h_{r_{z_i}}(x_i) \geq \tau_h$. Then we can control the numerator of $H_{\ell_{z_i}}(x_i)$ by

$$\sum_k h_k(x_i) R_{k\ell_{z_i}} \geq h_{r_{z_i}}(x_i) R_{r_{z_i}\ell_{z_i}} \geq \frac{\tau_\theta \tau_{\mathcal{C}} \tau_h}{2L} n \tag{B.17}$$

To control its denominator, using $\theta_j \leq 1$ and $h_k(x_i) \leq 1$, we have

$$\sum_{\ell'} \sum_k h_k(x_i) R_{k\ell'} \leq \sum_{\ell'} \sum_k R_{k\ell} = \frac{1}{2} \sum_{j=1}^n \theta_j \leq \frac{1}{2} n. \tag{B.18}$$

Combining (B.17) and (B.18) and the definition of $H_\ell(x_i)$, we obtain $H_{\ell_{z_i}}(x_i) \geq 2\tau_\rho$. Finally, by definition (2.82), on $\mathcal{R}$, we have

$$\rho_{i\ell_{z_i}} \geq H_{\ell_{z_i}}(x_i) - \frac{4K}{\tau_\rho} \sqrt{\frac{\log n}{n}},$$

which together with the assumption on $\frac{\log n}{n}$ gives the desired result. □

Proof of Lemma 16. Conditioning on $\mathcal{F}$, the quantities $\mathcal{G}_k$, $\bar{\rho}_{k\ell}$, $(d_i, i \in S_2)$ and d_+^k are fixed. Moreover, by (2.80) we have $X_{i\ell} \,|\, \mathcal{F} \sim \mathrm{Bin}(d_i, \rho_{i\ell})$. Then, by Proposition 4,

$$\mathbb{P}^{\mathcal{F}}\Big(\Big| \sum_{i \in \mathcal{G}_k} (X_{i\ell} - d_i \rho_{i\ell}) \Big| \geq \sqrt{2vu} + \frac{u}{3} \Big) \leq 2e^{-u}$$

for any $v \geq \mathrm{var}(\sum_{i \in \mathcal{G}_k} X_{i\ell})$ and $\mathbb{P}^{\mathcal{F}}$ denote the probability conditional on $\mathcal{F}$. We have $\mathrm{var}(\sum_{i \in \mathcal{G}_k} X_{i\ell}) = \sum_{i \in \mathcal{G}_k} d_i \rho_{i\ell}(1 - \rho_{i\ell}) \leq d_+^k/4$. Taking $v = d_+^k/4$, $u = 2 \log n$, we have

$$\mathbb{P}^{\mathcal{F}}\left(|\widetilde{\Delta}_{k\ell}| \geq \sqrt{\frac{\log n}{d_+^k}} + \frac{2 \log n}{3 d_+^k} \right) \leq 2n^{-2}.$$

From (2.84), on event $\mathcal{A}$, we have $d_+^k \geq 0.064 \tau_{\mathcal{C}} C_8 n \nu_n \geq \log n$ for all $k \in [K]$, where the second inequality is by assumption. It follows that

$$\sqrt{\frac{\log n}{d_+^k}} + \frac{2 \log n}{3 d_+^k} \leq 2 \sqrt{\frac{\log n}{d_+^k}} \leq \frac{8}{\sqrt{\tau_{\mathcal{C}} C_8}} \sqrt{\frac{\log n}{n \nu_n}}.$$

Therefore, $\mathbb{P}^{\mathcal{F}}(\mathcal{B}^c \cap \mathcal{A}) = \mathbb{P}^{\mathcal{F}}(\mathcal{B}^c) 1_{\mathcal{A}} \leq 2KLn^{-2} \leq 2Ln^{-1}$ by union bound and since on

$\mathcal{F}$, the event $\mathcal{A}$ is deterministic. The result follows by taking the expectation to both sides. □

Proof of Lemma 17. We use an idea similar to the one used in Lemma 6 in the proof of Proposition 2. We note that δ plays a similar role in both proofs, and $\bar{\rho}_{k\ell}$ and $\widetilde{\rho}_{k\ell}$ here play the role of $p_{k\ell}$ and $\widetilde{p}_{k\ell}$ there. Let

$$\tau_d := \frac{\omega_n}{\max_{i\in S_2} d_i}. \tag{B.19}$$

Combining (2.83) and (2.85), on $\mathcal{A}$, we have

$$\tau_d \geq \tau_{\mathcal{C}} C_8/9. \tag{B.20}$$

By Lemma 15, on $\Gamma \cap \mathcal{R}$, we have $\bar{\rho}_{k\ell_k} \geq \tau_\rho$ for all $k \in [K]$. Hence, τ_ρ plays the role of $\underline{p}$ in Proposition 2.

Then, to apply Lemma 6, we need $\alpha_n \leq \tau_d\tau_\rho/2$ and $\delta \leq \tau_\rho/2$, with τ_ρ. By (B.20), the first condition is satisfied on $\mathcal{A}$, if $\alpha_n \leq \tau_{\mathcal{C}}\tau_\rho C_8/18$, which holds by assumption. Then, the equivalent of Lemma 6 in this proof implies that on $\mathcal{B} \cap \mathcal{M}_n \cap (\Gamma \cap \mathcal{R} \cap \mathcal{A})$, we have $|\widehat{\Delta}_{k\ell_k}| \leq \hat{\delta} \cdot \widetilde{\rho}_{k\ell_k}$ for all $k \in [K]$, where

$$\hat{\delta} := \frac{6}{\underline{\rho}\,\tau_d}\alpha_n \leq \frac{54}{\tau_\rho\tau_{\mathcal{C}}C_8}\alpha_n.$$

□

Proof of Lemma 18. The proof is similar to that of Lemma 16 to which we refer for more details. We have $X_{i\ell} \,|\, \mathcal{F} \sim \mathrm{Bin}(d_i, \rho_{i\ell})$. Hence, by Proposition 8,

$$\mathbb{P}^{\mathcal{F}}\Big(|\xi_{i\ell} - \rho_{i\ell}| \geq \sqrt{\frac{\log n}{d_i}} + \frac{2\log n}{3d_i}\Big) \leq 2n^{-2}.$$

Recalling (2.83), on $\mathcal{A}$, we have $d_i \geq 0.16C_8\nu_n$ for $i \in S_2$ and by assumption $\log n \leq$

$0.16C_8\nu_n$. Hence, on $\mathcal{A}$,

$$\sqrt{\frac{\log n}{d_i}} + \frac{2\log n}{3d_i} \le 5\sqrt{\frac{\log n}{C_8\nu_n}}$$

By union bound over $\ell \in [L]$, we obtain $\mathbb{P}^{\mathcal{F}}(\mathcal{E}^c \cap \mathcal{A}) \le 2Ln^{-1}$. The result then follows by taking the expectation to both sides. □

Proof of Lemma 19. Let $\mathbb{V}((a_i),(p_i))$ be the variance of a random variable that takes values a_i with probability p_i, that is,

$$\mathbb{V}((a_i),(p_i)) = \sum_i p_i\Big(a_i - \sum_j p_j a_j\Big)^2 = \frac{1}{2}\sum_{i,j} p_i p_j (a_i - a_j)^2.$$

Note that $\varpi_k = \mathbb{V}((\rho_{i\ell_k}),(d_i/d_+^k))$. The next step is to show that in the definition of ϖ_k, we can replace $\rho_{i\ell_k}$ with $H_{\ell_k}(x_i)$ and d_i/d_+^k with deterministic quantities.

