BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

# BAYESIAN STOCHASTIC BLOCKMODELS FOR COMMUNITY DETECTION IN NETWORKS AND COMMUNITY-STRUCTURED COVARIANCE SELECTION

by

**LIJUN PENG**

Master of Science, Boston University, 2012

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2015

Approved by

First Reader

Luis Carvalho, PhD
Professor of Mathematics & Statistics

Second Reader

Eric Kolaczyk, PhD
Professor of Mathematics & Statistics

Third Reader

Dino Christenson, PhD
Professor of Political Science

## Acknowledgments

I would like to express my special appreciation and thanks to my advisor Professor Luis Carvalho. He has been a tremendous mentor for me in research and a great friend for me in life. I would like to thank him for inspiring my research, offering me valuable guidance and encouraging me in the face of difficulties. The very first class I attended in my PhD study was taught by Professor Luis Carvalho. He opened the door of curiosity in statistics and made doing research so enjoyable for me. I have been influenced greatly by his passion in exploring possibilities and devotion to scientific research.

I would also like to thank my committee members, Professor Dino Christenson, Professor Uri Eden, Professor Eric Kolaczyk and Professor Mark Kramer for serving as my committee members and offering me valuable suggestions. I obtained statistical knowledge on networks which is of great help in my dissertation writing from the class taught by Professor Eric Kolaczyk. I also want to thank him for offering helpful suggestions in submitting work on my first research project. I am grateful to Professor Dino Christenson from the Department of Political Science. He encouraged inter-discipline research and taught me to be an open-minded collaborator.

I also want to thank Professor Mark Kramer for the data he offered and for the discussions we had on covariance selection. My thanks also go to my friends— Han Han, Ian Johnston and Yaonan Zhang—who provided me with helpful discussions and suggestions.

Special thanks go to my family. Words cannot express how grateful I am to them for their unlimited love and support that encouraged me to strive toward my goal.

# BAYESIAN STOCHASTIC BLOCKMODELS FOR COMMUNITY DETECTION IN NETWORKS AND COMMUNITY-STRUCTURED COVARIANCE SELECTION

(Order No. )

**LIJUN PENG**

Boston University, Graduate School of Arts and Sciences, 2015

Major Professor: Luis Carvalho, PhD, Professor of Mathematics & Statistics

## ABSTRACT

Networks have been widely used to describe interactions among objects in diverse fields. Given the interest in explaining a network by its structure, much attention has been drawn to finding clusters of nodes with dense connections within clusters but sparse connections between clusters. Such clusters are called *communities*, and identifying such clusters is known as *community detection*. Here, to perform community detection, I focus on *stochastic blockmodels* (SBM), a class of statistically-based generative models. I present a flexible SBM that represents different types of data as well as node attributes under a Bayesian framework. The proposed models explicitly capture community behavior by guaranteeing that connections are denser within communities than between communities.

First, I present a degree-corrected SBM based on a logistic regression formulation to model binary networks. To fit the model, I obtain posterior samples via Gibbs sampling based on Pólya-Gamma latent variables. I conduct inference based on a novel, canonically mapped centroid estimator that formally addresses label non-identifiability and captures representative community assignments. Next, to accommodate large-scale datasets, I further extend the degree-corrected SBM to a broader family of generalized linear models with group correction terms. To conduct exact inference efficiently, I develop an iteratively-reweighted least squares procedure that implicitly updates sufficient statistics on the network to obtain *maximum a posteriori* (MAP) estimators. I demonstrate the proposed

model and estimation on simulated benchmark networks and various real-world datasets.

Finally, I develop a Bayesian SBM for community-structured *covariance selection*. Here, I assume that the data at each node are Gaussian and a latent network where two nodes are not connected if their observations are conditionally independent given observations of other nodes. Under the context of biological and social applications, I expect that this latent network shows a block dependency structure that represents community behavior. Thus, to identify the latent network and detect communities, I propose a hierarchical prior in two levels: a spike-and-slab prior on off-diagonal entries of the concentration matrix for variable selection and a degree-corrected SBM to capture community behavior. I develop an efficient routine based on ridge regularization and MAP estimation to conduct inference.

