Statistical Network Analysis: Beyond Block Models

by

Yuan Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2016

Doctoral Committee:

Professor Elizaveta Levina, Co-Chair
Professor Ji Zhu, Co-Chair
Assistant Professor Ambuj Tewari
Professor Mark Newman

A C K N O W L E D G M E N T S

First and foremost, I am indebted to my advisors, Professor Elizaveta Levina and Professor Ji Zhu, for their uncountable guidance and support over the five years of my PhD study, without whom I could not have come so far in my academic career. I am especially grateful for the generous freedom they offered me in pursuing challenges I am enthusiastic about. They set the role models of real scholars and great mentors. I feel very lucky to have them as my advisors.

I thank Professor Ambuj Tewari for being on my preliminary exam and dissertation committee and his warm encouragement, which was especially supportive at the early stage of my research. I thank Professor Yves Atchade for his guidance on improving my teaching and his very kind support for my job applications. I would like to thank Professor Mark Newman for agreeing to be on my dissertation committee.

It has been my privilege to work with my peer PhD students in the Levina-Zhu research group – learning from whom is among the most enjoyable experiences of my PhD life. I thank the wonderful staff in the Department of Statistics at the University of Michigan for their always reliable assistance to my work and study.

I thank Alice Xingwei Lu and Kam Chung Wong for being great friends, without whom it would have been much harder through the difficult times in the past, and I thank all my friends for being the warm company when I am over 11,000 kilometers away from home.

Last but not least, I owe the most to my parents. It has always been so refreshing to recall that they will forever love and be proud of me, no matter I am doing good or bad.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABSTRACT

Statistical Network Analysis: Beyond Block Models
by
Yuan Zhang

Co-Chairs: Professor Elizaveta Levina and Professor Ji Zhu

Network data represent connections between units of analysis and lead to many interesting research questions with diverse applications. In this thesis, we focus on inferring the structure underlying an observed network, which can be thought of as a noisy random realization of the unobserved true structure. Different applications focus on different types of underlying structure; one question of broad interest is finding a community structure, with communities typically defined as groups of nodes that share similar connectivity patterns. One common and widely used model for describing a community structure in a network is the stochastic block model. This model has attracted a lot of attention because of its tractable theoretical properties, but it is also well known to oversimplify the structure observed in real world networks and often does not fit the data well. Thus there has been a recent push to expand the stochastic block model in various ways to make it closer to what we observe in the real world, and this thesis makes several contributions to this effort.

We first study the problem of detecting communities in the presence of additional node features. Many existing methods detect communities based only on the observed edges between nodes, but in many networks, additional information on node features is available. Recent methods for community detection that incorporate node features typically either depend heavily on correct model specification, which is hard to verify, and/or do not attempt to perform feature selection. Including features related to communities can improve community detection, but including unrelated features amounts to adding noise to the data and can lead to substantial reductions in accuracy. In this thesis, we propose a model-free joint criterion for community detection with node features, with the ability to select only relevant features. We show that the underlying new community detection criterion has appropriate theoretical performance guarantees and the method is effective on both simulated and real networks.

Another direction we explore in this thesis is modeling and detecting overlapping communities. While community detection is commonly formulated as a partition problem, in practice communities in networks tend to overlap. Developing a good model for overlapping communities has been a challenge, due to identifiability issues and computational costs, although a number of special cases have been addressed. We propose a novel overlapping model that generalizes the stochastic block model and includes many of the previously studied overlapping models as special cases. The model is flexible and general but maintains identifiability and interpretability of parameters. We propose a fast algorithm to fit this model, establish its consistency, and demonstrate the method outperforms a large number of benchmarks on both simulated and real data examples.

The final contribution of this thesis is a novel method to estimate edge probabilities from a single observed network, a task closely related to the so-called graphon estima-

tion problem. The stochastic block model is able to infer this underlying edge probability matrix from a single observation by assuming the underlying probability function (the graphon) consists of constant blocks; we deal with the much more general case of piecewise Lipschitz continuous functions. Our estimator leverages a core technique of classical nonparametric statistics, neighborhood averaging, solving the challenge of defining suitable neighborhoods on networks. The method is fast and accurate, and adapts to a large range of different graphon families. We also show that it achieves the best theoretical error rate among currently known polynomial time methods for this problem.

CHAPTER 1

Introduction

This thesis focuses on statistical network analysis. Network-structured data arise in a wide range of areas; examples include social networks, communications, gene regulatory networks, brain imaging, recommender systems and so on. Analysis of network data plays an important role in many applications, including understanding social structures, disease diagnosis, marketing, and even design of parallel computing algorithms. I investigated several inference problems in statistical network analysis.

Community detection in networks

Communities are groups of nodes that have similar patterns of connection to other nodes. In many networks, nodes from the same community have a higher level of connectivity within themselves than average. Communities are present in many real world networks and usually carry meaningful interpretations, corresponding, for example, to real-life social circles, or genes and proteins with similar functions (Resnick et al., 1997; Zhang, 2009; Chamberlain, 1998). My work in this area focused on expanding the capabilities of community detection methods along two directions: utilizing additional node information and detecting overlapping communities.

Community detection using node features

Most existing methods detect communities based only on the observed edges between nodes, but in many networks, additional information on node features is available (Steglich et al., 2006; Snijders et al., 2006; Hummon et al., 1990). The question then arises whether we can combine these two sources of data to improve community detection. Many models that describe the network and the node features jointly have been proposed (Yang et al., 2013; Xu et al., 2012; Newman and Clauset, 2015), but their effectiveness typically relies heavily on correct model specification. Model-free algorithmic methods, such as those proposed by Viennet (2012); Binkiewicz et al. (2014) and Cheng et al. (2011), are usually based on the simple intuition that nodes in the same community have similar feature (network homophily). However, most existing methods ignore the fact that along with node

features helpful for community detection many datasets include many irrelevant ones, and including nuisance features usually jeopardizes community detection.

In Zhang et al. (2015a), we proposed a novel model-free criterion for community detection in weighted networks. To incorporate node features, we model edge weights as $W_{ij}z = W(f_i, f_j; \beta)$, where f_i is the feature of node i and β is a vector of coefficient that controls the influence of individual node features on community detection. We allow different communities to have different βs, and thus some features may be relevant in the formation of some communities but not others, and we learn the weights β from data simultaneously with estimating the community structure. We proved the consistency of our estimator and demonstrated its excellent empirical performance on simulated and data examples. As an example, in a lawyer friendship network (Lazega, 2001), our method discovered that type of practice (litigation or corporate), age, and years with the firm are more relevant to the lawyers social circles than gender and the law school from which they graduated.

Overlapping community detection

While community detection is commonly formulated as a partition problem, in practice communities in networks commonly overlap. For example, in a social network people may become friends because they are neighbors, classmates, colleagues, and so on; these are examples of overlapping communities. Developing a good model to describe overlapping communities has long been a challenge, for a number of reasons; in general, it is difficult to disentagle whether a connection between two results from the large number of communities they have in common, or from a higher status of a node that results in higher probability of connections to all the nodes. Most existing models (Airoldi et al., 2008; Latouche et al., 2009; Ball et al., 2011) address special cases, and even then identifiability is sometimes a challenge. Algorithmic methods (Lancichinetti et al., 2010; Gregory, 2010; Wang et al., 2011; Gillis and Vavasis, 2014) usually rely on local searches for significant communities and may perform poorly in presence of high degree nodes. The computational cost is another commonly encountered obstacle for many methods in both categories.

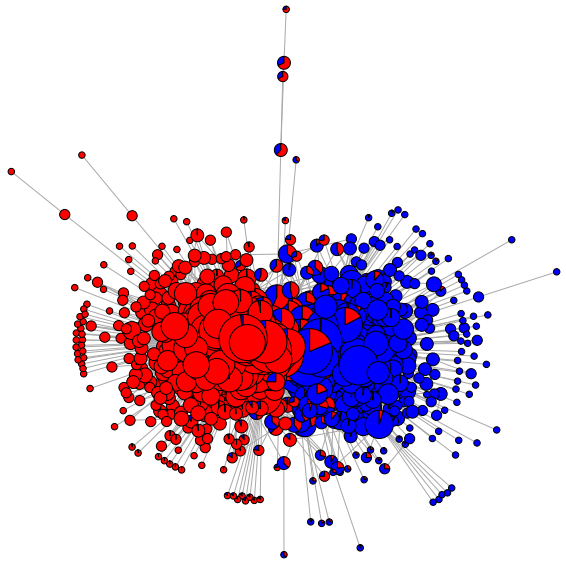


Figure 1.1: Political blog networks. Red: conservative; blue: liberal. The pie charts represent the tendency towards each community. The size of the node is proportional to log-degree.

In Zhang et al. (2014), we proposed a novel overlapping community model, with the goal of keeping it flexible, identifiable, interpretable, and computationally efficient. In our model, each node is mapped to a latent space position and communities correspond to clusters in the latent space. We let the nonnegative weighted linear combinations of cluster centers represent overlapping community memberships, with weights corresponding to the degree of association with a particular community. Our model is flexible in that it allows continuous community memberships, so that we can estimate whether a node belongs to a community strongly or weakly. Our model is identifiable under weak conditions and allows for heterogeneous node degrees. Our model can be seen as a generalization of several existing overlapping community models, including Latouche et al. (2009) and Karrer and Newman (2011), and the random dot-product model (Nickel, 2007; Young and Scheinerman, 2007), but none of them have as much flexebility while retaining intepretability. We also designed an efficient algorithm to fit the model, employing a variant of regularized spectral clustering (Von Luxburg, 2007; Qin and Rohe, 2013) to find cluster centers and replacing the commonly used K-means clustering with K-medians clustering, resulting in an asymptotically unbiased estimator for the overlapping model. Figure 1.1 shows our estimates of overlapping community memberships for the political blog network (Adamic and Glance, 2005).

Network edge probability estimation

While communities are useful in representing many real networks, more general ques-

tions about the underlying mechanism that generated the network and identifying probable missing or incorrectly recorded links are of interest, and are increasingly drawing the attention of statisticians. In this project, we designed a general estimator for network edge probabilities for a network with a more general underlying structure than typical community models allow.

Aldous (1981) and Hoover (1979) showed that in exchangeable networks (those where the order of nodes carries no information), edge probabilities can be represented by

$$P_{ij} = f(\xi_i, \xi_j) \tag{1.0.1}$$

where $\xi_i \sim \text{Uniform}[0,1]$'s and f is a function called the network graphon. The network adjacency matrix A is then assumed to have independent Bernoulli entries with $P(A_{ij} = 1) = P_{ij}$ the probability of an edge between nodes i and j. Problem (1.0.1) can be viewed as nonparametric regression with unknown design (Gao et al., 2014), with regularity in P induced by imposing smoothness conditions on f. The difficulty is that f and ξ_i's are in general unidentifiable (Diaconis and Janson, 2007) – bit it is still possible, and much more meaningful in practice, to estimate the probability matrix P. In this chapter, we focused on estimating P under the assumption that f is piecewise Lipschitz.

Several approaches have been proposed to estimate f and/or P. Step-functions approximations, including Wolfe and Olhede (2013); Olhede and Wolfe (2014); Choi and Wolfe (2014); Choi (2015) and Gao et al. (2014), usually achieve good error rates but require optimization over all possible node partitions, which is NP-hard in principle. The "sort-and-smooth" (SAS) methods such as Chan and Airoldi (2014) and Yang et al. (2014), focus on graphons with strictly monotone expected node degrees and depend crucially on this rather strong condition. The universal singular value thresholding (USVT) (Chatterjee, 2014), a general matrix completion and denoising tool, can also be used to estimate P with a comparatively loose error bound. Our key insight is that the effectiveness of most methods depends on choosing a good neighborhood for each node, so that averaging over $A_{i'j'}$'s for i' and j' in the neighborhood of i and j yields a good estimator for P_{ij}. The question is how to define a good neighborhood that both leads to a good error rate and can be learned efficiently.

Our recent paper (Zhang et al., 2015b) proposes a novel neighborhood smoothing method for estimating P with a general structure. We define the neighborhood for each node based on the ℓ_2 distance between the "graphon slices" of nodes, represented by the rows of A. This dissimilarity measure is distinct from the distance between the latent ξ_i's, and, for example, for networks generated from the stochastic block model, represents

a more meaningful difference between nodes. We designed an efficient algorithm to select the neighborhood for each node among its neighbors according to the graphon slice distance. Our method is almost tuning-free, numerically robust, computationally efficient, and allows parallelization. We showed that our estimator achieves the best error rate among existing methods that do not rely on optimizing over all node partitions and are thus computationally feasible. Our method can accurately estimate a wide variety of network structures and predict missing edges well when applied to the link prediction problem.

CHAPTER 2

Community detection in networks with node features

2.1 Introduction

Community detection is a fundamental problem in network analysis, extensively studied in a number of domains – see Rogers and Kincaid (1981) and Schlitt and Brazma (2007) for some examples of applications. A number of approaches to community detection are based on probabilistic models for networks with communities, such as the stochastic block model Holland et al. (1983), the degree-corrected stochastic block model Karrer and Newman (2011), and the latent factor model Hoff (2007). Other approaches work by optimizing a criterion measuring the strength of community structure in some sense, often through spectral approximations. Examples include normalized cuts Shi and Malik (2000), modularity Newman and Girvan (2004b); Newman (2006), and many variants of spectral clustering, e.g., Qin and Rohe (2013).

Many of the existing methods detect communities based only on the network adjacency matrix. However, we often have additional information on the nodes (node features), and sometimes edges as well, for example, Steglich et al. (2006), Snijders et al. (2006) and Hummon et al. (1990). In many networks the distribution of node features is correlated with community structure McAuley and Leskovec (2012), and thus a natural question is whether we can improve community detection by using the node features. Several generative models for jointly modeling the edges and the features have been proposed, including the network random effects model Hoff (2003), the embedding feature model Zanghi et al. (2010), the latent variable model Handcock et al. (2007), the discriminative approach Yang et al. (2009), the latent multi-group membership graph model M. Kim (2012), the social circles model for ego networks McAuley and Leskovec (2012), the communities from edge structure and node attributes (CESNA) model Yang et al. (2013), the Bayesian Graph Clustering (BAGC) model Xu et al. (2012), the topical communities and personal interest

(TCPI) model Hoang and Lim (2014) and the modified stochastic block model Newman and Clauset (2015). The latter paper was written after this work was completed, and while its goals are somewhat similar to ours by also learning the relationship between the features and the network from data, it is very different in that it postulates a model connecting them in a particular way. Most of these models are designed for specific feature types, and their effectiveness depends heavily on the correctness of model specification. Model-free approaches include weighted combinations of the network and feature similarities Viennet (2012); Binkiewicz et al. (2014), attribute-structure mining Silva et al. (2012), simulated annealing clustering Cheng et al. (2011), and compressive information flow Smith et al. (2014). Most methods in this category use all the features in the same way without determining which ones influence the community structure and which do not, and lack flexibility in how to balance the network information with the information coming from its node features, which do not always agree. Including irrelevant node features can only hurt community detection by adding in noise, while selecting features that by themselves cluster strongly may not correspond to features that correlate with the community structure present in the adjacency matrix.

In this chapter, we propose a new joint community detection criterion that uses both the network adjacency matrix and the node features. The idea is that by properly weighing edges according to feature similarities on their end nodes, we strengthen the community structure in the network thus making it easier to detect. Rather than using all available features in the same way, we learn which features are most helpful in identifying the community structure from data. Intuitively, our method looks for an agreement between clusters suggested by two data sources, the adjacency matrix and the node features. Numerical experiments on simulated and real networks show that our method performs well compared to methods that use either the network alone or the features alone for clustering, as well as to a number of benchmark joint detection methods.

2.2 The joint community detection criterion

Our method is designed to look for assortative community structure, that is, the type of communities where nodes are more likely to connect to each other if they belong to the same community, and thus there are more edges within communities than between. This is a very common intuitive definition of communities which is incorporated in many community detection criteria, for example, modularity Newman (2006). Our goal is to use such a community detection criterion based on the adjacency matrix alone, and add feature-based edge weights to improve detection. Several criteria using the adjacency matrix alone are

available, but having a simple criterion linear in the adjacency matrix makes optimization much more feasible in our particular situation, and we propose a new criterion which turns out to work particularly well for our purposes. Let A denote the adjacency matrix with $A_{ij} = 0$ if there is no edge between nodes i and j, and otherwise $A_{ij} > 0$ which can be either 1 for unweighted networks or the edge weight for weighted networks. The community detection criterion we start from is a very simple analogue of modularity, to be maximized over all possible label assignments e:

$$R(e) = \sum_{k=1}^{K} \frac{1}{|\mathcal{E}_k|^\alpha} \sum_{i,j \in \mathcal{E}_k} A_{ij} \ . \tag{2.2.1}$$

Here e is the vector of node labels, with $e_i = k$ if node i belongs to community k, for $k = 1, \ldots, K$, $\mathcal{E}_k = \{i : e_i = k\}$, and $|\mathcal{E}_k|$ is the number of nodes in community k. We assume each node belongs to exactly one community, and the number of communities K is fixed and known. Rescaling by $|\mathcal{E}_k|^\alpha$ is designed to rule out trivial solutions that put all nodes in the same community, and $\alpha > 0$ is a tuning parameter. When $\alpha = 2$, the criterion is approximately the sum of edge densities within communities, and when $\alpha = 1$, the criterion is the sum of average "within community" degrees, which both intuitively represent community structure. This criterion can be shown to be consistent under the stochastic block model by checking the conditions of the general theorem in Bickel and Chen (2009).

The ideal use of features with this criterion would be to use them to up-weigh edges within communities and down-weigh edges between them, thus enhancing the community structure in the observed network and making it easier to detect. However, node features may not be perfectly correlated with community structure, different communities may be driven by different features, as pointed out by McAuley and Leskovec (2012), and features themselves may be noisy. Thus we need to learn the impact of different features on communities as well as balance the roles of the network itself and its features. Let f_i denote the p-dimensional feature vector of node i. We propose a *joint community detection criterion* (JCDC),

$$R(e, \beta; w_n) = \sum_{k=1}^{K} \frac{1}{|\mathcal{E}_k|^\alpha} \sum_{i,j \in \mathcal{E}_k} A_{ij} W(f_i, f_j, \beta_k; w_n) \tag{2.2.2}$$

where α is a tuning parameter as in (2.2.1), $\beta_k \in \mathbb{R}^p$ is the coefficient vector that defines the impact of different features on the kth community, and $\beta := \{\beta_1, \ldots, \beta_K\}$. The criterion is then maximized over both e and β. Having a different β_k for each k allows us to learn the roles different features may play in different communities. The balance between the

information from A and $F := \{f_1, \ldots, f_n\}$ is controlled by w_n, another tuning parameter which in general may depend on n.

For the sake of simplicity, we model the edge weight $W(f_i, f_j, \beta_k; w_n)$ as a function of the node features f_i and f_j via a p-dimensional vector of their similarity measures $\phi_{ij} = \phi(f_i, f_j)$. The choice of similarity measures in ϕ depends on the type of f_i (for example, on whether the features are numerical or categorical) and is determined on a case by case basis; the only important property is that ϕ assigns higher values to features that are more similar. Note that this trivially allows the inclusion of edge features as well as node features, as long as they are converted to some sort of similarity. To eliminate potential differences in units and scales, we standardize all ϕ_{ij} along each feature dimension. Finally, the function W should be increasing in $\langle \phi_{ij}, \beta \rangle$, which can be viewed as the "overall similarity" between nodes, and for optimization purposes it is convenient to take W to be concave. Here we use the exponential function,

$$w_{ijk} = W(f_i, f_j, \beta_k; w_n) = w_n - e^{-\langle \phi_{ij}, \beta_k \rangle} \tag{2.2.3}$$

One can use other functions of similar shapes, for example, the logit exponential function, which we found empirically to perform similarly.

2.3 Estimation

The joint community detection criterion needs to be optimized over both the community assignments e and the feature parameters β. Using block coordinate descent, we optimize JCDC by alternately optimizing over the labels with fixed parameters and over the parameters with fixed labels, and iterating until convergence.

2.3.1 Optimizing over label assignments with fixed weights

When parameters β are fixed, all edge weights w_{ijk}'s can be treated as known constants. It is infeasible to search over all n^K possible label assignments, and, like many other community detection methods, we rely on a greedy label switching algorithm to optimize over e, specifically, the tabu search Glover (1986), which updates the label of one node at a time. Since our criterion involves the number of nodes in each community $|\mathcal{E}_k|$, no easy spectral approximations are available. Fortunately, our method allows for a simple local approximate update which does not require recalculating the entire criterion. For a given node i

considered for label switching, the algorithm will assign it to community k rather than l if

$$\frac{S_{kk} + 2S_{i\leftrightarrow k}}{(|\mathcal{E}_k| + 1)^\alpha} + \frac{S_{ll}}{|\mathcal{E}_l|^\alpha} > \frac{S_{kk}}{|\mathcal{E}_k|^\alpha} + \frac{S_{ll} + 2S_{i\leftrightarrow l}}{(|\mathcal{E}_l| + 1)^\alpha}\,, \tag{2.3.1}$$

where S_{kk} is twice the total edge weights in community k, and $S_{i\leftrightarrow k}$ is the sum of edge weights between node i and all the nodes in $\mathcal{E}_k$. When $|\mathcal{E}_k|$ and $|\mathcal{E}_l|$ are large, we can ignore $+1$ in the denominators, and (2.3.1) becomes

$$\frac{S_{i\leftrightarrow k}}{|\mathcal{E}_k|} \cdot \frac{|\mathcal{E}_k|^{1-\alpha}}{|\mathcal{E}_l|^{1-\alpha}} > \frac{S_{i\leftrightarrow l}}{|\mathcal{E}_l|}\,, \tag{2.3.2}$$

which allows for a "local" update for the label of node i without calculating the entire criterion. This also highlights the impact of the tuning parameter α: when $\alpha = 1$, the two sides of (2.3.2) can be viewed as averaged weights of all edges connecting node i to communities $\mathcal{E}_k$ and $\mathcal{E}_l$, respectively. Then our method assigns node i to the community with which it has the strongest connection. When $\alpha \neq 1$, the left hand side of (2.3.2) is multiplied by a factor $(|\mathcal{E}_k|/|\mathcal{E}_l|)^{1-\alpha}$. Suppose $|\mathcal{E}_k|$ is larger than $|\mathcal{E}_l|$; then choosing $0 < \alpha < 1$ indicates a preference for assigning a node to the larger community, while $\alpha > 1$ favors smaller communities. A detailed numerical investigation of the role of α is provided in the Supplemental Material.

The edge weights involved in (2.3.2) depend on the tuning parameter w_n. When $\beta = 0$, all weights are equal to $w_n - 1$. On the other hand, $w_{ijk} \leq w_n$ for all values of β. Therefore, $w_n/(w_n - 1)$ is the maximum amount by which our method can reweigh an edge. When w_n is large, $w_n/(w_n - 1) \approx 1$, and thus the information from the network structure dominates. When w_n is close to 1, the ratio is large and the feature-driven edge weights have a large impact. See the Supplemental Material for more details on the choice of w_n.

While the tuning parameter w_n controls the amount of influence features can have on community detection, it does not affect the estimated parameters β for a fixed community assignment. This is easy to see from rearranging terms in (2.2.2):

$$R(e, \beta; w_n) = w_n \sum_{k=1}^{K} \frac{1}{|\mathcal{E}_k|^\alpha} \sum_{i,j \in \mathcal{E}_k} A_{ij} - g(e, A, \beta, \phi) \tag{2.3.3}$$

where the function g does not depend on w_n. Note that the term containing w_n does not depend on β.

2.3.2 Optimizing over weights with fixed label assignments

Since we chose a concave edge weight function (2.2.3), for a given community assignment e the joint criterion is a concave function of β_k, and it is straightforward to optimize over β_k by gradient ascent. The role of β_k is to control the impact of different features on each community. One can show by a Taylor-series type expansion around the maximum (details omitted) and also observe empirically that for our method, the estimated $\hat{\beta}_k$'s are correlated with the feature similarities between nodes in community k. In other words, our method tends to produce a large estimated $\hat{\beta}_k^{(\ell)}$ for a feature with high similarity values $\phi_{ij}^{(\ell)}$'s for $i, j \in \mathcal{E}_k$. However, in the extreme case, the optimal $\hat{\beta}_k^{(\ell)}$ can be $+\infty$ if all $\phi_{ij}^{(\ell)}$'s are positive in community k or $-\infty$ if all $\phi_{ij}^{(\ell)}$'s are negative (recall that similarities are standardized, so this cannot happen in all communities). To avoid these extreme solutions, we subtract a penalty term $\lambda\|\beta\|_1$ from the criterion (2.2.2) while optimizing over β. We use a very small value of λ ($\lambda = 10^{-5}$ everywhere in the chapter) which safeguards against numerically unstable solutions but has very little effect on other estimated coefficients.

2.4 Consistency

The proposed JCDC criterion (2.2.2) is not model-based, but under certain models it is asymptotically consistent. We consider the setting where the network A and the features F are generated independently from a stochastic block model and a uniformly bounded distribution, respectively. Let $\mathbb{P}(A_{ij} = 1) = \rho_n P_{c_i c_j}$ where ρ_n is a factor controling the overall edge density and $c = (c_1, \ldots, c_n)$ is the vector of true labels. Assume the following regularity conditions hold:

1. There exist global constants M_ϕ and M_β, such that $\|\phi_{ij}\|_2 \leq M_\phi$ and $\|\beta_k\|_2 \leq M_\beta$ for all k, and the tuning parameter w_n satisfies $\log w_n > M_\phi M_\beta$.

2. Let $\mathcal{C}_k := \{i : c_i = k\}$. There exists a global constant π_0 such that $|C_k| \geq \pi_0 n > 0$ for all k.

3. For all $1 \leq k < l \leq K$, $2(K-1)P_{kl} < \min(P_{kk}, P_{ll})$.

Condition 1 states that node feature similarities are uniformly bounded. This is a mild condition in many applications as the node features are often themselves uniformly bounded. In practice, for numerical stability the user may want to standardize node features and discard individual features with very low variance, before calculating the corresponding similarities ϕ. Condition 2 guarantees communities do not vanish asymptotically. Condition 3 enforces assortativity. Since the estimated labels e are only defined

up to an arbitrary permutation of communities, we measure the agreement betwee e and c by $d(e,c) = \min_{\sigma\in\mathcal{P}_K} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\sigma(e_i) \neq c_i)$, where $\mathcal{P}_K$ is the set of all permutations of $\{1,\ldots,K\}$.

Theorem 1 (Consistency of JCDC)**.** *Under conditions 1, 2 and 3, if* $n\rho_n \to \infty$, $w_n\rho_n \to \infty$*, and the parameter* α *satisfies*

$$\frac{\max_{k,l} 2(K-1)P_{kl}}{\min_{k,l}(P_{kk}, P_{ll})} \leq \alpha \leq 1 \tag{2.4.1}$$

then we have, for any fixed $\delta > 0$*,*

$$\mathbb{P}\left(d\left(\arg\max_{e}(\max_{\beta} R(e,\beta;w_n)), c\right) > \delta\right) \to 0\,. \tag{2.4.2}$$

The proof is given in the Supplemental Material.

2.5 Simulation studies

We compare JCDC to three representative benchmark methods which use both the adjacency matrix and the node features: CASC (Covariate Assisted Spectral Clustering, Binkiewicz et al. (2014)), CESNA (Communities from Edge Structure and Node Attributes, Yang et al. (2013)), and BAGC (BAyesian Graph Clustering, Xu et al. (2012)). In addition, we also include two standard methods that use either the network adjacency alone (SC, spectral clustering on the Laplacian regularized with a small constant $\tau = 1e-7$, as in Amini et al. (2013)), or the node features alone (KM, K-means performed on the p-dimensional node feature vectors, with 10 random initial starting values). We generate networks with $n = 150$ nodes and $K = 2$ communities of sizes 100 and 50 from the degree-corrected stochastic block model as follows. The edges are generated independently with probability $\theta_i\theta_j p$ if nodes i and j are in the same community, and $r\theta_i\theta_j p$ if nodes i and j are in different communities. We set $p = 0.1$ and vary r from 0.25 to 0.75. We set 5% of the nodes in each community to be "hub" nodes with the degree correction parameter $\theta_i = 10$, and for the remaining nodes set $\theta_i = 1$. All resulting products are thresholded at 0.99 to ensure there are no probability values over 1. These settings result in the average expected node degree ranging approximately from 22 to 29.

For each node i, we generate $p = 2$ features, with one "signal" feature related to the community structure and one "noise" feature whose distribution is the same for all nodes. The "signal" feature follows the distribution $N(\mu, 1)$ for nodes in community 1 and $N(-\mu, 1)$ for nodes in community 2, with μ varying from 0.5 to 2 (larger μ corresponds

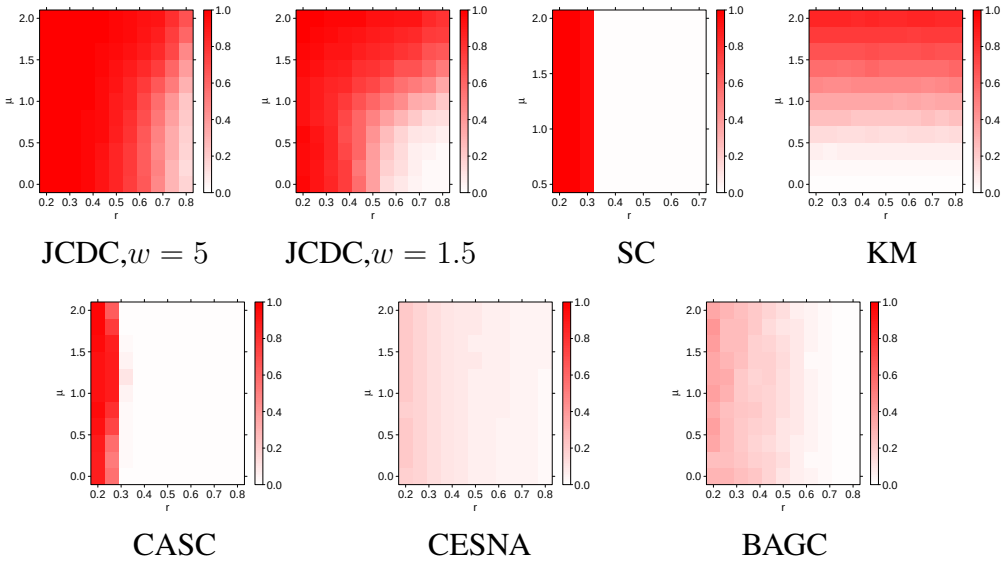


Figure 2.1: Performance of different methods measured by normalized mutual information as a function of r (out-in probability ratio) and μ (feature signal strength).

to stronger signal). For use with CESNA, which only allows categorical node features, we discretize the continuous node features by partitioning the real line into 20 bins using the $0.05, 0.1, \ldots, 0.95$-th quantiles. For the JCDC, based on the study of the tuning parameters in the Supplemental Material, we use $\alpha = 1$ and compare two values of w_n, $w_n = 1.5$ and $w_n = 5$. Finally, agreement between the estimated communities and the true community labels is measured by normalized mutual information, a measure commonly used in the network literature which ranges between 0 (random guessing) and 1 (perfect agreement). For each configuration, we repeat the experiments 30 times, and record the average NMI over 30 replications.

Figure 2.1 shows the heatmaps of average NMI for all methods under these settings, as a function of r and μ. As one would expect, the performance of spectral clustering (c), which uses only the network information, is only affected by r (the larger r is, the harder the problem), and the performance of K-means (d), which uses only the features, is only affected by μ (the larger μ is, the easier the problem). JCDC is able to take advantage of both network and feature information by estimating the coefficients β from data, and its performance only deteriorates when neither is informative. The informative features are more helpful with a larger value of w (a), and conversely uninformative features affect perfomance slightly more with a lower value of w (b), but this effect is not strong. CASC (e) appears to inherit the sharp phase transition from spectral clustering, which forms the basis of CASC; the sharp transition is perhaps due to different community sizes and hub nodes, which are both challenging to spectral clustering; CESNA (f) and BAGC (g) do not

perform as well overall, with BAGC often clustering all the hub nodes into one community.

2.6 Data applications

2.6.1 The world trade network

The world trade network De Nooy et al. (2011) connects 80 countries based on the amount of trade of metal manufactures between them in 1994, or when not available for that year, in 1993 or 1995. Nodes are countries and edges represent positive amount of import and/or export between the countries. Each country also has three categorical features: the continent (Africa, Asia, Europe, N. America, S. America, and Oceania), the country's structural position in the world system in 1980 (core, strong semi-periphery, weak semi-periphery, periphery) and in 1994 (core, semi-periphery, periphery). Figures 2.2 (a) to (c) show the adjacency matrix rearranged by sorting the nodes by each of the features. The partition by continent (Figure 2.2(a)) clearly shows community structure, whereas the other two features show hubs (core status countries trade with everyone), and no assortative community structure. We will thus compare partitions found by all the competing methods to the continents, and omit the three Oceania countries from further analysis because no method is likely to detect such a small community. The two world position variables ('80 and '94) will be used as features, treated as ordinal variables.

The results for all methods are shown in Figure 2.2, along with NMI values comparing the detected partition to the continents. All methods were run with the true value $K = 5$.

Table 2.1: Feature coefficients $\hat{\beta}_k$ estimated by JCDC with $w = 5$. Best match is determined by majority vote.

Community	Best match	Position '80	Position '94
blue	Europe	0.000	0.143
red	Asia	0.314	0.127
green	Europe	0.017	0.204
cyan	N. America	0.107	0.000
purple	S. America	0.121	0.000

The result of spectral clustering agrees much better with the continents than that of K-means, indicating that the community structure in the adjacency matrix is closer to the continents that the structure contained in the node features. JCDC obtains the highest NMI value, CASC performs similarly to spectral clustering, whereas CESNA and BAGC both

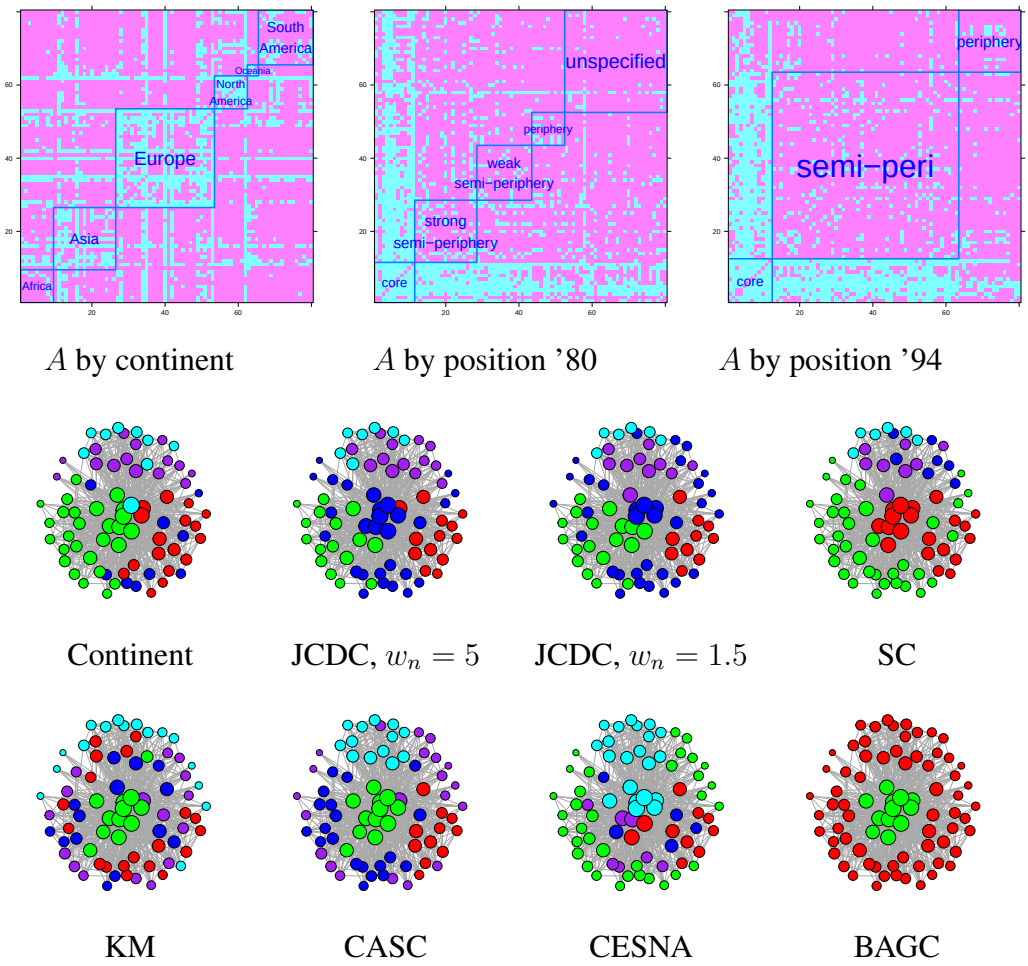


Figure 2.2: (a)-(c): the adjacency matrix ordered by different node features; (d) network with nodes colored by continent (taken as ground truth); blue is Africa, red is Asia, green is Europe, cyan is N. America and purple is S. America. (e)-(k) community detection results from different methods; colors are mated to (d) in the best way possible.

fail to recover the continent partition. Note that no method was able to estimate Africa well, likely due to the disassortative nature of its trade seen in Figure 2.2 (a). Figure 2.2 (e) indicates that JCDC estimated N. America, S. America and Asia with high accuracy, but split Europe into two communities, since it was run with $K = 5$ and could not pick up Africa due to its disassortative structure. Table 2.1 contains the estimated feature coefficients, suggesting that in 1980 the "world position" had the most influence on the connections formed by Asian countries, whereas in 1994 world position mattered most in Europe.