Lemma 28. *For $a_i, b_i \in \mathbb{R}$, with $\max_i |a_i - b_i| \le \varepsilon$, we have $|a_i - a_j| \ge |b_i - b_j| - 2\varepsilon$.*

Proof of Lemma 28. Assume $|a_1 - b_1| \le \varepsilon$ and $|a_2 - b_2| \le \varepsilon$. Then,

$$\begin{aligned}
|a_1 - a_2| &= |(b_1 - b_2) + (a_1 - b_1) - (a_2 - b_2)| \\
&\ge |b_1 - b_2| - |(a_1 - b_1) - (a_2 - b_2)| \\
&\ge |b_1 - b_2| - |a_1 - b_1| - |a_2 - b_2| \\
&\ge |b_1 - b_2| - 2\varepsilon.
\end{aligned}$$

□

Let us define

$$\zeta_n := \frac{4\tau_\theta L}{C_8}\sqrt{\frac{\log n}{n}} = \frac{4L}{\tau_{\mathcal{C}}\tau_h\tau_\theta}\sqrt{\frac{\log n}{n}}. \tag{B.21}$$

so that on $\mathcal{R}$, we have (see (2.82))

$$|\rho_{i\ell} - H_\ell(x_i)| \le \zeta_n \quad \forall i \in S_2,\ \forall \ell \in [L].$$

Then, by Lemma 28, for all $k \in [K]$,

$$|\rho_{i\ell_k} - \rho_{j\ell_k}| \geq |H_{\ell_k}(x_i) - H_{\ell_k}(x_j)| - 2\zeta_n.$$

Using the fact that $a \geq b - c$ implies $a^2 \geq \frac{1}{2}b^2 - c^2$ for $b \geq 0$, we have

$$(\rho_{i\ell_k} - \rho_{j\ell_k})^2 \geq \frac{1}{2}[H_{\ell_k}(x_i) - H_{\ell_k}(x_j)]^2 - 4\zeta_n^2$$

Let $s_k := |\mathcal{G}_k|$ Recall that $d_+^k = \sum_{i \in \mathcal{G}_k} d_i$. Then, on $\mathcal{A}$, we have

$$\frac{d_i}{d_+^k} \geq \frac{0.16 C_8 \nu_n}{0.96 \nu_n s_k} = \frac{C_8}{6 s_k}.$$

It follows that on $\mathcal{A} \cap \mathcal{R}$, we have

$$\begin{aligned}
\varpi_k &= \frac{1}{2} \sum_{i,j \in \mathcal{G}_k} \frac{d_i}{d_+^k} \frac{d_j}{d_+^k} (\rho_{i\ell_k} - \rho_{j\ell_k})^2 \\
&\geq \frac{1}{2} \frac{C_8^2}{36 s_k^2} \sum_{i,j \in \mathcal{G}_k} \Big(\frac{1}{2}[H_{\ell_k}(x_i) - H_{\ell_k}(x_j)]^2 - 4\zeta_n^2 1\{i \neq j\} \Big) \\
&\geq \frac{1}{4} \frac{C_8^2}{36} \Big[\frac{1}{4\binom{s_k}{2}} \sum_{i,j \in \mathcal{G}_k} [H_{\ell_k}(x_i) - H_{\ell_k}(x_j)]^2 - 8\zeta_n^2 1\{i \neq j\} \Big]
\end{aligned} \tag{B.22}$$

since by (2.76) $s_k \geq 2$ and hence $4\binom{s_k}{2} \geq s_k^2$. The first term above is proportinal to a U-statistic providing an estimate of the variance of $H_{\ell_k}(x), x \sim \mathbb{Q}_k$ based on an i.i.d. sample $x_i \sim \mathbb{Q}_k, i \in \mathcal{G}_k$ (assuming that S_2 is fixed). An argument using the Hansen–Wright inequality shows that such a quantity is concentrated around its mean, which is the population variance. We use the following result from [Kaz+17], with slight modifications:

Proposition 7 (Corollary 3 in [Kaz+17]). *Let $w = (w_1, \ldots, w_m) \in \mathbb{R}^m$ be a random vector with independent components w_i which satisfy $\|w_i - \mathbb{E}w_i\|_{\psi_2} \leq K$. Let*

$$\mathrm{imp}(w) := \frac{1}{\binom{m}{2}} \sum_{1 \leq i,j \leq m} \frac{1}{4}(w_i - w_j)^2$$

be the empirical variance of w. Then, there is a universal constant $c > 0$ such that for $u \geq 0$,

$$\mathbb{P}\Big(\operatorname{imp}(w) - \mathbb{E}\operatorname{imp}(w) < -K^2 u \Big) \leq \exp\big\{ -c\,(m-1)\min(u, u^2) \big\}. \tag{B.23}$$

We note the alternative expression $\operatorname{imp}(w) = \binom{m}{2}^{-1} \sum_{1 \leq i < j \leq m} \frac{1}{2}(w_i - w_j)^2$. In the context of Proposition 7, if $w_1, \ldots, w_m$ are i.i.d., then

$$\mathbb{E}\operatorname{imp}(w) = \frac{1}{2}\mathbb{E}(w_1 - w_2)^2 = \operatorname{var}(w_1).$$

Since $H_{\ell_k}(\cdot)$ is bounded in $[0,1]$, we have $\|H_{\ell_k}(x_i) - \mathbb{E}H_{\ell_k}(x_i)\|_{\psi_2} \leq 1$. Recall that $\vartheta_{k\ell} = \operatorname{var}(H_\ell(x))$ when $x \sim \mathbb{Q}_k$. Then, conditional on $\mathcal{F}_0 = \sigma(S_1)$ so that $\mathcal{G}_k$ is fixed, we have for $i, j \in \mathcal{G}_k$ and $i \neq j$,

$$\frac{1}{2}\mathbb{E}[H_{\ell_k}(x_i) - H_{\ell_k}(x_j)]^2 = \vartheta_{k\ell_k}$$

Applying the Proposition 7, we obtain, for $u \in [0,1]$,

$$\mathbb{P}^{\mathcal{F}_0}\Big(\frac{1}{4\binom{s_k}{2}} \sum_{i,j \in \mathcal{G}_k} [H_{\ell_k}(x_i) - H_{\ell_k}(x_j)]^2 < \vartheta_{k\ell_k} - u \Big) \leq e^{-c(s_k - 1)u^2}$$

On $\mathcal{A}$, $s_k - 1 \geq s_k/2 \geq 0.2\tau_{\mathcal{C}} n$. Take $u = u_n := \sqrt{\log n/(\tau_{\mathcal{C}} n)}$. By the scaling assumption $\log n/n \leq \tau_{\mathcal{C}}$, we have $u_n \leq 1$, hence

$$\mathbb{P}^{\mathcal{F}_0}\Big(\frac{1}{4\binom{s_k}{2}} \sum_{i,j \in \mathcal{G}_k} [H_{\ell_k}(x_i) - H_{\ell_k}(x_j)]^2 < \vartheta_{k\ell_k} - u_n \Big) 1_{\mathcal{A} \cap \mathcal{R}} \leq n^{-c_1}$$

where $c_1 = 0.2c$. Combining with (B.22)

$$\mathbb{P}^{\mathcal{F}_0}\big(\frac{144}{C_8^2}\varpi_k + 4\zeta_n^2 < \vartheta_{k\ell_k} - u_n \big) 1_{\mathcal{A} \cap \mathcal{R}} \leq n^{-c_1}$$

Taking the union bound and removing the conditioning, we get

$$\mathbb{P}\big(\big\{\exists k \in [K],\ \frac{144}{C_8^2}\varpi_k + 4\zeta_n^2 < \vartheta_{k\ell_k} - u_n\big\} \cap \mathcal{A} \cap \mathcal{R}\big) \le Kn^{-c_1}$$

Let us call the first event above $\mathcal{H}^c$. Then, on $\mathcal{H}$, we have

$$\frac{144}{C_8^2}\varpi_k \ge \vartheta_{k\ell_k} - (4\zeta_n^2 + u_n), \quad \forall k \in [K]. \tag{B.24}$$

We have, using assumption $\zeta_n \le 1$,

$$4\zeta_n^2 + u_n \le 4\zeta_n + u_n \le \frac{16L}{\tau_{\mathcal{C}}\tau_h\tau_\theta}\sqrt{\frac{\log n}{n}} + \sqrt{\frac{\log n}{\tau_{\mathcal{C}} n}} \le \frac{18L}{\tau_{\mathcal{C}}\tau_h\tau_\theta}\sqrt{\frac{\log n}{n}} \tag{B.25}$$

since $\tau_{\mathcal{C}} \le 1$. □

B.4 Other Technical Results

Lemma 29. *Assume that Z is a random variable taking values $z_1, \ldots, z_R$ with probabilities $\hat{\beta}_1, \ldots, \hat{\beta}_R$ respectively. Then,* $\mathrm{var}(Z) \ge \frac{1}{2}\hat{\beta}_1\hat{\beta}_2(z_1 - z_2)^2$.