# Contents

# List of Tables

# List of Figures

# List of Abbrevs

| | |
|---|---|
| CP .. | Core-Periphery |
| FG .. | Fast-Greedy |
| KN . | Karrer & Newman |
| LP .. | Label Propagation |
| MAP | Maximum A Posteriori |
| ML . | Multi-Level |
| MST | Maximum Spanning Tree |
| NMI | Normalized Mutual Information |
| PPL | Posterior Predictive Loss |
| SBM | Stochastic Blockmodel |
| WT . | Walktrap |
| ZIP . | Zero-Inflated Poisson |

# Chapter 1

# Introduction

Networks can be used to describe interactions among objects in diverse fields such as physics (Newman, 2006), biology (Hancock et al., 2010), and especially social sciences (Zachary, 1977; Adamic and Glance, 2005). In network theory, objects are represented by *nodes* and their interactions by *edges*. There are many ways to define *communities* in a network, with different connectivity properties or according to biological/social functions of nodes. In this dissertation, I follow the standard definition and consider clusters of nodes that share many edges between them but that, in contrast, do not interact often with nodes in other clusters as communities.

## 1.1 Community Detection

This characterization follows a traditional approach in social sciences that aims at discerning the structure of a network according to relationship patterns among "actors", e.g. friendship or collaboration. These interaction patterns may reflect "assortativity", a concept that originated in the ecological and epidemiological literature (Albert and Barabási, 2002): it refers to the tendency of nodes to associate with other similar nodes in a network. Among measures of similarity, the degree of a node is of great interest in the study of assortativity in networks (Newman, 2002, 2003; Vázquez, 2003), that is, degree assortative networks usually show a preference for high-degree nodes to connect with other high-degree nodes. We expect in some applications that actors exercise assortativity and prefer to group themselves according to similarity or kinship in communities, and so communities are *dense* in within-

group associations but *sparse* in between-group associations. Community detection has sparked great interest in many fields where recent applications aim at characterizing the structure of a network by detecting its communities.

### 1.1.1 Prior Work

There is a large body of literature in community detection, given its significance and interest. Traditional methods include graph partitioning (Kernighan and Lin, 1970; Barnes, 1982), hierarchical clustering (Hastie et al., 2001), and spectral clustering (Donath and Hoffman, 1973; Von Luxburg, 2007; Rohe et al., 2011); while these methods are heuristic and thus suitable for large networks, they do not address directly community detection but aim instead at partitioning the network according to edge densities between groups and thus identifying connection "bottlenecks".

The concept of *modularity* better captures community structure by also taking within-group edge densities into account (Newman and Girvan, 2004; Newman, 2006). Optimization methods based on modularity can then be used to detect communities, but since modularity optimization is NP-complete (Brandes et al., 2007), interest lies mostly in approximated methods such as the greedy method (Newman, 2004) and extremal optimization (Duch and Arenas, 2005; Bickel and Chen, 2009). However, there are still drawbacks: methods based on modularity may fail in detecting small communities and thus exhibit a "resolution limit" (Fortunato and Barthelemy, 2007).

Latent space network models (Hoff et al., 2002), latent variable models (Hoff et al., 2005), and latent position cluster models (Handcock et al., 2007) assume that the probability of an interaction depends on node-specific latent factors such as the distance between two nodes in an unobserved continuous "social space"; these models are generalizations of exponential random graph models [ERGMs; see (Robins et al., 2007)] where community structure is assumed from cluster structure in the latent space. Krivitsky (2012) generalized ERGMs to valued networks whose ties are counts. A more thorough review on the related methods to network community detection can be found in the paper by Parthasarathy

et al. (2011).

### 1.1.2 Stochastic Blockmodels

There are many other methods to mention as in the review by Parthasarathy et al. (2011), but we focus on *parametric statistical* approaches where inference on community structure is based on an assumed model of association. The motivation is that since there are many possible community configurations, that is, assignment of actors to communities, we want to not only infer communities, but also assess how likely each configuration is according to the model.