2.6.2 The lawyer friendship network

The second dataset we consider is a friendship network of 71 lawyers in a New England corporate law firm Lazega (2001). Seven node features are available: status (partner or associate), gender, office location (Boston, Hartford, or Providence, a very small office with only two non-isolated nodes), years with the firm, age, practice (litigation or corporate) and law school attended (Harvard, Yale, University of Connecticut, or other). Categorical features with M levels are represented by $M - 1$ dummy indicator variables. Figures 2.3 (a)-(g) show heatmap plots of the adjacency matrix with nodes sorted by each feature, after eliminating six isolated nodes. Partition by status (Figure 2.3(a)) shows a strong assortative structure, and so does partition by office (Figure 2.3(c)) restricted to Boston and Hartford, but the small Providence office does not have any kind of structure. Thus we chose the status partition as a reference point for comparisons, though other partitions are certainly also meaningful.

Communities estimated by different methods are shown in Figure 2.3 (i)-(o), all run with $K = 2$. Spectral clustering and K-means have equal and reasonably high NMI values, indicating that both the adjacency matrix and node features contain community information. JCDC obtains the highest NMI value, with $w_n = 5$ performing slightly better than $w_n = 1.5$. CASC improves upon spectral clustering by using the feature information, with NMI just slightly lower than that of JCDC with $w_n = 1.5$. CESNA and BAGC have much lower NMI values, possibly because of hub nodes, or because they detect communities corresponding to something other than status.

The estimated feature coefficients are shown in Table 2.2. Office location, years with the firm, and age appear to be the features most correlated with the community structure of status, for both partners and associates, which is natural. Practice, school, and gender are less important, though it may be hard to estimate the influence of gender accurately since there are relatively few women in the sample.

Table 2.2: Feature coefficients $\hat{\beta}_k$, JCDC with $w_n = 5$.

Comm.	gender	office	years	age	practice	school
partner	0.290	0.532	0.212	0.390	0.095	0.000
associate	0.012	0.378	0.725	0.320	0.118	0.097

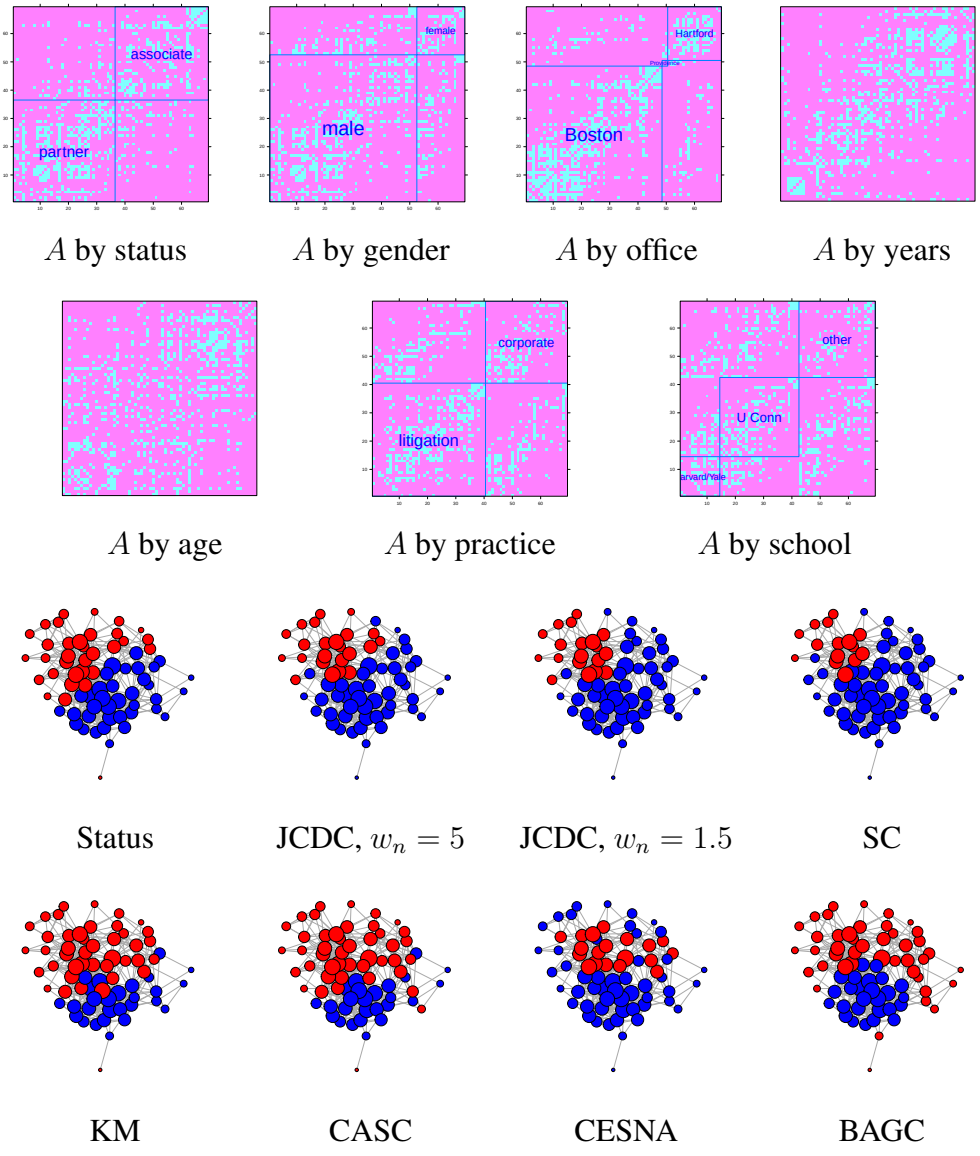


Figure 2.3: (a)-(g): adjacency matrix with nodes sorted by features; (h): network with nodes colored by status (blue is partner, red is associate); (i)-(n): community detection results from different methods.

2.7 Discussion

Our method incorporates feature-based weights into a community detection criterion, improving detection compared to using just the adjacency matrix or the node features alone, if the cluster structure in the features is related to the community structure in the adjacency matrix. It has the ability to estimate coefficients for each feature within each community and thus learn which features are correlated with the community structure. This ability guards against including noise features which can mislead community detection. The community detection criterion we use is designed for assortative community structure, with more connections within communities than between, and benefits the most from using features that have a similar clustering structure.

This work can be extended in several directions. Variation in node degrees, often modeled via the degree-corrected stochastic block model Karrer and Newman (2011) which regards degrees as independent of community structure, may in some cases be correlated with node features, and accounting for degree variation jointly with features can potentially further improve detection. Another useful extension is to overlapping communities. One possible way to do that is to optimize each summand in JCDC (2.2.2) separately and in parallel, which can create overlaps, but would require careful initialization. Statistical models that specify exactly how features are related to community assignments and edge probabilities can also be useful, though empirically we found no such standard models that could compete with the non-model-based JCDC on real data. This suggests that more involved and perhaps data-specific modeling will be necessary to accurately describe real networks, and some of the techniques we proposed, such as community-specific feature coefficients, could be useful in that context.

CHAPTER 3

Detecting overlapping communities in networks using spectral methods

3.1 Introduction

The problem of community detection in networks has been actively studied in several distinct fields, including physics, computer science, statistics, and the social sciences. Its applications include understanding social interactions of people (Zachary, 1977; Resnick et al., 1997) and animals (Lusseau et al., 2003), discovering functional regulatory networks of genes (Bolouri and Davidson, 2010; Zhang, 2009) and even designing parallel computing algorithms (Chamberlain, 1998; Hendrickson and Kolda, 2000). Community detection is in general a challenging task. The challenges include defining what a community is (commonly taken to be a group of nodes that have more connections to each other than to the rest of the network, although other types of communities are not unusual), formulating realistic and tractable statistical models of networks with communities, and designing fast scalable algorithms for fitting such models.

In this paper, we focus on network models with overlapping communities, with nodes potentially belonging to more than one community at a time. This is common in real-world networks (Palla et al., 2005; Pizzuti, 2009), and yet most literature to date has focused on partitioning the network into non-overlapping communities, with some notable exceptions discussed below. Our goal is to design an overlapping community model that is flexible, interpretable, and computationally feasible. We will thus focus on models which can be fitted by spectral methods, one of the most scalable tools for fitting non-overlapping community models available to date.

We start with a brief review of relevant work in community detection for non-overlapping communities, which mainly falls into one of two broad categories: algorithmic methods, based on optimizing some criterion reflecting desirable properties of a partition over all

possible partitions (see Fortunato (2010) for a review), and model fitting, where a generative model with communities is postulated for the network and its parameters are estimated from the observed adjacency matrix (see Goldenberg et al. (2010) for a review). Perhaps the most popular and best studied generative model for community detection is the stochastic block model (SBM) (Holland and Leinhardt, 1981; Holland et al., 1983). The SBM views the $n \times n$ network adjacency matrix $\boldsymbol{A}$, defined by $A_{ij} = 1$ if there is an edge between i and j and 0 otherwise, as a random graph with independent Bernoulli-distributed edges. The Bernoulli probabilities for the edges depend on the node labels c_i which take values in $\{1, \ldots, K\}$, and the $K \times K$ matrix $\boldsymbol{B}$ containing the probabilities of edges forming between different communities. The node labels can be represented by an $n \times K$ binary community membership matrix $\boldsymbol{Z}$ with exactly one "1" in each row, $Z_{ik} = \mathbf{1}[c_i = k]$ for all i, k. Then the probabilities of edges are given by $\boldsymbol{W} \equiv \mathbb{E}(\boldsymbol{A}) = \boldsymbol{Z}\boldsymbol{B}\boldsymbol{Z}^T$. Thus in this model, a node's label determines its behavior entirely, and thus all nodes in the same community are "stochastically equivalent", and in particular have the same expected degree. This is known to be often violated in practice, due to commonly present "hub" nodes with many more connections than other nodes in their community. The degree-corrected stochastic block model (DCSBM) (Karrer and Newman, 2011) was proposed to address this limitation, which multiplies the probability of an edge between nodes i and j by the product of node-specific positive "degree parameters" $\theta_i\theta_j$. Both SBM and DCSBM can be consistently estimated by maximizing the likelihood (Bickel and Chen, 2009; Zhao et al., 2012), but directly optimizing the likelihood over all label assignments is not computationally feasible. A number of faster algorithms for fitting these models have been proposed in recent years, including pseudo-likelihood (Amini et al., 2013), belief propagation (Decelle et al., 2011), spectral approximations to the likelihood (Newman, 2013; Le et al., 2014), spectral clustering on eigenvector ratios to fit DCSBM (Jin, 2015), and generic spectral clustering (Von Luxburg, 2007), used by many and analyzed, for example, in Rohe et al. (2011) and Sarkar and Bickel (2013). It was further shown that regularization improves on spectral clustering substantially (Amini et al., 2013; Chaudhuri et al., 2012), and its theoretical properties have been further analyzed by Qin and Rohe (2013) and Joseph and Yu (2013). While for specific likelihoods one can develop methods that are both fast and more accurate than spectral clustering, such as the pseudo-likelihood (Amini et al., 2013), in general spectral methods remain the most scalable option available.

While the majority of the existing models and algorithms for community detection focus on discovering non-overlapping communities, there has been a growing interest in exploring the overlapping scenario, although both extending the existing models to the overlapping case and developing brand new models remain challenging. Like methods for

non-overlapping community detection, most existing approaches for detecting overlapping communities can be categorized as either algorithmic or model-based methods. For a comprehensive review, see Xie et al. (2013). Model-based methods focus on specifying how node community memberships determine edge probabilities. For example, the overlapping stochastic block model (OSBM) (Latouche et al., 2009) extends the SBM by allowing the entries of the membership matrix $\boldsymbol{Z}$ to be independent Bernoulli variables, thus allowing multiple "1"s in one row, or all "0"s. The mixed membership stochastic block model (Airoldi et al., 2008) draws membership vectors $\boldsymbol{Z}_{i\cdot}$ from a Dirichlet prior. The membership vector is drawn again to generate every edge, instead of being fixed for the node, so the community membership for node i varies depending on which node j it is interacting with. The "colored edges" model (Ball et al., 2011), sometimes referred to as the Ball-Karrer-Newman model or BKN, allows continuous community membership by relaxing the binary $\boldsymbol{Z}$ to a matrix with non-negative entries (with some normalization constraints for identifiability), and discarding the matrix $\boldsymbol{B}$. The Bayesian nonnegative matrix factorization model (Psorakis et al., 2011) is related to the model but with notable differences.

Algorithmic methods for overlapping community detection mostly rely on local greedy searches and intuitive criteria. Current approaches include detecting each community separately by maximizing a local measure of goodness of the estimated community (Lancichinetti et al., 2011) and updating an initial estimate of the community membership by neighborhood vote (Gregory, 2010). Local methods typically rely heavily on a good starting value. Global algorithmic approaches include computing a non-negative matrix factorization approximation to the adjacency matrix and extracting a binary membership matrix from one of the factors (Wang et al., 2011; Gillis and Vavasis, 2014). Many heuristic methods do not take heterogeneous node degrees into account, and we found empirically they can perform poorly in the presence of hubs (see Section 3.5).

In this chapter, we propose a new generative model for overlapping communities, the overlapping continuous community assignment model (OCCAM). It allows a node to belong to different communities to a different extent, via the membership vector $\boldsymbol{Z}_{i\cdot}$ with non-negative entries which represent how strongly a node is associated with various communities. We also allow arbitrary degree distributions in a manner similar to the DCSBM, and retain the $K \times K$ matrix $\boldsymbol{B}$ which allows to interpret connections between communities and compare them. All the model parameters (membership vectors, degree corrections, and community-level connectivity) are identifiable under certain constraints which we will state explicitly. We also develop a fast spectral algorithm to fit OCCAM. Typically, spectral clustering projects the adjacency matrix or its Laplacian onto the K leading eigenvectors representing the nodes' latent positions, and performs K-means in that lower-dimensional

space to estimate community memberships. Our key insight here is that when the nodes come from a mixture of clusters (as they would with multiple community memberships), K-means has no chance of recovering the cluster centers correctly; but as long as there are enough pure nodes in each community, K-medians will still be able to identify the cluster centers correctly by ignoring the "mixed" nodes on the boundaries. We show that our method produces asymptotically consistent parameter estimates as the number of nodes grows as long as there are enough pure nodes and the network is not too sparse. We also employ a simple regularization scheme, since it is by now well known that regularizing spectral clustering substantially improves its performance, especially in sparse networks (Chaudhuri et al., 2012; Amini et al., 2013; Qin and Rohe, 2013). We provide an explicit rate for the regularization parameter, implied by our consistency analysis, and show that the overall performance is robust to the choice of the constant multiplier in the regularization parameter as long as the rate is specified correctly.

The rest of the chapter is organized as follows. We introduce the model and discuss parameter identifiability in Section 3.2, present the two-stage spectral clustering algorithm in Section 3.3, and state consistency results and describe the choice of the regularization parameter in Section 3.4. Some simulation results are presented in Section 3.5, where we investigate robustness of our method to the choice of regularization parameter and compare it to a number of benchmark methods for overlapping community detection. We apply the proposed method to a large number of real social ego-networks (networks consisting of all friends of one or several users) from Facebook, Twitter, and GooglePlus in Section 3.6. Section 3.7 concludes the chapter with a brief discussion of contributions, limitations, and future work. All proofs are given in the supplemental materials.

3.2 The overlapping continuous community assignment model

3.2.1 The model

Recall that we represent the network by its $n \times n$ adjacency matrix $\boldsymbol{A}$, a binary symmetric matrix with $\{A_{ij}, i < j\}$ independent Bernoulli variables and $\boldsymbol{W} \equiv \mathbb{E}(\boldsymbol{A})$. We will assume that $\boldsymbol{W}$ has the form

$$\boldsymbol{W} = \alpha_n \boldsymbol{\Theta} \boldsymbol{Z} \boldsymbol{B} \boldsymbol{Z}^T \boldsymbol{\Theta} \,. \tag{3.2.1}$$

We call this formulation the Overlapping Continuous Community Assignment Model (OCCAM). The factor α_n is a global scaling factor that controls the overall edge probability, and the only component that depends on n. As is commonly done in the literature, for theoretical analysis we will let $\alpha_n \to 0$ at a certain rate, otherwise the network becomes

completely dense as $n \to \infty$. The $n \times n$ diagonal matrix $\boldsymbol{\Theta} = \mathrm{diag}(\theta_1, \ldots, \theta_n)$ contains non-negative degree correction terms that allow for heterogeneity in the node degrees, in the same fashion as under the DCSBM. We will later assume that θ_i's are generated from a fixed distribution $\mathcal{F}_\Theta$ which does not depend on n. The $n \times K$ community membership matrix $\boldsymbol{Z}$ is the primary parameter of interest; the i-th row $\boldsymbol{Z}_{i\cdot}$ represents node i's propensities towards each of the K communities. We assume $Z_{ik} \geq 0$ for all i, k, and $\|\boldsymbol{Z}_{i\cdot}\|_2 = 1$ for identifiability. Formally, a node is "pure" if $Z_{ik} = 1$ for some k. Later, we will also assume that the rows $\boldsymbol{Z}_{i\cdot}$'s are generated independently from a fixed distribution $\mathcal{F}_Z$ that does not depend on n. Finally, the $K \times K$ matrix $\boldsymbol{B}$ represents (scaled) probabilities of connections between pure nodes of all communities. Since we are already using α_n and $\boldsymbol{\Theta}$, we constrain all diagonal elements of $\boldsymbol{B}$ to be 1 for identifiability. Other constraints are also needed to make the model fully identifiable; we will discuss them in Section 3.2.2.

Note that the general form (3.2.1) can, with additional constraints, incorporate many of the other previously proposed models as special cases. If all nodes are pure and $\boldsymbol{Z}$ has exactly one "1" in each row, we get DCSBM; if we further assume all θ_i's are equal, we have the regular SBM. If the constraint $\|\boldsymbol{Z}_{i\cdot}\|_2 = 1$ is removed and the entries of $\boldsymbol{Z}$ are required to be 0 or 1, and all θ_i's are equal, we have the OSBM of Latouche et al. (2009). Alternatively, if we set $\boldsymbol{B} = \boldsymbol{I}$, we have the "colored edges" model of Ball et al. (2011). Our model is also related to the random dot product model (RDPM) (Nickel, 2007; Young and Scheinerman, 2007), which stipulates $\boldsymbol{W} = \boldsymbol{X}_0 \boldsymbol{X}_0^T$ for some (usually low-rank) $\boldsymbol{X}_0$. This is true for our model if $\boldsymbol{B}$ is semi-positive definite, since then we can uniquely define $\boldsymbol{X}_0 = \sqrt{\alpha_n}\boldsymbol{\Theta}\boldsymbol{Z}\boldsymbol{B}^{1/2}$. OCCAM is thus more general than all of these models, and yet is fully identifiable and interpretable.

3.2.2 Identifiability

The parameters in (3.2.1) obviously need to be constrained to guarantee identifiability of the model. All models with communities, including the SBM, are considered identifiable if they are identifiable up to a permutation of community labels. To show the interplay between the model parameters, we first state identifiability conditions treating all of α_n, $\boldsymbol{\Theta}$, $\boldsymbol{Z}$, and $\boldsymbol{B}$ as constant parameters, and then discuss what happens if $\boldsymbol{\Theta}$ and $\boldsymbol{Z}$ are treated as random variables as we do in the asymptotic analysis. The following conditions are sufficient for identifiability:

I1 $\boldsymbol{B}$ is full rank and strictly positive definite, with $B_{kk} = 1$ for all k.

I2 All $Z_{ik} \geq 0$, $\|\boldsymbol{Z}_{i\cdot}\|_2 = 1$ for all $i = 1, \ldots, n$, and there is at least one "pure" node in every community, i.e., for each $k = 1, \ldots, K$, there exists at least one i such that

$Z_{ik} = 1$.

I3 The degree parameters $\theta_1, \ldots, \theta_n$ are all positive and $n^{-1}\sum_{i=1}^{n}\theta_i = 1$.

Theorem 2. *If conditions* (I1)*,* (I2) *and* (I3) *hold, the model is identifiable, i.e., if a given probability matrix* $\boldsymbol{W}$ *corresponds to a set of parameters* $(\alpha_n, \boldsymbol{\Theta}, \boldsymbol{Z}, \boldsymbol{B})$ *through* (3.2.1)*, these parameters are unique up to a permutation of community labels.*

The proof of Theorem 2 is given in the supplemental materials. In general, identifiability is non-trivial to establish for most overlapping community models, since, roughly speaking, an edge between two nodes can be explained by either their common memberships in many of the same communities, or the high probability of edges between their two different communities, a problem that does not occur in the non-overlapping case. Among previously proposed models, the OSBM was shown to be identifiable (Latouche et al., 2009), but their argument does not extend to our model since they only considered $\boldsymbol{Z}$ with binary entries. The identifiability of the BKN model was not discussed by Ball et al. (2011), but it is relatively straightforward (though still non-trivial) to show that it is identifiable as long as there are pure nodes in each community.

While Theorem 2 makes the model in (3.2.1) well defined, it is also common practice in the community detection literature to treat some of the model components as random quantities. For example, Holland et al. (1983) treat community labels under the SBM as sampled from a multinomial distribution, and Zhao et al. (2012) treat the degree parameters θ_i's in DCSBM as sampled from a general discrete distribution. For our consistency analysis, treating θ_i's and $\boldsymbol{Z}_{i\cdot}$'s as random significantly simplifies conditions and allows for an explicit choice of rate for the tuning parameter τ_n, which will be defined in Section 3.3. We will thus treat $\boldsymbol{\Theta}$ and $\boldsymbol{Z}$ as random and independent of each other for the purpose of theory, assuming that the rows of $\boldsymbol{Z}$ are independently generated from a distribution $\mathcal{F}_Z$ on the unit sphere, and θ_i's are i.i.d. from a distribution $\mathcal{F}_\Theta$ on positive real numbers. The conditions I2 and I3 are then replaced with the following two conditions, respectively:

RI2 $\mathcal{F}_Z = \pi_p\mathcal{F}_p + \pi_o\mathcal{F}_o$ is a mixture of a multinomial distribution $\mathcal{F}_p$ on K categories for pure nodes and an arbitrary distribution $\mathcal{F}_o$ on $\{\boldsymbol{z} \in \mathbb{R}^K : \boldsymbol{z}_k \geq 0, \|\boldsymbol{z}\|_2 = 1\}$ for nodes in the overlaps, and $\pi_p > 0$.

RI3 $\mathcal{F}_\Theta$ is a probability distribution on $(0, \infty)$ satisfying $\int_0^\infty t\, d\mathcal{F}_\Theta(t) = 1$.

The distribution $\mathcal{F}_o$ can in principle be any distribution on the positive quadrant of the unit sphere. For example, one could first specify that with probability $\pi_{k_1,\ldots,k_m}$, node i belongs to communities $\{k_1, \ldots, k_m\}$, and then set $Z_{ik} = \frac{1}{\sqrt{m}}\mathbf{1}(k \in \{k_1, \ldots, k_m\})$.

Alternatively, one could generate values for the m non-zero entries of $\boldsymbol{Z}_{i\cdot}$ from an m-dimensional Dirichlet distribution, and set the rest to 0.

Here we emphasize that the conditions guaranteeing identifiability is an indispensable part of our model. Ideas similar to (3.2.1) previously appeared in the literature, see, for example, Appendix C of Ball et al. (2011). However, without proper identifiability conditions, the parameters are not meaningful.

3.3 A spectral algorithm for fitting the model

The primary goal of fitting this model is to estimate the membership matrix $\boldsymbol{Z}$ from the observed adjacency matrix $\boldsymbol{A}$, although other parameters may also be of interest. Since computational scalability is one of our goals, we focus on algorithms based on spectral decompositions, one of the most scalable approaches available. Recall that spectral clustering typically works by first representing all data points (the n nodes) by an $n \times K$ matrix $\boldsymbol{X}$ consisting of leading eigenvectors of a matrix derived from the data, which we call $\boldsymbol{G}$ for now, and then applying K-means clustering to the rows of $\boldsymbol{X}$. For example, under the SBM, the matrix $\boldsymbol{G}$ should be chosen to have eigenvectors $\boldsymbol{X}$ that approximate the eigenvectors $\boldsymbol{X}_0$ of $\boldsymbol{W} = \mathbb{E}(\boldsymbol{A})$ as closely as possible, since the eigenvectors of $\boldsymbol{W}$ are piecewise constant and contain all the community information. A naive choice $\boldsymbol{G} = \boldsymbol{A}$ is intuitively appealing, though it has been shown in practice and in theory (Sarkar and Bickel, 2013) that the graph Laplacian of $\boldsymbol{A}$, i.e., $\boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, where $\boldsymbol{D} = diag(\boldsymbol{A}\mathbb{1})$, is a better choice, or, for sparse graphs, different regularized versions of $\boldsymbol{L}$ (Amini et al., 2013; Chaudhuri et al., 2012; Qin and Rohe, 2013; Joseph and Yu, 2013). An additional step of normalizing the rows of $\boldsymbol{X}$ before performing K-means is often appropriate if the underlying model is assumed to be the degree-corrected stochastic blockmodel (Qin and Rohe, 2013).

Regardless of the matrix chosen to estimate the eigenvectors of $\boldsymbol{W}$, the key difference between the regular SBM under which spectral clustering is usually studied and our model is that under the SBM there are only K unique rows in $\boldsymbol{X}_0$, and thus K-means can be expected to accurately cluster the rows of $\boldsymbol{X}$, which is a noisy version of $\boldsymbol{X}_0$. Under our model, the rows of $\boldsymbol{X}_0$ are linear combinations of the "pure" rows corresponding to "centers" of the K communities. Thus even if we could recover $\boldsymbol{X}_0$ exactly, K-means is not expected to work, and it is in fact straightforward to show that the K-means algorithm does not recover the positions of pure nodes correctly unless non-pure nodes either vanish in proportion or converge to pure nodes' latent positions as n grows (proof omitted here as it is not needed for our main argument). The key idea of our algorithm is to replace K-means with K-medians clustering: if the proportion of pure nodes is not too low, then

the latent positions of the cluster centers can still be recovered correctly, and therefore the coefficients of mixed nodes can be estimated accurately by projecting onto the pure nodes. Other details of the algorithm involve regularization and normalization that are necessary for dealing with sparse networks and heterogeneous degrees.

Our algorithm for fitting the OCCAM takes as input the adjacency matrix $\boldsymbol{A}$ and a regularization parameter $\tau_n > 0$ which we use to regularize the estimated latent node positions directly. This is easier to handle technically than regularizing the Laplacian, and we will give an explicit rate for τ_n that guarantees asymptotic consistency in Section 3.4. The algorithm proceeds as follows:

1. Compute $\hat{\boldsymbol{U}}_A \hat{\boldsymbol{L}}_A \hat{\boldsymbol{U}}_A^T$, where $\hat{\boldsymbol{L}}_A$ is the $K \times K$ diagonal matrix containing the K leading eigenvalues of $\boldsymbol{A}$, and $\hat{\boldsymbol{U}}_A$ is the $n \times K$ matrix containing the corresponding eigenvectors. While the true $\boldsymbol{W} = \mathbb{E}(\boldsymbol{A})$ is positive definite, in practice some of the eigenvalues of $\boldsymbol{A}$ may be negative; if that happens, we truncate them to 0. Let $\hat{\boldsymbol{X}} \equiv \hat{\boldsymbol{U}}_A \hat{\boldsymbol{L}}_A^{1/2}$ be the estimated latent node positions.

2. Compute $\hat{\boldsymbol{X}}^*$, a normalized and regularized version of $\hat{\boldsymbol{X}}$, the rows of which are given by $\hat{\boldsymbol{X}}^*_{i\cdot} = \frac{1}{\|\hat{\boldsymbol{X}}_{i\cdot}\|_2 + \tau_n} \hat{\boldsymbol{X}}_{i\cdot}$.

3. Perform K-medians clustering on the rows of $\hat{\boldsymbol{X}}^*$ and obtain K estimated cluster centers $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K \in \mathbb{R}^K$, i.e.,

$$\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K\} = \arg \min_{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K} \frac{1}{n} \sum_{i=1}^{n} \min_{\boldsymbol{s} \in \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K\}} \left\| \hat{\boldsymbol{X}}^*_{i\cdot} - \boldsymbol{s} \right\|_2 \tag{3.3.1}$$

   Form the $K \times K$ matrix $\hat{\boldsymbol{S}}$ with rows equal to the estimated cluster centers $\hat{\boldsymbol{s}}_1, \ldots, \boldsymbol{s}_K$.

4. Project the rows of $\hat{\boldsymbol{X}}^*$ onto the span of $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K$, i.e., compute the matrix $\hat{\boldsymbol{X}}^* \hat{\boldsymbol{S}}^{-1}$ and normalize its rows to have norm 1 to obtain the estimated community membership matrix $\hat{\boldsymbol{Z}}$.

This algorithm can also be used to obtain other types of community assignments. For example, to obtain binary rather than continuous community membership, we can threshold each element of $\hat{\boldsymbol{Z}}$ to obtain $\hat{Z}^0_{ik} = \mathbf{1}(\hat{Z}_{ik} > \delta_K)$ (see Section 3.5 and Section 3.6). To obtain assignments to non-overlapping communities, we can set $\hat{c}_i = \arg\max_{1 \le k \le K} \hat{Z}_{ik}$.

3.4 Asymptotic consistency

3.4.1 Main result

In this section, we show consistency of our algorithm for fitting the OCCAM as the number of nodes n and possibly the number of communities K increase. For the theoretical analysis, we treat $\boldsymbol{Z}$ and $\boldsymbol{\Theta}$ as random variables, as was done by Zhao et al. (2012). We first state regularity conditions on the model parameters.

A1 The distribution $\mathcal{F}_\Theta$ is supported on $(0, M_\theta)$, and for all $\delta > 0$ satisfies $\delta^{-1} \int_0^\delta d\mathcal{F}_\Theta(t) \leq C_\theta$, where $M_\theta > 0$ and $C_\theta > 0$ are global constants.

A2 Let λ_0 and λ_1 be the smallest and the largest eigenvalues of $\mathbb{E}[\theta_i^2 \boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot} \boldsymbol{B}]$, respectively. Then there exist global constants $M_{\lambda_0} > 0$ and $M_{\lambda_1} > 0$ such that $K\lambda_0 \geq M_{\lambda_0}$ and $\lambda_1 \leq M_{\lambda_1}$.

A3 There exists a global constant $m_B > 0$ such that $\lambda_{\min}(\boldsymbol{B}) \geq m_B$.

A key ingredient of our algorithm is the K-medians clustering, and consistency of K-medians requires its own conditions on clusters being well separated in the appropriate metric. The *sample* loss function for K-medians is defined by

$$\mathcal{L}_n(\boldsymbol{Q}; \boldsymbol{S}) = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq k \leq K} \|\boldsymbol{Q}_{i\cdot} - \boldsymbol{S}_{k\cdot}\|_2$$

where $\boldsymbol{Q} \in \mathbb{R}^{n \times K}$ is a matrix whose rows $\boldsymbol{Q}_{i\cdot}$ are vectors to be clustered, and $\boldsymbol{S} \in \mathbb{R}^{K \times K}$ is a matrix whose rows $\boldsymbol{S}_{k\cdot}$ are cluster centers.

Assuming the rows of $\boldsymbol{Q}$ are i.i.d. random vectors sampled from a distribution $\mathcal{G}$, we similarly define the *population* loss function for K-medians by

$$\mathcal{L}(\mathcal{G}; \boldsymbol{S}) = \int \min_{1 \leq k \leq K} \|\boldsymbol{x} - \boldsymbol{S}_{k\cdot}\|_2 d\mathcal{G}.$$

Finally we define the Hausdorff distance, which is used here to measure the dissimilarity between two sets of cluster centers. Specifically, for $\boldsymbol{S}, \boldsymbol{T} \in \mathbb{R}^{K \times K}$, let $D_H(\boldsymbol{S}, \boldsymbol{T}) = \min_\sigma \max_k \|\boldsymbol{S}_{k\cdot} - \boldsymbol{T}_{\sigma(k)\cdot}\|_2$, where σ ranges over all permutations of $\{1, \ldots, K\}$.

Define $\boldsymbol{X}_{i\cdot} = \theta_i \boldsymbol{Z}_{i\cdot} \boldsymbol{B}^{1/2}$ and $\boldsymbol{X}^*_{i\cdot} = \|\boldsymbol{X}_{i\cdot}\|_2^{-1} \boldsymbol{X}_{i\cdot} = \|\boldsymbol{Z}_{i\cdot} \boldsymbol{B}^{1/2}\|_2^{-1} \boldsymbol{Z}_{i\cdot} \boldsymbol{B}^{1/2}$, and let $\mathcal{F}$ denote the distribution of $\boldsymbol{X}^*_{i\cdot}$. If the distribution $\mathcal{F}$ of these linear combinations puts enough probability mass on the pure nodes (rows of $\boldsymbol{B}^{1/2}$), the rows of $\boldsymbol{B}^{1/2}$ will be recovered by K-medians clustering and then the $\boldsymbol{Z}_{i\cdot}$'s be recovered via projection. Bearing this in mind, we assume the following condition on $\mathcal{F}$ holds:

B Let $\boldsymbol{S}_{\mathcal{F}} = \arg\min_{\boldsymbol{S}} \mathcal{L}(\mathcal{F}; \boldsymbol{S})$ be the global minimizer of the population K-medians loss function $\mathcal{L}(\mathcal{F}; \boldsymbol{S})$. Then $\boldsymbol{S}_{\mathcal{F}} = \boldsymbol{B}^{1/2}$ up to a row permutation. Further, there exists a global constant M such that, for all $\boldsymbol{S}$, $\mathcal{L}(\mathcal{F}; \boldsymbol{S}) - \mathcal{L}(\mathcal{F}; \boldsymbol{S}_{\mathcal{F}}) \geq MK^{-1} D_H(\boldsymbol{S}, \boldsymbol{S}_{\mathcal{F}})$.