Proof. Let Z' be an independent copy of Z. Then $\mathrm{var}(Z) = \frac{1}{2}\mathbb{E}(Z - Z')^2$, and (Z, Z') takes value (z_1, z_2) with probability $\hat{\beta}_1\hat{\beta}_2$. The result follows. □

Lemma 30. *Let $X \sim \mathrm{Poi}(\lambda)$. Then, for any $t \in (0, 1]$,*

$$\mathbb{P}(|X - \lambda| \ge t\lambda) \le 2\exp(-\lambda t^2/4).$$

Proof of Lemma 30. Fix $t \in (0, 1]$. For $\theta \in (0, 1.79]$, by the Chernoff bound,

$$\mathbb{P}(X - \lambda \ge t\lambda) \le e^{-\theta t\lambda}\mathbb{E}[e^{(X-\lambda)\theta}] = e^{-\theta t\lambda}\exp(\lambda(e^\theta - 1 - \theta)) \le e^{\lambda\theta^2 - \theta t\lambda}$$

using $e^\theta - 1 - \theta \le \theta^2$ when $\theta \le 1.79$. Since $\lambda\theta^2 - \theta t\lambda$ attains its minimum at $\theta = t/2 \le 1$,

we obtain $\mathbb{P}(X - \lambda \geq t\lambda) \leq \exp(-\lambda t^2/4)$. On the other hand,

$$\mathbb{P}(\lambda - X \geq t\lambda) \leq e^{-\theta t\lambda}\mathbb{E}[e^{(\lambda - X)\theta}] = e^{-\theta t\lambda}\exp(\lambda(e^{-\theta} - 1 + \theta)) \leq e^{\lambda\theta^2/2 - \theta t\lambda}$$

using $e^{-\theta} - 1 + \theta \leq \theta^2/2$ for $\theta \geq 0$. Since $\lambda\theta^2/2 - \theta t\lambda$ attains its smallest value at $\theta = t \leq 1$, we get $\mathbb{P}(\lambda - X \geq t\lambda) \leq \exp(-\lambda t^2/2)$, finishing the proof. □

Proposition 8 (Giné and Nickl [GN15] Theorem 3.1.7)**.** *Let $S = \sum_{i=1}^n X_i$ where $\{X_i\}$ are independent random variables with $|X_i - \mathbb{E}X_i| \leq c$ for all i. Let $v \geq \mathrm{var}(S)$. Then, for all $u \geq 0$,*

$$\mathbb{P}\Big(|S - \mathbb{E}S| \geq \sqrt{2vu} + \frac{cu}{3}\Big) \leq 2e^{-u}.$$

In particular, if $S \sim \mathrm{Bin}(n, p)$, then we can take $v = \mathbb{E}[S] \geq \mathrm{var}(S)$. Letting $\widehat{p} = S/n$, the result gives $\mathbb{P}\big(|\widehat{p} - p| \geq \sqrt{\frac{2pu}{n}} + \frac{u}{3n}\big) \leq 2e^{-u}$.

Chapter C

Remaining Proofs in Chapter 3

C.1 Proofs of Propositions

C.1.1 Proof of Propostion 3

For $\alpha \in [0, \pi/2)$, consider a constellation of points in $\mathbb{R}^2$ at locations $a_1 = (\varepsilon \sin\alpha - \delta, \varepsilon\cos\alpha)$, $a_2 = (-\varepsilon\sin\alpha - \delta, -\varepsilon\cos\alpha)$, $b_1 = (-\varepsilon\sin\alpha, \varepsilon\cos\alpha)$ and $b_2 = (\varepsilon\sin\alpha, -\varepsilon\cos\alpha)$. Assume that $n/4$ of the data points are on each of the points a_1, a_2, b_1 and b_2. Assume that data points in $\{a_1, a_2\}$ form cluster 1 and points in $\{b_1, b_2\}$ form cluster 2. That is, this is the true cluster labels as specified by an external source. The true cluster centers are then at locations $\xi_1^* = (-\delta, 0)$ and $\xi_2^* = (0, 0)$. We also have $(\frac{1}{n}\sum_i ||x_i - \xi_{z_i}^*||^2)^{1/2} = \varepsilon$ for true cluster labels $\{z_i\}$. Now take $\delta = \varepsilon\sin\alpha$. Figure C.1 shows the geometry of this construction.

To show the result, it is enough to use Theorem 6 with properly chosen (fake) centers on the above dataset. In particular, we are going to show that a 2-factor k-means algorithm has a small misclassification rate with respect to a new clustering that puts points $\{a_1, b_1\}$ in one cluster and $\{a_2, b_2\}$ in another cluster. Consider "fake" centers $\xi_1^{**} = (a_1 + b_1)/2$ and $\xi_2^{**} = (a_2 + b_2)/2$. Then, the new separation is $\delta^* = 2\varepsilon\cos\alpha$ and the new deviation can be taken to be $\varepsilon^* = \delta/2 + \varepsilon\sin\alpha = (3/2)\varepsilon\sin\alpha$ guaranteeing that $(\frac{1}{n}\sum_i \|x_i - \xi_{y_i}^{**}\|^2)^{1/2} \leq \varepsilon^*$ where $\{y_i\}$ are labels relative to the new clustering.

Applying Theorem 6 with $\kappa = p = 2$, $c = 2.1$ and $\pi_{\min} = 1/2$, as long as $\delta^*/\varepsilon^* \geq 9 > 3\sqrt{2}c$, the misclassification rate to the new clustering is bounded above as $\text{Miss}^* \leq$

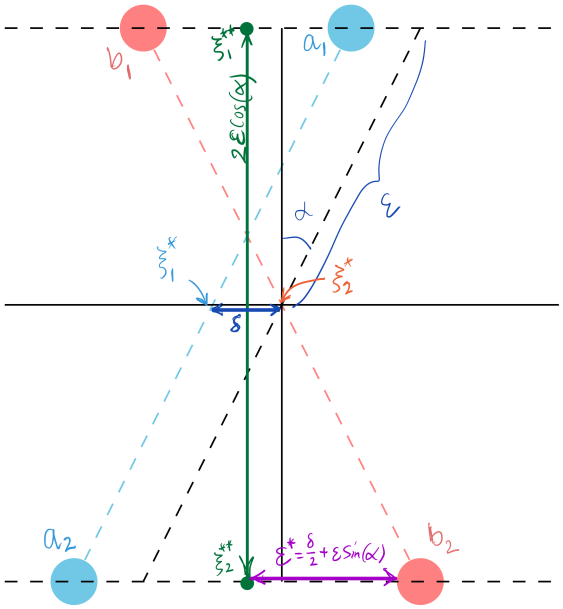


Figure C.1: The geometry of the dataset in Proposition 3

$80(\varepsilon^*/\delta^*)^2$. We have $\varepsilon^*/\delta^* = (3/4)\tan\alpha$. Thus, for $\alpha \le \tan^{-1}(4/27)$ we have $\text{Miss}^* \le 45(\tan\alpha)^2$ w.r.t. to clustering $\{\{a_1, b_1\}, \{a_2, b_2\}\}$. Hence, w.r.t. the original clustering, $\frac{1}{2} \ge \text{Miss} \ge \frac{1}{2} - 45(\tan\alpha)^2$, which can be made arbitrarily close to $\frac{1}{2}$ by choosing α small enough.

To see the last step above, let q_1, q_2, q_3, q_4 be the fractions of misclassified nodes from each of the four categories a_1, a_2, b_1, b_2, w.r.t. to the new clustering (i.e., $\{y_i\}$). The above argument shows that $\frac{1}{4}(q_1 + q_2 + q_3 + q_4) \le 45(\tan\alpha)^2$. The misclassification rate to the original clustering (i.e., $\{z_i\}$) is then

$$\text{Miss} = \frac{1}{n}\Big(\frac{n}{4}(1 - q_{i_1}) + \frac{n}{4}(1 - q_{i_2})\Big) = \frac{1}{2} - \frac{1}{4}(q_{i_1} + q_{i_2}) \ge \frac{1}{2} - 45(\tan\alpha)^2$$

where $\{i_1, i_2\}$ is a pair of distinct elements from $\{1, 2, 3, 4\}$. This proves the lower bound. The upper bound $\text{Miss} \le 1/2$ always holds due to the minimization over permutations in the definition of the misclassification rate.

Since for $\beta \in (0, 1/2)$, $\sqrt{\beta/45} \le 4/27$, we only need $\alpha \le \tan^{-1}(\sqrt{\beta/45})$ to have $\frac{1}{2} \ge \text{Miss} \ge \frac{1}{2} - \beta$. Recalling that $\delta/\varepsilon = \sin\alpha$, this shows that one can take $c_2(\beta) = \sin(\tan^{-1}(\sqrt{\beta/45}))$ in the statement of the lower bound.