The first endeavors in such parametric models—albeit not in community detection—are the $p_1$ exponential family models due to Holland and Leinhardt (1981). These models follow a log-linear formulation (Fienberg and Wasserman, 1981) with parameters that are related to in- and out-degrees and edge densities. A common modeling choice is to treat actors as behaving similarly given their respective communities. This structural equivalence assumption is at the core of *blockmodels* (Lorrain and White, 1971). Later, these models were extended to incorporate actor and group parameters (Fienberg et al., 1985; Tallberg, 2005; Daudin et al., 2008). Wang and Wong (1987) further adapted the models to consider a block structure through *stochastic blockmodels* [SBMs (Holland et al., 1983; Anderson et al., 1992)], yielding $p_1$ blockmodels. Zanghi et al. (2010), Mariadassou et al. (2010) and Vu et al. (2013) proposed scalable approximate variational approaches based on modified version of those $p_1$ (block)models.

SBMs explore a simpler model structure where the probability of an association between two actors depends on the communities to which they belong, that is, two actors within the same group are stochastically equivalent (Batagelj et al., 2005). Karrer and Newman (2011) developed an SBM with *degree-correction*, where the degree distribution of nodes within each community can be heterogeneous. Celisse et al. (2012), Choi et al. (2012) and Bickel et al. (2013) addressed the asymptotic inference in SBM using maximum likelihood and variational approaches. More flexible SBM were obtained by adopting a hierarchical

Bayesian setup that regards probabilities of association as random and group membership as latent variables (Snijders and Nowicki, 1997; Nowicki and Snijders, 2001; Hofman and Wiggins, 2008). As in all latent mixture models, label non-identifiability is a known problem since multiple label assignments yield the same partition into communities; ultimately, we only care if two actors are in the same community or in different communities. It is also possible to incorporate node attributes in the model (Kim and Leskovec, 2011; Fosdick and Hoff, 2013) and to allow actors to belong to more than one community (Airoldi et al., 2008).

## 1.2 Covariance Selection

In the community detection problem stated above the network is given as data, which is not always the case in practice. Under this circumstance, observations per node instead of connections between nodes are collected. The lack of information in connections leads to interest in identifying edges in a network via the covariance structure of the observed data. Among such models the most widely used one is Gaussian graphical model, especially when the observations are normally distributed.

### 1.2.1 Background in Gaussian Graphical Models

Gaussian graphical models (Dempster, 1972) have been widely used to describe conditional independence between components of a random vector (Whittaker, 2009). In a Gaussian graphical model, we associate to each component $X_i$ of a Gaussian random vector $X$ a node in a graph $G = (V, E)$, and two nodes $i$ and $j$ are *not* connected in $G$ if and only if their corresponding components are conditionally independent given all the other components, that is, if the partial correlation between $X_i$ and $X_j$ is zero. In other words, $(i, j) \notin E$ if and only if $\rho_{X_i, X_j \,|\, X_{V \setminus \{i,j\}}} = 0$. Equivalently, if $C$ is the *concentration* matrix of $X$, that is, the inverse variance of $X$, then, since the partial correlation $\rho_{X_i, X_j \,|\, X_{V \setminus \{i,j\}}} \propto C_{ij}$, $(i, j) \notin E$ if and only if $C_{ij} = 0$. Thus, inferring conditional in-

dependence is equivalent to estimating null entries in a concentration matrix (Lauritzen, 1996).

### 1.2.2 Prior Work

There are several difficulties in identifying the conditional independence of Gaussian variables. The first challenge is the "large $p$, small $n$" regimen that stems from inferring a large concentration matrix based on relatively few observations. Under this regimen, the sample covariance matrix is singular and thus cannot be inverted to directly compute the concentration matrix. A commonly used alternative is to include some form of regularization, such as graph lasso (Dempster, 1972; Meinshausen and Bühlmann, 2006; Friedman et al., 2008). The second challenge is developing an effective procedure to identify the corresponding network based on the estimated concentration matrix. Traditional variable selection approaches such as stepwise regression and those especially adapted to Gaussian graphical models (Drton and Perlman, 2004) offer a way to check the significance of estimated partial correlations. However, determining edges of networks and inferring concentration matrices are performed separately. A potential drawback of these methods is that once a "bad" model is selected, the estimation of concentration matrices is unreliable as a result. To remedy this problem, efficient approaches to jointly perform model selection and graph estimation have been proposed (Yuan and Lin, 2007; Yuan, 2008). However, none of the methods mentioned above has taken community structure into account when estimating the conditional independence.