Condition B essentially states that the population K-medians loss function, which is determined by $\mathcal{F}$, has a unique minimum at the right place and there is curvature around the minimum.

Theorem 3 (Main theorem)**.** *Assume that the identifiability conditions I1, RI2, RI3 and regularity conditions A1-A3, B hold. If* $n^{1-\alpha_0}\alpha_n \to \infty$ *for some* $0 < \alpha_0 < 1$*,* $K = O(\log n)$*, and the tuning parameter is set to*

$$\tau_n = C_\tau \frac{\alpha_n^{0.2} K^{1.5}}{n^{0.3}} \tag{3.4.1}$$

where C_τ *is a constant, then the estimated community membership matrix* $\hat{\boldsymbol{Z}}$ *is consistent in the sense that*

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} \|\hat{\boldsymbol{Z}} - \boldsymbol{Z}\|_F \leq C(n^{1-\alpha_0}\alpha_n)^{-\frac{1}{5}} \right) \geq 1 - P(n, \alpha_n, K) \tag{3.4.2}$$

where C *is a global constant, and* $P(n, \alpha_n, K) \to 0$ *as* $n \to \infty$*.*

Remark: The condition $n^{1-\alpha_0}\alpha_n \to \infty$ is slightly stronger than $n\alpha_n \to \infty$, which was required for weak consistency of non-overlapping community detection with fixed K using likelihood or modularities by Bickel and Chen (2009), Zhao et al. (2012), and others, and which is in fact necessary under the SBM (Mossel et al., 2014). The rate at which K is allowed to grow works out to be $K = (n\alpha_n)^\delta$ for a small δ (see details in the supplemental materials), which is slower than the rates of K allowed in previous work that considered a growing K (Rohe et al., 2011; Choi et al., 2012). However, these results are not really comparable since we are facing additional challenges of overlapping communities and estimating a continuous rather than a binary membership matrix.

3.4.2 Example: checking conditions

The planted partition model is a widely studied special case which we use to illustrate our conditions and their interpretation. Let $\boldsymbol{B} = (1-\rho)\boldsymbol{I}_K + \rho \mathbf{1}\mathbf{1}^T$, $0 \leq \rho < 1$, where $\boldsymbol{I}_K$ is the $K \times K$ identity matrix and $\mathbf{1}$ is a column vector of all ones. Then $\boldsymbol{B}^{1/2}$ is a $K \times K$ matrix with diagonal entires $K^{-1}\left(\sqrt{(K-1)\rho + 1} + (K-1)\sqrt{1-\rho}\right)$ and off-diagonal

entires $K^{-1}\left(\sqrt{(K-1)\rho+1}-\sqrt{1-\rho}\right)$. We restrict the overlap to two communities at a time and generate the rows of the community membership matrix $\boldsymbol{Z}$ by

$$\boldsymbol{Z}_{i\cdot} = \begin{cases} \boldsymbol{e}_k,\ 1 \le k \le K & \text{w. prob. } \pi^{(1)}\ , \\ \frac{1}{\sqrt{2}}(\boldsymbol{e}_k + \boldsymbol{e}_l),\ 1 \le k < l \le K & \text{w. prob. } \pi^{(2)}\ , \end{cases} \tag{3.4.3}$$

where $\boldsymbol{e}_k$ is a row vector that contains a one in the kth position and zeros elsewhere, and $K\pi^{(1)} + \frac{1}{2}K(K-1)\pi^{(2)} = 1$. We set $\theta_i \equiv 1$ for all i, therefore conditions RI2 and RI3 hold.

For a $K \times K$ matrix of the form $(a-b)\boldsymbol{I}_K + b\mathbf{1}\mathbf{1}^T$, $a, b > 0$, the largest eigenvalue is $a+(K-1)b$ and all other eigenvalues are $a-b$. Thus $\lambda_{\max}(\boldsymbol{B}) = 1+(K-1)\rho$, $\lambda_{\min}(\boldsymbol{B}) = 1-\rho$, and conditions I1 and A3 hold. To verify condition A2, note $\mathbb{E}[\theta_i^2 \boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}\boldsymbol{B}] = \mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}]\boldsymbol{B}$, and since

$$\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot} = \begin{cases} \boldsymbol{e}_k^T \boldsymbol{e}_k,\ 1 \le k \le K & \text{w. prob. } \pi^{(1)}, \\ \frac{1}{2}(\boldsymbol{e}_k + \boldsymbol{e}_l)^T(\boldsymbol{e}_k + \boldsymbol{e}_l),\ 1 \le k < l \le K & \text{w. prob. } \pi^{(2)}, \end{cases}$$

we have $\mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}] = \left(\pi^{(1)} + \frac{K-2}{2}\pi^{(2)}\right)\boldsymbol{I}_K + \frac{\pi^{(2)}}{2}\mathbf{1}\mathbf{1}^T$. Therefore,

$$\lambda_{\max}(\mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}]) = \pi^{(1)} + (K-1)\pi^{(2)} \le \frac{2}{K}$$
$$\lambda_{\min}(\mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}]) = \pi^{(1)} + \frac{K-2}{2}\pi^{(2)} \ge \frac{1}{2K}$$

Since $\lambda_{\max}(\mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}]\boldsymbol{B}) \le \lambda_{\max}(\mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}])\lambda_{\max}(\boldsymbol{B})$ and $\lambda_{\min}(\mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}]\boldsymbol{B}) \ge \lambda_{\min}(\mathbb{E}[\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}])\lambda_{\min}(\boldsymbol{B})$, condition A2 holds.

It remains to check condition B. Given $\boldsymbol{x} \in \mathbb{R}^K$ with $\|\boldsymbol{x}\|_2 = 1$, for any $\boldsymbol{S}$, let $\boldsymbol{s}(\boldsymbol{x})$ and $\boldsymbol{s}_{\mathcal{F}}(\boldsymbol{x})$ be the best approximations to $\boldsymbol{x}$ in ℓ_2 norm among the rows of $\boldsymbol{S}$ and $\boldsymbol{S}_{\mathcal{F}}$ respectively.

Then we have

$$\begin{aligned}
\mathcal{L}(\mathcal{F};\boldsymbol{S}) - \mathcal{L}(\mathcal{F};\boldsymbol{B}^{1/2}) &= \left\{\pi^{(1)} D_H(\boldsymbol{S},\boldsymbol{B}^{1/2}) + \int_{\boldsymbol{x}\neq(\boldsymbol{B}^{1/2})_{k\cdot},1\leq k\leq K} \|\boldsymbol{x}-\boldsymbol{s}(\boldsymbol{x})\|_2 d\mathcal{F}\right\} \\
&\quad - \left\{\int_{\boldsymbol{x}\neq(\boldsymbol{B}^{1/2})_{k\cdot},1\leq k\leq K} \|\boldsymbol{x}-\boldsymbol{s}_{\mathcal{F}}(\boldsymbol{x})\|_2 d\mathcal{F}\right\} \\
&\geq \pi^{(1)} D_H(\boldsymbol{S},\boldsymbol{B}^{1/2}) - \int_{\boldsymbol{x}\neq(\boldsymbol{B}^{1/2})_{k\cdot},1\leq k\leq K} \|\boldsymbol{s}(\boldsymbol{x})-\boldsymbol{s}_{\mathcal{F}}(\boldsymbol{x})\|_2 d\mathcal{F} \\
&\geq \pi^{(1)} D_H(\boldsymbol{S},\boldsymbol{B}^{1/2}) - \int_{\boldsymbol{x}\neq(\boldsymbol{B}^{1/2})_{k\cdot},1\leq k\leq K} D_H(\boldsymbol{S},\boldsymbol{B}^{1/2}) d\mathcal{F} \\
&= \left(\pi^{(1)} - \frac{K(K-1)}{2}\pi^{(2)}\right) D_H(\boldsymbol{S},\boldsymbol{B}^{1/2}) \\
&= \left((K+1)\pi^{(1)} - 1\right) D_H(\boldsymbol{S},\boldsymbol{B}^{1/2}) \qquad (3.4.4)
\end{aligned}$$

We then see that in order for B to hold, i.e., for the RHS of (3.4.4) to be non-negative and equal to zero only when $D_H(\boldsymbol{S},\boldsymbol{S}_{\mathcal{F}}) = 0$, we need

$$\pi^{(1)} > \frac{1}{K+1}\left(1+\frac{M}{K}\right). \qquad (3.4.5)$$

This gives a precise condition on the proportion of pure nodes for this example. In general, the proportion of pure nodes cannot always be expressed explicitly other than through condition B.

3.5 Evaluation on synthetic networks

Our experiments on synthetic networks focus on two issues: the choice of constant in the regularization parameter τ_n, and comparisons of OCCAM to other overlapping community detection methods. Since many other methods only output binary membership vectors, we use a performance measure based on binary overlapping membership vectors. Following Lancichinetti et al. (2009), we measure performance by an extended version of the *normalized variation of information* (exNVI). Consider two binary random vectors $\boldsymbol{\Gamma} = (\Gamma_1,\ldots,\Gamma_K)$ and $\hat{\boldsymbol{\Gamma}} = (\hat{\Gamma}_1,\ldots,\hat{\Gamma}_K)$, which indicate whether a node belongs to com-

munity k in the true and estimated communities, respectively. Define

$$\begin{aligned}
\bar{H}(\hat{\Gamma}_l|\Gamma_k) &= \frac{H(\hat{\Gamma}_l|\Gamma_k)}{H(\hat{\Gamma}_k)}\ , \text{ where} \\
H(\Gamma_k) &= -\sum_z \mathbb{P}(\Gamma_k = z)\log \mathbb{P}(\Gamma_k = z), \\
H(\hat{\Gamma}_l|\Gamma_k) &= H(\Gamma_k, \hat{\Gamma}_l) - H(\Gamma_k), \text{ and} \\
H(\Gamma_k, \hat{\Gamma}_l) &= -\sum_{z,\hat{z}} \mathbb{P}(\Gamma_k = z, \hat{\Gamma}_l = \hat{z})\log \mathbb{P}(\Gamma_k = z, \hat{\Gamma}_l = \hat{z}).
\end{aligned} \tag{3.5.1}$$

where $H(\Gamma_k)$, $H(\hat{\Gamma}_l|\Gamma_k)$ and $H(\Gamma_k, \hat{\Gamma}_l)$ are commonly called individual, conditional and joint entropies. It can be seen that $\bar{H}(\hat{\Gamma}_l|\Gamma_k)$ takes values between 0 and 1, with 0 corresponding to $\hat{\Gamma}_l$ and Γ_k being independent and 1 to a perfect match. We then define the overall exNVI between $\mathbf{\Gamma}$ and $\hat{\mathbf{\Gamma}}$ to be

$$\bar{H}(\mathbf{\Gamma}, \hat{\mathbf{\Gamma}}) = 1 - \min_{\sigma} \frac{1}{2K}\sum_{k=1}^{K}\left[\bar{H}(\hat{\Gamma}_{\sigma(k)}|\Gamma_k) + \bar{H}(\Gamma_k|\hat{\Gamma}_{\sigma(k)})\right] \tag{3.5.2}$$

where σ ranges over all permutations on $\{1, \ldots, K\}$. We also define the sample versions of all the quantities in (3.5.1) with probabilities replaced with frequencies, e.g., $\hat{H}(\Gamma_k) = -\sum_{z=0}^{1} |\{i : \Gamma_{ik} = z\}|/n \cdot \log\left(|\{i : \Gamma_{ik} = z\}|/n\right)$, etc.

3.5.1 Choice of constant for the regularization parameter

The regularization parameter τ_n is defined by (3.4.1), up to a constant, as a function of n, K, and the unobserved α_n. Absorbing a constant factor into C_τ, we estimate α_n by

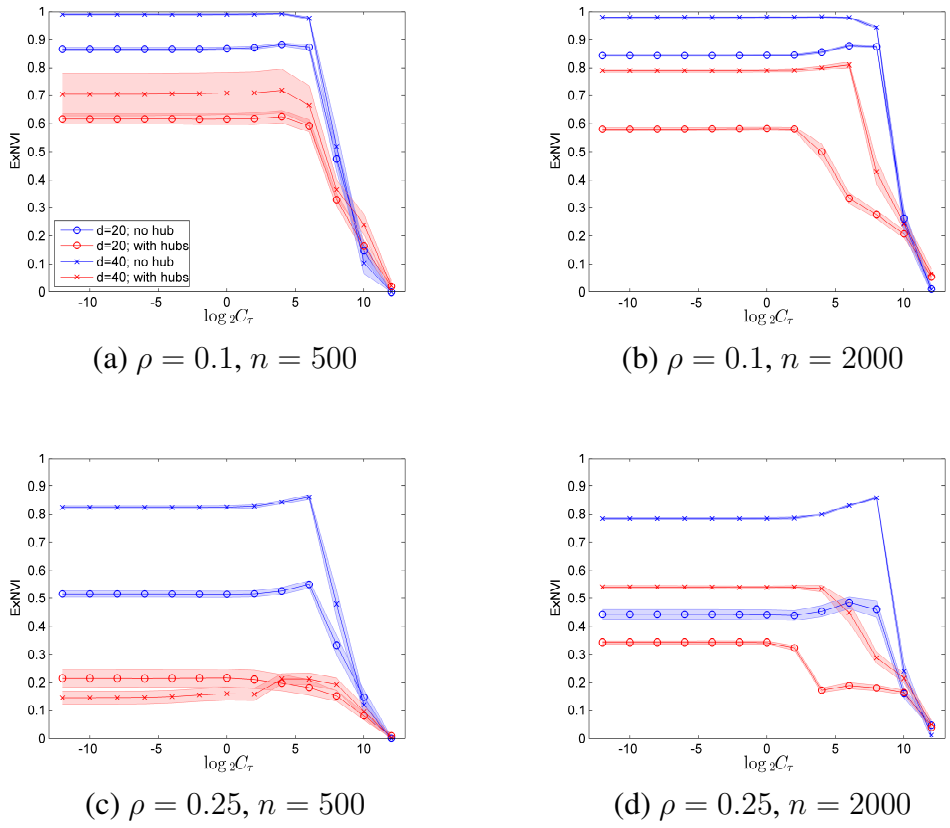
$$\hat{\alpha}_n = \frac{\sum_{i\neq j} A_{ij}}{n(n-1)K} \tag{3.5.3}$$

and investigate the effect of the constant C_τ empirically.

For this simulation, we generate networks with $n = 500$ or 2000 nodes with $K = 3$ communities. We consider two settings for θ_i's: (1) $\theta_i = 1$ for all i (no hubs), and (2) $\mathbb{P}(\theta_i = 1) = 0.8$ and $\mathbb{P}(\theta_i = 20) = 0.2$ (20% hub nodes). We generate $\boldsymbol{Z}$ as follows: for $1 \le k_1 < \ldots < k_m \le K$, we assign $n \cdot \pi_{k_1 \cdots k_m}$ nodes to the intersection of communities $k_1, \ldots, k_m$, and for each node i in this set we set $Z_{ik} = m^{-1/2}\mathbf{1}(k \in \{k_1, \ldots, k_m\})$. Let $\pi_1 = \pi_2 = \pi_3 = \pi^{(1)}$, $\pi_{12} = \pi_{13} = \pi_{23} = \pi^{(2)}$, $\pi_{123} = \pi^{(3)}$ and set $(\pi^{(1)}, \pi^{(2)}, \pi^{(3)}) = (0.3, 0.03, 0.01)$. Finally, we choose α_n so that the expected average node degree $\bar{d}$ is either 20 or 40. We vary the constant factor C_τ in (3.4.1) in the range $\{2^{-12}, 2^{-10}, \ldots, 2^{10}, 2^{12}\}$.

To use exNVI, we convert both the estimated $\hat{\boldsymbol{Z}}$ and $\boldsymbol{Z}$ to a binary overlapping community assignment by thresholding its elements at $1/K$. The results, shown in Figure 3.1, indicate that the performance of OCCAM is stable over a wide range of the constant factor ($2^{-12} - 2^5$), and degrades only for very large values of C_τ. Based on this empirical evidence, we recommend setting

$$\tau_n = 0.1 \frac{\hat{\alpha}_n^{0.2} K^{1.5}}{n^{0.3}} \,. \tag{3.5.4}$$



(a) $\rho = 0.1$, $n = 500$ (b) $\rho = 0.1$, $n = 2000$

(c) $\rho = 0.25$, $n = 500$ (d) $\rho = 0.25$, $n = 2000$

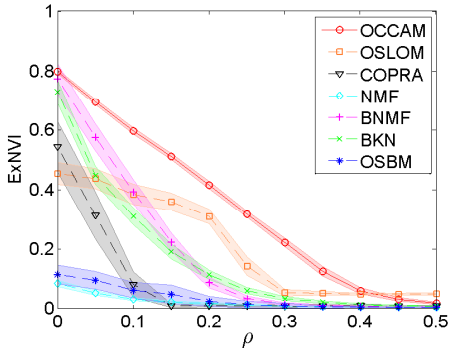
Figure 3.1: Performance of OCCAM measured by exNVI as a function of C_τ.

3.5.2 Comparison to benchmark methods

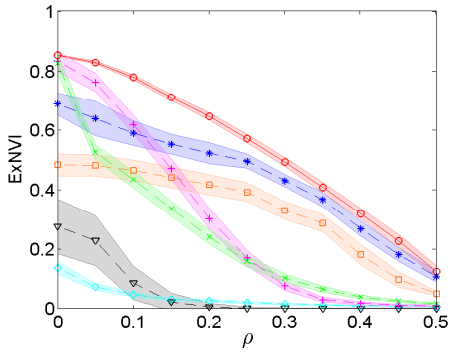
To compare OCCAM to other methods for overlapping community detection, we fix $n = 500$ and use the same settings for K, $\boldsymbol{Z}$, θ_i's and α_n as in Section 3.5.1. We set $B_{kk'} = \rho$ for $k \neq k'$, with $\rho = 0, 0.05, 0.10, \ldots, 0.5$, and set $(\pi^{(1)}, \pi^{(2)}, \pi^{(3)})$ to be either $(0.3, 0.03, 0.01)$ or $(0.25, 0.07, 0.04)$. The regularization parameter τ_n is set to the recommended value (3.5.4), and detection performance is measured by exNVI.

We compare OCCAM to both algorithmic methods and model-based methods that can be thought of as special cases of our model. Algorithmic methods we compare include the order statistics local optimization method (OSLOM) by Lancichinetti et al. (2011), the community overlap propagation algorithm (COPRA) by Gregory (2010), the nonnegative matrix factorization (NMF) on $\boldsymbol{A}$ computed via the algorithm of Gillis and Vavasis (2014), and the Bayesian nonnegative matrix factorization (BNMF) (Psorakis et al., 2011). Model-based methods we compare are two special cases of our model, the BKN overlapping community model (Ball et al., 2011) and the overlapping stochastic blockmodel (OSBM) (Latouche et al., 2009). For methods that produce continuous community membership values, thresholding was applied for the purpose of comparisons. For OCCAM and BNMF, where the membership vector is constrained to have norm 1, we use the threshold of $1/K$; for NMF, where there are no such constraints to guide the choice of threshold, we simply use a small positive number 10^{-3}; and for BKN, we follow the scheme suggested by the authors and assign node i to community k if the estimated number of edges between i and nodes in community k is greater than 1. For each parameter configuration, we repeat the experiment 200 times. Results are shown in Figure 3.2.

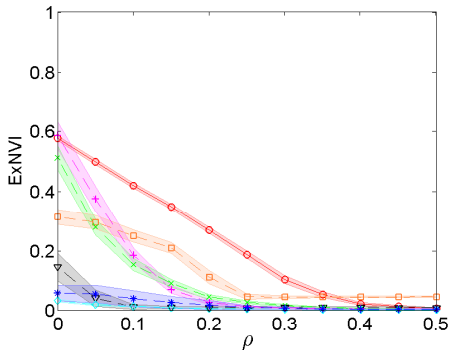
As one might expect, all methods degrade as (1) the between-community edge probability approaches the within-community edge probability (i.e., ρ increases); (2) the overlap between communities increases; and (3) the average node degree decreases. In all cases, OCCAM performs best, but we should also keep in mind that the networks were generated from the OCCAM model. BKN and BNMF perform well when ρ is small but degrade much faster than OCCAM as ρ increases, possibly because they require shared community memberships for nodes to be able to connect, thus eliminating connections between pure nodes from different communities; NMF requires this too. OSLOM detects communities by locally modifying initial estimates, and when ρ increases beyond a certain threshold, the connections between pure nodes blur the "boundaries" between communities and lead OSLOM to assign all nodes to all communities. COPRA, a local voting algorithm, is highly sensitive to ρ for the same reasons as OSLOM, and additionally suffers from numerical instability that sometimes prevents convergence. OSBM performs well under the homogeneous node degree setting (when all $\theta_i = 1$), where OSBM correctly specifies the data generating mechanism, but its performance degrades quickly in the presence of hubs. Overall, in this set of simulations OCCAM has a clear advantage over its less flexible competitors.



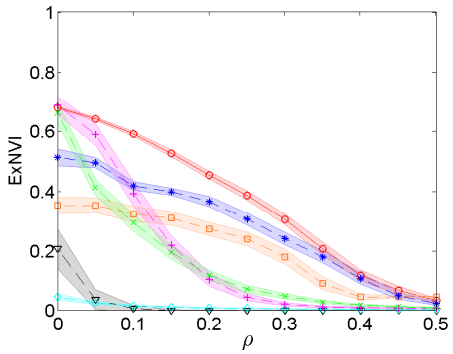
(a) A, $d = 20$, with hub nodes



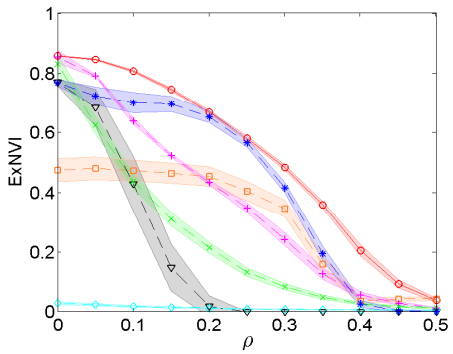
(b) A, $d = 40$, with hub nodes



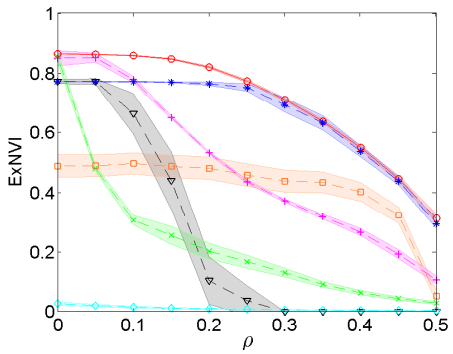
(c) B, $d = 20$, with hub nodes



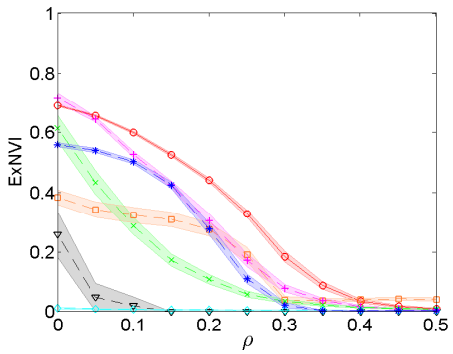
(d) B, $d = 40$, with hub nodes



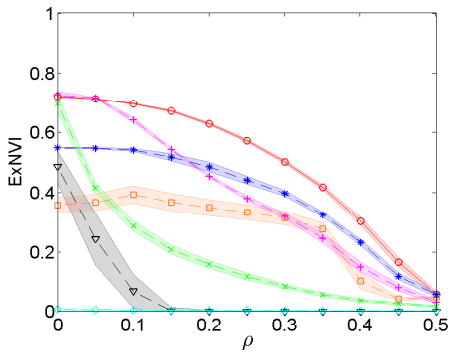
(e) A, $d = 20$, no hub nodes



(f) A, $d = 40$, no hub nodes



(g) B, $d = 20$, no hub nodes



(h) B, $d = 40$, no hub nodes

Figure 3.2: A: $(\pi^{(1)}, \pi^{(2)}, \pi^{(3)}) = (0.3, 0.03, 0.03)$; B:$(\pi^{(1)}, \pi^{(2)}, \pi^{(3)}) = (0.25, 0.07, 0.04)$

3.6 Application to SNAP ego-networks

The ego network datasets (Leskovec and Mcauley, 2012) contain more than 1000 ego-networks from Facebook, Twitter and GooglePlus. In an ego network, all the nodes are friends of one central user, and the friendship groups or circles (depending on the platform) set by this user can be used as ground truth communities. This dataset was introduced by Leskovec and Mcauley (2012), who also proposed an algorithm for overlapping community detection, which we will refer to as ML. We did not include this method in simulation studies because it uses additional node features which all other algorithms under comparison do not; however, we include it in comparisons in this section. Before comparing the methods, we carried out some pre-processing to make sure the test cases do in fact have a substantial community structure. First, we "cleaned" each network by (1) dropping nodes that are not assigned to any community; (2) dropping isolated nodes; (3) dropping communities whose pure nodes are less than 10% of the network size. Note that step (3) is done iteratively, i.e., after dropping the smallest community that does not meet this criterion, we inspect all remaining communities again and continue until either all communities meet the criterion or only one community remains. After this process is complete, we select cleaned networks that (a) contain at least 30 nodes; (b) have at least 2 communities; and (c) have Newman-Girvan modularities (Newman and Girvan, 2004a) on the true communities of no less than 0.05, indicating some assortative community structure is present. These three rules eliminated 19, 45 and 28 networks respectively of the 132 GooglePlus networks, 455, 236 and 99 networks respectively out of 973 Twitter networks, and (b) eliminated 3 out of 10 Facebook networks. The remaining 40 GooglePlus networks, 183 Twitter networks, and 7 Facebook networks were used in all comparisons, using exNVI to measure performance.

To get a better sense of what the different social networks look like and how different characteristics potentially affect performance, we report the following summary statistics for each network: (1) density $\sum_{ij} A_{ij}/(n(n-1))$, i.e., the overall edge probability; (2) average node degree d; (3) the coefficient of variation of node degrees (the standard deviation divided by the mean) σ_d/d, which measures the amount of heterogeneity in the node degrees; (4) the proportion of overlapping nodes r_o; (5) Newman-Girvan modularity. Even though modularity was defined for non-overlapping communities, it still reflects the strength of the community structure in the networks in this dataset, which only have a modest amount of overlaps. We report the means and standard deviations of these measures for each of the social networks in Table 3.1. Note that Facebook and Gplus networks tend to be larger than Twitter networks, while Twitter networks tend to be denser, with more homogeneous degrees as reflected by σ_d/d, though their smaller size makes these measures

less reliable.

To compare methods, we report the average performance over each of the social platforms and the corresponding standard deviation in Table 3.2. We also report the mean pairwise difference between OCCAM and each of the other methods, along with its standard deviation in Table 3.3.

Table 3.1: Mean (SD) of summary statistics for ego-networks

	#Networks	n	K	Density	d	σ_d/d	r_o	Modularity
Facebook	7	224	3.3	0.137	28	0.644	0.030	0.418
	-	(221)	(0.8)	(0.046)	(29)	(0.145)	(0.021)	(0.148)
Gplus	40	414	2.3	0.170	53	1.035	0.057	0.171
	-	(330)	(0.5)	(0.109)	(34)	(0.471)	(0.077)	(0.109)
Twitter	183	62	2.8	0.264	15	0.595	0.036	0.204
	-	(31)	(0.9)	(0.264)	(8)	(0.148)	(0.055)	(0.119)

Table 3.2: Mean (SD) of exNVI for all methods.

	OCCAM	OSLOM	COPRA	NMF	BNMF	BKN	OSBM	ML
Facebook	**0.576**	0.212	0.394	0.314	0.500	0.474	0.473	0.133
	(0.116)	(0.068)	(0.115)	(0.079)	(0.094)	(0.107)	(0.114)	(0.033)
Gplus	**0.503**	0.126	0.114	0.293	0.393	0.357	0.333	0.175
	(0.038)	(0.017)	(0.036)	(0.036)	(0.046)	(0.030)	(0.039)	(0.023)
Twitter	**0.451**	0.208	0.232	0.212	0.437	0.346	0.348	0.200
	(0.021)	(0.012)	(0.023)	(0.013)	(0.021)	(0.017)	(0.017)	(0.010)

Table 3.3: Mean (SD) of pairwise differences in exNVI between OCCAM and other methods.

	vs OSLOM	vs COPRA	vs NMF	vs BNMF	vs BKN	vs OSBM	vs ML
Facebook	0.363	0.182	0.261	0.075	0.101	0.102	0.443
	(0.093)	(0.082)	(0.071)	(0.072)	(0.053)	(0.032)	(0.134)
Gplus	0.377	0.389	0.210	0.110	0.146	0.171	0.328
	(0.037)	(0.037)	(0.040)	(0.038)	(0.020)	(0.028)	(0.042)
Twitter	0.243	0.219	0.239	0.014	0.105	0.103	0.251
	(0.020)	(0.019)	(0.016)	(0.012)	(0.012)	(0.011)	(0.024)

As in simulation studies, we observe that OCCAM outperforms other methods. Gplus networks on average have the most heterogeneous node degrees and thus are challenging

for COPRA and OSBM, while OCCAM is relatively robust to node degree heterogeneity. Further, Gplus networks tend to have higher proportions of overlapping nodes than Facebook networks; this creates difficulties for all methods. Empirically, we also found that OSLOM and COPRA are prone to convergence to degenerate community assignments, assigning all nodes to one community. NMF, BNMF and BKN often create substantial overlaps compared to other methods, likely because they do not allow connections between pure nodes from different communities. The results suggest that OCCAM works well when the overlap is not large even when modularity is relatively low, while other methods are more sensitive to modularity, which measures the strength of an assortative community structure. On the other hand, large overlaps between communities cause the performance of OCCAM to deteriorate, which is consistent with our theoretical results. ML is not readily comparable to others since it uses both network information and node features when fitting the model, and one would expect it do to better since it makes use of more information; however, using node features that are uncorrelated with the community structure can in fact worsen community detection, which may explain its poor performance on some of the networks.

A fair comparison of computing times is difficult because the methods compared here are implemented in different languages. Qualitatively, we can say that the most expensive part of OCCAM is the K-medians clustering, which involves gradient descent, and is about one order of magnitude slower than NMF. The computational cost of OCCAM is comparable to that of BNMF, BKN and COPRA, and is at least two orders of magnitude less than that of OSLOM, OSBM and ML.

3.7 Discussion

This chapter makes two major contributions, the model and the algorithm. The model we proposed for overlapping communities, OCCAM, is identifiable, interpretable, and flexible; it addresses limitations of several earlier approaches by allowing continuous community membership, allowing for pure nodes from different communities to be connected, and accommodating heterogeneous node degrees. Our goals in designing an algorithm to fit the model were scalability and of course accuracy, and therefore we made a number of modifications to spectral clustering to deal with the overlaps, most importantly replacing K-means with K-medians. Empirically we found the algorithm is a lot faster than most of its competitors, and it performs well on both synthetic and real networks. We also showed estimation consistency under conditions that articulate the appropriate setting for our method – the overlaps are not too large and the network is not too sparse (the latter

being a general condition for all community detection consistency, and the former specific to our method).

In addition to its many advantages, our method has a number of limitations. The upper bound on the amount of overlap is a restriction, expressed by implicit condition B, which may not be easy to verify except in special cases. It is clear, however, that some limit on the amount of overlap is necessary for any model to be identifiable. Like all other spectral clustering based methods, OCCAM works best when communities have roughly similar sizes; this is implied by condition B which implicitly excludes communities of size $o(n/K)$ as n and K grow. Further, our model only applies to assortative communities, in other words, requires the matrix of probabilities B to be positive definite. This constraint seems to be unavoidable if the model is to be identifiable.

Like the vast majority of existing community detection methods, we assume that the number of communities K is given as input to the algorithm. There has been some very recent work on choosing K by hypothesis testing (Bickel and Sarkar, 2015) or a BIC-type criterion (Saldana et al., 2014) for the non-overlapping case; testing these methods and adapting them to the overlapping case is a topic for future work which is outside the scope of this manuscript but is an interesting topic. Another interesting and difficult challenge is detecting communities in the presence of "outliers" that do not belong to any community, considered by Zhao et al. (2011) and Cai and Li (2015). Our algorithm may be able to do this with additional regularization. Finally, incorporating node features when they are available into overlapping community detection is another challenging task for future, since the features may introduce both additional useful information and additional noise.

CHAPTER 4

Estimating network edge probabilities by neighborhood smoothing

4.1 Introduction

Statistical network analysis spans a wide range of disciplines (network science, statistics, physics, computer science, sociology, and others) and an equally wide range of applications and analysis tasks (community detection, link prediction, etc). In this chapter, we study the problem of inferring the generative mechanism of an undirected network based on a single realization of the network. The data consist of the network adjacency matrix $A \in \{0,1\}^{n\times n}$, where n is the number of nodes, and $A_{ij} = A_{ji} = 1$ if there is an edge between nodes i and j. We assume the observed adjacency matrix A is generated from an underlying probability matrix P, so that for $i \leq j$, A_{ij}'s are independent Bernoulli(P_{ij}) trials, and P_{ij}'s are edge probabilities

It is obviously impossible to estimate P from a single realization of A unless one assumes some form of structure in P. When the network is expected to have communities, arguably the most popular assumption is that of the stochastic block model, where each node belongs to one of K blocks and the probability of an edge between two nodes is determined by the blocks the nodes belong to. In this case, the $n \times n$ matrix P is parametrized by the $K \times K$ matrix of within- and between-block edge probabilities, and thus it is possible to estimate P from a single realization. The main challenge in fitting the stochastic block model is estimating the blocks themselves, and that has been the focus of the literature, see for example Bickel and Chen (2009); Rohe et al. (2011); Amini et al. (2013); Saade et al. (2014) and a technical report by Guédon and Vershynin (arXiv:1411.4686) . Once the blocks are estimated, P can be estimated efficiently by a plug-in moment estimator. Many extensions and alternatives to the stochastic block model have been proposed to model networks with communities, see Hoff (2008); Airoldi et al. (2008); Karrer and Newman (2011); Cai and Li (2015) and a technical report by Zhang et al (arXiv:1412.3432) , but their properties are

generally only known under the correctly specified model with communities, and here we are interested in estimating P for more general networks.

A general representation for the matrix P for unlabeled networks, where any permutation of nodes defines the same network goes back to Aldous (1981) and Hoover (1979). Formally, a network is *exchangeable*, that is, for any permutation π of the set $\{1, \ldots, n\}$, the distribution of edges is invariant under permutations of node labels. That is, if the adjacency matrix $A = [A_{ij}]$ is drawn from the probability matrix P as described above (which we write as $A \sim P$), then for any permutation π,

$$\left[A_{\pi(i)\pi(j)}\right] \sim P \, . \tag{4.1.1}$$

Aldous and Hoover showed that an exchangeable network always admits the following representation:

Definition 4 (Aldous-Hoover representation)**.** *For any network satisfying* (4.1.1)*, there exists a function* $f : [0,1] \times [0,1] \to [0,1]$ *and a set of i.i.d. random variables* $\xi_i \sim$ *Uniform* $[0,1]$*, such that*

$$P_{ij} = f(\xi_i, \xi_j) \, . \tag{4.1.2}$$

Following the literature, we call f the graphon function. Unfortunately, f in this representation is neither unique nor identifiable, see a technical report by Diaconis and Janson (arXiv:0712.2749), since for any measure-preserving one-to-one transformation $\sigma : [0,1] \to [0,1]$, both $f(\sigma(\cdot), \sigma(\cdot))$ and $f(\cdot, \cdot)$ yield the same distribution of A. An identifiable and unique canonical representation can be defined if one requires $g(u) = \int_0^1 f(u,v)\mathrm{d}v$ to be non-decreasing (Bickel and Chen, 2009), and it was shown that f and ξ_i's are jointly identifiable when $g(u)$, which can be interpreted as expected node degree, is strictly monotone , see Chan and Airoldi (arXiv:1402.1888) . This assumption is strong and excludes the stochastic block model.