For the claim regarding perfect recovery with $L = 4$ clusters, take $\xi_1^{**} = a_1$, $\xi_2^{**} = b_1$,

$\xi_3^{**} = a_2$ and $\xi_4^{**} = b_2$ and apply Theorem 1, noting that $\delta^* = \min_{i \neq j} \|\xi_i^{**} - \xi_j^{**}\| > 0$ while we can take $\varepsilon^* = 0$.

C.1.2 Proof of Proposition 4

Let m_k be the mean of $\mathbb{Q}_k$. Then, $\mathbb{P}(|t_i - m_{z_i}| > t) \leq 2e^{-t^2/2\sigma^2}$. Let $M = \sqrt{6\sigma^2 \log n}$. By union bound, with probability $\geq 1 - 2n^{-2}$ we have $|t_i - m_{z_i}| \leq M$ for all $i \in [n]$. We can cover the set $[-M, M] \subset \mathbb{R}$, with $L' = M/\varepsilon$ 1-D balls of radius ε. (Without loss of generality, we assume that L' is an integer for simplicity.) Let $\mathcal{T} = \{\tau_1, \ldots, \tau_{L'}\}$ one such cover and note that $m_k + \mathcal{T}$ is an ε-cover of $m_k + [-M, M]$. Let $\pi_k : \mathbb{R} \to (m_k + \mathcal{T})$ be the projection from $\mathbb{R}$ onto $m_k + \mathcal{T}$. Then, $\|\gamma_{z_i}(t_i) - \gamma_{z_i}(\pi_{z_i}(t_i))\| \leq \rho |t_i - \pi_{z_i}(t_i)| \leq \rho\varepsilon$, assuming that $\varepsilon \leq 1/\rho$.

Let $z_i' := \operatorname{argmin}_{\ell' \in [L']} |t_i - (m_{z_i} + \tau_{\ell'})|$ so that $\pi_{z_i}(t_i) = m_{z_i} + \tau_{z_i'}$. Then let $L_n = KL'$ and fix a bijection $\phi : [L_n] \to [K] \times [L']$ and define the labels $\widetilde{z}_i = \phi^{-1}(z_i, z_i')$. Also consider the map $\omega_0 : [K] \times [L'] \to [K]$ given by $\omega_0(k, \ell') = k$ and set $\widetilde{\omega} := \omega_0 \circ \phi$ which is a surjective map from $[L_n]$ to $[K]$ satisfying $\widetilde{\omega}(\widetilde{z}_i) = z_i$. For $\ell \in [L_n]$ with $\phi(\ell) = (k, \ell')$, define $\widetilde{\xi}_\ell = \gamma_k(m_k + \tau_{\ell'})$. Then, we have $\widetilde{\xi}_{\widetilde{z}_i} = \gamma_{z_i}(m_{z_i} + \tau_{z_i'}) = \gamma_{z_i}(\pi_{z_i}(t_i))$, hence the above argument gives $\|\gamma(t_i) - \widetilde{\xi}_{\widetilde{z}_i}\| \leq \rho\varepsilon$. It is also clear that the the separation condition (3.11) is satisfied since by construction if $\widetilde{\omega}(\ell_1) \neq \widetilde{\omega}(\ell_2)$ with $\phi(\ell_1) = (k_1, \ell_1')$ and $\phi(\ell_2) = (k_2, \ell_2')$, then $k_1 \neq k_2$ hence $\widetilde{\xi}_{\ell_1}$ and $\widetilde{\xi}_{\ell_2}$ lie on different manifolds (on $\mathcal{C}_{k_1}$ and $\mathcal{C}_{k_2}$). It follows that conclusion (3.12) of Theorem 7 holds for $p = 2$ and, say, $c = 3$ but with ε replaced with $\rho\varepsilon$. Take $\varepsilon = (c_1\sqrt{n})^{-1}$ for constant c_1 to be determined. Let $c_2 = 3\rho(1 + \kappa)/\delta$. As long as $n\pi_{\min} > (c_2/c_1)^2$, the separation condition in (3.12) is satisfied and we have $\mathrm{Mis}(z, \widehat{z}) \leq K(c_2/c_1)^2/n$. Hence, as long as $c_1 > \sqrt{K}c_2$, we will have $\mathrm{Mis}(z, \widehat{z}) < 1/n$ which implies $\mathrm{Mis}(z, \widehat{z}) = 0$. We also need to satisfy $\varepsilon < 1/\rho$ that is $c_1 \geq \rho/\sqrt{n}$. Taking $c_1 = \sqrt{K}c_2 + \rho$ satisfies all the required constraints on c_1. The required number of clusters is

$$L_n = KL' = KM/\varepsilon \leq 3K\sigma c_1 \sqrt{n \log n},$$

which proves the result with $C = 3K\sigma c_1$. Note that since $c_2/c_1 < 1$ and $n\pi_{\min} \geq 1$, the condition $n\pi_{\min} > (c_2/c_1)^2$ is automatically satisfied. The proof is complete.

C.1.3 Proof of Proposition 5

The proof follows that of Proposition 4. We only highlight the differences. When $z_i = k$, by Lemma 31, $\|t_i - m_k\|$ is sub-gaussian with parameter $\leq c_0\sigma\sqrt{r_k}$ for some universal constant $c_0 > 0$. Thus, we have $\mathbb{P}(\|t_i - m_k\| \geq t) \leq 2e^{-c_0 t^2/(r_k\sigma^2)}$. Let $M = \sqrt{3c_0 r\sigma^2 \log n}$. By union bound, with probability at least $1 - 2n^{-2}$, we have $\|t_i - m_{z_i}\| \leq M$ for all $i \in [n]$. The ε-cover has to be constructed for $\{u : \|u\| \leq M\}$ in the ℓ_2 norm, which can be done with a net of size at most $L' = (1 + 2M/\varepsilon)^r$. Take $\varepsilon = (c_1 n^{1/p})^{-1}$ and let $c_2 = 3\rho(1+\kappa)/\delta$. As long as $n\pi_{\min} > (c_2/c_1)^p$, the separation condition in (3.12) is satisfied and we have $\mathrm{Mis}(z, \widehat{z}) \leq K(c_2/c_1)^p/n$. Hence, as long as $c_1 > K^{1/p}c_2$, we will have $\mathrm{Mis}(z, \widehat{z}) < 1/n$ which implies $\mathrm{Mis}(z, \widehat{z}) = 0$. We also need to satisfy $\varepsilon < 1/\rho$ that is $c_1 \geq \rho/\sqrt{n}$. Taking $c_1 = K^{1/p}c_2 + \rho$ satisfies all the required constraints on c_1. The required number of clusters is

$$\begin{aligned} L_n = KL' = K(1 + 2M/\varepsilon)^r &= K(1 + 2c_1\sqrt{3c_0 r\sigma^2}n^{1/p}\sqrt{\log n})^r \\ &\leq C(n^{1/p}\sqrt{\log n})^r \end{aligned}$$

for $C = K(2 + 2c_1\sqrt{3c_0 r\sigma^2})^r$. Here, we have used $1 \leq 2n^{1/p}\sqrt{\log n}$ for $n \geq 2$. Note that since $c_2/c_1 < 1$ and $n\pi_{\min} \geq 1$, the condition $n\pi_{\min} > (c_2/c_1)^p$ is automatically satisfied. The proof is complete.

C.2 Proof of Lemmas

Proof of Lemma 20. The proof is based on the connection between the sub-gaussian and sub-exponential norm. First recall the sub-gaussian norm of X is defined as $\|X\|_{\psi_2} = \sup_{u \in S^{d-1}} \|u^T X\|_{\psi_2}$, where $\|\cdot\|_{\psi_2}$ denotes the sub-gaussian norm of a random variable and S^{d-1} the unit sphere in $\mathbb{R}^d$. Alternatively, we can define a sub-gaussian vector with parameter σ, as a random vector satisfying $\mathbb{P}(|u^T X| \geq t) \leq 2\exp(-\frac{t^2}{2\sigma^2})$ for all $u \in S^{d-1}$ and $t \geq 0$. We will have $\sigma \asymp \|X\|_{\psi_2}$. We also use $\|\cdot\|_{\psi_1}$ for the sub-exponential norm of a random variable. For any random variable, we have $\|Y^2\|_{\psi_1} = \|Y\|_{\psi_2}^2$ [Ver18, Lemma 2.7.6].

Below we apply this fact with $Y = \|X\| = (\sum_{i=1}^d X_i^2)^{1/2}$, leading to the following useful lemma.