## 1.3 Challenges

Despite the increasing interest and study in characterizing the network structure, community detection is arguably still an open problem in network analysis.

(1) The first challenge in developing a useful approach to community detection is the flexibility in modeling node attributes, for example "degrees", and edge attributes,

such as "counts on edges". Additional information in nodes and edges leads to more reliable estimation (see the example in Section 3.1).

(2) Even though the models reviewed above are flexible enough to identify social block structure, they might fail to actually recognize communities. Degree-correction is not enough to accurately characterize assortative community behavior (see the examples in Section 2.6.1).

(3) While many model-based solutions, such as degree-corrected SBMs, are available for relatively small networks, these approaches become computationally intractable for large-scale networks. Besides, those methods lack the flexibility in efficiently modeling node attributes.

(4) Commonly used heuristic methods mentioned in Section 1.1.1 are feasible for large-scale networks. However, they do not address directly community detection but aim instead at partitioning the network according to edge densities between groups and thus fail to consider the within-group interactions. As a result, they may suffer from limited ability to detect small communities and thus exhibit a "resolution limit".

## 1.4 Contributions and Organizations of the Dissertation

This dissertation is intended to contribute to identifying a network and detecting its community structure. More specifically, I make the following contributions:

(1) I present a Bayesian SBM that explicitly captures the assortative community behavior via informative prior specification and displays flexibility in modeling different types of data and node attributes.

(2) I develop computational methods to make inference under the proposed Bayesian SBM from both sampling and optimization prospectives.

- I propose a novel centroid estimator that accounts for label non-identifiability issues via Gibbs sampling based on a data augmentation strategy.

- I propose a MAP estimator based on an iteratively-reweighted least squares procedure and an active set method.

(3) To make the inference more efficient on large-scale networks, I adopt the following two strategies:

- I develop a graph generalized linear model (GGLM) procedure tailored for graphs. GGLM implicitly computes sufficient statistics rather than generating responses and design matrices.
- I introduce group corrections that stem from degree assortativity and centrality. In particular, group corrections can be interpreted to represent different assortative structure such as core-periphery structure. Moreover, group corrections reduce to degree corrections if each node is considered as a group by itself.

(4) I propose a latent formulation to covariance selection and a Bayesian approach that jointly identifies the underlying network and detects its community structure. To identify the latent network, I develop a Bayesian ridge-regularized covariance selection that specifies a spike-and-slab prior.

The organization of the remainder of this dissertation is as follows.

In Chapter 2, I present a classical degree-corrected SBM and estimation adapted to Bayesian point of view. The proposed Bayesian degree-corrected model is based on a logistic regression formulation with degree correction terms, and explicitly captures the community behavior via prior specification. I further adopt a data augmentation strategy with latent Pólya-Gamma variables to obtain posterior samples. In the estimation aspect, I propose a principled, canonically mapped centroid estimator that formally addresses label non-identifiability and captures representative community assignments. I demonstrate the proposed model and estimation on real-world as well as simulated benchmark networks.

In Chapter 3, I discuss more general models that can be applied to large-scale data of various types. I generalize the SBM in Chapter 2 by (i) considering a broader fam-

ily of generalized regression models instead of the logistic regression; (ii) making corrections for groups of nodes instead of each individual node; (iii) modeling within-community interactions rather than between-community interactions and (iv) adopting optimization approaches instead of sampling approaches to make inference. I demonstrate the group-corrected SBM on an amicus curiae (literally "friend of the court") network of count data. In addition, I show its flexibility in modeling large-scale networks through both simulation study and case study based on real-world online social networks.