In practice, the main purpose of estimating the function f is to estimate P, and thus identifiability of f or lack thereof may not matter as long as P itself can be estimated. On the other hand, in practice people care more about P than f. It is shown that the measure-preserving map σ is the only source of non-identifiability (Hoover (1979) and Diaconis and Janson (arXiv:0712.2749)) . A technical report by Wolfe and Ohlede (arXiv:1309.5936) and Choi and Wolfe (2014) proposed estimating f up to measure-preserving transformation σ via step-function approximations based on fitting the stochastic block model with a larger number of blocks K. This approximation does not assume the network itself follows the block model, and some theoretical guarantees have been obtained under more general models. In related work, Olhede and Wolfe (2014) proposed to approximate the graphon with

"network histograms", that is, stochastic block models with many blocks of equal size, akin to histogram bins. Another method to compute a network histogram was proposed in a technical report by Amini and Levina (arXiv:1406.5647) , as an application of their semi-definite programming approach to fitting block models with equal size blocks. Quite recently, Gao et al. (2014) established the minimax error rate for estimating P and proposed a least squares type estimator to achieve this rate, which obtains the estimated probability P by averaging the adjacency matrix elements within a given block partition. A similar estimator was proposed in a technical report by Choi (arXiv:1507.06352) , applicable also to non-smooth graphons. However, these methods are in principle computationally infeasible since they require an exhaustive enumeration of all possible block partitions. Also Newman and Peixoto (2015) proposed to estimate the latent node positions using an EM method. Despite the lack of identifiabiliaty, we conjecture that such a method may produce a good stochastic blockmodel approximations under proper conditions. A technical report by Cai et al (arXiv:1412.2129) proposed an iterative algorithm to fit a stochastic blockmodel and approximate the graphon, but the error rate of this method is unknown for general graphons. A Bayesian approach using block priors proposed in a technical report by Gao et al (arXiv:1506.02174) achieves the minimax error rate adaptively, but it still requires evaluating the posterior likelihood over all possible block partitions to obtain the posterior mode or the expectation for the probability matrix.

Other recent efforts on graphon estimation focus on the case of monotone node degrees, which make the graphon identifiable. The sort and smooth methods as in a technical report by Chan and Airoldi (arXiv:1402.1888) and Yang et al. (2014) estimate the graphon under this assumption by first sorting nodes by their degrees and then smoothing the matrix A locally to estimate edge probabilities. The monotone degree assumption is crucial for the success of these methods, and as we later show in this chapter the sort and smooth methods perform poorly when it does not hold. Finally, general matrix denoising methods can be applied to this problem if one considers A to be a noisy version of its expectation P; a good general representative of this class of methods is the universal singular value thresholding approach of Chatterjee (2014). Since this is a general method, we cannot expect its error rate to be especially competitive for this specific problem, and indeed its mean squared error rate is slower than the cubic root of the minimax rate.

In this thesis, we propose a novel computationally efficient method for probability matrix estimation based on neighborhood smoothing, for piecewise Lipschitz graphon functions. The key to this method is adaptive neighborhood selection, which allows us to avoid making strong assumptions such as monotone node degrees. A node's neighborhood consists of nodes with similar rows in the adjacency matrix, which intuitively correspond to

nodes with similar values of the latent node positions ξ_i.To the best of our knowledge, our estimator achieves the best error rate among existing computationally feasible methods. Computationally, the estimator allows easy parallelization. The size of the neighborhood is controlled by a tuning parameter, similar to bandwidth in nonparametric regression; the rate of this bandwidth parameter is determined from theory, and we show empirically the method is robust to the choice of the constant. Experiments on synthetic networks demonstrate our method performs very well under a wide range of graphon models, including low rank and full rank, with monotone degrees and without, and so on . We also test the performance of our method on the link prediction problem, using both synthetic and real networks.

4.2 The neighborhood smoothing estimator and its error rate

4.2.1 Neighborhood smoothing for edge probability estimation

Our goal is to estimate the probabilities P_{ij} from the observed network adjacency matrix A, where A_{ij} is drawn from Bernoulli(P_{ij}) and all A_{ij}'s are independent. While $P_{ij} = f(\xi_i, \xi_j)$, where ξ_i's are latent, our goal is to estimate P for the single realization of ξ_i's that gave rise to the data, rather than the function f. We think of f as a fixed unknown smooth function on $[0, 1]^2$, with formal smoothness assumptions to be stated later on. Let $e_{ij} = e_{ij}(P_{ij})$ denote the Bernoulli error and omit its dependence on P. We can then write

$$A_{ij} = P_{ij} + e_{ij} = f(\xi_i, \xi_j) + e_{ij}. \tag{4.2.1}$$

Formulation (4.2.1) resembles a nonparametric regression problem, but with the important difference that ξ_i's are not observed. This has important consequences, for example, assuming further smoothness in f beyond order one does not improve the minimax error rate when estimating P (Gao et al., 2014). The idea of our method is to apply neighborhood smoothing, which would be a natural approach had the latent variables ξ_i's been observed. Intuitively, if we had a set $\mathcal{N}_i$ of neighbors of a node i, in the sense that $\mathcal{N}_i = \{i' : P_{i'\cdot} \approx P_{i\cdot}\}$, where $P_{i\cdot}$ represents the i-th row of P, then we could estimate $P_{i\cdot}$ by averaging $A_{i'\cdot}$ over $i' \in \mathcal{N}_i$. Postponing the question of how to select $\mathcal{N}_i$ until Section

4.2.2, we can define a general form of the neighborhood smoothing estimator by

$$\hat{P}_{ij} = \frac{1}{2}\left(\frac{\sum_{i' \in \mathcal{N}_i} A_{i'j}}{|\mathcal{N}_i|} + \frac{\sum_{j' \in \mathcal{N}_j} A_{ij'}}{|\mathcal{N}_j|}\right). \tag{4.2.2}$$

It is immediately evident that $\hat{P}$ is symmetric if A is symmetric, although it can be applied to either directed or undirected networks. For simplicity, in this chapter we focus on undirected networks. A natural alternative is to average over $\mathcal{N}_i \times \mathcal{N}_j$, but (4.2.2) allows vectorization and is thus more computationally efficient. Our estimator can also be viewed as a relaxation of step function approximations such as Olhede and Wolfe (2014). In step function approximations, the neighborhood for each node is the nodes from its block, so the neighborhoods for two nodes from the same block are very similar, and the blocks have to be estimated first; in contrast, neighborhood smoothing provides for more flexible neighborhoods that are different from node to node, and an efficient way to select the neighborhood, which we will discuss next.

4.2.2 Neighborhood selection

Selecting the neighborhood $\mathcal{N}_i$ in (4.2.2) is at the core of our method. Since we estimate $P_{i\cdot}$ by averaging over $A_{i'\cdot}$ for $i' \in \mathcal{N}_i$, good neighborhood candidates i' should have $f(\xi_{i'}, \cdot)$ close to $f(\xi_i, \cdot)$, which implies $P_{i'\cdot}$ close to $P_{i\cdot}$. We use the ℓ_2 distance between graphon slices to quantify this, defining

$$d(i, i') = \|f(\xi_i, \cdot) - f(\xi_{i'}, \cdot)\|_2 = \left\{\int_0^1 |f(\xi_i, v) - f(\xi_{i'}, v)|^2 \, \mathrm{d}v\right\}^{1/2} \tag{4.2.3}$$

While one may consider more general ℓ_p or other distances, the ℓ_2 distance is particularly easy to work with when it comes to theory. For the purpose of neighborhood selection, it is not necessary to estimate $d(i, i')$; it suffices to provide a tractable upper bound. For integrable functions g_1 and g_2 defined on $[0, 1]$, define $\langle g_1, g_2\rangle = \int_0^1 g_1(u)g_2(u)\mathrm{d}u$. Then we can write

$$d^2(i, i') = \langle f(\xi_i, \cdot), f(\xi_i, \cdot)\rangle + \langle f(\xi_{i'}, \cdot), f(\xi_{i'}, \cdot)\rangle - 2\langle f(\xi_i, \cdot), f(\xi_{i'}, \cdot)\rangle. \tag{4.2.4}$$

The third term in (4.2.4) can be estimated by $2\langle A_{i\cdot}, A_{i'\cdot}\rangle/n$, where $A_{i\cdot}$ and $A_{i'\cdot}$ are nearly independent (up to a single duplicated entry due to symmetry). The first two terms in (4.2.4) are more difficult since $\langle A_{i\cdot}, A_{i\cdot}\rangle/n$ is not a good estimator for $\langle f(\xi_i, \cdot), f(\xi_i, \cdot)\rangle$. Here we present the intuition and provide a full theoretical justification in Theorem 6. For simplicity,

assume for now f is Lipschitz with a Lipschitz constant of 1. The idea is to use nodes with graphon slices similar to i and i' to make the terms in the inner product distinct graphon slices. With high probability, for each i, we can find $\tilde{i} \neq i$ such that $|\xi_{\tilde{i}} - \xi_i| \leq e_n$, where $e_n = o(1)$ is the error rate to be specified later. Then we have $\|f(\xi_i, \cdot) - f(\xi_{\tilde{i}}, \cdot)\|_2 \leq e_n$, and we can approximate $\langle f(\xi_i, \cdot), f(\xi_i, \cdot)\rangle$ by $\langle f(\xi_i, \cdot), f(\xi_{\tilde{i}}, \cdot)\rangle$, where the latter can now be estimated by $\langle A_{i\cdot}, A_{\tilde{i}\cdot}\rangle/n$. The same technique can be used to approximate the second term in (4.2.4), but all these approximations depend on the unknown ξ's. To deal with this, we rearrange the terms in (4.2.4) as follows:

$$\begin{aligned} d^2(i,i') &= \langle f(\xi_i,\cdot) - f(\xi_{i'},\cdot), f(\xi_i,\cdot)\rangle - \langle f(\xi_i,\cdot) - f(\xi_{i'},\cdot), f(\xi_{i'},\cdot)\rangle \\ &\leq |\langle f(\xi_i,\cdot) - f(\xi_{i'},\cdot), f(\xi_{\tilde{i}},\cdot)\rangle| + |\langle f(\xi_i,\cdot) - f(\xi_{i'},\cdot), f(\xi_{\tilde{i'}},\cdot)\rangle| + 2e_n \\ &\leq 2\max_{k\neq i,i'} |\langle f(\xi_i,\cdot) - f(\xi_{i'},\cdot), f(\xi_k,\cdot)\rangle| + 2e_n \end{aligned} \tag{4.2.5}$$

The inner product on the right side of (4.2.5) can be estimated by

$$\tilde{d}^2(i,i') = \max_{k\neq i,i'} |\langle A_{i\cdot} - A_{i'\cdot}, A_{k\cdot}\rangle| \,/n\,. \tag{4.2.6}$$

Intuitively, the neighborhood $\mathcal{N}_i$ should consist of i's with small $\tilde{d}(i,i')$. To formalize this, let $q_i(h)$ denote the h-th sample quantile of the set $\left\{\tilde{d}(i,i') : i' \neq i\right\}$, where h is a tuning parameter, and set

$$\mathcal{N}_i = \left\{i' \neq i : \tilde{d}(i,i') \leq q_i(h)\right\} \tag{4.2.7}$$

where for notational simplicity we suppress the dependence of $\mathcal{N}_i$ on h. Thresholding at a quantile rather than at some absolute value is convenient since real networks vary in their average node degrees and other parameters, which leads to very different values and distributions of $\tilde{d}$. Empirically, thresholding at a quantile shows significant advantage in stability and performance compared to an absolute threshold. The choice of h will be guided by both theory in Section 4.2.3 , which suggests the order of h, and empirical performance which suggests the constant factor. See Appendix for more details.

An important feature of this definition is that the neighborhood admits nodes with similar graphon slices, but not necessarily similar ξ's. For example, in the stochastic block model, all nodes from the same block would be equally likely to be included in each other's neighborhoods, regardless of their ξ's. Even though we use ξ_i and $\xi_{i'}$ to motivate (4.2.5), we always work with the function values $f(\xi_i, \xi_j)$'s and never attempt to estimate the ξ_i or f by themselves. This sharply contrasts with the approaches of Chan and Airoldi 2014 (arXiv:1402.1888) and Yang et al. (2014), and gives us a substantial computational advantage as well as much more flexibility in assumptions.

4.2.3 Consistency of the neighborhood smoothing estimator

We study the theoretical properties of our estimator for a family of piecewise Lipschitz graphon functions, defined as follows.

Definition 5 (Piecewise Lipschitz graphon family)**.** *For any $\delta, L > 0$, let $\mathcal{F}_{\delta;L}$ denote a family of piecewise Lipschitz graphon functions $f : [0,1]^2 \to [0,1]$ such that: (i) there exists an integer $K \geq 1$ and a sequence $0 = x_0 < \ldots < x_K = 1$ satisfying $\min_{0\leq s\leq K-1}(x_{s+1} - x_s) \geq \delta$, and (ii) both $|f(u_1, v) - f(u_2, v)| \leq L|u_1 - u_2|$ and $|f(u, v_1) - f(u, v_2)| \leq L|v_1 - v_2|$ hold for all $u, u_1, u_2 \in [x_s, x_{s+1}]$, $v, v_1, v_2 \in [x_t, x_{t+1}]$ and $0 \leq s, t \leq K - 1$.*

Then we have the following error rate bound.

Theorem 6. *Assume that L is a global constant and $\delta = \delta(n)$ depends on n, satisfying $\lim_{n\to\infty} \delta/\left(\frac{\log n}{n}\right)^{1/2} \to \infty$. Then the estimator $\hat{P}$ defined in* (4.2.2)*, with neighborhood $\mathcal{N}_i$ defined in* (4.2.7) *and $h = C\left(\frac{\log n}{n}\right)^{1/2}$ for any global constant $C \in (0, 1]$, satisfies*

$$\max_{f\in\mathcal{F}_{\delta;L}} \text{pr}\left\{\frac{1}{n^2}\|\hat{P} - P\|_F^2 \geq C_1\left(\frac{\log n}{n}\right)^{1/2}\right\} \leq n^{-C_2} \tag{4.2.8}$$

where C_1 and C_2 are global constants.

To the best of our knowledge, this is the best error rate available to date among non-combinatorial cost graphon estimation methods. The minimax error rate $\log n/n$ established by Gao et al. (2014) has (so far) only been achieved by methods that require combinatorial optimizations or evaluations, including Gao et al. (2014), Klopp et al. (2015) and a technical report by Klopp et al (arXiv:1507.04118) . The rate $(\log n/n)^{1/2}$ was also previously only achieved by combinatorial methods, including Wolfe and Ohlede (arXiv:1309.5936) and Olhede and Wolfe (2014). Among computationally efficient methods, the best error rate we are aware of is achieved by singular value thresholding proposed in Chatterjee (2014) at $n^{-1/3}$ (Theorem 2.7). Additionally, the sort-and-smooth method proposed by Chan and Airoldi 2014 (arXiv:1402.1888) achieves the minimax error rate under the strong assumption that f has strictly monotone expected node degrees $d_f(v) = \int_0^1 f(u, v)\mathrm{d}u$,

4.3 Probability matrix estimation on synthetic networks

4.3.1 Comparison with benchmarks

In this section we evaluate the performance of our estimator on two tasks, estimating the probability matrix and link prediction, using synthetic networks. We generate the networks from the four graphons listed in Table 4.1, selected to have different features in different combinations (monotone degrees or not, low rank or not, etc). These graphons (represented by the corresponding probability matrix P) are also pictured in the first panels of Figures 4.1 – 4.4. For all networks, we use $n = 2000$ nodes to generate P from the function f.

Table 4.1: Synthetic graphons

Graphon	Function $f(u, v)$	Monotone degrees	Rank	Local structure
1	$k/(K+1)$ if $u, v \in ((k-1)/K, k/K)$, $0.3/(K+1)$ otherwise; $K = \lfloor \log n \rfloor$	Yes	$\lfloor \log n \rfloor$	No
2	$\sin\left(5\pi(u+v-1)+1\right)/2 + 0.5$	No	3	No
3	$1 - \left[1 + \exp\left\{15\big(0.8\lvert u-v\rvert\big)^{4/5} - 0.1\right\}\right]^{-1}$	No	Full	No
4	$(u^2+v^2)/3\cos\left(1/\left(u^2+v^2\right)\right) + 0.15$	No	Full	Yes

In this experiment, we compare the performance of a number of popular benchmarks for estimating P. From the general matrix denoising methods, we selected the widely used method of universal singular value thresholding (Chatterjee, 2014) to include in the comparison. We also compare to the sort and smooth methods of Chan and Airoldi 2014 (arXiv:1402.1888) and Yang et al. (2014). These two methods are similar, with the difference that the latter method employs singular value thresholding to denoise the network as a pre-processing step. We also include two step function approximations based on fitting a stochastic block model. One is the oracle stochastic blockmodel, where the blocks are formed based on the actual values of the latent ξ_i's. This is obviously not a method that can be implemented in practice, but we use it as the gold standard of what can be achieved with an stochastic blockmodel-based step function approximation. The practical version of this we compare to is step function approximation based on a stochastic blockmodel fit by regularized spectral clustering (Qin and Rohe, 2013). Any other algorithm for fitting the stochastic blockmodel can be used to estimate the blocks; for example, Olhede and Wolfe (2014) used a local updating algorithm initialized with spectral clustering to compute their network histograms. Here we chose regularized spectral clustering because of its speed and good empirical performance. For both stochastic blockmodel-based approximations, we set the number of blocks to $n^{1/2}$, as proposed by Olhede and Wolfe (2014). A recent independent method proposed in a technical report by Airoldi in 2015, which

we shall cite as "Airoldi (2015)", proposes a stochastic blockmodel approximation, which is an evolution of the method in Airoldi et al. (2013) in that it works with a single adjacency matrix as input. It defined a dissimilarity measure between each node pair (i, i') as $\sum_{k \neq i,i'} |\langle A_{i\cdot} - A_{k\cdot}, A_{k\cdot} \rangle|$, which coincides that in an earlier version of our work. Then it builds blocks in a collective fashion by starting with one not-yet-clustered node and admitting all nodes whose dissimilarity measures with the node is below a threshold Δ. Under this approach, nodes from the same block can be viewed as neighbors to each other, and the dissimilarity measure that our method uses (4.2.6) leads to a better guaranteed error rate than that in Airoldi (2015). The strategy of thresholding by quantile of our method is also advantageous in efficiency and stability, whereas choosing a proper threshold Δ can be challenging in practice.

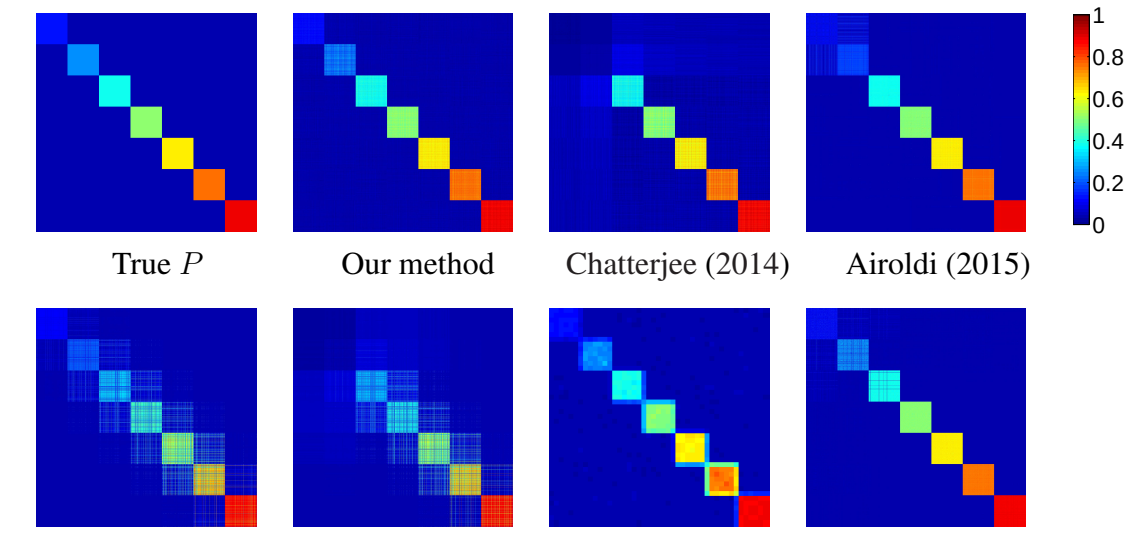


Figure 4.1: Estimated probability matrices for Graphon 1.

Figure 4.1 shows the results for Graphon 1. The network contains $\lfloor \log n \rfloor = 7$ blocks with different within-block edge probabilities, which all dominate the low between-block probability. The best results are obtained by our method and the two stochastic blockmodel methods (one of which is the oracle), which is expected given that the data are in fact generated from a stochastic block model. The two sort-and-smooth methods correctly estimate the main blocks because the blocks have different expected degrees, but they suffer from boundary effects due to smoothing over the entire region. In contrast, our method, which determines smoothing neighborhoods based on similarities of graphon slices, does not suffer from such boundary effects at all. Chatterjee (2014) does a good job on blocks with larger expected degrees, but thresholds away sparser blocks; this defect is inherited by the

method of Yang et al. (2014), which relies on Chatterjee (2014) as pre-processing. Airoldi (2015) performs similarly to our method but with a slightly lower resolution at sparser blocks, perhaps due to the dissimilarity measure it uses.

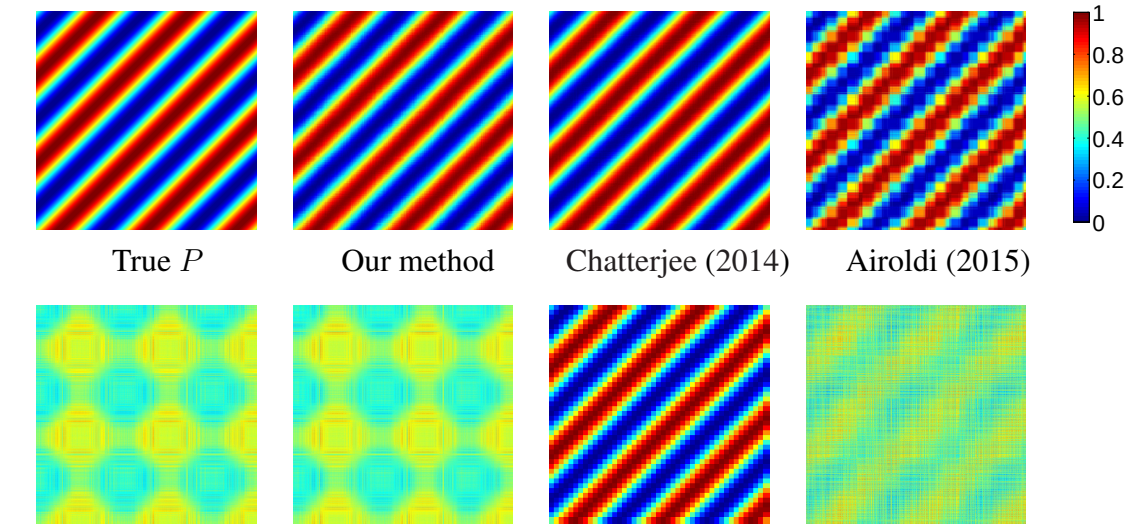


Figure 4.2: Estimated probability matrices for Graphon 2.

Figure 4.2 shows the estimation results for Graphon 2. This graphon lacks node degree monotonicity, and thus sort-and-smooth methods do not work here at all. Spectral clustering also performs poorly since the $n^{1/2}$ eigenvectors it uses turn out to be too noisy. Airoldi (2015) and stochastic blockmodel oracle method gives a grainy but reasonable approximation to P, and the best results are obtained by our method and by Chatterjee (2014), which is expected to work well here since this is a low rank matrix.

Figure 4.3 shows the estimation results for Graphon 3. Here the probabilities drop off sharply away from the diagonal, and our method captures the main structures but suffers some boundary effects due to smoothing. Nonetheless, it still provides the best approximation, apart from the oracle. Chatterjee (2014) does not perform well because this is not a low rank matrix; spectral clustering, on the other hand, does fine, because there are many non-zero eigenvalues and the $n^{1/2}$ eigenvectors used in spectral clustering contain meaningful information. Airoldi (2015) roughly recovers the region of high values but misses its smooth boundaries. The sort and smooth methods fail since all node expected degrees are almost the same and the sorting produces nothing but noise.

Finally, graphon 4 shown in Figure 4.4 is difficult to estimate for all methods. The graphon is full rank but with eigenvalues at different scales, and the adjacency matrix tends to have a spectrum very different from the probability matrix. Therefore, this is a very

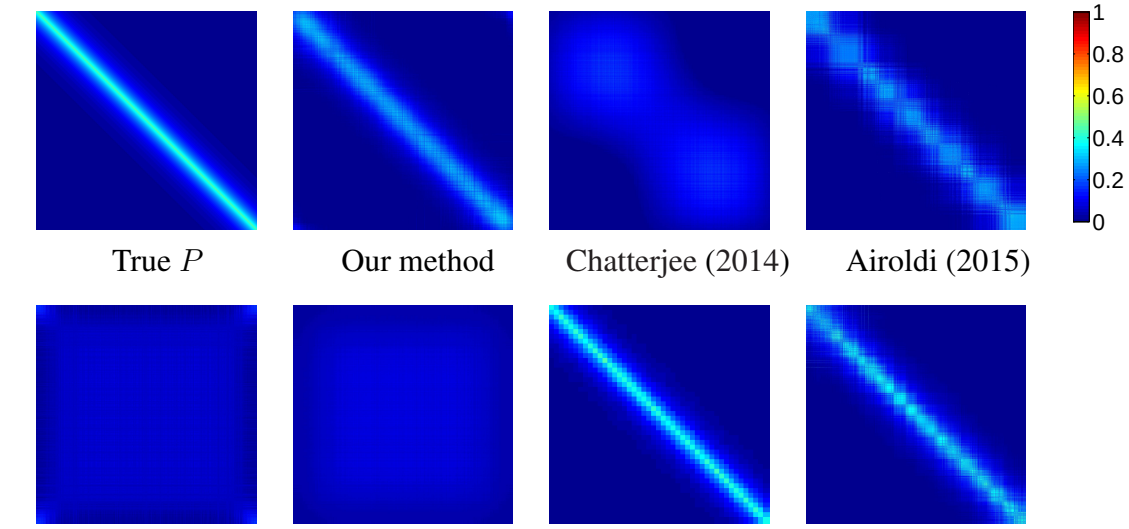


Figure 4.3: Estimated probability matrices for Graphon 3.

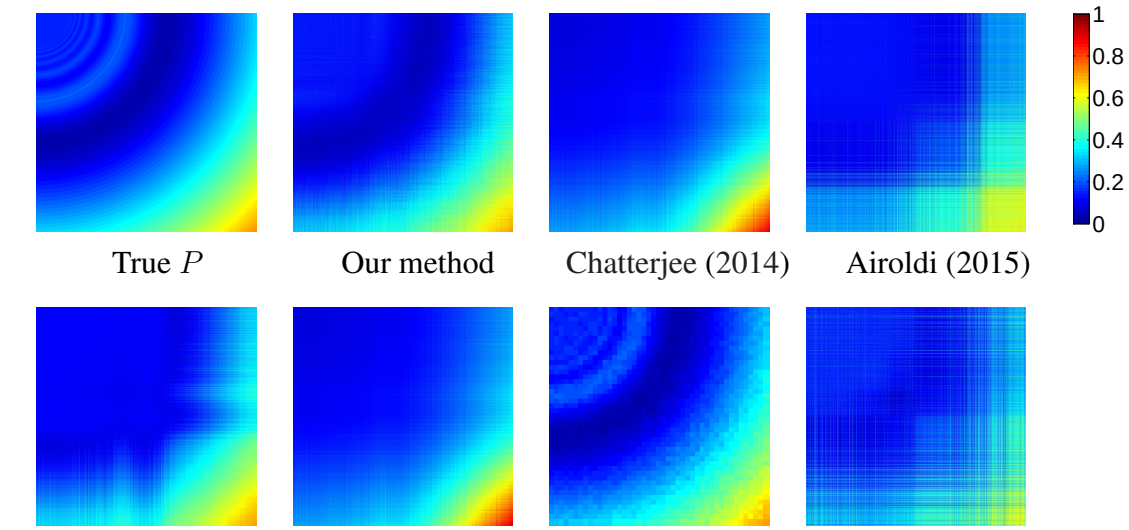


Figure 4.4: Estimated probability matrices for Graphon 4.

difficult setting for singular value thresholding and spectral clustering. The degrees are monotone for nodes with $\xi \in [0.5, 1]$ but not for $\xi \in [0, 0.5]$, so this graphon is also difficult for the sort and smooth methods, which completely miss the structure in the top left corner of the matrix. Our method successfully picks up the global structure, including non-monotone degrees, though it misses the local variations in the top left corner, as do all other methods except for the oracle approximation. This illustrates a limitation of our

method resulting from selecting neighbors based on global similarity of graphon slices, which may miss their local differences.

Table 4.2 shows the mean squared errors of all methods on the four graphons. Our method is among the top two performs with all graphons, even taking the stochastic block-model oracle method into account.

Table 4.2: Mean squared errors ($\times 10^{-3}$)
30 experiments, top two performers are bolded

	Graphon 1	Graphon 2	Graphon 3	Graphon 4
Our method	**0.229(0.002)**	**0.709(0.003)**	**1.075(0.008)**	**1.004(0.008)**
Chan & Airoldi (2014)	7.693(0.112)	116.388(0.281)	8.858(0.009)	1.992(0.007)
Yang et al (2014)	9.019(0.153)	116.407(0.406)	9.146(0.009)	3.218(0.003)
stochastic blockmodel spectral	**0.276(0.014)**	110.49(0.621)	1.238(0.018)	7.811(0.111)
stochastic blockmodel oracle	3.009(0.000)	2.617(0.001)	**0.210(0.001)**	**0.111(0.002)**
Chatterjee (2014)	1.677(0.001)	**0.353(0.003)**	6.351(0.004)	3.211(0.002)
Airoldi (2015)	0.320(0.048)	9.148(0.26)	3.124(0.779)	4.313(0.234)

Overall, the results in this section show that various previously proposed methods can perform very well when their assumptions hold (which may be monotone degrees or low rank or an underlying block model), but they fail when these assumptions are not satisfied. Our method is the only one among those compared that can perform well in a large range of scenarios, because it learns the structure from data via neighborhood selection instead of imposing a priori structural assumptions.

4.4 Application to link prediction

Evaluating the performance of probability matrix estimation methods on real networks directly is difficult, since the true probability matrix is unknown. To assess the practical utility of our method, we apply it to the link prediction problem, a practical task that relies on estimating the probability matrix. In this context, we think of the true adjacency matrix A^{true} as unobserved, with binary edges drawn independently according to the probability matrix P, also unobserved. The observed adjacency matrix is defined by $A^{\text{obs}}_{ij} = M_{ij}A^{\text{true}}_{ij}$, where unobserved independent M_{ij}'s $\sim$ Bernoulli$(1-p)$ indicate whether edges are missing and p is the unknown missing rate. A link prediction method usually produces a nonnegative score matrix $\hat{A}$, whose elements represent the estimated propensity of a node pair to form an edge.

We measure link prediction performance by the receiver operating characteristic (Receiver operating characteristic) curve defined as follows. For each $t > 0$, we define the

false positive rate r_{FP} and the true positive rate r_{TP} by

$$r_{\text{FP}}(t) = \sum_{ij} 1\left[\hat{A}_{ij} > t, A_{ij}^{\text{true}} = 0, M_{ij} = 0\right] / \sum_{ij} \left[A_{ij}^{\text{true}} = 0, M_{ij} = 0\right]$$

and

$$r_{\text{TP}}(t) = \sum_{ij} 1\left[\hat{A}_{ij} > t, A_{ij}^{\text{true}} = 1, M_{ij} = 0\right] / \sum_{ij} 1\left[\hat{A}_{ij} = 1, M_{ij} = 0\right]$$

. Then varying t we obtain the Receiver operating characteristic curve.

In this section we include three additional benchmark methods that produce score matrices rather than estimated probability matrices. One standard benchmark is to use the Jaccard index $\langle A_{i\cdot}, A_{j\cdot}\rangle / \{(\sum_k A_{ik})(\sum_k A_{jk})\}$ as the score, see for example Lichtenwalter et al. (2010). The method proposed in a technical report by Zhao et al (arXiv:1301.7047) solves an optimization problem to obtain $\hat{A}_{ij}$ which encourages similar node pairs to have similar predicted scores. The PropFlow algorithm proposed by Lichtenwalter et al. (2010) uses the probability for a random walk starting at one node to reach another node within a certain number of steps as the propensity score. We first compare all methods on simulated networks generated from the graphons in Table 4.1. We set $n = 2000$ and $p = 10\%$.

Figure 2 in the Appendix shows the Receiver operating characteristic curves for four graphons. Most differences between the methods compared in Section 4.3.1 can be understood from Figures 4.1 to 4.4. Overall, the methods based on graphon estimation outperform score-based methods. Our method outperforms all other methods on this task, producing an Receiver operating characteristic curve very close to that based on the true probability matrix P.

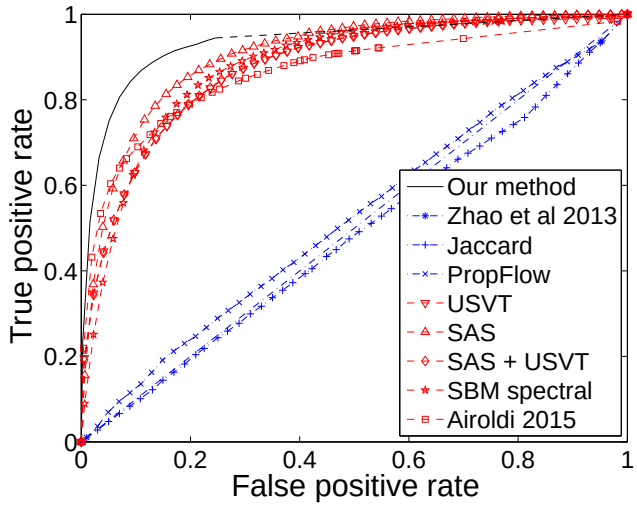


Figure 4.5: Receiver operating characteristic curve for link prediction on the political blogs network. 10% of edges are missing at random.

We also applied our method to the political blogs network (Adamic and Glance, 2005) and compared it to benchmarks. This network consists of 1222 blogs, manually labeled as 586 liberal blogs and 636 conservative blogs, and the network clearly shows two communities corresponding to these two groups. It also has quite heterogeneous node degrees , that is, some nodes are hubs. We removed 10% of edges at random and then calculated the Receiver operating characteristic curve for predicting the missing links, shown in Figure 4.5. Again, methods based on estimating the probability matrix performed much better than the scoring methods, and our method had the best overall performance. Sort and smooth methods slightly outperformed spectral clustering and Chatterjee (2014), perhaps due to the presence of hubs.