Lemma 31. *Assume that $X \in \mathbb{R}^d$ is a sub-gaussian random vector with parameter σ. Then, $\|X\|$ is sub-gaussian with parameter $\lesssim \sigma\sqrt{d}$. In fact, for some universal constant $C > 0$,*

$$\|\|X\|\|_{\psi_2} \le C\sigma\sqrt{d}, \quad \|\|X\|^2\|_{\psi_1} \le C^2\sigma^2 d.$$

Proof. We have $\|\|X\|^2\|_{\psi_1} \le \sum_{i=1}^d \|X_i^2\|_{\psi_1} = \sum_{i=1}^d \|X_i\|_{\psi_2}^2 \le dC^2\sigma^2$, for some universal constant $C^2 > 0$. The first inequality is the triangle inequality for $\|\cdot\|_{\psi_1}$ and the second by the equivalence of the sub-gaussian norm and sub-gaussian parameter. Next, we note that $\|\|X\|\|_{\psi_2} = \sqrt{\|\|X\|^2\|_{\psi_1}}$ and the result follows. □

Then, by Lemma 31, $\|w_i\|^2/d$ is sub-exponential with sub-exponential norm $\lesssim \sigma_i^2$. By the Bernstein inequality for sub-exponential variables [Ver18, Corollary 2.8.3],

$$\mathbb{P}\Big(\frac{1}{n}\Big(\sum_{i=1}^n \frac{\|w_i\|^2}{d} - \alpha_i^2\Big) > t\Big) \le \exp\Big(-cn\min\Big(\frac{t^2}{\sigma_{\max}^4}, \frac{t}{\sigma_{\max}^2}\Big)\Big).$$

Let $t = \bar{\alpha}_n$, and recall that $\bar{\alpha}_n^2/\sigma_{\max}^2 \le C$. Then the result follows. □

Proof of Lemma 21. Recall that $\widehat{\xi}$ is the output of ALG for L clusters. Let $\widehat{\xi}^{(K)}$ be the output of the ALG for K clusters. Then, since $L \ge K$,

$$\widehat{Q}(\widehat{\xi}) \le \widehat{Q}(\widehat{\xi}^{(K)}) \le \kappa\, \widehat{Q}_{\min}^{(K)}, \quad \text{where} \quad \widehat{Q}_{\min}^{(K)} := \min_{\xi \in \mathcal{X}^K} \widehat{Q}(\xi).$$

The first inequality is by the monotonicity of ALG and the second by its constant-factor approximation property. Since by assumption $\xi^* \in \mathcal{X}^K$, we have

$$\widehat{Q}_{\min}^{(K)} \le \widehat{Q}(\xi^*) \le \Big(\frac{1}{n}\sum_{i=1}^n \|x_i - \xi_{z_i}^*\|^p\Big)^{1/p} \le \varepsilon.$$

It follows that $\widehat{Q}(\widehat{\xi}) \le \kappa\varepsilon$. Recalling (3.14) and noting that $\widehat{Q}(\widehat{\xi}) = \big(\frac{1}{n}\sum_{i=1}^n \|x_i - \widehat{\xi}_{\widehat{z}_i}\|^p\big)^{1/p}$,

we have

$$
\begin{aligned}
Q(\widehat{\xi};\mu^*) = \Big(\frac{1}{n}\sum_{i=1}^{n}\min_{\ell\in[L]}\|\xi^*_{z_i}-\widehat{\xi}_\ell\|^p\Big)^{1/p} &\le \Big(\frac{1}{n}\sum_{i=1}^{n}\|\xi^*_{z_i}-\widehat{\xi}_{\widehat{z}_i}\|^p\Big)^{1/p} \\
&\le \Big(\frac{1}{n}\sum_{i=1}^{n}\|x_i-\xi^*_{z_i}\|^p\Big)^{1/p} + \Big(\frac{1}{n}\sum_{i=1}^{n}\|x_i-\widehat{\xi}_{\widehat{z}_i}\|^p\Big)^{1/p} \\
&\le \varepsilon + \kappa\varepsilon \qquad \text{(C.1)}
\end{aligned}
$$

where the second line is the triangle inequality in the aforementioned $L^p(\nu_n, \mathcal{X})$ space. The proof is complete. □

Proof of Lemma 22. Consider the partition of the space by the Voronoi cells of $\xi = (\xi_\ell)$. Assume first that there is a Voronoi cell that contains at least two distinct elements of ξ^*, e.g., $\xi^*_{k_1}$ and $\xi^*_{k_2}$, with $k_1 \neq k_2$, both belonging to the Voronoi cell of ξ_{ℓ_0}. That is, $\min_\ell \|\xi^*_k - \xi_\ell\| = \|\xi^*_k - \xi^*_{\ell_0}\|$ for $k = k_1, k_2$. As $\|\xi^*_{k_1} - \xi^*_{k_2}\| \le \|\xi^*_{k_2} - \xi^*_{\ell_0}\| + \|\xi^*_{k_1} - \xi^*_{\ell_0}\|$, at least one of the $k = k_1, k_2$ satisfy $\|\xi^*_k - \xi^*_{\ell_0}\| \ge \|\xi^*_{k_1} - \xi^*_{k_2}\|/2$, and assume this is true for $k = k_1$, we have

$$
Q(\xi;\mu^*) \ge \pi_{\min}^{1/p}\|\xi^*_{k_1} - \xi_{\ell_0}\| \ge \frac{\pi_{\min}^{1/p}}{2}\|\xi^*_{k_1} - \xi^*_{k_2}\| \ge \frac{\pi_{\min}^{1/p}}{2}\delta.
$$

Otherwise, each Voronoi cell of ξ contains at most one element of ξ^*. On the other hand, each element of ξ^* belongs to at least one Voronoi cell of ξ, since the union of Voronoi cells is the whole space. It follows that there are K distinct Voronoi cells of ξ, each of which contains exactly one element of ξ^*. Thus, there is an injective map $\sigma : [K] \to [L]$ such that ξ^*_k belongs to Voronoi cell of $\xi_{\sigma(k)}$, that is, $\min_\ell \|\xi^*_k - \xi_\ell\| = \|\xi^*_k - \xi_{\sigma(k)}\|$. Then,

$$
Q(\xi;\mu^*) \ge \pi_{\min}^{1/p}\Big(\sum_{k=1}^{K}\|\xi^*_k - \xi_{\sigma(k)}\|^p\Big)^{1/p} \ge \pi_{\min}^{1/p}\, d_p(\xi,\xi^*).
$$

The proof is complete. □

Proof of Lemma 23. By assumption, there exists an injective map $\sigma : [K] \to [L]$ such

that

$$\max_{k\in K} \|\xi_k^* - \widehat{\xi}_{\sigma(k)}\| \leq \gamma.$$

Then, σ is invertible on $\mathrm{Im}(\sigma) := \{\sigma(k) : k \in [K]\}$, with an inverse denoted as σ^{-1}. We obtain

$$\|\xi^*_{\sigma^{-1}(\ell)} - \widehat{\xi}_\ell\| \leq \gamma, \quad \forall \ell \in \mathrm{Im}(\sigma). \tag{C.2}$$

First assume that $\widehat{z}_i \in \mathrm{Im}(\sigma)$. We prove that $\sigma(z_i) = \widehat{z}_i$ by contradiction. Suppose that $\sigma(z_i) \neq \widehat{z}_i$. Then, we show that $\|x_i - \widehat{\xi}_{\sigma(z_i)}\| < \|x_i - \widehat{\xi}_{\widehat{z}_i}\|$ contradicting $\widehat{z}_i = \operatorname{argmin}_\ell \|x_i - \widehat{\xi}_\ell\|^2$. By the triangle inequality

$$\|x_i - \widehat{\xi}_{\sigma(z_i)}\| \leq \|x_i - \xi^*_{z_i}\| + \|\widehat{\xi}_{\sigma(z_i)} - \xi^*_{z_i}\| \leq \eta + \gamma. \tag{C.3}$$

Since $\widehat{z}_i \in \mathrm{Im}(\sigma)$ and $\sigma(z_i) \neq \widehat{z}_i$, we have $\sigma^{-1}(\widehat{z}_i) \neq z_i$. By (C.2), $\|\widehat{\xi}_{\widehat{z}_i} - \xi^*_{\sigma^{-1}(\widehat{z}_i)}\| \leq \gamma$. Therefore,

$$\begin{aligned}
\|x_i - \widehat{\xi}_{\widehat{z}_i}\| &\geq \|\widehat{\xi}_{\widehat{z}_i} - \xi^*_{z_i}\| - \|x_i - \xi^*_{z_i}\| \\
&\geq \|\xi^*_{z_i} - \xi^*_{\sigma^{-1}(\widehat{z}_i)}\| - \|\widehat{\xi}_{\widehat{z}_i} - \xi^*_{\sigma^{-1}(\widehat{z}_i)}\| - \eta \\
&\geq \delta - \gamma - \eta.
\end{aligned} \tag{C.4}$$

Since by assumption $\delta > 2\gamma + 2\eta$, the claimed contradiction follows by combining (C.3) and (C.4). Hence, we have $\sigma(z_i) = \widehat{z}_i$ when $\widehat{z}_i \in \mathrm{Im}(\sigma)$. Define $\omega(\cdot) = \sigma^{-1}(\cdot)$ on $\mathrm{Im}(\sigma) \subset [L]$. Then, ω satisfies $\omega(\widehat{z}_i) = z_i$ whenever $\widehat{z}_i \in \mathrm{Im}(\sigma)$. This finishes proof for the case $L = K$.