4.5 Discussion

In this chapter, we proposed a computationally feasible method to estimate the matrix of edge probabilities from a single network realization under the assumption of a piecewise Lipschitz graphon, with a competitive mean squared error rate and good empirical performance. The main advantage of our method is the adaptive neighborhood choice which allows for good performance under many different conditions; it is also computationally efficient, very easy to implement, and essentially tuning free. The main limitation of our method is in the piecewise Lipschitz condition, which may lead it to miss small-scale local structures and over-smooth occasionally. Our method does not achieve the minimax error rate, and whether this rate can be achieved by any polynomial time method is, to the best of our knowledge, an open problem. Going forward, a major challenge is to relax the unrealistic assumption of independent edges and resulting exchangeability, extending the model to better describe real world networks.

CHAPTER 5

Future work

I am interested in further exploring the network edge probability estimation problem. Currently, a gap remains between the best error rate achieved by polynomial time methods (Zhang et al., 2015b) and the statistical minimax error rate (Gao et al., 2014). The question whether this gap can be closed remains open, without even a plausible conjecture at this point. Known results for related topics conflict: for the problem of detecting a small hidden block of nodes with elevated edge probabilities in a network, a computational barrier exists that prevents the minimax rate from being achieved by any polynomial time algorithm (Ma and Wu, 2015; Cai et al., 2015), while for the community detection problem, a rate-optimal algorithm has been proposed (Gao et al., 2015). These results provide motivation but are not directly applicable to the edge probability estimation problem.

While my projects so far have focused on static networks, I am becoming interested in exploring dynamic networks. Many models have been proposed to describe the mechanisms for dynamic formation of networks (Durante et al., 2015; Sewell and Chen, 2014; Perry and Wolfe, 2013), but most of them are context-specific. One fundamental question is whether a general model, like the Aldous-Hoover representation for static networks, can be established. I am currently working on extending the Aldous-Hoover representation to dynamic settings, which is difficult due to more complicated identifiability issues. It is also of interest to understand the asymptotic properties of parameter estimation in dynamic networks. In static networks, many network parameters can be consistently estimated under appropriate modeling assumptions as the number of nodes grows to infinity (Zhao et al., 2012); while in dynamic networks, sometimes the network size does not grow, but more network snapshots can be observed over time (Sarkar et al., 2012). New tools and insights are needed to explore the interplay between the number of nodes and the number of time points in parameter estimation. Finally, a number of nonparametric tests and estimation procedures have been developed for static networks (Bickel and Chen, 2009; Lei, 2014), but hardly any dynamic analogues exist. In another ongoing project, I am studying the

problem of testing and estimating the dependence between network edges observed at different time points. Snapshots of the network at different time points are often assumed independent, but this assumption is unrealistic in practice. So far I have developed a nonparametric test for dependence between two networks which share the same set of nodes; in general the problem is more difficult.

In the future, I will continue to focus on developing useful new methods, understanding their theoretical behaviors, and applying them to realistic and important practical problems. My goal is to develop methods that are versatile and adaptive in general settings, with a focus on computational efficiency. I am also interested in developing applied and interdisciplinary collaborations.

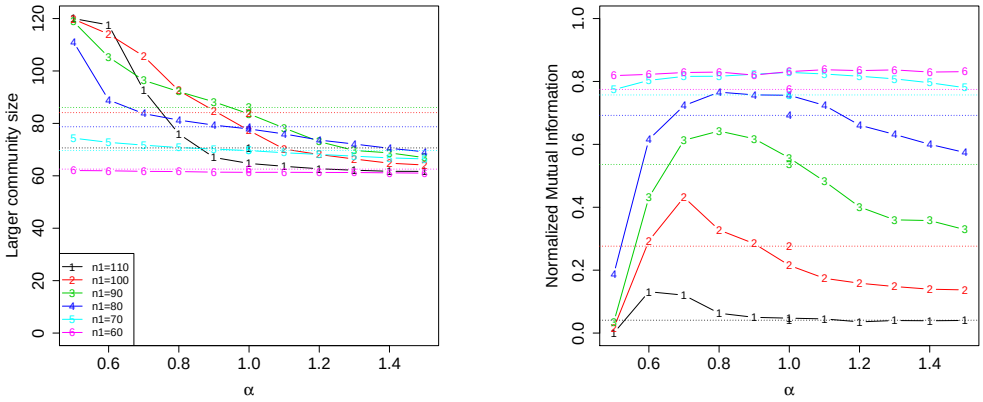
APPENDIX A

Appendix for "Community detection in networks with node features"

A.1 Choice of tuning parameters

The JCDC method involves two user-specified tuning parameters, α and w_n. In this section, we investigate the impact of these tuning parameters on community detection results via numerical experiments.

First we study the impact of α, which determines the algorithm's preference for larger or smaller communities. We study its effect on the estimated community size as well as on the accuracy of estimated community labels. We generate data from a stochastic block model with $n = 120$ nodes and $K = 2$ communities of sizes n_1 and $n_2 = n - n_1$. We set the within-community edge probabilities to 0.3 and between-community edge probabilities to 0.15, and vary n_1 from 60 to 110. Since α is not related to feature weights, we set features to a constant, resulting in unweighted networks. The results are averaged over 50 replications and shown in Figure A.1.



Size of the larger estimated community Community detection accuracy

Figure A.1: (a) The size of the larger estimated community as a function of the tuning parameter α. (b) Estimation accuracy measured by NMI as a function of the tuning parameter α. Solid lines correspond to JCDC and horizontal dotted lines correspond to spectral clustering, which does not depend on α.

We report the size of the larger estimated community in Figure A.1(a), and the accuracy of community detection as measured by normalized mutual information (NMI) in Figure A.1(b). For comparison, we also record the results from spectral clustering (horizontal lines in Figure A.1), which do not depend on α. When communities are balanced ($n_1 = n_2 = 60$), JCDC performs well for all values of α, producing balanced communities and uniformly outperforming spectral clustering in terms of NMI. In general, larger values of α in JCDC result in more balanced communities, while smaller α's tend to produce a large and a small community. In terms of community detection accuracy, Figure A.1(b) shows that the JCDC method outperforms spectral clustering over a range of values of α, and this range depends on how unbalanced the communities are. For simplicity and ease of interpretation, we set $\alpha = 1$ for all the simulations and data analysis reported in the main manuscript; however, it can be changed by the user if information about community sizes is available.

Next, we investigate the impact of w_n, which controls the influence of features. To study the trade-off between the two sources of information (network and features), we generate two different community partitions. Specifically, we consider two communities of sizes n_1 and n_2, with $n_1 + n_2 = n = 120$. We generate two label vectors c^A and c^F, with $c_i^A = 1$ for $i = 1, \ldots, n_1$ and $c_i^A = 2$ for $i = n_1 + 1, \ldots, n$, while the other label vector has $c_i^F = 1$ for $i = 1, \ldots, n_2$ and $c_i^F = 2$ for $i = n_2 + 1, \ldots, n$. Then the edges are generated from the stochastic block model based on c^A, and the node features are generated based on

c^F. We generate two node features: one feature is sampled from the distribution $N(\mu, 1)$ if $c_i^F = 1$ and $N(0, 1)$ if $c_i^F = 2$; the other feature is sampled from $N(0, 1)$ if $c_i^F = 1$ and $N(-\mu, 1)$ if $c_i^F = 2$. We fix $\mu = 3$ and set $\alpha = 1$, as discussed above. We set the within- and between-community edge probabilities to 0.3 and 0.15, respectively, same as in the previous simulation, and vary the value of w_n from 1.1 to 10. Finally, we look at the the agreement between the estimated communities $\hat{e}$ and c_A and c_F, as measured by normalized mutual information. The results are shown in Figure A.2.

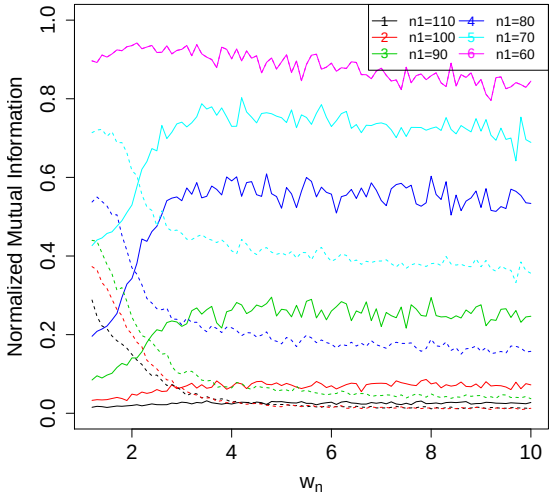


Figure A.2: MNI between the estimated community structure $\hat{e}$ and the network community structure c_A (solid lines) and the feature community structure c_F (dotted lines). Note that when $n_1 = n_2 = 60$, $c^A = c^F$, so the solid and dotted lines coincide.

As we expect, smaller values of w_n give more influence to features and thus the estimated community structure agrees better with c^F than with c^A. As w_n increases, the estimated $\hat{e}$ becomes closer to c^A. In the manuscript, we compare two values of w_n, 1.5 and 5.

A.2 Proofs

We start with summarizing notation. Let $\mathcal{E}_1, \ldots, \mathcal{E}_K$ be the estimated communities corresponding to the label vector e, and $\mathcal{C}_1, \ldots, \mathcal{C}_K$ the true communities corresponding to the label vector c. Recall we estimate e by maximizing the criterion R over e and β, where

$$R(e, \beta; w_n) = \sum_{k=1}^{K} \frac{1}{|\mathcal{E}_k|^\alpha} \sum_{i,j \in \mathcal{E}_k} A_{ij} W(\phi_{ij}, \beta_k; w_n) \ ,$$

and define

$$\hat{e} = \arg\max_{e} \left(\max_{\beta} R(e, \beta; w_n) \right) ,$$

where $\hat{e}$ and the corresponding $\hat{\beta}$ are defined up to a permutation of community labels. Recall that we assumed A and F are conditionally independent given c and defined R^0, the "population version" of R, as

$$R^0(e, \beta; w_n) = \sum_{k=1}^{K} \frac{1}{|\mathcal{E}_k|^\alpha} \sum_{i,j \in \mathcal{E}_k} \rho_n P_{c_i c_j} \mathbb{E}[W(\phi_{ij}, \beta_k; w_n)] .$$

The expectation in R^0 is taken with respect to the distribution of node features, which determine the similarities ϕ_{ij}.

Lemma 7. *Under conditions 1 and 2, if $w_n \rho_n \to \infty$ and $0 < \alpha \leq 2$, we have*

$$\max_{e,\beta} \frac{|R(e, \beta; w_n) - R^0(e, \beta; w_n)|}{w_n \rho_n n^{2-\alpha}} = O_p \left(\frac{1}{\sqrt{w_n \rho_n}} \right) .$$

Proof of Lemma 11. We first bound the difference between R and R^0 for fixed e and β. By Hoeffding's inequality and the fact that $2[n/2] \geq n - 1$, where $[x]$ is the integer part of x, we have

$$\mathbb{P} \left\{ \left| \frac{1}{|\mathcal{E}_k|^2} \sum_{i,j \in \mathcal{E}_k} \left(A_{ij} W(\phi_{ij}, \beta_k; w_n) - \rho_n P_{c_i c_j} \mathbb{E}[W(\phi_{ij}, \beta_k; w_n)] \right) \right| > t \right\} \leq 2 \exp \left(-(|\mathcal{E}_k| - 1) t^2 \right) .$$

Taking $t = w_n \rho_n n^{2-\alpha} |\mathcal{E}_k|^{\alpha - 2} \delta$ and applying the union bound, we have

$$\begin{aligned}
&\mathbb{P} \left(\frac{|R(e, \beta; w_n) - R^0(e, \beta; w_n)|}{w_n \rho_n n^{2-\alpha}} > K\delta \right) \\
\leq &\sum_{k=1}^{K} \mathbb{P} \left\{ \frac{\left| \sum_{i,j \in \mathcal{E}_k} \left(A_{ij} W(\phi_{ij}, \beta_k; w_n) - \rho_n P_{c_i c_j} \mathbb{E}\left[W(\phi_{ij}, \beta_k; w_n) \right] \right) \right|}{w_n \rho_n |\mathcal{E}_k|^\alpha n^{2-\alpha}} \geq \delta \right\} \\
\leq &\sum_{k=1}^{K} 2 \exp \left\{ -(|\mathcal{E}_k| - 1) w_n^2 \rho_n^2 n^{4-2\alpha} |\mathcal{E}_k|^{2\alpha - 4} \delta^2 \right\} \leq 2K \exp \left\{ -(\pi_0 n - 1) w_n^2 \rho_n^2 \delta^2 \right\} .
\end{aligned}$$

Next, we take the uniform bound over β. Consider the set

$$B_\epsilon = \left\{ \left(\frac{s_1 \epsilon}{\sqrt{p}}, \ldots, \frac{s_p \epsilon}{\sqrt{p}} \right), s_1, \ldots, s_p \in \left\{ 0, \pm 1, \ldots, \pm \left[\frac{M_\beta \sqrt{p}}{\epsilon} \right], \pm \frac{M_\beta \sqrt{p}}{\epsilon} \right\} \right\} .$$

It is straightforward to verify that B_ϵ is an ϵ-net on $[-M_\beta, M_\beta]^p$, the space of β_k's. For each

β_k, let $\beta(\beta_k, B_\epsilon)$ be the best approximation to β_k in B_ϵ. Then

$$\begin{aligned}
&\max_{\beta_k} |W(\phi_{ij}, \beta_k; w_n) - W(\phi_{ij}, \beta(\beta_k, B_\epsilon); w_n)| \\
\leq &\max_{\beta_k} \left| \frac{\partial W}{\partial \beta_k}(\phi_{ij}, \beta_k; w_n) \right| |\beta_k - \beta(\beta_k, B_\epsilon)| \\
\leq &2M_\phi M_\beta \exp(M_\phi M_\beta)\epsilon \leq 2M_\phi M_\beta w_n \epsilon
\end{aligned}$$

Therefore, choosing $\epsilon = \frac{\rho_n \delta}{4M_\phi M_\beta}$, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{\beta} \frac{|R(e,\beta;w_n) - R^0(e,\beta;w_n)|}{w_n \rho_n n^{2-\alpha}} > K\delta \right) \\
\leq &\sum_{k=1}^K \mathbb{P}\left\{ \max_{\beta_k} \frac{\left|\sum_{i,j\in\mathcal{E}_k} (A_{ij} - \rho_n P_{c_i c_j}) W(\phi_{ij}, \beta_k; w_n)\right|}{w_n \rho_n |\mathcal{E}_k|^\alpha n^{2-\alpha}} > \delta \right\} \\
\leq &\sum_{k=1}^K \mathbb{P}\left\{ \max_{\beta_k} \frac{\sum_{i,j\in\mathcal{E}_k} |A_{ij} - \rho_n P_{c_i c_j}| |W(\phi_{ij}, \beta_k; w_n) - W(\phi_{ij}, \beta(\beta_k, B_\epsilon); w_n)|}{w_n \rho_n |\mathcal{E}_k|^\alpha n^{2-\alpha}} > \frac{\delta}{2} \right\} \\
&+ \sum_{k=1}^K \mathbb{P}\left\{ \max_{\beta_0 \in B_\epsilon} \frac{\left|\sum_{i,j\in\mathcal{E}_k} \left(A_{ij} - \rho_n P_{c_i c_j}\right) W(\phi_{ij}, \beta_0; w_n)\right|}{w_n \rho_n |\mathcal{E}_k|^\alpha n^{2-\alpha}} > \frac{\delta}{2} \right\} \\
\leq &K\mathbb{P}\left(\frac{|\mathcal{E}_k|^{2-\alpha} \cdot 2M_\phi M_\beta \epsilon}{\rho_n n^{2-\alpha}} \geq \frac{\delta}{2} \right) + 2K|B_\epsilon| \exp\left\{-(\pi_0 n - 1) w_n^2 \rho_n^2 \delta^2/4\right\} \\
\leq &0 + 2K\left(\frac{4M_\phi M_\beta^2 \sqrt{p}}{\rho_n \delta} + 3 \right)^p \exp\left\{-(\pi_0 n - 1) w_n^2 \rho_n^2 \delta^2/4\right\} \ ,
\end{aligned}$$

where the first term becomes 0 because of the choice of ϵ and $|\mathcal{E}_k| < n$. Finally, taking a union bound over all possible community assignments, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{e,\beta} \frac{|R(e,\beta;w_n) - R^0(e,\beta;w_n)|}{w_n \rho_n n^{2-\alpha}} > K\delta \right) \\
\leq &2K^{n+1}\left(\frac{4M_\phi M_\beta^2 \sqrt{p}}{\rho_n \delta} + 3 \right)^p \exp\left\{-(\pi_0 n - 1) w_n^2 \rho_n^2 \delta^2/4\right\} \\
\leq &2K \exp\left[-\pi_0 n w_n^2 \rho_n^2 \delta^2/8 + n \log K + p \log\{C_1/(\rho_n \delta)\}\right]
\end{aligned}$$

where $C_1 := 4M_\phi M_\beta^2 \sqrt{p}$. Taking $\delta = 1/\sqrt{w_n \rho_n}$ completes the proof of Lemma 11. □

We now proceed to investigate the "population version" of our criterion, R^0. Define $U \in \mathbb{R}^{K\times K}$ by $U_{kl} = \sum_{i=1}^n 1[e_i = k, c_i = l]/n$, and let D be a diagonal $K \times K$ matrix with $\pi_1, \ldots, \pi_K$ on the diagonal, where $\pi_k = \sum_{i=1}^n 1[c_i = k]/n$ is the fraction of nodes in community $\mathcal{C}_k$. Roughly speaking, U is the confusion matrix between e and c, and $U = DO$

for a permutation matrix O means the estimation is perfect. Define

$$g(U) = \sum_{k=1}^{K} \frac{\sum_{l=1}^{K}\sum_{l'=1}^{K} U_{kl}U_{kl'}P_{ll'}}{\left(\sum_{a=1}^{K} U_{ka}\right)^{\alpha}} .$$

Each estimated community assignment e induces a unique $U = U(e)$. It is not difficult to verify that

$$g\left(U(e)\right) = \sum_{k=1}^{K} \frac{\sum_{i,j\in\mathcal{E}_k} P_{c_i c_j}}{|\mathcal{E}_k|^{\alpha} n^{2-\alpha}} .$$

Lemma 8. *Under conditions 1 and 2, there exists a constant C_2 such that*

$$\max_{e,\beta} \left| \frac{R^0(e,\beta;w_n)}{w_n\rho_n n^{2-\alpha}} - g\left(U(e)\right) \right| \le \frac{C_2}{w_n} .$$

Proof of Lemma 12. By definition, we have

$$\max_{e,\beta} \left| \frac{R^0(e,\beta;w_n)}{w_n\rho_n n^{2-\alpha}} - g\left(U(e)\right) \right| = \max_{e} \sum_{k=1}^{K} \max_{\beta_k} \sum_{i,j\in\mathcal{E}_k} P_{c_i c_j} \frac{\mathbb{E}[\exp(-\langle\phi_{ij},\beta_k\rangle)]}{|\mathcal{E}_k|^{\alpha} w_n n^{2-\alpha}}$$
$$\le \max_{e} \sum_{k=1}^{K} \sum_{i,j\in\mathcal{E}_k} \frac{\exp(M_\phi M_\beta)}{|\mathcal{E}_k|^{\alpha} w_n n^{2-\alpha}} \max_{kl} P_{kl} \le \frac{K\exp(M_\phi M_\beta)}{w_n \pi_0^{2-\alpha}} \max_{kl} P_{kl} = \frac{C_2}{w_n} ,$$

where $C_2 := K\pi_0^{\alpha-2}\exp(M_\phi M_\beta)\max_{kl} P_{kl}$, and the two inequalities follow from conditions 1 and 2, respectively. □

Lemma 9. *Under condition 3, if $\alpha \in [\max_{1\le k<l\le K} 2(K-1)P_{kl}/\min(P_{kk},P_{ll}), 1]$, then for all U satisfying $\sum_{k=1}^{K} U_{kl} = \pi_l$ for $1 \le k \le K$, $g(U)$ is uniquely maximized at $U = DO$ for $O \in \mathcal{O}_K$, where $\mathcal{O}_K$ denotes the set of $K \times K$ permutation matrices.*

Proof of Lemma 9. We have

$$\begin{aligned}
&g(D) - g(U) \\
=&\sum_{l=1}^{K} \left(\sum_{k=1}^{K} U_{kl}\right)^{2-\alpha} P_{ll} - \sum_{k=1}^{K} \frac{\sum_{l=1}^{K} U_{kl}^2 P_{ll} + \sum_{l=1}^{K}\sum_{l'\neq l} U_{kl}U_{kl'}P_{ll'}}{\left(\sum_{a=1}^{K} U_{ka}\right)^{\alpha}} \\
=&\sum_{l=1}^{K} \left\{ \left(\sum_{k=1}^{K} U_{kl}\right)^{2-\alpha} - \sum_{k=1}^{K} \frac{U_{kl}^2}{\left(\sum_{a=1}^{K} U_{ka}\right)^{\alpha}} \right\} P_{ll} - \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{l'\neq l} \left\{ \frac{U_{kl}U_{kl'}}{\left(\sum_{a=1}^{K} U_{ka}\right)^{\alpha}} \right\} P_{ll'}
\end{aligned} \tag{A.2.1}$$

For $0 < \alpha \leq 1$, since $U_{kl} \geq 0$ for all k and l, we have $\left(\sum_{k=1}^K U_{kl}\right)^{2-\alpha} \geq \sum_{k=1}^K U_{kl}^{2-\alpha}$. By mid-value theorem, there exists $\xi_{kl} \in \left(0, \sum_{a\neq l} U_{ka}\right)$, such that

$$\left(\sum_{a=1}^K U_{ka}\right)^{\alpha} - U_{kl}^{\alpha} = \alpha \left(\sum_{a\neq l} U_{ka}\right) / \left(U_{kl} + \xi^{kl}\right)^{1-\alpha} \geq \alpha \left(\sum_{a\neq l} U_{ka}\right) / \left(\sum_{a=1}^K U_{ka}\right)^{1-\alpha} . \tag{A.2.2}$$

Finally, we will need the following inequality: for $0 < \alpha \leq 2$ and $x, y \geq 0$ satisfying $x + y \leq u$,

$$x^{2-\alpha}(u-x) + y^{2-\alpha}(u-y) \geq xyu^{1-\alpha} . \tag{A.2.3}$$

For $x = y = 0$, equality holds. To verify (A.2.3) when $0 < x + y \leq u$, dividing by $u^{3-\alpha}$ we have

$$\frac{x^{2-\alpha}(u-x) + y^{2-\alpha}(u-y) - xyu^{1-\alpha}}{u^{3-\alpha}} = \left(\frac{x}{u}\right)^{2-\alpha}\left(1-\frac{x}{u}\right) + \left(\frac{y}{u}\right)^{2-\alpha}\left(1-\frac{y}{u}\right) - \frac{xy}{u^2}$$
$$\geq \left(\frac{x}{u}\right)^{2}\left(1-\frac{x}{u}\right) + \left(\frac{y}{u}\right)^{2}\left(1-\frac{y}{u}\right) - \frac{xy}{u^2} = \left\{\left(\frac{x}{u}\right)^{2} + \left(\frac{y}{u}\right)^{2} - \frac{xy}{u^2}\right\}\left(1-\frac{x+y}{u}\right) \geq 0 .$$

The first inequality above implies that a necessary condition for equality to hold in (A.2.3) is $xy = 0$.

We now lower bound the first term on the right hand side of (A.2.1).

$$\begin{aligned}
&\sum_{l=1}^K \left\{ \left(\sum_{k=1}^K U_{kl}\right)^{2-\alpha} - \sum_{k=1}^K \frac{U_{kl}^2}{\left(\sum_{a=1}^K U_{ka}\right)^{\alpha}} \right\} P_{ll} \\
\geq &\sum_{l=1}^K \sum_{k=1}^K \frac{U_{kl}^{2-\alpha}\left\{\left(\sum_{a=1}^K U_{ka}\right)^{\alpha} - U_{kl}^{\alpha}\right\}}{\left(\sum_{a=1}^K U_{ka}\right)^{\alpha}} P_{ll} \\
\geq &\sum_{l=1}^K \sum_{k=1}^K \frac{U_{kl}^{2-\alpha}\left(\sum_{a\neq l} U_{ka}\right)}{\sum_{a=1}^K U_{ka}} \alpha P_{ll} \geq \sum_{l=1}^K \sum_{k=1}^K \frac{U_{kl}^{2-\alpha}\left(\sum_{a\neq l} U_{ka}\right)}{\sum_{a=1}^K U_{ka}} \sum_{l'\neq l} 2P_{ll'} \\
= &\sum_{k=1}^K \left\{ \sum_{l=1}^K \sum_{l'\neq l} \frac{U_{kl}^{2-\alpha}\left(\sum_{a\neq l} U_{ka}\right) P_{ll'}}{\sum_{a=1}^K U_{ka}} + \sum_{l'=1}^K \sum_{l\neq l'} \frac{U_{kl'}^{2-\alpha}\left(\sum_{a\neq l'} U_{ka}\right) P_{ll'}}{\sum_{a=1}^K U_{ka}} \right\} \\
= &\sum_{k=1}^K \sum_{l=1}^K \sum_{l'\neq l} \frac{U_{kl}^{2-\alpha}\left(\sum_{a\neq l} U_{ka}\right) + U_{kl'}^{2-\alpha}\left(\sum_{a\neq l'} U_{ka}\right)}{\sum_{a=1}^K U_{ka}} P_{ll'} \\
\geq &\sum_{k=1}^K \sum_{l=1}^K \sum_{l'\neq l} \frac{U_{kl}U_{kl'}}{\left(\sum_{a=1}^K U_{ka}\right)^{\alpha}} P_{ll'} ,
\end{aligned} \tag{A.2.4}$$

where the last equality is obtained by applying (A.2.3) with $x = U_{kl}$, $y = U_{kl'}$ and $u = \sum_{a=1}^{K} U_{ka}$. Plugging (A.2.4) into (A.2.1), we have

$$g(D) - g(U) \geq 0 \, .$$

It remains to show that equality holds only if $U = DO$ for some $O \in \mathcal{O}_K$. Note that the last inequality in (A.2.4) is obtained from (A.2.3), where equality holds only when $xy = 0$. The corresponding condition for equality to hold in (A.2.4) is thus $U_{kl}U_{kl'} = 0$ for all k, l and l'. Therefore, for each k, there is only one l such that $U_{kl} \neq 0$, i.e., $U = DO$ for some $O \in \mathcal{O}_K$. □

Proof of Theorem 1. By Lemma 11 and Lemma 12, we have

$$\max_{e,\beta} \left| \left\{ \frac{R(e,\beta;w_n)}{w_n\rho_n n^{2-\alpha}} - g\left(U(e)\right) \right\} \right| = O_p\left(\frac{1}{\sqrt{w_n\rho_n}} \right) \, . \tag{A.2.5}$$

It is straightforward to verify that, for any e, $2d(e,c) = \min_{O\in\mathcal{O}_K} \|U(e) - DO\|_1$, where $\|Q\|_1 = \sum_{k=1}^{K}\sum_{l=1}^{K} |Q_{kl}|$. Take a sequence of decreasing positive numbers $x_n \to 0$ and define

$$y_n = \max_{U: g(D)-g(U)\leq x_n} \min_{O\in\mathcal{O}_K} \|U - DO\|_1 \tag{A.2.6}$$

We now show, by contradiction, that $x_n \to 0$ implies $y_n \to 0$. First, note that y_n is non-increasing. Now if $y_0 = \lim_{n\to\infty} y_n > 0$, by compactness of the set $\mathcal{U}_{y_0} = \{U : \min_{O\in\mathcal{O}_K} \|U - DO\|_1 \geq y_0\}$ and continuity of the function g, the supremum of $g(U)$ over $U \in \mathcal{U}_{y_0}$, which equals $g(D)$, is attained in $\mathcal{U}_{y_0}$. This contradicts Lemma 9.

Now let $x_n = 1/\sqrt[4]{w_n\rho_n}$. By assumption of Theorem 1, $x_n \to 0$, which yields $y_n \to 0$. Also $x_n/\left(1/\sqrt{w_n\rho_n}\right) = \sqrt[4]{w_n\rho_n} \to \infty$, so by (A.2.5) we have

$$\mathbb{P}\left[\left\{ \left| \frac{R(\hat{e},\hat{\beta};w_n)}{w_n\rho_n n^{2-\alpha}} - g\left(U(\hat{e})\right) \right| > \frac{x_n}{2} \right\} \bigcup \left\{ \left| \frac{R(c,\beta;w_n)}{w_n\rho_n n^{2-\alpha}} - g\left(D\right) \right| > \frac{x_n}{2} \right\} \right] \to 0 \, . \tag{A.2.7}$$

Now, the event

$$\left| \frac{R(\hat{e},\hat{\beta};w_n)}{w_n\rho_n n^{2-\alpha}} - g\left(U(\hat{e})\right) \right| \leq \frac{x_n}{2} \text{ and } \left| \frac{R(c,\beta;w_n)}{w_n\rho_n n^{2-\alpha}} - g\left(D\right) \right| \leq \frac{x_n}{2}$$

implies that $g(D) - g(U(\hat{e})) \leq \frac{R(c,\beta;w_n)}{w_n \rho_n n^{2-\alpha}} - \frac{R(\hat{e},\hat{\beta};w_n)}{w_n \rho_n n^{2-\alpha}} + x_n \leq x_n$. So we have

$$\mathbb{P}\left(g(D) - g(U(\hat{e})) \leq x_n\right) \to 1 \tag{A.2.8}$$

and

$$2d(\hat{e}, c) = \min_{O \in \mathcal{O}_K} \|U(\hat{e}) - DO\|_1 \leq \max_{U: g(D) - g(U) \leq x_n} \min_{O \in \mathcal{O}_K} \|U - DO\|_1 = y_n \to 0.$$

□

APPENDIX B

Appendix for "Detecting overlapping communities in networks using spectral methods"

B.1 Appendix

B.1.1 Proof of identifiability

Proof of Theorem 2. We start with stating a Lemma of Tang et al. (2013):

Lemma 10 (Lemma A.1 of Tang et al. (2013))**.** *Let $\boldsymbol{Y}_1, \boldsymbol{Y}_2 \in \mathbb{R}^{n\times d}$, $d < n$, be full rank matrices and $\boldsymbol{G}_1 = \boldsymbol{Y}_1\boldsymbol{Y}_1^T$, $\boldsymbol{G}_2 = \boldsymbol{Y}_2\boldsymbol{Y}_2^T$. Then there exists an orthonormal $\boldsymbol{O}$ such that*

$$\|\boldsymbol{Y}_1\boldsymbol{O} - \boldsymbol{Y}_2\|_F \leq \frac{\sqrt{d}\|\boldsymbol{G}_1 - \boldsymbol{G}_2\|(\sqrt{\|\boldsymbol{G}_1\|} + \sqrt{\|\boldsymbol{G}_2\|})}{\lambda_{\min}(\boldsymbol{G}_2)} \tag{B.1.1}$$

where $\lambda_{\min}(\cdot)$ is the smallest positive eigenvalue.

Lemma 10 immediately implies

Claim 1. *For two full rank matrices $\boldsymbol{H}_1$, $\boldsymbol{H}_2 \in \mathbb{R}^{n\times K}$ satisfying $\boldsymbol{H}_1\boldsymbol{H}_1^T = \boldsymbol{H}_2\boldsymbol{H}_2^T$, there exists an orthonormal matrix $\boldsymbol{O}_H$ such that $\boldsymbol{H}_1\boldsymbol{O}_H = \boldsymbol{H}_2$.*

Suppose parameters $(\alpha_{n,1}, \boldsymbol{\Theta}_1, \boldsymbol{Z}_1, \boldsymbol{B}_1)$ and $(\alpha_{n,2}, \boldsymbol{\Theta}_2, \boldsymbol{Z}_2, \boldsymbol{B}_2)$ generate the same $\boldsymbol{W}$. Then by Lemma 1, there exists an orthonormal matrix $\boldsymbol{O}_{12}$ such that

$$\alpha_{n,1}\boldsymbol{\Theta}_1\boldsymbol{Z}_1\boldsymbol{B}_1^{1/2}\boldsymbol{O}_{12} = \alpha_{n,2}\boldsymbol{\Theta}_2\boldsymbol{Z}_2\boldsymbol{B}_2^{1/2} \tag{B.1.2}$$

We then show that the indices for "pure" rows in $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ match up. More precisely, for $1 \leq k \leq K$, let $\mathcal{I}_k := \{i : row_i(\boldsymbol{Z}_1) = \boldsymbol{e}_k\}$. We show that $row_j(\boldsymbol{Z}_2)$, $j \in \boldsymbol{\mathcal{I}}_k$ are also pure nodes, i.e., there exists k' such that $\{j : row_j(\boldsymbol{Z}_2) = \boldsymbol{e}_{k'}\} = \boldsymbol{\mathcal{I}}_k$. It suffices to show that

there exists $i \in \boldsymbol{\mathcal{I}}_k$ such that $row_i(\boldsymbol{Z}_2)$ is pure, then the claim follows from the fact that all rows in $\boldsymbol{Z}_2$ with indices in I_k equal each other, since their counterparts in $\boldsymbol{Z}_1$ are equal. We prove this by contradiction: if $\{row_i(\boldsymbol{Z}_2), i \in \boldsymbol{\mathcal{I}}_k\}$ are not pure nodes, then for any $i \in \boldsymbol{\mathcal{I}}_k$, there exists $\{i_1, \ldots, i_K\} \subset \{1, \ldots, n\} - \boldsymbol{\mathcal{I}}_k$ and $\omega_1, \ldots, \omega_K \geq 0$ such that

$$row_i(\boldsymbol{Z}_2) = \sum_{k=1}^{K} \omega_k row_{i_k}(\boldsymbol{Z}_2) \tag{B.1.3}$$

By (B.1.2), this yields

$$row_i(\boldsymbol{Z}_1) = \sum_{k=1}^{K} \omega_k \frac{\alpha_{n,1}(\boldsymbol{\Theta}_1)_{i_k i_k}}{\alpha_{n,2}(\boldsymbol{\Theta}_2)_{i_k i_k}} row_{i_k}(\boldsymbol{Z}_1) \tag{B.1.4}$$

i.e. the ith row of $\boldsymbol{Z}_1$ can be expressed as a non-negative linear combination of at most K rows outside $\boldsymbol{\mathcal{I}}_k$, and thus $row_i(\boldsymbol{Z}_1)$ is not pure. Essentially we have shown the identifiability for all pure nodes. To show identifiability for the rest, take one pure node from each community as representative, i.e., let $\tilde{\boldsymbol{\mathcal{I}}} := \{j_1, \ldots, j_K\}$, where $j_k \in \boldsymbol{\mathcal{I}}_k$, $1 \leq k \leq K$. Let $\boldsymbol{Z}_1^K$ be the submatrix induced by concatenating rows of $\boldsymbol{Z}_1$ with indices in $\tilde{\boldsymbol{\mathcal{I}}}$, similarly define $\boldsymbol{Z}_2^K$, and let $\tilde{\boldsymbol{\Theta}}_1$ and $\tilde{\boldsymbol{\Theta}}_2$ be the corresponding submatrices of $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$. Note $\boldsymbol{Z}_1^K$ and $\boldsymbol{Z}_2^K$ are both order K permutations, which is an ambiguity allowed by our definition of identifiability, so we take $\boldsymbol{Z}_1^K = \boldsymbol{Z}_2^K = I$. By (B.1.2),

$$\alpha_{n,1}\tilde{\boldsymbol{\Theta}}_1 \boldsymbol{B}_1^{1/2} \boldsymbol{O}_{12} = \alpha_{n,2}\tilde{\boldsymbol{\Theta}}_2 \boldsymbol{B}_2^{1/2} \,. \tag{B.1.5}$$

By condition I1, both $\boldsymbol{B}_1^{1/2}\boldsymbol{O}_{12}$ and $\boldsymbol{B}_2^{1/2}$ have rows of norm 1, so $\alpha_{n,1} \cdot (\tilde{\boldsymbol{\Theta}}_1)_{kk} = \|row_k(\alpha_{n,1}\boldsymbol{\Theta}_1^K \boldsymbol{B}_1^{1/2}\boldsymbol{O}_{12})\|_2 = \|row_k(\alpha_{n,2}\boldsymbol{\Theta}_2^K \boldsymbol{B}_2^{1/2})\|_2 = \alpha_{n,1} \cdot (\tilde{\boldsymbol{\Theta}}_2)_{kk}$ and therefore $\alpha_{n,1}\tilde{\boldsymbol{\Theta}}_1 = \alpha_{n,2}\tilde{\boldsymbol{\Theta}}_2$. Then from (B.1.5) we have

$$\boldsymbol{B}_1^{1/2}\boldsymbol{O}_{12} = \boldsymbol{B}_2^{1/2} \tag{B.1.6}$$

Thus $\boldsymbol{B}_1 = \boldsymbol{B}_1^{1/2}\boldsymbol{O}_{12}(\boldsymbol{B}_1^{1/2}\boldsymbol{O}_{12})^T = \boldsymbol{B}_2$, and (B.1.2) implies $\alpha_{n,1}\boldsymbol{\Theta}_1 = \alpha_{n,2}\boldsymbol{\Theta}_2$ since all rows of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are normalized. This in turn implies $\alpha_{n,1} = \alpha_{n,2}$ by condition I3 and thus $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2$. Finally, plugging all of this back into (B.1.2) we have $\boldsymbol{Z}_1 = \boldsymbol{Z}_2$. □

B.1.2 Proof of consistency

Proof outline: The proof of consistency of $\hat{\boldsymbol{Z}}$ follows the steps of the algorithm: we first bound the difference between $\hat{\boldsymbol{X}}^*_{\boldsymbol{\tau_n}}$ and the row-normalized version of the true node positions $\boldsymbol{X}^*$ with high probability (Lemma 11); then bound the difference between $\hat{\boldsymbol{S}}$ and

the true community centers $\boldsymbol{S} = \boldsymbol{B}^{1/2}$ (Lemma 12) with high probability; these combine to give a bound on the difference between $\hat{\boldsymbol{Z}}$ and $\boldsymbol{Z}$ (Theorem 3).