Next, we define ω for $\ell_0 \notin \mathrm{Im}(\sigma)$. Since $\widehat{\xi}$ is an efficient solution, there exists at least one $i \in [n]$ such that $\widehat{z}_i = \ell_0$. When there is only one such i, we can just let $\omega(\ell_0) = \omega(\widehat{z}_i) = z_i$. When there are at least two data points x_i and x_j such that $\widehat{z}_i = \widehat{z}_j = \ell_0$, we are going to show, by contradiction, that their true cluster labels must be the same, i.e., $z_i = z_j$. Suppose that $z_i \neq z_j$, then we will show that $\|x_i - \widehat{\xi}_{\ell_0}\| > \|x_i - \widehat{\xi}_{\sigma(z_i)}\|$ which contradicts x_i being in the Voronoi cell of $\widehat{\xi}_{\ell_0}$. Inequality (C.3) still holds in this

case. Furthermore

$$\begin{aligned}\|x_i - \widehat{\xi}_{\ell_0}\| &\geq \|\widehat{\xi}_{\ell_0} - \xi^*_{z_i}\| - \|x_i - \xi^*_{z_i}\| \\ &\geq \|x_j - \xi^*_{z_i}\| - \|x_j - \widehat{\xi}_{\ell_0}\| - \eta \\ &\geq \|\xi^*_{z_i} - \xi^*_{z_j}\| - \|x_j - \xi^*_{z_j}\| - \|x_j - \widehat{\xi}_{\ell_0}\| - \eta \\ &\geq \delta - 2\eta - \|x_j - \widehat{\xi}_{\ell_0}\|. \end{aligned} \tag{C.5}$$

Since x_j is in the Voronoi cell of $\widehat{\xi}_{\ell_0}$ and $\ell_0 \notin \mathrm{Im}(\sigma)$, we have $\ell_0 \neq \sigma(z_j)$. Therefore,

$$\begin{aligned}\|x_j - \widehat{\xi}_{\ell_0}\| &\leq \|x_j - \widehat{\xi}_{\sigma(z_j)}\| \\ &\leq \|x_j - \xi^*_{z_j}\| + \|\widehat{\xi}_{\sigma(z_j)} - \xi^*_{z_j}\| \\ &\leq \eta + \gamma. \end{aligned} \tag{C.6}$$

Combining inequalities (C.3), (C.5) and (C.6) and using the assumption $\delta > 2\gamma + 4\eta$, we get

$$\|x_i - \widehat{\xi}_{\ell_0}\| \ \geq \ \delta - 3\eta - \gamma \ > \ \eta + \gamma \ \geq \ \|x_i - \widehat{\xi}_{\sigma(z_i)}\|$$

which is the claimed contradiction. Therefore, we can define ω on $[L] \setminus \mathrm{Im}(\sigma)$ so that $\omega(\widehat{z}_i) = z_i$ when $\widehat{z}_i \notin \mathrm{Im}(\sigma)$. Combining with the definition of ω on $\mathrm{Im}(\sigma)$, we have successfully constructed a surjective map $\omega : [L] \to [K]$ satisfying $\omega(\widehat{z}_i) = z_i$ for all $i \in [n]$. The proof is complete. □

Bibliography

[Abb17] E. Abbe. "Community detection and stochastic block models: recent developments". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 6446–6531.

[Abb18] E. Abbe. "Community Detection and Stochastic Block Models: Recent Developments". In: *Journal of Machine Learning Research* 18.177 (2018), pp. 1–86.

[Abb+20] E. Abbe, J. Fan, K. Wang, Y. Zhong, et al. "Entrywise eigenvector analysis of random matrices with low expected rank". In: *Annals of Statistics* 48.3 (2020), pp. 1452–1474.

[ABH16] E. Abbe, A. S. Bandeira, and G. Hall. "Exact Recovery in the Stochastic Block Model". In: *IEEE Transactions on Information Theory* 62.1 (2016), pp. 471–487.

[ABS10] P. Awasthi, A. Blum, and O. Sheffet. "Stability yields a PTAS for k-median and k-means clustering". In: *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE. 2010, pp. 309–318.

[AG05] L. A. Adamic and N. Glance. "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog". In: *Proceedings of the 3rd International Workshop on Link Discovery*. LinkKDD '05. Chicago, Illinois: Association for Computing Machinery, 2005, 36–43.

[AH10] S. Anders and W. Huber. "Differential expression analysis for sequence count data". In: *Genome Biology* 11.10 (2010), R106.

[AJD02] C. Ahn, S.-H. Jung, and A. Donner. "Application of an adjusted $\chi 2$ statistic to site-specific data in observational dental studies". In: *Journal of clinical periodontology* 29.1 (2002), pp. 79–82.

[AL18] A. A. Amini and E. Levina. "On semidefinite relaxations for the block model". In: *Ann. Statist.* 46.1 (Feb. 2018), pp. 149–179.

[Ami+13] A. A. Amini, A. Chen, P. J. Bickel, and E. Levina. "Pseudo-likelihood methods for community detection in large sparse networks". In: *Ann. Statist.* 41.4 (Aug. 2013), pp. 2097–2122.

[AR21] A. A. Amini and Z. S. Razaee. "Concentration of kernel matrices with application to kernel spectral clustering". In: *The Annals of Statistics* 49.1 (2021), pp. 531–556.

[AS15] E. Abbe and C. Sandon. "Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic BP, and the information-computation gap". In: *arXiv preprint arXiv:1512.09080* (2015).

[AV06] D. Arthur and S. Vassilvitskii. *k-means++: The advantages of careful seeding*. Tech. rep. Stanford, 2006.

[Awa+15] P. Awasthi et al. "Relax, no need to round: Integrality of clustering formulations". In: *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. 2015, pp. 191–200.

[BC09] P. J. Bickel and A. Chen. "A nonparametric view of network models and Newman–Girvan and other modularities". In: *Proceedings of the National Academy of Sciences* 106.50 (2009), pp. 21068–21073. eprint: `https://www.pnas.org/content/106/50/21068.full.pdf`.

[Bic+13] P. Bickel, D. Choi, X. Chang, and H. Zhang. "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels". In: *Ann. Statist.* 41.4 (Aug. 2013), pp. 1922–1943.

[BS16] P. J. Bickel and P. Sarkar. “Hypothesis testing for automated community detection in networks”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.1 (2016), pp. 253–273. eprint: `https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12117`.

[CAS17] V. Cohen-Addad and C. Schwiegelshohn. “On the local structure of stable clustering instances”. In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2017, pp. 49–60.

[CCT12] K. Chaudhuri, F. Chung, and A. Tsiatas. “Spectral clustering of graphs with general degrees in the extended planted partition model”. In: *Conference on Learning Theory*. 2012, pp. 35–1.

[CL18] K. Chen and J. Lei. “Network cross-validation for determining the number of communities in network data”. In: *Journal of the American Statistical Association* 113.521 (2018), pp. 241–251.

[CRV15] P. Chin, A. Rao, and V. Vu. “Stochastic Block Model and Community Detection in Sparse Graphs: A spectral algorithm with optimal rate of recovery”. In: ed. by P. Grünwald, E. Hazan, and S. Kale. Vol. 40. Proceedings of Machine Learning Research. Paris, France: PMLR, 2015, pp. 391–423.

[Dec+11] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. “Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications”. In: *Phys. Rev. E* 84 (6 2011), p. 066106.

[DPR08] J.-J. Daudin, F. Picard, and S. Robin. “A mixture model for random graphs”. In: *Statistics and computing* 18.2 (2008), pp. 173–183.

[EK10] N. El Karoui. “On information plus noise kernel random matrices”. In: *The Annals of Statistics* 38.5 (2010), pp. 3191–3216.

[FC18] Y. Fei and Y. Chen. “Hidden integrality of SDP relaxations for sub-Gaussian mixture models”. In: *Conference On Learning Theory*. PMLR. 2018, pp. 1931–1965.

[FC19] Y. Fei and Y. Chen. “Achieving the Bayes Error Rate in Stochastic Block Model by SDP, Robustly”. In: ed. by A. Beygelzimer and D. Hsu. Vol. 99. Proceedings of Machine Learning Research. Phoenix, USA: PMLR, 2019, pp. 1235–1269.

[Fis+13] D. E. Fishkind et al. “Consistent Adjacency-Spectral Partitioning for the Stochastic Block Model When the Model Parameters Are Unknown”. In: *SIAM Journal on Matrix Analysis and Applications* 34.1 (2013), pp. 23–39. eprint: `https://doi.org/10.1137/120875600`.

[Gao+17] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. “Achieving optimal misclassification proportion in stochastic block models”. In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 1980–2024.

[Gao+18] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. “Community detection in degree-corrected block models”. In: *Ann. Statist.* 46.5 (Oct. 2018), pp. 2153–2185.

[GBP19] J. Geng, A. Bhattacharya, and D. Pati. “Probabilistic Community Detection With Unknown Number of Communities”. In: *Journal of the American Statistical Association* 114.526 (2019), pp. 893–905. eprint: `https://doi.org/10.1080/01621459.2018.1458618`.

[GBW20] L. L. Gao, J. Bien, and D. Witten. “Selective inference for hierarchical clustering”. In: *arXiv preprint arXiv:2012.02936* (2020).

[GHW79] G. H. Golub, M. Heath, and G. Wahba. “Generalized cross-validation as a method for choosing a good ridge parameter”. In: *Technometrics* 21.2 (1979), pp. 215–223.

[GN15] E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.

[GV19] C. Giraud and N. Verzelen. “Partial recovery bounds for clustering with the relaxed K-means”. In: *Mathematical Statistics and Learning* 1.3 (2019), pp. 317–374.

[HGH08] D. R. Hunter, S. M. Goodreau, and M. S. Handcock. “Goodness of fit of social network models”. In: *Journal of the American Statistical Association* 103.481 (2008), pp. 248–258.

[HLL83] P. W. Holland, K. B. Laskey, and S. Leinhardt. “Stochastic blockmodels: First steps”. In: 1983.

[HO93] P. C. Hansen and D. P. O’Leary. “The use of the L-curve in the regularization of discrete ill-posed problems”. In: *SIAM journal on scientific computing* 14.6 (1993), pp. 1487–1503.

[HW08] J. M. Hofman and C. H. Wiggins. “Bayesian approach to network modularity”. In: *Physical review letters* 100.25 (2008), p. 258701.

[Igu+17] T. Iguchi, D. G. Mixon, J. Peterson, and S. Villar. “Probably certifiably correct k-means clustering”. In: *Mathematical Programming* 165.2 (2017), pp. 605–642.

[JAD01] S.-H. Jung, C. Ahn, and A. Donner. “Evaluation of an adjusted chi-square statistic as applied to observational studies involving clustered binary data”. In: *Statistics in medicine* 20.14 (2001), pp. 2149–2161.

[Jin15] J. Jin. “Fast community detection by score”. In: *Annals of Statistics* 43.1 (2015), pp. 57–89.

[JVZ20] T. Jiang, S. Vavasis, and C. W. Zhai. “Recovery of a Mixture of Gaussians by Sum-of-norms Clustering”. In: *Journal of Machine Learning Research* 21.225 (2020), pp. 1–16.

[JY16] A. Joseph and B. Yu. “Impact of regularization on spectral clustering”. In: *Ann. Statist.* 44.4 (Aug. 2016), pp. 1765–1791.

[Kan+04] T. Kanungo et al. “A local search approximation algorithm for k-means clustering”. In: *Computational Geometry* 28.2-3 (2004), pp. 89–112.

[Kar+16] V. Karwa et al. “Monte Carlo goodness-of-fit tests for degree corrected and related stochastic blockmodels”. In: *arXiv preprint arXiv:1612.06040* (2016).

[Kaz+17] J. Kazemitabar, A. Amini, A. Bloniarz, and A. S. Talwalkar. “Variable importance using decision trees”. In: *Advances in neural information processing systems* 30 (2017).

[Kim+17] P. K. Kimes, Y. Liu, D. Neil Hayes, and J. S. Marron. “Statistical significance for hierarchical clustering”. In: *Biometrics* 73.3 (2017), pp. 811–821.

[Kło20] M. A. Kłopotek. “On the Consistency of k-means++ algorithm”. In: *Fundamenta Informaticae* 172.4 (2020), pp. 361–377.

[KN11] B. Karrer and M. E. J. Newman. “Stochastic blockmodels and community structure in networks”. In: *Phys. Rev. E* 83 (1 Jan. 2011), p. 016107.

[Krz+13] F. Krzakala et al. “Spectral redemption in clustering sparse networks”. In: *Proceedings of the National Academy of Sciences* 110.52 (2013), pp. 20935–20940. eprint: `https://www.pnas.org/content/110/52/20935.full.pdf`.

[KSS04] A. Kumar, Y. Sabharwal, and S. Sen. “A simple linear time (1+/spl epsiv/)-approximation algorithm for k-means clustering in any dimensions”. In: *45th Annual IEEE Symposium on Foundations of Computer Science*. IEEE. 2004, pp. 454–462.

[LCX18] X. Li, Y. Chen, and J. Xu. “Convex relaxation methods for community detection”. In: *arXiv preprint arXiv:1810.00315* (2018).

[Lei16] J. Lei. “A goodness-of-fit test for stochastic block models”. In: *Ann. Statist.* 44.1 (Feb. 2016), pp. 401–424.

[Li+20a] T. Li et al. “Hierarchical community detection by recursive partitioning”. In: *Journal of the American Statistical Association* (2020), pp. 1–18.

[Li+20b] X. Li et al. “When do birds of a feather flock together? k-means, proximity, and conic programming”. In: *Mathematical Programming* 179.1 (2020), pp. 295–341.

[Lin02] T. Linder. “Learning-theoretic methods in vector quantization”. In: *Principles of nonparametric learning*. Springer, 2002, pp. 163–210.

[LL22] C. M. Le and E. Levina. “Estimating the number of communities by spectral methods”. In: *Electronic Journal of Statistics* 16.1 (2022), pp. 3315–3342.

[Llo82] S. Lloyd. “Least squares quantization in PCM”. In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[LLV17] C. M. Le, E. Levina, and R. Vershynin. “Concentration and regularization of random graphs”. In: *Random Structures & Algorithms* 51.3 (2017), pp. 538–561. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/rsa.20713`.

[LLZ20] T. Li, E. Levina, and J. Zhu. “Network cross-validation by edge sampling”. In: *Biometrika* 107.2 (Apr. 2020), pp. 257–276. eprint: `https://academic.oup.com/biomet/article-pdf/107/2/257/33218033/asaa006.pdf`.

[LR15a] J. Lei and A. Rinaldo. “Consistency of spectral clustering in stochastic block models”. In: *Ann. Statist.* 43.1 (Feb. 2015), pp. 215–237.

[LR15b] J. Lei and A. Rinaldo. “Consistency of spectral clustering in stochastic block models”. In: *Annals of Statistics* 43.1 (2015), pp. 215–237.

[LS19] S. Lattanzi and C. Sohler. “A better k-means++ algorithm via local search”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 3662–3671.

[LZ16] Y. Lu and H. H. Zhou. “Statistical and computational guarantees of Lloyd’s algorithm and its variants”. In: *arXiv preprint arXiv:1612.02099* (2016).

[LZ17] J. Lei and L. Zhu. “Generic Sample Splitting For Refined Community Recovery In Degree Corrected Stochastic Block Models”. In: *Statistica Sinica* 27.4 (2017), pp. 1639–1659.

[Mac67] J. MacQueen. “Some methods for classification and analysis of multivariate observations”. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA. 1967, pp. 281–297.