Lemma 11. *Assume conditions A1, A2 and A3 hold. When* $\frac{\log n}{n\alpha_n} \to 0$ *and* $K = O(\log n)$*, there exists a global constant* C_1*, such that with the choice* $\tau_n = \frac{\alpha_n^{0.2} K^{1.5}}{n^{0.3}}$*, for large enough* n*, we have*

$$\mathbb{P}\left(\frac{\|\hat{\boldsymbol{X}}^*_{\tau_n} \boldsymbol{O}_{\hat{X}} - \boldsymbol{X}^*\|_F}{\sqrt{n}} \le \frac{C_1 K^{\frac{4}{5}}}{(n\alpha_n)^{\frac{1}{5}}} \right) \ge 1 - P_1(n, \alpha_n, K) \tag{B.1.7}$$

where $P_1(n, \alpha_n, K) \to 0$ *as* $n \to \infty$*, and* $\boldsymbol{O}_{\hat{X}}$ *is an orthonormal matrix depending on* $\hat{X}$*.*

Proof of Lemma 11. Define the population version of $\hat{\boldsymbol{X}}^*_{\boldsymbol{\tau_n}}$ as $\boldsymbol{X}^*_{\tau_n} \in \mathbb{R}^{n \times K}$, where $row_i(\boldsymbol{X}^*_{\tau_n}) := \frac{\boldsymbol{X}_{i\cdot}}{\|\boldsymbol{X}_{i\cdot}\|_2 + \tau}$. We first bound $\|\hat{\boldsymbol{X}}^*_{\boldsymbol{\tau_n}} \boldsymbol{O}_{\hat{X}} - \boldsymbol{X}^*_{\tau_n}\|_F$ for a certain orthonormal matrix $\boldsymbol{O}_{\hat{X}}$ and then the bias term $\|\boldsymbol{X}^*_{\tau_n} - \boldsymbol{X}^*\|_F$. Then the triangular inequality gives (B.1.7).

We now bound $\|\hat{\boldsymbol{X}}^*_{\boldsymbol{\tau_n}} \boldsymbol{O}_{\hat{X}} - \boldsymbol{X}^*_{\tau_n}\|_F$. For any orthonormal matrix $\boldsymbol{O}$,

$$\begin{aligned}
&\|row_i(\hat{\boldsymbol{X}}^*_{\tau_n} \boldsymbol{O} - \boldsymbol{X}^*_{\tau_n})\|_2 = \|row_i(\hat{\boldsymbol{X}}^*_{\tau_n})\boldsymbol{O} - row_i(\boldsymbol{X}^*_{\tau_n})\|_2 \\
&= \Big\| \frac{\hat{\boldsymbol{X}}_{i\cdot}\boldsymbol{O}}{\|\hat{\boldsymbol{X}}_{i\cdot}\|_2 + \tau_n} - \frac{\boldsymbol{X}_{i\cdot}}{\|\boldsymbol{X}_{i\cdot}\|_2 + \tau_n} \Big\|_2 = \Big\| \frac{\hat{\boldsymbol{X}}_{i\cdot}\boldsymbol{O}}{\|\hat{\boldsymbol{X}}_{i\cdot}\boldsymbol{O}\|_2 + \tau_n} - \frac{\boldsymbol{X}_{i\cdot}}{\|\boldsymbol{X}_{i\cdot}\|_2 + \tau_n} \Big\|_2 \\
=& \frac{\|\hat{\boldsymbol{X}}_{i\cdot}\boldsymbol{O}(\|\boldsymbol{X}_{i\cdot}\|_2 - \|\hat{\boldsymbol{X}}_{i\cdot}\boldsymbol{O}\|_2) + \|\hat{\boldsymbol{X}}_{i\cdot}\boldsymbol{O}\|_2(\hat{\boldsymbol{X}}_{i\cdot}\boldsymbol{O} - \boldsymbol{X}_{i\cdot}) + \tau_n(\hat{\boldsymbol{X}}_{i\cdot}\boldsymbol{O} - \boldsymbol{X}_{i\cdot})\|_2}{(\|\hat{\boldsymbol{X}}_{i\cdot}\boldsymbol{O}\|_2 + \tau_n)(\|\boldsymbol{X}_{i\cdot}\|_2 + \tau_n)} \\
\le& \frac{(2\|\hat{\boldsymbol{X}}_{i\cdot}\boldsymbol{O}\|_2 + \tau_n)\|\hat{\boldsymbol{X}}_{i\cdot}\boldsymbol{O} - \boldsymbol{X}_{i\cdot}\|_2}{(\|\hat{\boldsymbol{X}}_{i\cdot}\boldsymbol{O}\|_2 + \tau_n)(\|\boldsymbol{X}_{i\cdot}\|_2 + \tau_n)} \le \frac{2\|\hat{\boldsymbol{X}}_{i\cdot}\boldsymbol{O} - \boldsymbol{X}_{i\cdot}\|_2}{\|\boldsymbol{X}_{i\cdot}\|_2 + \tau_n} \le \frac{2\|\hat{\boldsymbol{X}}_{i\cdot}\boldsymbol{O} - \boldsymbol{X}_{i\cdot}\|_2}{\tau_n}.
\end{aligned}$$

Then

$$\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{\boldsymbol{X}}} - \boldsymbol{X}^*_{\tau_n}\|_F \le \sqrt{\sum_{i=1}^n \left(\frac{2}{\tau_n}\right)^2 \|\hat{\boldsymbol{X}}_{i\cdot}\boldsymbol{O}_{\hat{\boldsymbol{X}}} - \boldsymbol{X}_{i\cdot}\|_2^2} = \frac{2\|\hat{\boldsymbol{X}}\boldsymbol{O}_{\hat{\boldsymbol{X}}} - \boldsymbol{X}\|_F}{\tau_n}.$$

By Lemma 10, there exists an orthonormal matrix $\boldsymbol{O}_{\hat{\boldsymbol{X}}}$, such that

$$\begin{aligned}\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{\boldsymbol{X}}} - \boldsymbol{X}^*_{\tau_n}\|_F &\leq \frac{2\sqrt{K}\|\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^T - \boldsymbol{X}\boldsymbol{X}^T\|\left(\sqrt{\|\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^T\|} + \sqrt{\|\boldsymbol{X}\boldsymbol{X}^T\|}\right)}{\tau_n\lambda_{\min}(\boldsymbol{X}\boldsymbol{X}^T)} \\ &\leq \frac{2\sqrt{K}\|\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^T - \boldsymbol{X}\boldsymbol{X}^T\|\left(\sqrt{\|\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^T - \boldsymbol{X}\boldsymbol{X}^T\|} + 2\sqrt{\|\boldsymbol{X}\boldsymbol{X}^T\|}\right)}{\tau_n\lambda_{\min}(\boldsymbol{X}\boldsymbol{X}^T)} \\ &= \frac{2\sqrt{K}\|\boldsymbol{A} - \boldsymbol{W}\|\left(\sqrt{\|\boldsymbol{A} - \boldsymbol{W}\|} + 2\sqrt{\|\boldsymbol{W}\|}\right)}{\tau_n\lambda_{\min}(\boldsymbol{W})} \end{aligned} \tag{B.1.8}$$

where $\|\cdot\|$ denotes the operator norm. We then bound each term on the RHS of (B.1.8). To bound $\|\boldsymbol{A} - \boldsymbol{W}\|$, we mostly follow Tang et al. (2013). Let $\boldsymbol{U}$ and $\boldsymbol{U}$ be $n \times K$ matrices of the leading K eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{W}$ respectively, and define $\boldsymbol{\mathcal{P}}_{\boldsymbol{A}} := \hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^T$ and $\boldsymbol{\mathcal{P}}_{\boldsymbol{W}} := \boldsymbol{U}\boldsymbol{U}^T$, then $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{\mathcal{P}}_{\boldsymbol{W}}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\mathcal{P}}_{\boldsymbol{W}} = \boldsymbol{\mathcal{P}}_{\boldsymbol{W}}\boldsymbol{W}\boldsymbol{\mathcal{P}}_{\boldsymbol{W}}$, and similarly $\boldsymbol{A} = \boldsymbol{\mathcal{P}}_{\boldsymbol{A}}\boldsymbol{A}\boldsymbol{\mathcal{P}}_{\boldsymbol{A}}$. We have

$$\begin{aligned}\|\boldsymbol{A} - \boldsymbol{W}\| =& \|\boldsymbol{\mathcal{P}}_{\boldsymbol{A}}\boldsymbol{A}\boldsymbol{\mathcal{P}}_{\boldsymbol{A}} - \boldsymbol{\mathcal{P}}_{\boldsymbol{W}}\boldsymbol{W}\boldsymbol{\mathcal{P}}_{\boldsymbol{W}}\| \\ \leq& \|\boldsymbol{\mathcal{P}}_{\boldsymbol{A}}(\boldsymbol{A} - \boldsymbol{W})\boldsymbol{\mathcal{P}}_{\boldsymbol{A}}\| + \|(\boldsymbol{\mathcal{P}}_{\boldsymbol{A}} - \boldsymbol{\mathcal{P}}_{\boldsymbol{W}})\boldsymbol{W}\boldsymbol{\mathcal{P}}_{\boldsymbol{A}}\| + \|\boldsymbol{\mathcal{P}}_{\boldsymbol{A}}\boldsymbol{W}(\boldsymbol{\mathcal{P}}_{\boldsymbol{A}} - \boldsymbol{\mathcal{P}}_{\boldsymbol{W}})\| \\ &+ \|(\boldsymbol{\mathcal{P}}_{\boldsymbol{A}} - \boldsymbol{\mathcal{P}}_{\boldsymbol{W}})\boldsymbol{W}(\boldsymbol{\mathcal{P}}_{\boldsymbol{A}} - \boldsymbol{\mathcal{P}}_{\boldsymbol{W}})\| \\ \leq& \|\boldsymbol{A} - \boldsymbol{W}\| + 2\|\boldsymbol{\mathcal{P}}_{\boldsymbol{A}} - \boldsymbol{\mathcal{P}}_{\boldsymbol{W}}\|\|\boldsymbol{W}\| + \|\boldsymbol{\mathcal{P}}_{\boldsymbol{A}} - \boldsymbol{\mathcal{P}}_{\boldsymbol{W}}\|^2\|\boldsymbol{W}\| . \end{aligned} \tag{B.1.9}$$

By Appendix A.1 of Lei and Rinaldo (2013), we have

$$\|\boldsymbol{\mathcal{P}}_{\boldsymbol{A}} - \boldsymbol{\mathcal{P}}_{\boldsymbol{W}}\| \leq \|\boldsymbol{\mathcal{P}}_{\boldsymbol{A}} - \boldsymbol{\mathcal{P}}_{\boldsymbol{W}}\|_F \leq \frac{2\sqrt{2K}\|\boldsymbol{A} - \boldsymbol{W}\|}{\lambda_{\min}(\boldsymbol{W})} . \tag{B.1.10}$$

By Theorem 5.2 of Lei and Rinaldo (2013), when $\log n/(n\alpha_n) \to 0$ and θ_i's are uniformly bounded by a constant M_θ, there exists constant C_{r,M_θ} depending on r, such that with probability $1 - n^{-r}$

$$\|\boldsymbol{A} - \boldsymbol{W}\| \leq C_r\sqrt{n\alpha_n} . \tag{B.1.11}$$

Since M_θ is a global constant in our setting, we write $C_r := C_{r,M_\theta}$.

In order to bound $\|\hat{\boldsymbol{X}}^*_{\boldsymbol{\tau_n}}\boldsymbol{O}_{\hat{X}} - \boldsymbol{X}^*_{\tau_n}\|_F$, it remains to bound the maximum and minimum eigenvalues of $\boldsymbol{W}$. We will show that the eigenvalues of $(n\alpha_n)^{-1}\boldsymbol{W}$ converge to those of

$\mathbb{E}[\theta_1^2 \boldsymbol{Z}_{1\cdot}^T \boldsymbol{Z}_{1\cdot}]\boldsymbol{B}$, which is strictly positive definite: for any $v \in \mathbb{R}^K$,

$$v^T \mathbb{E}[\boldsymbol{Z}_{1\cdot} \boldsymbol{Z}_{1\cdot}^T] \geq \sum_{k=1}^{K} \mathbb{P}(1 \in \mathcal{C}_k) \cdot v^T \boldsymbol{e}_k \boldsymbol{e}_k^T v \geq 0 \,,$$

where $\mathcal{C}_k$ denotes the set of nodes in community k and $\boldsymbol{e}_k$ denotes the vector the kth element equal to 1 and all others being 0. Equality holds only when all $v^T \boldsymbol{e}_k \boldsymbol{e}_k^T v = v_k^2 = 0$, i.e. $v = 0$.

Claim 2. *Assume that $\theta_i > 0$ for all i, and both $\boldsymbol{Z}$ and B are full rank. Let λ_0 and λ_1 denote the smallest and largest eigenvalues of $\mathbb{E}[\theta_1^2 \boldsymbol{Z}_{1\cdot}^T \boldsymbol{Z}_{1\cdot}]B$. Then*

$$\mathbb{P}\left(\left| \frac{\lambda_{\max}(\boldsymbol{W})}{n\alpha_n} - \lambda_1 \right| > \epsilon \right) \leq 2K^2 \exp\left(- \frac{\frac{1}{2} n\epsilon^2}{M_\theta^4 K^3 + \frac{1}{3} M_\theta^2 K \sqrt{K}\epsilon} \right) \tag{B.1.12}$$

$$\mathbb{P}\left(\left| \frac{\lambda_{\min}(\boldsymbol{W})}{n\alpha_n} - \lambda_0 \right| > \epsilon \right) \leq 2K^2 \exp\left(- \frac{\frac{1}{2} n\epsilon^2}{M_\theta^4 K^3 + \frac{1}{3} M_\theta^2 K \sqrt{K}\epsilon} \right) \tag{B.1.13}$$

Proof of Claim 2. For $k = 1, \ldots, K$, let λ_k denote the kth largest eigenvalue of $\boldsymbol{W}$, then

$$\begin{aligned} \lambda_k\left(\frac{\boldsymbol{W}}{n\alpha_n} \right) &= \lambda_k\left(\frac{\boldsymbol{\Theta Z B Z}^T \boldsymbol{\Theta}}{n} \right) = \lambda_k\left(\frac{\boldsymbol{B}^{1/2} \boldsymbol{Z}^T \boldsymbol{\Theta}^2 \boldsymbol{Z} \boldsymbol{B}^{1/2}}{n} \right) \\ &= \lambda_k\left(\frac{\boldsymbol{Z}^T \boldsymbol{\Theta}^2 \boldsymbol{Z B}}{n} \right) = \lambda_k\left(\frac{1}{n} \sum_{i=1}^{n} \theta_i^2 \boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot} \boldsymbol{B} \right) \end{aligned}$$

where the second equality is due to the fact that $\boldsymbol{X}\boldsymbol{X}^T$ and $\boldsymbol{X}^T\boldsymbol{X}$ share the same K leading eigenvalues ($\boldsymbol{X} = \sqrt{\alpha_n}\boldsymbol{\Theta Z} \boldsymbol{B}^{1/2}$). The third equality holds because $\boldsymbol{B}^{1/2}$ is full rank. To show (B.1.13), it suffices to show that

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^{n} \theta_i^2 \boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot} \boldsymbol{B} - \mathbb{E}[\theta_1^2 \boldsymbol{Z}_{1\cdot}^T \boldsymbol{Z}_{1\cdot} \boldsymbol{B}] \right\| > \epsilon \right) \leq 2 \exp\left(- \frac{\frac{1}{2} n\epsilon^2}{M_\theta^4 K^3 + \frac{1}{3} M_\theta^2 K \sqrt{K}\epsilon} \right) \tag{B.1.14}$$

For any $k, l \in \{1, \ldots, K\}$, $\{\theta_i^2 (\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot} \boldsymbol{B})_{kl}\}_i$ are an iid sequence uniformly bounded by $M_\theta^2 \sqrt{K}$ with mean $\left(\mathbb{E}[\theta_i^2 (\boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot} \boldsymbol{B})] \right)_{kl}$. By Bernstein's inequality,

$$\mathbb{P}\left(\left| \left(\frac{1}{n} \sum_{i=1}^{n} \theta_i^2 \boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot} \boldsymbol{B} - \mathbb{E}[\theta_1^2 \boldsymbol{Z}_{1\cdot}^T \boldsymbol{Z}_{1\cdot} \boldsymbol{B}] \right)_{kl} \right| > \epsilon \right) \leq 2 \exp\left(- \frac{\frac{1}{2} n\epsilon^2}{M_\theta^4 K + \frac{1}{3} M_\theta^2 \sqrt{K}\epsilon} \right) \,.$$

By the union bound and $\|A\| \leq \|A\|_F$, we have

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\theta_i^2 \boldsymbol{Z}_{i\cdot}^T \boldsymbol{Z}_{i\cdot}\boldsymbol{B} - \mathbb{E}[\theta_1^2 \boldsymbol{Z}_{1\cdot}^T \boldsymbol{Z}_{1\cdot}\boldsymbol{B}]\right\| > K\epsilon\right) \leq 2K^2 \exp\left(-\frac{\frac{1}{2}n\epsilon^2}{M_\theta^4 K + \frac{1}{3}M_\theta^2\sqrt{K}\epsilon}\right).$$

Replacing ϵ by ϵ/K completes the proof of Claim 2. □

We now return to the proof of Lemma 11 and complete the bound on $\|\hat{\boldsymbol{X}}^*_{\boldsymbol{\tau_n}}\boldsymbol{O}_{\hat{X}} - \boldsymbol{X}^*_{\tau_n}\|_F$. Taking ϵ to be $\frac{\lambda_1}{2}$ and $\frac{\lambda_0}{2}$ respectively in (B.1.12) and (B.1.13), by Claim 2, $\|W\| \leq \frac{3}{2}n\alpha_n\lambda_1 \leq \frac{3}{2}M_{\lambda_1}n\alpha_n K$ and $\lambda_{\min}(\boldsymbol{W}) \geq \frac{1}{2}n\alpha_n\lambda_0 \geq \frac{1}{2}M_{\lambda_0}n_\alpha$ hold with probability:

$$\begin{aligned}
&1 - 2K^2\exp\left(-\frac{\frac{1}{8}n\lambda_0^2}{M_\theta^4K^3 + \frac{1}{6}M_\theta^2K\sqrt{K}\lambda_0}\right) + 2K^2\exp\left(-\frac{\frac{1}{8}n\lambda_1^2}{M_\theta^4K^3 + \frac{1}{6}M_\theta^2K\sqrt{K}\lambda_1}\right)\\
\geq &1 - 4K^2\exp\left(-\frac{\frac{1}{8}n{M_{\lambda_0}}^2}{M_\theta^4K^5 + \frac{1}{6}M_\theta^2K^{5/2}M_{\lambda_0}}\right)
\end{aligned}$$

Plugging this, together with (B.1.10) and (B.1.10), back into (B.1.9), we have

$$\begin{aligned}
\|\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^T - \boldsymbol{X}\boldsymbol{X}^T\| &\leq \|\boldsymbol{A} - \boldsymbol{W}\|\left(1 + \frac{4\sqrt{2K}\|\boldsymbol{W}\|}{\lambda_{\min}(\boldsymbol{W})} + \frac{8K\|\boldsymbol{A}-\boldsymbol{W}\|\|\boldsymbol{W}\|}{(\lambda_{\min}(\boldsymbol{W}))^2}\right)\\
&\leq C_r\sqrt{n\alpha_n}\left(1 + \frac{12\sqrt{2K}M_{\lambda_1}}{M_{\lambda_0}} + \frac{48K^2C_rM_{\lambda_1}}{{M_{\lambda_0}}^2\sqrt{n\alpha_n}}\right) \qquad \text{(B.1.15)}
\end{aligned}$$

with probability at least $1-4K^2\exp\left(-\frac{\frac{1}{8}n{M_{\lambda_0}}^2}{M_\theta^4K^5+\frac{1}{6}M_\theta^2K^{5/2}M_{\lambda_0}}\right)-n^{-r}$. Then plugging (B.1.15) and Claim 2 into (B.1.8), we have

$$
\begin{aligned}
&\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}} - \boldsymbol{X}^*_{\tau_n}\|_F \\
\leq &\frac{2\sqrt{K}}{\tau_n}\frac{\|\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^T - \boldsymbol{X}\boldsymbol{X}^T\|\left(\sqrt{\|\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^T - \boldsymbol{X}\boldsymbol{X}^T\|} + 2\sqrt{\|\boldsymbol{X}\boldsymbol{X}^T\|}\right)}{\lambda_{\min}(\boldsymbol{X}\boldsymbol{X}^T)} \\
\leq &\frac{2\sqrt{K}}{\tau_n}\frac{C_r\sqrt{n\alpha_n}\left(1 + \frac{12\sqrt{2K}KM_{\lambda_1}}{M_{\lambda_0}} + \frac{48K^2C_rM_{\lambda_1}}{M_{\lambda_0}\sqrt{n\alpha_n}}\right)}{n\alpha_n\frac{M_{\lambda_0}}{2K}} \\
&\cdot\left(\left[C_r\sqrt{n\alpha_n}\left(1 + \frac{12\sqrt{2K}KM_{\lambda_1}}{M_{\lambda_0}} + \frac{48K^2C_rM_{\lambda_1}}{M_{\lambda_0}\sqrt{n\alpha_n}}\right)\right]^{\frac{1}{2}} + \sqrt{6M_{\lambda_1}n\alpha_n}\right) \\
= &\frac{4C_r\sqrt{K}}{\tau_nM_{\lambda_0}}\left(1 + \frac{12\sqrt{2}M_{\lambda_1}K\sqrt{K}}{M_{\lambda_0}} + \frac{48C_rM_{\lambda_1}K^2}{{M_{\lambda_0}}^2\sqrt{n\alpha_n}}\right) \\
&\cdot\left(\left[C_r\left(\frac{1}{\sqrt{n\alpha_n}} + \frac{12\sqrt{2}M_{\lambda_1}}{M_{\lambda_0}}\frac{K\sqrt{K}}{\sqrt{n\alpha_n}} + \frac{48C_rM_{\lambda_1}}{{M_{\lambda_0}}^2}\frac{K^2}{n\alpha_n}\right)\right]^{\frac{1}{2}} + \sqrt{6M_{\lambda_1}}\right) \qquad \text{(B.1.16)}
\end{aligned}
$$

By assumption $K = O(\log(n))$, we have $\frac{K^3}{n\alpha_n} \to 0$, thus for large enough n, the following inequalities that simplify (B.1.16) hold:

$$
\begin{aligned}
\frac{(24 - 12\sqrt{2})M_{\lambda_1}K\sqrt{K}}{M_{\lambda_0}} - 1 &\geq \frac{48C_rM_{\lambda_1}K^2}{{M_{\lambda_0}}^2\sqrt{n\alpha_n}} \\
(3 - \sqrt{6})\sqrt{M_{\lambda_1}} &\geq \left[C_r\left(\frac{1}{\sqrt{n\alpha_n}} + \frac{12\sqrt{2}M_{\lambda_1}}{M_{\lambda_0}}\frac{K\sqrt{K}}{\sqrt{n\alpha_n}} + \frac{48C_rM_{\lambda_1}}{{M_{\lambda_0}}^2}\frac{K^2}{n\alpha_n}\right)\right]^{\frac{1}{2}},
\end{aligned}
$$

and we have

$$
\text{RHS of (B.1.16)} \leq \tilde{C}_r \cdot \frac{K^2}{\tau_n} \qquad \text{(B.1.17)}
$$

where the constant $\tilde{C}_r := \frac{288C_r{M_{\lambda_1}}^{3/2}}{{M_{\lambda_0}}^2}$, which for simplicity we will continue to write as C_r. This completes the bound on $\|\hat{\boldsymbol{X}}^*_{\boldsymbol{\tau_n}}\boldsymbol{O}_{\hat{X}} - \boldsymbol{X}^*_{\tau_n}\|_F$.

The second part of the proof requires a bound on $\|\boldsymbol{X}^*_{\tau_n} - \boldsymbol{X}^*\|_F$. From the defintion of $\boldsymbol{X}^*_{\tau_n}$, we can write

$$
\|\boldsymbol{X}^*_{\tau_n} - \boldsymbol{X}^*\|_F^2 = \sum_{i=1}^{n}\left(\frac{\tau_n/\sqrt{\alpha_n}}{\|\boldsymbol{X}_{i\cdot}\|_2/\sqrt{\alpha_n} + \tau_n/\sqrt{\alpha_n}}\right)^2. \qquad \text{(B.1.18)}
$$

Since $\frac{\|X_{\cdot}\|_2}{\sqrt{\alpha_n}} = \theta_i\|\boldsymbol{Z}_{i\cdot}\boldsymbol{B}^{1/2}\|_2 = \theta_i\sqrt{\boldsymbol{Z}_{i\cdot}\boldsymbol{B}\boldsymbol{Z}_{i\cdot}^T} \geq \theta_i\sqrt{\lambda_{\min}(\boldsymbol{B})} \geq \theta_i\sqrt{m_B} > 0$, by assump-

tion, for $\epsilon \in (0, \epsilon_0)$, we have $\mathbb{P}\left(\frac{\|\boldsymbol{X}_{i\cdot}\|_2}{\sqrt{\alpha_n}} < \epsilon\sqrt{m_B}\right) \leq \mathbb{P}(\theta_i < \epsilon) \leq C_\theta \epsilon$. Therefore, for any $\epsilon \in (0, \epsilon_0)$, we have

$$\mathbb{E}\left[\left(\frac{\tau_n/\sqrt{\alpha_n}}{\|\boldsymbol{X}_{i\cdot}\|_2/\sqrt{\alpha_n} + \tau_n/\sqrt{\alpha_n}}\right)^2\right] \leq C_\theta\epsilon + (1 - C_\theta\epsilon)\left(\frac{\tau_n/\sqrt{\alpha_n}}{\epsilon\sqrt{m_B} + \tau_n/\sqrt{\alpha_n}}\right)^2 \quad \text{(B.1.19)}$$

By assumption, $\tau_n/\sqrt{\alpha_n} \to 0$, so for large enough n such that $\tau_n/\sqrt{\alpha_n} < \epsilon_0^{3/2}$, taking $\epsilon := (\tau_n/\sqrt{n})^{2/3} < \epsilon_0$, we have

$$\begin{aligned}\text{LHS of (B.1.19)} &\leq C_\theta(\tau_n/\sqrt{\alpha_n})^{2/3} + (1 - C_\theta(\tau_n/\sqrt{\alpha_n})^{2/3})\left(\frac{(\tau_n/\sqrt{\alpha_n})^{1/3}}{m_B + (\tau_n/\sqrt{\alpha_n})^{1/3}}\right)^2 \\ &\leq \left(C_\theta + m_B^{-1}\right)(\tau_n/\sqrt{\alpha_n})^{2/3}\end{aligned}$$

Then for any $\delta > 0$, we have

$$\begin{aligned}&\mathbb{P}\left(\frac{\|\boldsymbol{X}^*_{\tau_n} - \boldsymbol{X}^*\|_F^2}{n} - (C_\theta + m_B^{-1})(\tau_n/\sqrt{\alpha_n})^{2/3} > \delta\right) \\ \leq&\mathbb{P}\left(\frac{\|\boldsymbol{X}^*_{\tau_n} - \boldsymbol{X}^*\|_F^2}{n} - \mathbb{E}\left[\frac{\|\boldsymbol{X}^*_{\tau_n} - \boldsymbol{X}^*\|_F^2}{n}\right] > \delta\right) \leq \exp\left(-\frac{\frac{1}{2}\delta^2 n}{1 + \frac{1}{3}\delta}\right) \quad \text{(B.1.20)}\end{aligned}$$

where the second inequality is Bernstein's inequality plus the fact that each summand in the numerator of (B.1.18) is uniformly bounded by 1 with an expectation bounded by $\left(C_\theta + m_B^{-1}\right)(\tau_n/\sqrt{\alpha_n})^{2/3}$.

We can now complete the proof of Lemma 11. Combining (B.1.17) and (B.1.20) yields

$$\begin{aligned}&\mathbb{P}\left(\frac{\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}} - \boldsymbol{X}^*\|_F}{\sqrt{n}} \leq \frac{C_r K^2}{\tau_n\sqrt{n}} + \delta + (C_\theta + m_B^{-1})(\tau_n/\sqrt{\alpha_n})^{2/3}\right) \\ \geq&1 - P_1(n, \alpha_n, K; r) \quad \text{(B.1.21)}\end{aligned}$$

The optimal τ_n that minimizes the RHS of the inequality inside the probability is $\tau_n = \frac{\alpha_n^{0.2}K^{1.5}}{n^{0.3}}$ – here for simplicity we drop the constant factor in τ_n, the effect of which we evaluated empirically in Section 3.5. Plugging this into (B.1.21) and taking $\delta = K(n\alpha_n)^{-\frac{1}{5}}$ and denote $C_1 := \left(\frac{2}{3}(C_\theta + m_B^{-1})C_r^{\frac{2}{3}}\right)^{\frac{3}{5}} + 1 + \left(\frac{3}{2}C_r(C_\theta + m_B^{-1})^{\frac{3}{2}}\right)^{\frac{2}{5}}$ and $P_1(n, \alpha_n, K; r) := n^{-r} - 4\exp\left(-\frac{\frac{1}{8}{M_{\lambda_0}}^2 n}{M_\theta^4 K^5 + \frac{1}{6}M_\theta^2 M_{\lambda_0}K^{\frac{5}{2}}}\right) - 2\exp\left(-\frac{\frac{1}{2}K^{\frac{8}{5}}n^{\frac{3}{5}}\alpha_n^{-\frac{2}{5}}}{1 + \frac{1}{3}K^{\frac{4}{3}}(n\alpha_n)^{\frac{1}{5}}}\right)$, we obtain Lemma 11. Note that since we are free to choose and fix r, we can drop the dependence on it from P_1, as we did in the statement of Lemma 11. □

The next step is to show the convergence of the estimated cluster centers $\hat{\boldsymbol{S}}$ to the

population cluster centers $\boldsymbol{S}_{\mathcal{F}}$.

Lemma 12. *Recall that $\mathcal{F}$ denotes the popualtion distribution of the rows of $\boldsymbol{X}^*$ and let $\hat{\boldsymbol{S}} \in \arg\min_S \mathcal{L}_n(\hat{\boldsymbol{X}}^*_{\tau_n}; \boldsymbol{S})$ and $S_{\mathcal{F}} \in \arg\min_S \mathcal{L}(\mathcal{F}; \boldsymbol{S})$. Assume that conditions A1, A2, A3 and B hold. Then if $\frac{\log n}{n\alpha_n} \to 0$ and $K = O(\log n)$, for large enough n we have*

$$\mathbb{P}\left(D_H(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}}; \boldsymbol{S}_{\mathcal{F}}) \leq \frac{C_2 K^{\frac{9}{5}}}{(n\alpha_n)^{\frac{1}{5}}} \right) \leq 1 - P_1(n, \alpha_n, K) - P_2(n, \alpha_n, K) \tag{B.1.22}$$

where C_2 is a global constant, $P_2(n, \alpha_n, K) \to 0$ as $n \to \infty$ and $D_H(\cdot, \cdot)$ is as defined in condition B.

Proof of Lemma 12. Since the rows of $\hat{\boldsymbol{X}}^*_{\tau_n}$ and $\boldsymbol{X}^*$ have l_2 norms bounded by 1, the sample space of $\mathcal{F}$ is uniformly bounded in the unit l_2 ball. Following the argument of Pollard et al. (1981), we show that all cluster centers estimated by K-medians fall in the l_2 ball centered at origin with radius 3, which we denote as $\mathcal{R}$. Otherwise, if there exists an estimated cluster center $\boldsymbol{s}$ outside $\mathcal{R}$, it is at least distance 2 away from any point assigned to its cluster. Therefore, moving $\boldsymbol{s}$ to an arbitrary point inside the unit ball yields an improvement in the loss function since any two points inside the unit ball are at most distance 2 away from each other.

We first show the uniform convergence of $\mathcal{L}_n(\hat{\boldsymbol{X}}^*_{\boldsymbol{\tau_n}}\boldsymbol{O}_{\hat{X}}; \boldsymbol{S})$ to $\mathcal{L}(\mathcal{F}; \boldsymbol{S})$ and then show the optimum of $\mathcal{L}_n(\hat{\boldsymbol{X}}^*_{\boldsymbol{\tau_n}}\boldsymbol{O}_{\hat{X}}; \boldsymbol{S})$ is close to that of $\mathcal{L}(\mathcal{F}; \boldsymbol{S})$. Let

$$\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}} := \arg\min_S \mathcal{L}_n(\hat{\boldsymbol{X}}^*_{\boldsymbol{\tau_n}}\boldsymbol{O}_{\hat{X}}; \boldsymbol{S})$$

We start with showing that

$$\sup_{\boldsymbol{S}\subset\mathcal{R}} |\mathcal{L}_n(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}}; \boldsymbol{S}) - \mathcal{L}_n(\boldsymbol{X}^*; \boldsymbol{S})| \leq \frac{\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}} - \boldsymbol{X}^*\|_F}{\sqrt{n}} . \tag{B.1.23}$$

To prove (B.1.23), take any $\boldsymbol{s} \in \mathcal{R}$. For each i, let $\hat{\boldsymbol{s}}$ and $\boldsymbol{s}$ be (possibly identical) rows in $\boldsymbol{S}$ that are closest to $(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}})_{i\cdot}$ and $\boldsymbol{X}^*_{i\cdot}$ respectively in l_2 norm. We have

$$\|\boldsymbol{X}^*_{i\cdot} - \boldsymbol{s}\|_2 - \|(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}})_{i\cdot} - \hat{\boldsymbol{s}}\|_2 \leq \|(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}})_{i\cdot} - \boldsymbol{X}^*_{i\cdot}\|_2$$

and similarly, $\|(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}})_{i\cdot} - \hat{\boldsymbol{s}}\|_2 - \|\boldsymbol{X}^*_{i\cdot} - \boldsymbol{s}\|_2 \leq \|\boldsymbol{X}^*_{i\cdot} - (\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}})_{i\cdot}\|_2$. Thus $|\|(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}})_{i\cdot} - \hat{\boldsymbol{s}}\|_2 - \|\boldsymbol{X}^*_{i\cdot} - \boldsymbol{s}\|_2| \leq \|(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}})_{i\cdot} - \boldsymbol{X}^*_{i\cdot}\|_2$. Combining this inequalities for all rows, we

have

$$|\mathcal{L}_n(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}};\boldsymbol{S}) - \mathcal{L}_n(\boldsymbol{X}^*;\boldsymbol{S})| = \Big|\frac{1}{n}\sum_{i=1}^{n}\Big(\|(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}})_{i\cdot} - \hat{\boldsymbol{s}}\|_2 - \|\boldsymbol{X}^*_{i\cdot} - \boldsymbol{s}\|_2\Big)\Big|$$

$$\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}})_{i\cdot} - \boldsymbol{X}^*_{i\cdot}\|_2^2} = \frac{1}{\sqrt{n}}\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}} - \boldsymbol{X}^*\|_F\,. \tag{B.1.24}$$

Then since that (B.1.24) holds for any $\boldsymbol{S}$, the uniform bound (B.1.23) follows.

For simplicity, we introduce the notation "$\boldsymbol{S} \subset \mathcal{R}$", by which we mean that the rows of a matrix $\boldsymbol{S}$ belong to the set $\mathcal{R}$. We now derive the bound for $\sup_{\boldsymbol{S}\subset\mathcal{R}}|\mathcal{L}_n(\boldsymbol{X}^*;\boldsymbol{S}) - \mathcal{L}(\mathcal{F};\boldsymbol{S})|$, which, without taking the supremum, is easily bounded by Bernstein's inequality. To tackle the uniform bound, we employ an ϵ-net (see, for example, Haussler and Welzl (1986)). There exists an ϵ-net $\mathcal{R}_\epsilon$, with size $|\mathcal{R}_\epsilon| \leq C_{\mathcal{R}}\frac{K}{\epsilon}\log\frac{K}{\epsilon}$, where $C_{\mathcal{R}}$ is a global constant. For any $\tilde{\boldsymbol{S}} \subset \mathcal{R}_\epsilon$, $\tilde{\boldsymbol{S}} \in \mathbb{R}^{K\times K}$, notice that $\min_{1\leq k\leq K}\|\boldsymbol{X}^*_{i\cdot} - \tilde{\boldsymbol{S}}_{k\cdot}\|_2$ is a random variable uniformly bounded by 6 with expectation $\mathcal{L}(\mathcal{F};\tilde{\boldsymbol{S}})$ for each i. Therefore, by Bernstein's inequality, for any $\delta > 0$ we have

$$\mathbb{P}(|\mathcal{L}_n(\boldsymbol{X}^*;\tilde{\boldsymbol{S}}) - \mathcal{L}(\mathcal{F};\tilde{\boldsymbol{S}})| > \delta) \leq \exp\left(-\frac{\frac{1}{2}n\delta^2}{4R_M^2 + \frac{2}{3}R_M\delta}\right) = \exp\left(-\frac{n\delta^2}{72+4\delta}\right) \tag{B.1.25}$$

The number of all such $\tilde{\boldsymbol{S}} \subset \mathcal{R}_\epsilon$ is bounded by

$$\Big|\{\tilde{\boldsymbol{S}} : \tilde{\boldsymbol{S}} \subset \mathcal{R}_\epsilon\}\Big| = \binom{C_{\mathcal{R}}\frac{K}{\epsilon}\log\frac{K}{\epsilon}}{K} \leq \left(C_{\mathcal{R}}\frac{K}{\epsilon}\log\frac{K}{\epsilon}\right)^K$$

By the union bound, we have

$$\mathbb{P}\left(\sup_{\tilde{\boldsymbol{S}}\in\mathcal{R}_\epsilon}\Big|\mathcal{L}(\boldsymbol{X}^*;\tilde{\boldsymbol{S}}) - \mathcal{L}(\mathcal{F};\tilde{\boldsymbol{S}})\Big| > \delta\right) < \left(C_{\mathcal{R}}\frac{K}{\epsilon}\log\frac{K}{\epsilon}\right)^K \exp\left(-\frac{n\delta^2}{72+4\delta}\right) \tag{B.1.26}$$

The above shows the uniform convergence of the loss functions for $\tilde{\boldsymbol{S}}$ from the ϵ-net $\mathcal{R}_\epsilon$. We then expand it to the uniform convergence of all $\boldsymbol{S} \subset \mathcal{R}$. For any $\boldsymbol{S} \subset \mathcal{R}$, there exists $\tilde{\boldsymbol{S}} \subset \mathcal{R}_\epsilon$, such that both $\mathcal{L}_n(\cdot;\boldsymbol{S})$ and $\mathcal{L}(\cdot;\boldsymbol{S})$ can be well approximated by $\mathcal{L}_n(\cdot;\tilde{\boldsymbol{S}})$ and $\mathcal{L}(\cdot;\tilde{\boldsymbol{S}})$ respectively. To emphasize the dependence of $\tilde{\boldsymbol{S}}$ on $\boldsymbol{S}$, we write $\tilde{\boldsymbol{S}} = \tilde{\boldsymbol{S}}(\boldsymbol{S})$.

Formally, we now prove the following.

$$\sup_{\boldsymbol{S}\subset\mathcal{R}} |\mathcal{L}_n(\boldsymbol{X}^*;\boldsymbol{S}) - \mathcal{L}_n(\boldsymbol{X}^*;\tilde{\boldsymbol{S}}(\boldsymbol{S}))| < \epsilon \tag{B.1.27}$$

$$\sup_{\boldsymbol{S}\subset\mathcal{R}} |\mathcal{L}(\mathcal{F};\boldsymbol{S}) - \mathcal{L}(\mathcal{F};\tilde{\boldsymbol{S}}(\boldsymbol{S}))| < \epsilon \tag{B.1.28}$$

To prove (B.1.27) and (B.1.28), for any $\boldsymbol{S} \subset \mathcal{R}$, let $\tilde{\boldsymbol{S}} \subset \mathcal{R}_\epsilon$ be a matrix formed by concatenating the points in $\mathcal{R}_\epsilon$ that best approximate the rows in $\boldsymbol{S}$. Notice that $\tilde{\boldsymbol{S}}$ formed such way may contain less than K rows. In this case, we arbitrarily pick points in $\mathcal{R}_\epsilon$ to enlarge $\tilde{\boldsymbol{S}}$ to K rows. For any $x \in \mathbb{R}^K$, let $\boldsymbol{s}_0$ be the best approximation to x among the rows of $\boldsymbol{S}$ and $\tilde{\boldsymbol{s}}_0$ be the best approximation to $\boldsymbol{s}_0$ among the rows of $\tilde{\boldsymbol{S}}$; let $\tilde{\boldsymbol{s}}_1$ be the best approximation to x among the rows of $\tilde{\boldsymbol{S}}$ and let $\boldsymbol{s}_1$ be the point among the rows of $\boldsymbol{S}$ that is best approximated by $\tilde{\boldsymbol{s}}_1$. Since $\|\boldsymbol{x}-\boldsymbol{s}_0\|_2 \le \|\boldsymbol{x}-\boldsymbol{s}_1\|_2 \le \|\boldsymbol{x}-\boldsymbol{s}_1\|_2 + \|\boldsymbol{s}_1-\tilde{\boldsymbol{s}}_1\|_2 \le \|\boldsymbol{x}-\tilde{\boldsymbol{s}}_1\|_2 + \epsilon$, and similarly, $\|\boldsymbol{x}-\tilde{\boldsymbol{s}}_1\|_2 \le \|\boldsymbol{x}-\tilde{\boldsymbol{s}}_0\|_2 \le \|\boldsymbol{x}-\boldsymbol{s}_0\|_2 + \epsilon$, we have

$$\left| \min_{1\le k\le K} \|\boldsymbol{x}-\boldsymbol{S}_{k\cdot}\|_2 - \min_{1\le k\le K} \|\boldsymbol{x}-\tilde{\boldsymbol{S}}_{k\cdot}\|_2 \right| = \left| \|\boldsymbol{x}-\boldsymbol{s}_0\|_2 - \|\boldsymbol{x}-\tilde{\boldsymbol{s}}_1\|_2 \right| \le \epsilon \tag{B.1.29}$$

which implies (B.1.27) and (B.1.28).

Combining (B.1.23), (B.1.26), (B.1.27) and (B.1.28), we have shown that with probability $P_2(n,\epsilon,\delta) := 1 - \left(C_{\mathcal{R}}\frac{K+2}{\epsilon}\log\frac{K+2}{\epsilon}\right)^K \exp\left(-\frac{n\delta^2}{72+4\delta}\right)$,

$$\begin{aligned}
&\sup_{\boldsymbol{S}\subset\mathcal{R}} \left| \mathcal{L}(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}};\boldsymbol{S}) - \mathcal{L}(\mathcal{F};\boldsymbol{S}) \right| \\
\le &\sup_{\boldsymbol{S}\subset\mathcal{R}} |\mathcal{L}(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}};\boldsymbol{S}) - \mathcal{L}(\boldsymbol{X}^*;\boldsymbol{S})| + \sup_{\boldsymbol{S}\subset\mathcal{R}} |\mathcal{L}(\boldsymbol{X}^*;\boldsymbol{S}) - \mathcal{L}(\boldsymbol{X}^*;\tilde{\boldsymbol{S}}(\boldsymbol{S}))| \\
&+ \sup_{\tilde{\boldsymbol{S}}\subset\mathcal{R}_\epsilon} \left| \mathcal{L}(\boldsymbol{X}^*;\tilde{\boldsymbol{S}}) - \mathcal{L}(\mathcal{F};\tilde{\boldsymbol{S}}) \right| + \sup_{\boldsymbol{S}\subset\mathcal{R}} |\mathcal{L}(\mathcal{F};\boldsymbol{S}) - \mathcal{L}(\mathcal{F};\tilde{\boldsymbol{S}}(\boldsymbol{S}))| \\
&\le \frac{\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}} - \boldsymbol{X}^*\|_F}{\sqrt{n}} + \delta + 2\epsilon
\end{aligned} \tag{B.1.30}$$

Finally, we use (B.1.30) to bound $D_H(\hat{\boldsymbol{S}}, \boldsymbol{S}_{\mathcal{F}})$. Note that

$$\begin{aligned}
&\mathcal{L}(\mathcal{F};\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}}) - \mathcal{L}(\mathcal{F};\boldsymbol{S}_{\mathcal{F}}) \le |\mathcal{L}(\mathcal{F};\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}}) - \mathcal{L}(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}};\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}})| \\
&+ (\mathcal{L}(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}};\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}}) - \mathcal{L}(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}};\boldsymbol{S}_{\mathcal{F}})) + |\mathcal{L}(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}};\boldsymbol{S}_{\mathcal{F}}) - \mathcal{L}(\mathcal{F};\boldsymbol{S}_{\mathcal{F}})| \\
&\le 2\sup_{\boldsymbol{S}\subset\mathcal{R}} \left| \mathcal{L}(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}};\boldsymbol{S}) - \mathcal{L}(\mathcal{F};\boldsymbol{S}) \right|.
\end{aligned}$$

Taking $\delta = \epsilon = \frac{K^{\frac{4}{5}}}{(n\alpha_n)^{\frac{1}{5}}}$, define $P_2(n,\alpha_n,K) := P_2(n,\epsilon,\delta)$ with plug-in values of δ and ϵ.

To summarize, we have that with probability at least $1 - P_1(n, \alpha_n, K; r) - P_2(n, \alpha_n, K)$, the following holds:

$$\begin{aligned} D_H(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}}, \boldsymbol{S}_{\mathcal{F}}) &\leq (MK^{-1})^{-1}(\mathcal{L}(\mathcal{F}, \hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}}) - \mathcal{L}(\mathcal{F}, \boldsymbol{S}_{\mathcal{F}})) \\ &\leq 2K/M \sup_{\boldsymbol{S}\subset\mathcal{R}} \left|\mathcal{L}(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}}; \boldsymbol{S}) - \mathcal{L}(\mathcal{F}; \boldsymbol{S})\right| \\ &\leq 2K/M \left(\frac{\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}} - \boldsymbol{X}^*\|_F}{\sqrt{n}} + \delta + 2\epsilon\right) \\ &\leq \frac{(C_1+3)K^{\frac{9}{5}}}{M(n\alpha_n)^{\frac{1}{5}}} =: \frac{C_2 K^{\frac{9}{5}}}{M(n\alpha_n)^{\frac{1}{5}}} \end{aligned}$$

where we let $C_2 := (C_1 + 3)/M$. This concludes the proof of Lemma 12

□

Proof of the main result (Theorem 3). Without loss of generality, we assume that the rows of $\hat{\boldsymbol{S}}$ and $\boldsymbol{S}_{\mathcal{F}}$ are aligned in the sense that $\|\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}} - \boldsymbol{S}_{\mathcal{F}}\|_2 = D_H(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}}, \boldsymbol{S}_{\mathcal{F}})$. We denote the unnormalized projection coefficients of $\hat{\boldsymbol{X}}^*_{\boldsymbol{\tau_n}}$ onto $\hat{\boldsymbol{S}}$ by $\hat{\boldsymbol{Y}}$ and $\boldsymbol{X}^*$ onto $\boldsymbol{S}_{\mathcal{F}} = \boldsymbol{B}^{1/2}$ by $\boldsymbol{Y}$. We have:

$$Y = \boldsymbol{X}^*(\boldsymbol{B}^{1/2})^T\boldsymbol{B}^{-1} = \boldsymbol{X}^*\boldsymbol{S}_{\mathcal{F}}^T(\boldsymbol{S}_{\mathcal{F}}\boldsymbol{S}_{\mathcal{F}}^T)^{-1} \tag{B.1.31}$$

$$\hat{Y} = \hat{\boldsymbol{X}}^*_{\boldsymbol{\tau_n}}\hat{\boldsymbol{S}}^T(\hat{\boldsymbol{S}}\hat{\boldsymbol{S}}^T)^{-1} = \hat{\boldsymbol{X}}^*_{\boldsymbol{\tau_n}}\boldsymbol{O}_{\hat{X}}(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}})^T(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}}(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}})^T)^{-1} \tag{B.1.32}$$

Recall that $\boldsymbol{X}_{i\cdot} = \boldsymbol{Z}_{i\cdot}\boldsymbol{S}_{\mathcal{F}}$, thus $\boldsymbol{X}^*_{i\cdot} = \frac{\boldsymbol{Z}_{i\cdot}\boldsymbol{S}_{\mathcal{F}}}{\|\boldsymbol{Z}_{i\cdot}\boldsymbol{S}_{\mathcal{F}}\|_2}$, and

$$\begin{aligned} \|\boldsymbol{Y}_{i\cdot}\|_2 = \boldsymbol{X}^*_{i\cdot}\boldsymbol{S}_{\mathcal{F}}^T(\boldsymbol{S}_{\mathcal{F}}\boldsymbol{S}_{\mathcal{F}}^T)^{-1} &= \frac{\|\boldsymbol{Z}_{i\cdot}\|_2}{\|\boldsymbol{Z}_{i\cdot}\boldsymbol{S}_{\mathcal{F}}\|_2} = \frac{1}{\|\sum_{i=1}^K Z_{ik}(\boldsymbol{S}_{\mathcal{F}})_{k\cdot}\|_2} \\ &\geq \frac{1}{\sum_{k=1}^K |Z_{ik}|\|(\boldsymbol{S}_{\mathcal{F}})_{k\cdot}\|_2} = \frac{1}{\sum_{k=1}^K |Z_{ik}|} \geq \frac{1}{\|\boldsymbol{Z}_{i\cdot}\|_2\sqrt{K}} = \frac{1}{\sqrt{K}}\,. \end{aligned}$$

The difference between the row-normalized projection coefficients $\boldsymbol{Z}$ and $\hat{\boldsymbol{Z}}$ can be bounded by the difference between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$, since

$$\|\hat{\boldsymbol{Z}}_{i\cdot} - \boldsymbol{Z}_{i\cdot}\|_2 = \left\|\frac{\hat{Y}_{i\cdot}\|\boldsymbol{Y}_{i\cdot}\|_2 - \boldsymbol{Y}_{i\cdot}\|\hat{\boldsymbol{Y}}_{i\cdot}\|_2}{\|\hat{\boldsymbol{Y}}_{i\cdot}\|_2\|\boldsymbol{Y}_{i\cdot}\|_2}\right\|_2 \leq \frac{2\|\hat{\boldsymbol{Y}}_{i\cdot} - \boldsymbol{Y}_{i\cdot}\|_2}{\|\boldsymbol{Y}_{i\cdot}\|_2}\,. \tag{B.1.33}$$

Then we have

$$
\begin{aligned}
&\frac{\|\hat{\boldsymbol{Z}}-\boldsymbol{Z}\|_2}{\sqrt{n}} \le \frac{2\|\hat{\boldsymbol{Y}}-\boldsymbol{Y}\|_F\sqrt{K}}{\sqrt{n}} \\
=&2\sqrt{\frac{K}{n}}\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}}(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}})^T(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}}(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}})^T)^{-1}-\boldsymbol{X}^*\boldsymbol{S}_{\mathcal{F}}^T(\boldsymbol{S}_{\mathcal{F}}\boldsymbol{S}_{\mathcal{F}}^T)^{-1}\|_F \\
\le&2\sqrt{\frac{K}{n}}\Big\{\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}}(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}})^T\left(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}}(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}})^T)^{-1}-(\boldsymbol{S}_{\mathcal{F}}\boldsymbol{S}_{\mathcal{F}}^T)^{-1}\right)\|_F \\
&+\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}}((\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}})^T-\boldsymbol{S}_{\mathcal{F}}^T)(\boldsymbol{S}_{\mathcal{F}}\boldsymbol{S}_{\mathcal{F}}^T)^{-1}\|_F \\
&+\|\left(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}}-\boldsymbol{X}^*\right)\boldsymbol{S}_{\mathcal{F}}^T(\boldsymbol{S}_{\mathcal{F}}\boldsymbol{S}_{\mathcal{F}}^T)^{-1}\|_F\Big\} =: I_1+I_2+I_3\,, \qquad \text{(B.1.34)}
\end{aligned}
$$

where

$$
\begin{aligned}
I_1 :=& 2\sqrt{\frac{K}{n}}\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}}(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}})^T\left((\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}}(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}})^T)^{-1}-(\boldsymbol{S}_{\mathcal{F}}\boldsymbol{S}_{\mathcal{F}}^T)^{-1}\right)\|_F \\
\le& 2\sqrt{\frac{K}{n}}\left(\|\boldsymbol{X}^*\|_F+\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}}-\boldsymbol{X}^*\|_F\right) \\
&\cdot\left(\|\boldsymbol{S}_{\mathcal{F}}\|_F+\|\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}}-\boldsymbol{S}_{\mathcal{F}}\|_F\right)\|(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}}(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}})^T)^{-1}-(\boldsymbol{S}_{\mathcal{F}}\boldsymbol{S}_{\mathcal{F}}^T)^{-1}\|_F\,, \\
I_2 :=& 2\sqrt{\frac{K}{n}}\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}}((\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}})^T-\boldsymbol{S}_{\mathcal{F}}^T)(\boldsymbol{S}_{\mathcal{F}}\boldsymbol{S}_{\mathcal{F}}^T)^{-1}\|_F \\
\le& 2\sqrt{\frac{K}{n}}\left(\|\boldsymbol{X}^*\|_F+\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}}-\boldsymbol{X}^*\|_F\right)\|\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}}-\boldsymbol{S}_{\mathcal{F}}\|_F\|(\boldsymbol{S}_{\mathcal{F}}\boldsymbol{S}_{\mathcal{F}}^T)^{-1}\|_F \\
I_3 :=& 2\sqrt{\frac{K}{n}}\|\left(\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}}-\boldsymbol{X}^*\right)\boldsymbol{S}_{\mathcal{F}}^T(\boldsymbol{S}_{\mathcal{F}}\boldsymbol{S}_{\mathcal{F}}^T)^{-1}\|_F \\
\le& 2\sqrt{\frac{K}{n}}\|\hat{\boldsymbol{X}}^*_{\tau_n}\boldsymbol{O}_{\hat{X}}-\boldsymbol{X}^*\|_F\|\boldsymbol{S}_{\mathcal{F}}\|_F\|(\boldsymbol{S}_{\mathcal{F}}\boldsymbol{S}_{\mathcal{F}}^T)^{-1}\|_F
\end{aligned}
$$

The term $\|(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}}(\hat{\boldsymbol{S}}\boldsymbol{O}_{\hat{X}})^T)^{-1}-(\boldsymbol{S}_{\mathcal{F}}\boldsymbol{S}_{\mathcal{F}}^T)^{-1}\|_F$ is bounded by the following claim.

Claim 3. *For two $K\times K$ matrices $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ such that $\|\boldsymbol{V}_2\|_F=\sqrt{K}$, $\|\boldsymbol{V}_1-\boldsymbol{V}_2\|_2\le\epsilon$, and $\lambda_{\min}(\boldsymbol{V}_2)>0$, we have*

$$
\begin{aligned}
\|(\boldsymbol{V}_1\boldsymbol{V}_1^T)^{-1}-(\boldsymbol{V}_2\boldsymbol{V}_2^T)^{-1}\|_F \le& K^2\left(\lambda_{\min}(\boldsymbol{V}_2\boldsymbol{V}_2^T)-K(2+\epsilon)\epsilon\right)^{-1} \\
&\cdot(\lambda_{\min}(\boldsymbol{V}_2\boldsymbol{V}_2^T))^{-1}(2+\epsilon)\epsilon \qquad \text{(B.1.35)}
\end{aligned}
$$

Proof of Claim 3. First,

$$\begin{aligned}
&\|\boldsymbol{V}_1\boldsymbol{V}_1^T - \boldsymbol{V}_2\boldsymbol{V}_2^T\|_F \le \|\boldsymbol{V}_1(\boldsymbol{V}_1 - \boldsymbol{V}_2)^T + (\boldsymbol{V}_1 - \boldsymbol{V}_2)\boldsymbol{V}_2^T\|_F \\
&\le (\|\boldsymbol{V}_1\|_F + \|\boldsymbol{V}_2\|_F)\|\boldsymbol{V}_1 - \boldsymbol{V}_2\|_F \le (\|\boldsymbol{V}_1 - \boldsymbol{V}_2\|_F + 2\|\boldsymbol{V}_2\|_F)\|\boldsymbol{V}_1 - \boldsymbol{V}_2\|_F \\
&\le (\sqrt{K}\epsilon + 2\sqrt{K}\epsilon)\sqrt{K}\epsilon = (2+\epsilon)K\epsilon\,.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\|(\boldsymbol{V}_1\boldsymbol{V}_1^T)^{-1} - (\boldsymbol{V}_2\boldsymbol{V}_2^T)^{-1}\|_F \le \|(\boldsymbol{V}_1\boldsymbol{V}_1^T)^{-1}\|_F\|(\boldsymbol{V}_2\boldsymbol{V}_2^T)^{-1}\|_F\|\boldsymbol{V}_1\boldsymbol{V}_1^T - \boldsymbol{V}_2\boldsymbol{V}_2^T\|_F \\
\le&\sqrt{K}(\lambda_{\min}(\boldsymbol{V}_1\boldsymbol{V}_1^T))^{-1}\sqrt{K}(\lambda_{\min}(\boldsymbol{V}_2\boldsymbol{V}_2^T))^{-1}\|\boldsymbol{V}_1\boldsymbol{V}_1^T - \boldsymbol{V}_2\boldsymbol{V}_2^T\|_F \\
\le&K\left(\lambda_{\min}(\boldsymbol{V}_2\boldsymbol{V}_2^T) - \|\boldsymbol{V}_1\boldsymbol{V}_1^T - \boldsymbol{V}_2\boldsymbol{V}_2^T\|_F\right)^{-1}(\lambda_{\min}(\boldsymbol{V}_2\boldsymbol{V}_2^T))^{-1}\|\boldsymbol{V}_1\boldsymbol{V}_1^T - \boldsymbol{V}_2\boldsymbol{V}_2^T\|_F \\
\le&K(\lambda_{\min}(\boldsymbol{V}_2\boldsymbol{V}_2^T) - K(2+\epsilon)\epsilon)^{-1}(\lambda_{\min}(\boldsymbol{V}_2\boldsymbol{V}_2^T))^{-1}K(2+\epsilon)\epsilon
\end{aligned}$$

□

We are now ready to bound $I_1+I_2+I_3$. For large enough n such that $\max\{C_1, C_2\}\frac{K^{\frac{13}{10}}}{(n\alpha_n)^{\frac{1}{5}}} < \frac{1}{2}$ and $\frac{C_2K^{\frac{9}{5}}}{(n\alpha_n)^{\frac{1}{5}}} < \min\{1, \frac{m_B}{6K}\}$, with probability at least $1 - P_1(n, \alpha_n, K; r) - P_2(n, \alpha_n, K)$, we have:

$$\begin{aligned}
I_1 \le& 2\sqrt{K}\left(1 + \frac{C_1K^{\frac{4}{5}}}{(n\alpha_n)^{\frac{1}{5}}}\right)\left(\sqrt{K} + \frac{C_2K^{\frac{9}{5}}}{(n\alpha_n)^{\frac{1}{5}}}\right) \\
&\cdot K^2\left(m_B - K\left(2 + \frac{C_2K^{\frac{9}{5}}}{(n\alpha_n)^{\frac{1}{5}}}\right)\frac{C_2K^{\frac{9}{5}}}{(n\alpha_n)^{\frac{1}{5}}}\right)^{-1} \cdot m_B^{-1}\left(2 + \frac{C_2K^{\frac{9}{5}}}{(n\alpha_n)^{\frac{1}{5}}}\right)\frac{C_2K^{\frac{9}{5}}}{(n\alpha_n)^{\frac{1}{5}}} \\
\le& 2\sqrt{K}\cdot\frac{3}{2}\cdot\frac{3}{2}\sqrt{K}\cdot K^2\cdot\frac{2}{m_B^2}\cdot\frac{3C_2K^{\frac{9}{5}}}{(n\alpha_n)^{\frac{1}{5}}} = \frac{C_{I_1}K^{\frac{43}{10}}}{(n\alpha_n)^{\frac{1}{5}}} \qquad \text{(B.1.36)}
\end{aligned}$$

where $C_{I_1} := \frac{27C_2}{m_B^2}$, and

$$I_2 \le 2\sqrt{K}\left(1 + \frac{C_1K^{\frac{4}{5}}}{(n\alpha_n)^{\frac{1}{5}}}\right)\cdot\frac{C_2K^{\frac{9}{5}}}{(n\alpha_n)^{\frac{1}{5}}}\frac{\sqrt{K}}{m_B} \le 2\sqrt{K}\cdot\frac{3}{2}\cdot\frac{C_2K^{\frac{9}{5}}}{(n\alpha_n)^{\frac{1}{5}}}\frac{\sqrt{K}}{m_B} = \frac{C_{I_2}K^{\frac{14}{5}}}{(n\alpha_n)^{\frac{1}{5}}} \qquad \text{(B.1.37)}$$

where $C_{I_2} := \frac{3C_2}{m_B}$, and

$$I_3 \le 2\sqrt{K}\cdot\frac{C_1K^{\frac{4}{5}}}{(n\alpha_n)^{\frac{1}{5}}}\cdot\sqrt{K}\cdot\frac{\sqrt{K}}{m_B} = \frac{C_{I_3}K^{\frac{23}{10}}}{(n\alpha_n)^{\frac{1}{5}}}\,. \qquad \text{(B.1.38)}$$

where $C_{I_3} := \frac{2C_1}{m_B}$. Plugging (B.1.36), (B.1.37) and (B.1.38) back to (B.1.34), we have

$$\frac{\|\hat{Z} - Z\|_F}{\sqrt{n}} \leq (C_{I_1} + C_{I_2} + C_{I_3}) \frac{K^{\frac{43}{10}}}{(n\alpha_n)^{\frac{1}{5}}} \leq C_3 (n^{1-\alpha_0}\alpha_n)^{-\frac{1}{5}} \tag{B.1.39}$$

with probability at least $1 - P(n, \alpha_n, K; r)$, where $P(n, \alpha_n, K; r) := P_1(n, \alpha_n, K; r) - P_2(n, \alpha_n, K)$, and $C_3 := (C_{I_1} + C_{I_2} + C_{I_3}) \Big/ \left(\sup_n K^{\frac{43}{10}} n^{-\alpha_0}\right)$. This completes the proof.

□

APPENDIX C

Appendix for "Estimating network edge probabilities by neighborhood smoothing"

C.1 Choosing the constant factor for the bandwidth

First, we need to choose the quantile cut-off parameter h which controls neighborhood selection. Theorem 6 gives the order of h, and the following numerical experiments empirically justify our choice of the constant factor. Figure C.1 shows the mean squared error curves for networks with $n = 2000$ nodes generated from the four graphons in Table 4.1, with the constant factor C varying in the range $\{2^{-3}, 2^{-2}, \ldots, 2^3\}$.

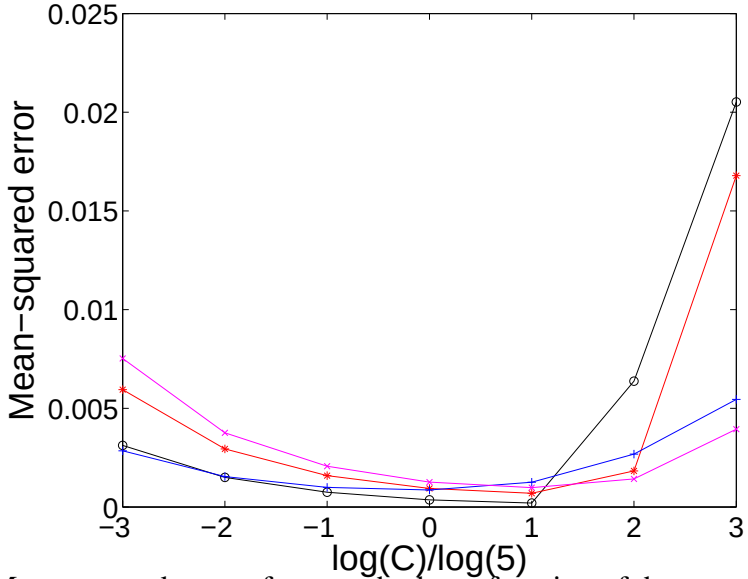


Figure C.1: Mean squared error of our method as a function of the constant C in the tuning parameter $h = C\sqrt{\frac{\log n}{n}}$.

Figure C.1 demonstrates that C in the range from 2^{-2} to 2 works equally well for all these very different graphons. This suggests empirically that the method is robust to the choice of C, and therefore we set $C = 1$ for the rest of the paper.

C.2 Receiver operating characteristic curves for link prediction simulations in Section 4.4

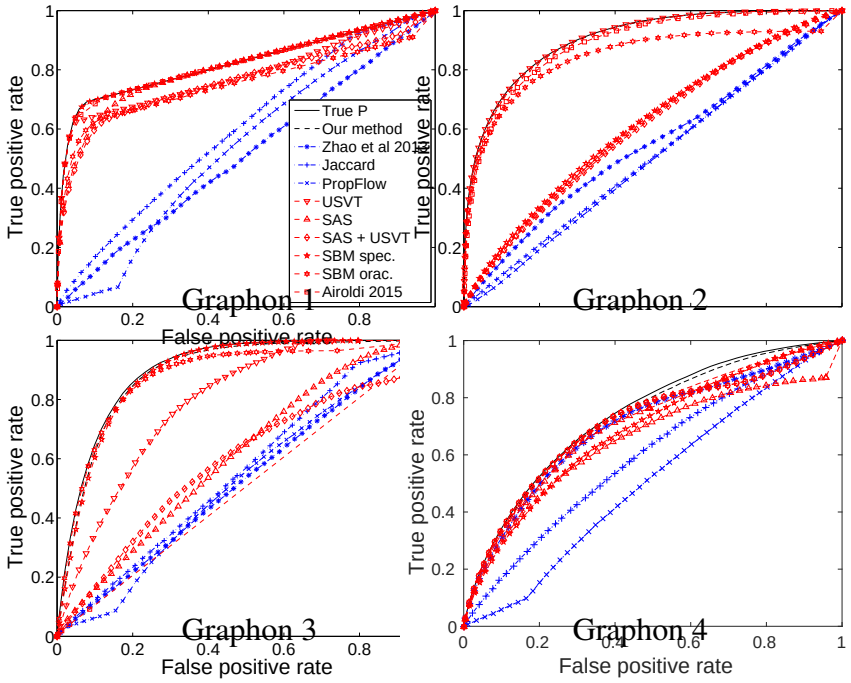


Figure C.2: ROC curves for link prediction of different methods under Graphons 1 to 4.

C.3 Proofs

of Theorem 6. For convenience, we start with summarizing notation and assumptions made in the main paper. Let $0 = x_0 < x_1 < \ldots < x_K = 1$, $I_k := [x_{k-1}, x_k)$ for $1 \le k \le K-1$ and $I_K = [x_{K-1}, X_K]$. Assume the graphon f is a bi-Lipschitz function on each of $I_k \times I_\ell$ for $1 \le k, \ell \le K$. Let L denote the maximum piece-wise bi-Lipschitz constant. Assume the number of pieces K may grow with n, as long as $\min_k |I_k| \big/ \big(\frac{\log n}{n}\big)^{1/2} \to \infty$.