[Mat00] J. Matoušek. "On approximate geometric k-clustering". In: *Discrete & Computational Geometry* 24.1 (2000), pp. 61–84.

[MNS16a] E. Mossel, J. Neeman, and A. Sly. "Belief propagation, robust reconstruction and optimal recovery of block models". In: *Ann. Appl. Probab.* 26.4 (Aug. 2016), pp. 2211–2256.

[MNS16b] E. Mossel, J. Neeman, and A. Sly. "Consistency thresholds for binary symmetric block models". In: *Electronic Journal of Probability* 21 (2016).

[MRS20] K. Makarychev, A. Reddy, and L. Shan. "Improved Guarantees for k-means++ and k-means++ Parallel". In: *Advances in Neural Information Processing Systems* 33 (2020).

[MS12] M. Mørup and M. N. Schmidt. "Bayesian Community Detection". In: *Neural Computation* 24.9 (2012). PMID: 22509971, pp. 2434–2456. eprint: `https://doi.org/10.1162/NECO_a_00314`.

[MSZ18] S. Ma, L. Su, and Y. Zhang. "Determining the Number of Communities in Degree-corrected Stochastic Block Models". In: *arXiv preprint arXiv:1809.01028* (2018).

[MVW16] D. G. Mixon, S. Villar, and R. Ward. "Clustering subgaussian mixtures with k-means". In: *2016 IEEE Information Theory Workshop (ITW)*. IEEE. 2016, pp. 211–215.

[NG04] M. E. Newman and M. Girvan. "Finding and Evaluating Community Structure in Networks". In: *Physical review. E, Statistical, nonlinear, and soft matter physics* 69 (Mar. 2004), p. 026113.

[NR16] M. E. Newman and G. Reinert. "Estimating the number of communities in a network". In: *Physical review letters* 117.7 (2016), p. 078301.

[NW15] A. Nellore and R. Ward. "Recovery guarantees for exemplar-based clustering". In: *Information and Computation* 245 (2015), pp. 165–180.

[PAL19] M. S. Paez, A. A. Amini, and L. Lin. “Hierarchical stochastic block model for community detection in multiplex networks”. In: *arXiv preprint arXiv:1904.05330* (2019).

[Pan+17] A. Panahi, D. Dubhashi, F. D. Johansson, and C. Bhattacharyya. “Clustering by sum of norms: Stochastic incremental algorithm, convergence and cluster recovery”. In: *International conference on machine learning*. PMLR. 2017, pp. 2769–2777.

[Pol81] D. Pollard. “Strong consistency of k-means clustering”. In: *The Annals of Statistics* (1981), pp. 135–140.

[Pol82] D. Pollard. “Quantization and the method of k-means”. In: *IEEE Transactions on Information theory* 28.2 (1982), pp. 199–205.

[PV18] S. L. van der Pas and A. W. van der Vaart. “Bayesian Community Detection”. In: *Bayesian Anal.* 13.3 (Sept. 2018), pp. 767–796.

[PW07] J. Peng and Y. Wei. “Approximating k-means-type clustering via semidefinite programming”. In: *SIAM journal on optimization* 18.1 (2007), pp. 186–205.

[QR13] T. Qin and K. Rohe. “Regularized spectral clustering under the degree-corrected stochastic blockmodel”. In: *Advances in neural information processing systems*. 2013, pp. 3120–3128.

[RCY11a] K. Rohe, S. Chatterjee, and B. Yu. “Spectral clustering and the high-dimensional stochastic blockmodel”. In: *The Annals of Statistics* 39.4 (2011), pp. 1878–1915.

[RCY11b] K. Rohe, S. Chatterjee, and B. Yu. “Spectral clustering and the high-dimensional stochastic blockmodel”. In: *Ann. Statist.* 39.4 (Aug. 2011), pp. 1878–1915.

[Ree04] J. F. Reed. “Adjusted chi-square statistics: application to clustered binary data in primary care”. In: *The Annals of Family Medicine* 2.3 (2004), pp. 201–203.

[Rio+17] M. A. Riolo, G. T. Cantwell, G. Reinert, and M. E. Newman. “Efficient method for estimating the number of communities in a network”. In: *Physical review e* 96.3 (2017), p. 032310.

[She10] I. Shevtsova. “An Improvement of Convergence Rate Estimates in the Lyapunov Theorem”. In: *Doklady Mathematics* 82 (Dec. 2010), pp. 862–864.

[SKZ14] A. Saade, F. Krzakala, and L. Zdeborová. “Spectral Clustering of graphs with the Bethe Hessian”. In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 406–414.

[SN97] T. A. Snijders and K. Nowicki. “Estimation and prediction for stochastic blockmodels for graphs with latent block structure”. In: *Journal of classification* 14.1 (1997), pp. 75–100.

[STY21] D. Sun, K.-C. Toh, and Y. Yuan. “Convex clustering: model, theoretical guarantee and efficient algorithm”. In: *Journal of Machine Learning Research* 22.9 (2021), pp. 1–32.

[Suw+16] S. Suwan et al. “Empirical Bayes estimation for the stochastic blockmodel”. In: *Electron. J. Statist.* 10.1 (2016), pp. 761–782.

[TMP12] A. L. Traud, P. J. Mucha, and M. A. Porter. “Social structure of Facebook networks”. In: *Physica A: Statistical Mechanics and its Applications* 391.16 (2012), pp. 4165 –4180.

[Tra+11] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter. “Comparing community structure to characteristics in online collegiate social networks”. In: *SIAM review* 53.3 (2011), pp. 526–543.

[Vaa98] A. W. v. d. Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[Ver18] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge University Press, 2018.

[WB17] Y. X. R. Wang and P. J. Bickel. “Likelihood-based model selection for stochastic block models”. In: *Ann. Statist.* 45.2 (Apr. 2017), pp. 500–528.

[Wei16] D. Wei. "A constant-factor bi-criteria approximation guarantee for k-means++". In: *Advances in Neural Information Processing Systems* 29 (2016), pp. 604–612.

[Yan+14a] X. Yan et al. "Model selection for degree-corrected block models". In: *Journal of Statistical Mechanics: Theory and Experiment* 2014.5 (2014), P05007.

[Yan+14b] X. Yan et al. "Model selection for degree-corrected block models". In: *Journal of Statistical Mechanics: Theory and Experiment* 2014.5 (May 2014), P05007.

[YFS18] M. Yuan, Y. Feng, and Z. Shang. *A likelihood-ratio type test for stochastic block models with bounded degrees*. 2018. arXiv: `1807.04426 [stat.ME]`.

[YP14] S.-Y. Yun and A. Proutiere. "Accurate community detection in the stochastic block model via spectral algorithms". In: *arXiv preprint arXiv:1412.7335* (2014).

[ZA19] Z. Zhou and A. A. Amini. "Analysis of spectral clustering algorithms for community detection: the general bipartite setting." In: *J. Mach. Learn. Res.* 20 (2019), pp. 47–1.

[ZA20a] L. Zhang and A. A. Amini. *Adjusted chi-square test for degree-corrected block models: Experiments in R*. `https://github.com/linfanz/nac-test`. 2020.

[ZA20b] Z. Zhou and A. A. Amini. "Optimal Bipartite Network Clustering." In: *Journal of Machine Learning Research* 21.40 (2020), pp. 1–68.

[Zhu+14] C. Zhu, H. Xu, C. Leng, and S. Yan. "Convex optimization procedure for clustering: Theoretical revisit". In: *Advances in Neural Information Processing Systems* 27 (2014), pp. 1619–1627.

[ZLZ12] Y. Zhao, E. Levina, and J. Zhu. "Consistency of community detection in networks under degree-corrected stochastic block models". In: *Ann. Statist.* 40.4 (Aug. 2012), pp. 2266–2292.

[ZM14] P. Zhang and C. Moore. "Scalable detection of statistically significant communities and hierarchies, using message passing for modularity". In: *Proceedings of the National Academy of Sciences* 111.51 (2014), pp. 18144–18149. eprint: https://www.pnas.org/content/111/51/18144.full.pdf.

[ZR18] Y. Zhang and K. Rohe. "Understanding regularized spectral clustering via graph conductance". In: *Advances in Neural Information Processing Systems*. 2018, pp. 10631–10640.

[ZZ20] A. Y. Zhang and H. H. Zhou. "Theoretical and computational guarantees of mean field variational inference for community detection". In: *Ann. Statist.* 48.5 (Oct. 2020), pp. 2575–2598.