For any $\xi \in [0,1]$, let $I(\xi)$ denote the I_k that contains ξ. Let $S_i(\Delta) = [\xi_i - \Delta, \xi_i + \Delta] \cap I(\xi_i)$ denote the neighborhood of ξ_i in which $f(x,y)$ is Lipschitz in $x \in S_i(\Delta)$ for any fixed y. Finally, recall our estimator is defined by

$$\hat{P}_{ij} = \frac{1}{2}\left(\frac{\sum_{i' \in \mathcal{N}_i} A_{i'j}}{|\mathcal{N}_i|} + \frac{\sum_{j' \in \mathcal{N}_j} A_{ij'}}{|\mathcal{N}_j|}\right)$$

We begin the proof of the main theorem with the following decomposition of the mean

squared error:

$$\frac{1}{n^2}\sum_{ij}(\hat{P}_{ij}-P_{ij})^2=\frac{1}{4n^2}\sum_{ij}\left\{\frac{\sum_{i'\in\mathcal{N}_i}(A_{i'j}-P_{ij})}{|\mathcal{N}_i|}+\frac{\sum_{j'\in\mathcal{N}_j}(A_{ij'}-P_{ij})}{|\mathcal{N}_j|}\right\}^2$$
$$\leq\frac{1}{n^2}\sum_{ij}\left[\frac{1}{2}\left\{\frac{\sum_{i'\in\mathcal{N}_i}((A_{i'j}-P_{i'j})+(P_{i'j}-P_{ij}))}{|\mathcal{N}_i|}\right\}^2\right.$$
$$\left.+\frac{1}{2}\left\{\frac{\sum_{j'\in\mathcal{N}_j}((A_{ij'}-P_{ij'})+(P_{ij'}-P_{ij}))}{|\mathcal{N}_j|}\right\}^2\right] \tag{C.3.1}$$

Next, we show how to bound the first term in (C.3.1); the second term can be handled similarly. Note that

$$\frac{1}{2}\left[\frac{\sum_{i'\in\mathcal{N}_i}\{(A_{i'j}-P_{i'j})+(P_{i'j}-P_{ij})\}}{|\mathcal{N}_i|}\right]^2\leq\left\{\frac{\sum_{i'\in\mathcal{N}_i}(A_{i'j}-P_{i'j})}{|\mathcal{N}_i|}\right\}^2$$
$$+\left\{\frac{\sum_{i'\in\mathcal{N}_i}(P_{i'j}-P_{ij})}{|\mathcal{N}_i|}\right\}^2=J_1(i,j)+J_2(i,j) \tag{C.3.2}$$

Our goal is to bound $\frac{1}{n^2}\sum_{ij}\{J_1(i,j)+J_2(i,j)\}$. First, we prove a lemma which estimates the proportion of nodes in a diminishing neighborhood of ξ_i's.

Lemma 13. *For arbitrary global constants $C_1,\tilde{C}_1>0$, define*

$$\Delta_n=\left(C_1+\sqrt{\tilde{C}_1+4}\right)\left(\frac{\log n}{n}\right)^{1/2}$$

. For n large enough so that $\left\{\frac{(\tilde{C}_1+4)\log n}{n}\right\}^{1/2}\leq 1$ and $\Delta_n<\min_k|I_k|/2$, we have

$$\mathrm{pr}\left\{\min_i\frac{|\{i'\neq i:\xi_{i'}\in S_i(\Delta_n)\}|}{n-1}\geq C_1\left(\frac{\log n}{n}\right)^{1/2}\right\}\geq 1-2n^{-\frac{\tilde{C}_1}{4}}. \tag{C.3.3}$$

of Lemma 13. For any $0<\epsilon\leq 1$ and n large enough to satisfy the assumptions, by Bernstein's inequality we have, for any i,

$$\mathrm{pr}\left(\left|\frac{|\{i'\neq i:\xi_{i'}\in S_i(\Delta_n)\}|}{n-1}-|S_i(\Delta_n)|\right|\geq\epsilon\right)$$
$$\leq 2\exp\left\{-\frac{\frac{1}{2}(n-1)\epsilon^2}{1+\frac{1}{3}\epsilon}\right\}\leq 2\exp\left(-\frac{1}{4}n\epsilon^2\right)$$

Taking a union bound over all i's gives

$$\mathrm{pr}\left(\max_i \left|\frac{|\{i' \neq i : \xi_{i'} \in S_i(\Delta_n)\}|}{n-1} - |S_i(\Delta_n)|\right| \geq \epsilon\right) \leq 2n \exp\left(-\frac{1}{4}n\epsilon^2\right).$$

Letting $\epsilon = \left\{\frac{(\tilde{C}_1+4)\log n}{n}\right\}^{1/2}$, we have

$$\mathrm{pr}\left[\max_i \left|\frac{|\{i' \neq i : \xi_{i'} \in S_i(\Delta_n)\}|}{n-1} - |S_i(\Delta_n)|\right| \geq \left\{\frac{(\tilde{C}_1+4)\log n}{n}\right\}^{1/2}\right] \leq 2n^{-\frac{\tilde{C}_1}{4}}. \tag{C.3.4}$$

Next we claim that either $[\xi_i - \Delta_n, \xi_i] \subseteq I(\xi_i)$ or $[\xi_i, \xi_i + \Delta_n] \subseteq I(\xi_i)$ holds for all i. If for some i the claim does not hold, by the definition of $I(\xi_i)$, we have $I(\xi_i) \subset [\xi_i - \Delta_n, \xi_i + \Delta_n]$. So we have $|I(\xi_i)| \leq 2\Delta_n$, but this contradicts the condition $\Delta_n < \min_k |I_k|/2$. The claim yields that $|S_i(\Delta_n)| \geq \Delta_n$. Finally, by (C.3.4), with probability $1 - 2n^{-\frac{\tilde{C}_1}{4}}$, we have

$$\begin{aligned}\min_i \frac{|\{i' \neq i : \xi_{i'} \in S_i(\Delta_n)\}|}{n-1} &\geq |S_i(\Delta_n)| - \left\{\frac{(\tilde{C}_1+4)\log n}{n}\right\}^{1/2} \\ &\geq \Delta_n - \left\{\frac{(\tilde{C}_1+4)\log n}{n}\right\}^{1/2} \geq C_1\left(\frac{\log n}{n}\right)^{1/2}\end{aligned}$$

This completes the proof of Lemma 13. □

We now continue with the proof of Theorem 6. Recall that we defined a measure of closeness of adjacency matrix slices in Section 4.2 as

$$\tilde{d}(i,i') = \max_{k \neq i,i'} |\langle A_{i\cdot} - A_{i'\cdot}, A_{k\cdot}\rangle| \,/n = \max_{k \neq i,i'} \left|(A^2/n)_{ik} - (A^2/n)_{jk}\right| .$$

The neighborhood $\mathcal{N}_i$ of node i consists of nodes (i')'s with $\tilde{d}(i,i')$ below the h-th quantile of $\{\tilde{d}(i,k)\}_{k\neq i}$. The next lemma shows two key properties of $\mathcal{N}_i$.

Lemma 14. *Suppose that we select the neighborhood $\mathcal{N}_i$ by thresholding at the lower h-th quantile of $\{\tilde{d}(i,k)\}_{k\neq i}$, where we set $h = C_0\big(\frac{\log n}{n}\big)^{1/2}$ with an arbitrary global constant C_0 satisfying $0 < C_0 \leq C_1$ for the C_1 from Lemma 13. Let $C_2, \tilde{C}_2 > 0$ be arbitrary global constants and assume $n \geq 6$ is large enough so that*

(i) All conditions on n in Lemma 13 are satisfied;

(ii) $\left\{\frac{(C_2+2)\log n}{n}\right\}^{1/2} \leq 1$;

(iii) $C_1 (n \log n)^{1/2} \geq 4$*; and*

(iv) $\frac{4}{n} \leq \left\{ \left(C_2 + \tilde{C}_2 + 2 \right)^{1/2} - (C_2 + 2)^{1/2} \right\} \left(\frac{\log n}{n} \right)^{1/2}$.

Then the neighborhood $\mathcal{N}_i$ has the following properties:

1. $|\mathcal{N}_i| \geq C_0 (n \log n)^{1/2}$.
2. *With probability* $1 - 2n^{-\frac{\tilde{C}_1}{4}} - 2n^{-\frac{C_2}{4}}$*, for all* i *and* $i' \in \mathcal{N}_i$*, we have*

$$\|P_{i'\cdot} - P_{i\cdot}\|_2^2 / n \leq \left[6L \left\{ C_1 + \left(\tilde{C}_2 + 4 \right)^{1/2} \right\}^{1/2} + 8 \left(C_2 + \tilde{C}_2 + 2 \right)^{1/2} \right] \left(\frac{\log n}{n} \right)^{1/2}$$

of Lemma 14. The first claim follows immediately from the choice of h and the definition of $\mathcal{N}_i$. To show the second claim, we start with concentration results. For any i, j such that $i \neq j$, we have

$$\begin{aligned} &\left| \left(A^2/n \right)_{ij} - \left(P^2/n \right)_{ij} \right| = \left| \sum_k \left(A_{ik} A_{kj} - P_{ik} P_{kj} \right) \right| \Big/ n \\ \leq &\frac{|\sum_{k \neq i,j} (A_{ik} A_{kj} - P_{ik} P_{kj})|}{n-2} \cdot \frac{n-2}{n} + \frac{|(A_{ii} + A_{jj}) A_{ij}| + |(P_{ii} + P_{jj}) P_{ij}|}{n} \\ \leq &\frac{|\sum_{k \neq i,j} (A_{ik} A_{kj} - P_{ik} P_{kj})|}{n-2} + \frac{4}{n} \end{aligned} \tag{C.3.5}$$

By Bernstein's inequality, for any $0 < \epsilon \leq 1$ and $n \geq 3$ we have

$$\mathrm{pr} \left(\frac{|\sum_{k \neq i,j} (A_{ik} A_{kj} - P_{ik} P_{kj})|}{n-2} \geq \epsilon \right) \leq 2 \exp \left\{ -\frac{\frac{1}{2}(n-2)\epsilon^2}{1 + \frac{1}{3}\epsilon} \right\} \leq 2 \exp \left(-\frac{1}{4} n \epsilon^2 \right).$$

Taking a union bound over all $i \neq j$, we have

$$\mathrm{pr} \left\{ \max_{i,j: i \neq j} \frac{|\sum_{k \neq i,j} (A_{ik} A_{kj} - P_{ik} P_{kj})|}{n-2} \geq \epsilon \right\} \leq 2n^2 \exp \left(-\frac{1}{4} n \epsilon^2 \right).$$

Then setting $\epsilon = \left\{ \frac{(C_2+2) \log n}{n} \right\}^{1/2}$ with n large enough so that $\epsilon \leq 1$, we have

$$\mathrm{pr} \left\{ \max_{i,j: i \neq j} \frac{|\sum_{k \neq i,j} (A_{ik} A_{kj} - P_{ik} P_{kj})|}{n-2} \geq \left\{ \frac{(C_2 + 2) \log n}{n} \right\}^{1/2} \right\} \leq 2n^{-\frac{C_2}{4}} \tag{C.3.6}$$

Combining (C.3.5) and (C.3.6), with probability $1 - 2n^{-\frac{C_2}{4}}$, the following holds

$$\max_{i,j:i\neq j} \left| \left(A^2/n\right)_{ij} - \left(P^2/n\right)_{ij} \right| \leq \left\{ \frac{(C_2+2)\log n}{n} \right\}^{1/2} + \frac{4}{n} \leq \left\{ \frac{(C_2+\tilde{C}_2+2)\log n}{n} \right\}^{1/2} \tag{C.3.7}$$

for n large enoug to satisfy ((iv)).

Next, we prove a useful inequality. For all i and any $\tilde{i}$ such that $\xi_{\tilde{i}} \in S_i(\Delta_n)$, we have

$$\left| \left(P^2/n\right)_{ik} - \left(P^2/n\right)_{\tilde{i}k} \right| = |\langle P_{i\cdot}, P_{k\cdot}\rangle - \langle P_{\tilde{i}\cdot}, P_{k\cdot}\rangle|/n \leq \|P_{i\cdot} - P_{\tilde{i}\cdot}\|_2 \|P_{k\cdot}\|_2/n \leq L\Delta_n \tag{C.3.8}$$

for all k, where the last inequality follows from

$$|P_{i'\ell} - P_{i\ell}| = |f(\xi_{i'}, \xi_\ell) - f(\xi_i, \xi_\ell)| \leq L|\xi_{i'} - \xi_i| \leq L\Delta_n$$

for all ℓ, and $\|P_{k\cdot}\|_2 \leq n^{1/2}$ for all k. Note that this holds for all k, including $k = i$ or $k = \tilde{i}$.

We are now ready to upper bound $\tilde{d}(i,i')$ for $i' \in \mathcal{N}_i$. We bound $\tilde{d}(i,i')$ via bounding $\tilde{d}(i,\tilde{i})$ for $\tilde{i}$ with $\xi_{\tilde{i}} \in S_i(\Delta_n)$. By (C.3.7) and (C.3.8), with probability $1 - 2n^{-\frac{C_2}{4}}$, we have

$$\begin{aligned}
\tilde{d}(i,\tilde{i}) &= \max_{k\neq i,\tilde{i}} |(A^2/n)_{ik} - (A^2/n)_{\tilde{i}k}| \\
&\leq \max_{k\neq i,\tilde{i}} |(P^2/n)_{ik} - (P^2/n)_{\tilde{i}k}| + 2\max_{i,j:i\neq j} |(A^2/n)_{ij} - (P^2/n)_{ij}| \\
&\leq L\Delta_n + 2\left\{ \frac{(C_2+\tilde{C}_2+2)\log n}{n} \right\}^{1/2}
\end{aligned} \tag{C.3.9}$$

Now since the fraction of nodes contained in $\left|\{\tilde{i} : \xi_{\tilde{i}} \in S_i(\Delta_n)\}\right|$ is at least h, this puts an upper bound on $\tilde{d}(i,i')$ for $i' \in \mathcal{N}_i$, since nodes in $\mathcal{N}_i$ have the lowest h fraction of values in $\{\tilde{d}(i,k)\}_k$. Setting Δ_n as in Lemma 13, by Lemma 13 and (C.3.7), with probability $1 - 2n^{-\frac{\tilde{C}_1}{4}} - 2n^{-\frac{C_2}{4}}$, for all i, at least $C_1\left(\frac{\log n}{n}\right)^{1/2}$ fraction of nodes $\tilde{i} \neq i$ satisfy both $\xi_{\tilde{i}} \in S_i(\Delta_n)$ and

$$\tilde{d}(i,\tilde{i}) \leq L\Delta_n + 2\left\{ \frac{(C_2+\tilde{C}_2+2)\log n}{n} \right\}^{1/2}. \tag{C.3.10}$$

Recall that $i' \in \mathcal{N}_i$ have the smallest $h = C_0\left(\frac{\log n}{n}\right)^{1/2} \leq C_1\left(\frac{\log n}{n}\right)^{1/2}$ fraction of $\tilde{d}(i,i')$'s.

Then (C.3.10) yields that

$$\tilde{d}(i,i') \le L\Delta_n + 2\left\{\frac{(C_2+\tilde{C}_2+2)\log n}{n}\right\}^{1/2} \tag{C.3.11}$$

holds for all i and all $i' \in \mathcal{N}_i$ simultaneously with probability $1-2n^{-\frac{\tilde{C}_1}{4}}-2n^{-\frac{C_2}{4}}$.

We are now ready to complete the proof of the second claim of Lemma 14. By Lemma 13, (C.3.7), (C.3.8) and (C.3.11), with probability $1-2n^{-\frac{\tilde{C}_1}{4}}-2n^{-\frac{C_2}{4}}$, the following holds. For n large enough such that $\min_i |\{i' : \xi_{i'} \in S_i(\Delta_n)\}| \ge C_1 (n\log n)^{1/2} \ge 4$ (by Lemma 13), for all i and $i' \in \mathcal{N}_i$ we can find $\tilde{i} \in S_i(\Delta_n)$ and $\tilde{i}' \in S_{i'}(\Delta_n)$ such that i, i', $\tilde{i}$ and $\tilde{i}'$ are different from each other. Then we have

$$\begin{aligned}
&\|P_{i\cdot}-P_{i'\cdot}\|_2^2/n = (P^2/n)_{ii} - (P^2/n)_{i'i} + (P^2/n)_{i'i'} - (P^2/n)_{ii'} \\
\le& \left|(P^2/n)_{ii} - (P^2/n)_{i'i}\right| + \left|(P^2/n)_{i'i'} - (P^2/n)_{ii'}\right| \\
\le& \left|(P^2/n)_{i\tilde{i}} - (P^2/n)_{i'\tilde{i}}\right| + \left|(P^2/n)_{i'\tilde{i}'} - (P^2/n)_{i\tilde{i}'}\right| + 4L\Delta_n \\
\le& \left|(A^2/n)_{i\tilde{i}} - (A^2/n)_{i'\tilde{i}}\right| + \left|(A^2/n)_{i'\tilde{i}'} - (A^2/n)_{i\tilde{i}'}\right| + 4\left\{\frac{(C_2+\tilde{C}_2+2)\log n}{n}\right\}^{1/2} + 4L\Delta_n \\
\le& 2\max_{k\neq i,i'}\left|(A^2/n)_{ik} - (A^2/n)_{i'k}\right| + 4\left\{\frac{(C_2+\tilde{C}_2+2)\log n}{n}\right\}^{1/2} + 4L\Delta_n \\
=& 2\tilde{d}(i,i') + 4\left\{\frac{(C_2+\tilde{C}_2+2)\log n}{n}\right\}^{1/2} + 4L\Delta_n \le 8\left\{\frac{(C_2+\tilde{C}_2+2)\log n}{n}\right\}^{1/2} + 6L\Delta_n \\
=& \left[6L\left\{C_1 + \left(\tilde{C}_2+4\right)^{1/2}\right\}^{1/2} + 8\left(C_2+\tilde{C}_2+2\right)^{1/2}\right]\left(\frac{\log n}{n}\right)^{1/2}
\end{aligned}$$

This completes the proof of Lemma 14. □

We are now ready to bound $\frac{1}{n^2}\sum_{ij}\{J_1(i,j)+J_2(i,j)\}$, which will complete the proof of Theorem 6. Note that we cannot simply bound each individual $J_1(i,j)$'s by Bernstein's inequality since $A_{i'j}$ is not independent of the event $i' \in \mathcal{N}_i$. Instead, we work with the

sum $\frac{1}{n}\sum_j J_1(i,j)$ and decompose it as follows.

$$\begin{aligned}\frac{1}{n}\sum_j J_1(i,j) &= \frac{1}{n|\mathcal{N}_i|^2}\sum_j\left\{\sum_{i'\in\mathcal{N}_i}(A_{i'j}-P_{i'j})\right\}^2\\ =&\frac{1}{n|\mathcal{N}_i|^2}\sum_j\left\{\sum_{i'\in\mathcal{N}_i}(A_{i'j}-P_{i'j})^2+\sum_{i'\in\mathcal{N}_i}\sum_{i''\neq i',i''\in\mathcal{N}_i}(A_{i'j}-P_{i'j})(A_{i''j}-P_{i''j})\right\}.\end{aligned} \tag{C.3.12}$$

The first term in (C.3.12) satisfies

$$\sum_j(A_{i'j}-P_{i'j})^2/n = \|A_{i'\cdot}-P_{i'\cdot}\|_2^2/n \le 1 \tag{C.3.13}$$

where the inequality is due to $|A_{i'j}-P_{i'j}|\le 1$ for all j. The second term in (C.3.12) can be bounded by

$$\begin{aligned}&\frac{1}{n|\mathcal{N}_i|^2}\sum_j\sum_{i'\in\mathcal{N}_i}\sum_{i''\neq i',i''\in\mathcal{N}_i}(A_{i'j}-P_{i'j})(A_{i''j}-P_{i''j})\le\\ &\quad\le\frac{1}{|\mathcal{N}_i|^2}\sum_{i',i''\in\mathcal{N}_i:i'\neq i''}\left|\frac{1}{n}\sum_j(A_{i'j}-P_{i'j})(A_{i''j}-P_{i''j})\right|\\ &\quad\le\frac{1}{|\mathcal{N}_i|^2}\sum_{i',i''\in\mathcal{N}_i:i'\neq i''}\left\{\frac{1}{n-2}\left|\sum_{j\neq i',i''}(A_{i'j}-P_{i'j})(A_{i''j}-P_{i''j})\right|\cdot\frac{n-2}{n}\right.\\ &\qquad\left.+\frac{|(A_{i'i''}-P_{i'i''})|\,|(A_{i'i'}-P_{i'i'}+A_{i''i''}-P_{i''i''})|}{n}\right\}\\ &\quad\le\frac{1}{|\mathcal{N}_i|^2}\sum_{i',i''\in\mathcal{N}_i:i'\neq i''}\left\{\frac{1}{n-2}\left|\sum_{j\neq i',i''}(A_{i'j}-P_{i'j})(A_{i''j}-P_{i''j})\right|+\frac{2}{n}\right\}.\end{aligned} \tag{C.3.14}$$

To bound the first term in (C.3.14), for any $i_1\neq i_2$ and $\epsilon>0$, by Bernstein's inequality we have

$$\mathrm{pr}\left\{\frac{1}{n-2}\left|\sum_{j\neq i_1,i_2}(A_{i_1j}-P_{i_1j})\,(A_{i_2j}-P_{i_2j})\right|\ge\epsilon\right\}\le 2\exp\left\{-\frac{\frac{1}{2}(n-2)\epsilon^2}{1+\frac{1}{3}\epsilon}\right\}\le 2n^2e^{-\frac{n\epsilon^2}{4}}.$$

Let $C_3,\tilde{C}_3>0$ be arbitrary global constants and let n be large enough so that $\frac{1}{C_0(n\log n)^{1/2}}+\frac{2}{n}\le\left\{\left(C_3+\tilde{C}_3+8\right)^{1/2}-(C_3+8)^{1/2}\right\}\left(\frac{\log n}{n}\right)^{1/2}$. First, taking $\epsilon=\left\{\frac{(C_3+8)\log n}{n}\right\}^{1/2}$ and

a union bound over all $i_1 \neq i_2$, we have

$$\mathrm{pr}\left[\max_{i_1,i_2,i_1\neq i_2} \frac{1}{n-2}\left|\sum_{j\neq i_1,i_2}\left(A_{i_1 j}-P_{i_1 j}\right)\left(A_{i_2 j}-P_{i_2 j}\right)\right| \geq \left\{\frac{(C_3+8)\log n}{n}\right\}^{1/2}\right] \leq 2n^{-\frac{C_3}{4}}. \tag{C.3.15}$$

Then plugging (C.3.13), (C.3.14) and (C.3.15) into (C.3.12) and combining with claim 1 of Lemma 14, with probability $1-2n^{-\frac{\tilde{C}_1}{4}}-2n^{\frac{C_2}{4}}-2n^{-\frac{C_3}{4}}$, for all i simultaneously, we have

$$\begin{aligned}
&\frac{1}{n}\sum_j J_1(i,j) \leq \frac{1}{|\mathcal{N}_i|^2}\sum_{i'\in\mathcal{N}_i}\left[1+(|\mathcal{N}_i|-1)\left(\left\{\frac{(C_3+8)\log n}{n}\right\}^{1/2}+\frac{2}{n}\right)\right] \\
\leq &\frac{1}{|\mathcal{N}_i|}+\left\{\frac{(8+C_3)\log n}{n}\right\}^{1/2}+\frac{2}{n} \leq \frac{1}{C_0\left(n\log n\right)^{1/2}}+\frac{2}{n}+\left\{\frac{(C_3+8)\log n}{n}\right\}^{1/2} \\
\leq &\left\{\frac{(C_3+\tilde{C}_3+8)\log n}{n}\right\}^{1/2}.
\end{aligned} \tag{C.3.16}$$

We now bound $\frac{1}{n^2}\sum_{ij} J_2(i,j)$. By Lemma 14, with probability $1-2n^{-\frac{\tilde{C}_1}{4}}-2n^{\frac{C_2}{4}}$, we have

$$\begin{aligned}
&\frac{1}{n^2}\sum_{ij} J_2(i,j) = \frac{1}{n}\sum_i\left\{\frac{1}{n}\sum_j J_2(i,j)\right\} = \frac{1}{n}\sum_i\left\{\frac{1}{n}\sum_j\left(\frac{\sum_{i'\in\mathcal{N}_i}(P_{i'j}-P_{ij})}{|\mathcal{N}_i|}\right)^2\right\} \\
\leq &\frac{1}{n}\sum_i\left\{\frac{\sum_{i'\in\mathcal{N}_i}\sum_j(P_{i'j}-P_{ij})^2/n}{|\mathcal{N}_i|}\right\} = \frac{1}{n}\sum_i\left\{\frac{\sum_{i'\in\mathcal{N}_i}\|P_{i'\cdot}-P_{i\cdot}\|_2^2/n}{|\mathcal{N}_i|}\right\} \\
\leq &\left[6L\left\{C_1+\left(\tilde{C}_2+4\right)^{1/2}\right\}^{1/2}+8\left(C_2+\tilde{C}_2+2\right)^{1/2}\right]\left(\frac{\log n}{n}\right)^{1/2},
\end{aligned} \tag{C.3.17}$$

where the first inequality is the Cauchy-Schwartz inequality and the second inequality follows from claim 2 of Lemma 14.

Combining (C.3.16) and (C.3.17) completes the proof of Theorem 6.

□

BIBLIOGRAPHY

Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *J. Machine Learning Research*, 9:1981–2014.

Airoldi, E. M., Costa, T. B., and Chan, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700.

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598.

Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122.

Ball, B., Karrer, B., and Newman, M. E. J. (2011). An efficient and principled method for detecting communities in networks. *Physical Review E*, 34:036103.

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

Bickel, P. J. and Sarkar, P. (2015). Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Binkiewicz, N., Vogelstein, J. T., and Rohe, K. (2014). Covariate assisted spectral clustering. *arXiv preprint arXiv:1411.2158*.

Bolouri, H. and Davidson, E. H. (2010). The gene regulatory network basis of the community effect, and analysis of a sea urchin embryo example. *Developmental biology*, 340(2):170–178.

Cai, T. T. and Li, X. (2015). Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *The Annals of Statistics*, 43(3):1027–1059.

Cai, T. T., Liang, T., and Rakhlin, A. (2015). Computational and statistical boundaries for submatrix localization in a large noisy matrix. *arXiv preprint arXiv:1502.01988*.

Chamberlain, B. L. (1998). Graph partitioning algorithms for distributing workloads of parallel computations. *University of Washington Technical Report UW-CSE-98-10*, 3.

Chan, S. H. and Airoldi, E. M. (2014). A consistent histogram estimator for exchangeable graph models. *arXiv preprint arXiv:1402.1888*.

Chatterjee, S. (2014). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214.

Chaudhuri, K., Chung, F., and Tsiatas, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research*, 2012:1–23.

Cheng, H., Zhou, Y., and Yu, J. X. (2011). Clustering large attributed graphs: A balance between structural and attribute similarities. *ACM Trans. Knowl. Discov. Data*, 5(2):12:1–12:33.

Choi, D. (2015). Co-clustering of nonsmooth graphons. *arXiv preprint arXiv:1507.06352*.

Choi, D. and Wolfe, P. J. (2014). Co-clustering separately exchangeable network data. *The Annals of Statistics*, 42(1):29–63.

Choi, D. S., Wolfe, P. J., and Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika*, page asr053.

De Nooy, W., Mrvar, A., and Batagelj, V. (2011). *Exploratory social network analysis with Pajek*, volume 27. Cambridge University Press.

Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106.

Diaconis, P. and Janson, S. (2007). Graph limits and exchangeable random graphs. *arXiv preprint arXiv:0712.2749*.

Durante, D., Yang, X., and Dunson, D. B. (2015). Locally adaptive dynamic networks. *arXiv preprint arXiv:1505.05668*.

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

Gao, C., Lu, Y., and Zhou, H. H. (2014). Rate-optimal graphon estimation. *arXiv preprint arXiv:1410.5837*.

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2015). Achieving optimal misclassification proportion in stochastic block model. *arXiv preprint arXiv:1505.03772*.

Gillis, N. and Vavasis, S. (2014). Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(4):698–714.

Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Comput. Oper. Res.*, 13(5):533–549.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233.

Gregory, S. (2010). Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 12(10):103018.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354.

Haussler, D. and Welzl, E. (1986). Epsilon-nets and simplex range queries. In *Proceedings of the Second Annual Symposium on Computational Geometry*, SCG '86, pages 61–71, New York, NY, USA. ACM.

Hendrickson, B. and Kolda, T. G. (2000). Graph partitioning models for parallel computing. *Parallel computing*, 26(12):1519–1534.

Hoang, T.-A. and Lim, E.-P. (2014). On joint modeling of topical communities and personal interest in microblogs. In *Social Informatics*, pages 1–16. Springer.

Hoff, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems 20*, pages 657–664.

Hoff, P. D. (2003). Random effects models for network data. In *Dynamic social network modeling and analysis: Workshop summary and papers*, pages 303–312. National Academies Press Washington, DC.

Hoff, P. D. (2007). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, Cambridge, MA.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association*, 76(373):33–50.

Hoover, D. N. (1979). Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 2.

Hummon, N. P., Doreian, P., and Freeman, L. C. (1990). Analyzing the structure of the centrality-productivity literature created between 1948 and 1979. *Science Communication*, 11(4):459–480.

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

Joseph, A. and Yu, B. (2013). Impact of regularization on spectral clustering. *arXiv preprint arXiv:1312.1733*.

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

Klopp, O., Tsybakov, A. B., and Verzelen, N. (2015). Oracle inequalities for network models and sparse graphon estimation. *arXiv preprint arXiv:1507.04118*.

Lancichinetti, A., Fortunato, S., and Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015.

Lancichinetti, A., Radicchi, F., and Ramasco, J. J. (2010). Statistical significance of communities in networks. *Phys. Rev. E*, 81(4):046110.

Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S. (2011). Finding statistically significant communities in networks. *PloS one*, 6(4):e18961.

Latouche, P., Birmelé, E., and Ambroise, C. (2009). Overlapping stochastic block models. *arXiv preprint arXiv:0910.2098*.

Lazega, E. (2001). *The collegial phenomenon: the social mechanisms of co-operation among peers in a corporate law partnership*. Oxford University Press.

Le, C. M., Levina, E., and Vershynin, R. (2014). Optimization via low-rank approximation for community detection in networks. *arXiv preprint arXiv:1406.0067*.

Lei, J. (2014). A goodness-of-fit test for stochastic block models. *arXiv preprint arXiv:1412.4857*.

Lei, J. and Rinaldo, A. (2013). Consistency of spectral clustering in sparse stochastic block models. *arXiv preprint arXiv:1312.2050*.

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

Lichtenwalter, R. N., Lussier, J. T., and Chawla, N. V. (2010). New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252.

Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., and Dawson, S. M. (2003). The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405.

M. Kim, J. L. (2012). Latent multi-group membership graph model. *International Conference on Machine Learning*.

Ma, Z. and Wu, Y. (2015). Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116.

McAuley, J. and Leskovec, J. (2012). Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems 25*, pages 548–556.

Mossel, E., Neeman, J., and Sly, A. (2014). Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*.

Newman, M. and Clauset, A. (2015). Structure and inference in annotated networks. *arXiv preprint arXiv:1507.04001*.

Newman, M. and Peixoto, T. P. (2015). Generalized communities in networks. *Physical review letters*, 115(8):088701.

Newman, M. E. and Girvan, M. (2004a). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103(23):8577–8582.

Newman, M. E. J. (2013). Spectral methods for network community detection and graph partitioning. *Physical Review E*, 88:042822.

Newman, M. E. J. and Girvan, M. (2004b). Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113.

Nickel, C. L. M. (2007). *Random dot product graphs: A model for social networks*. PhD thesis, Johns Hopkins University.

Olhede, S. C. and Wolfe, P. J. (2014). Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, 111(41):14722–14727.

Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818.

Perry, P. O. and Wolfe, P. J. (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B*, 75(5):821–849.

Pizzuti, C. (2009). Overlapped community detection in complex networks. In *GECCO*, volume 9, pages 859–866.

Pollard, D. et al. (1981). Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1):135–140.

Psorakis, I., Roberts, S., Ebden, M., and Sheldon, B. (2011). Overlapping community detection using Bayesian non-negative matrix factorization. *Physical Review E*, 83(6):066114.

Qin, T. and Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128.

Resnick, M. D., Bearman, P. S., Blum, R. W., Bauman, K. E., Harris, K. M., Jones, J., Tabor, J., Beuhring, T., Sieving, R. E., and Shew, M. (1997). Protecting adolescents from harm: findings from the National Longitudinal Study on Adolescent Health. *JAMA*, 278(10):823–832.

Rogers, E. M. and Kincaid, D. L. (1981). *Communication networks: Toward a new paradigm for research*. Free Press New York.

Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915.

Saade, A., Krzakala, F., and Zdeborová, L. (2014). Spectral clustering of graphs with the Bethe Hessian. In *Advances in Neural Information Processing Systems*, pages 406–414.

Saldana, D. F., Yu, Y., and Feng, Y. (2014). How many communities are there? arXiv:1412.1684.

Sarkar, P. and Bickel, P. J. (2013). Role of normalization in spectral clustering for stochastic blockmodels. *arXiv preprint arXiv:1310.1495*.

Sarkar, P., Chakrabarti, D., and Jordan, M. (2012). Nonparametric link prediction in dynamic networks. *arXiv preprint arXiv:1206.6394*.

Schlitt, T. and Brazma, A. (2007). Current approaches to gene regulatory network modelling. *BMC bioinformatics*, 8(Suppl 6):S9.

Sewell, D. K. and Chen, Y. (2014). Latent space models for dynamic networks. *Journal of the American Statistical Association*, page to appear.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Silva, A., Meira, Jr., W., and Zaki, M. J. (2012). Mining attribute-structure correlated patterns in large attributed graphs. *Proc. VLDB Endow.*, 5(5):466–477.

Smith, L. M., Zhu, L., Lerman, K., and Percus, A. G. (2014). Partitioning networks with node attributes by compressing information flow. *arXiv preprint arXiv:1405.4332*.

Snijders, T. A., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological methodology*, 36(1):99–153.

Steglich, C., Snijders, T. A. B., and West, P. (2006). Applying SIENA: An illustrative analysis of the coevolution of adolescents' friendship networks, taste in music, and alcohol consumption. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 2(1):48.

Tang, M., Sussman, D. L., Priebe, C. E., et al. (2013). Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41(3):1406–1430.

Viennet, E. (2012). Community detection based on structural and attribute similarities. In *ICDS 2012, The Sixth International Conference on Digital Society*, pages 7–12.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Wang, F., Li, T., Wang, X., Zhu, S., and Ding, C. (2011). Community discovery using nonnegative matrix factorization. *Data Mining and Knowledge Discovery*, 22(3):493–521.

Wolfe, P. J. and Olhede, S. C. (2013). Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*.

Xie, J., Kelley, S., and Szymanski, B. K. (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (CSUR)*, 45(4):43.

Xu, Z., Ke, Y., Wang, Y., Cheng, H., and Cheng, J. (2012). A model-based approach to attributed graph clustering. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 505–516. ACM.

Yang, J., McAuley, J., and Leskovec, J. (2013). Community detection in networks with node attributes. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 1151–1156. IEEE.

Yang, J. J., Han, Q., and Airoldi, E. M. (2014). Nonparametric estimation and testing of exchangeable graph models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33, pages 1060–1067.

Yang, T., Jin, R., Chi, Y., and Zhu, S. (2009). Combining link and content for community detection: a discriminative approach. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 927–936. ACM.

Young, S. J. and Scheinerman, E. R. (2007). Random dot product graph models for social networks. In *Algorithms and models for the web-graph*, pages 138–149. Springer.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473.

Zanghi, H., Volant, S., and Ambroise, C. (2010). Clustering based on random graph model embedding vertex features. *Pattern Recogn. Lett.*, 31(9):830–836.

Zhang, A. (2009). *Protein interaction networks: computational analysis*. Cambridge University Press.

Zhang, Y., Levina, E., and Zhu, J. (2014). Detecting overlapping communities in networks with spectral methods. *arXiv preprint arXiv:1412.3432*.

Zhang, Y., Levina, E., and Zhu, J. (2015a). Community detection in networks with node features. *arXiv preprint arXiv:1509.01173*.

Zhang, Y., Levina, E., and Zhu, J. (2015b). Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv:1509.08588*.

Zhao, Y., Levina, E., and Zhu, J. (2011). Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18):7321–7326.

Zhao, Y., Levina, E., and Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292.