In Chapter 4, I propose a Bayesian approach that jointly identifies the underlying network and detects its community structure. To identify the latent network, I develop a Bayesian ridge-regularized covariance selection that specifies a spike-and-slab prior. I offer a Bayesian approach for community detection that explicitly characterizes community behavior and a *maximum a posteriori* (MAP) estimator. I compare our ridge-regularized covariance selection to other commonly used methods on simulated benchmark networks and apply it to a real-world meta-genomic dataset of complex microbial biofilms.

# Chapter 2

# Bayesian Degree-Corrected SBM

## 2.1 Introduction

To tackle community detection stated in Section 1.1, we adopt a hierarchical Bayesian SBM where group labels are considered as random. We contend that a suitable prior specification is essential to accurately characterize assortative behavior, and thus that a Bayesian approach is essential to community detection as shown in the examples in Section 2.6.1.

Our results can be connected to the work of Nowicki and Snijders (2001), Karrer and Newman (2011) and Hofman and Wiggins (2008) but we make two important distinctions: (i) we capture community behavior by explicitly requiring the probability of within-community associations to be higher than that of between-community associations; and (ii) we address parameter and label non-identifiability issues directly by remapping configurations to a unique canonical space. The first point is important in light of the examples in Section 2.6.1. The second point allows us to sample from the posterior space of label configurations more efficiently and to formally define an estimator based on a meaningful loss function. Moreover, our model can be related to the work of Mariadassou et al. (2010) and Vu et al. (2013) as they are all based on exponential-family clustering frameworks, but our model is different from theirs in two respects besides the two points just mentioned: (i) we make exact inference by introducing latent variables, rather than adopting approximate variational approaches; and (ii) we add more flexibility by setting hyper-prior structure on model parameters.

The organization of this chapter is as follows.

(1) I propose a Bayesian degree-corrected SBM for community detection that explicitly characterizes community behavior. I discuss this new model and how to account for parameter non-identifiability in Section 2.2.

(2) I treat label non-identifiability issues by defining a canonical projection of the space of label configurations in Section 2.3.

(3) I develop an efficient posterior sampler by identifying good initial configurations through approximate mode finding and then exploring a Gibbs sampler based on a data augmentation strategy in Section 2.4.

(4) I propose a *remapped* centroid estimator for community inference in Section 2.5. This new estimator is based on Hamming loss and is arguably a good representative of a projected space of label configurations.

In Section 2.6 I show that the proposed method is efficient and able to fit medium-sized networks with thousands of nodes in reasonable time. Moreover, I show that our proposed estimator yields more reliable estimation when compared to the ML-based estimators. Finally, I offer some concluding remarks in Section 2.7.

## 2.2 A Bayesian Stochastic Blockmodel for Community Detection

Under our community detection setup we assume a *fixed* number of groups $K \geq 2$ and we are given, as data, a matrix $[A]_{ij}$ representing relationships between "actors" $i$ and $j$ in a network with $n > K$ nodes. We represent the assignment of actors to communities through $\sigma : \{1, \ldots, n\} \mapsto \{1, \ldots, K\}$, a vector of *labels*: $\sigma_i = k$ codes for the $i$-th individual belonging to the $k$-th community.

A simple SBM specifies that the probability of an edge between actors $i$ and $j$ depends

only on their labels $\sigma_i$ and $\sigma_j$, and that $\sigma$ follows a product multinomial distribution:

$$\begin{aligned} A_{ij} \,|\, \sigma, \theta &\overset{\text{ind}}{\sim} \mathsf{Bern}\big(\theta_{\sigma_i\sigma_j}\big), \qquad i, j = 1, \ldots, n, i < j, \\ \sigma_i &\overset{\text{iid}}{\sim} \mathsf{MN}(1; \pi), \qquad i = 1, \ldots, n, \end{aligned} \tag{2.1}$$

where $\pi$ is a vector of prior probabilities over $K$ labels, parameter $\theta_{kk}$ is the "within" probability of a relationship in community $k$, and $\theta_{kl}$ is the "between" probability of a relationship for communities $k$ and $l$, $k, l = 1, \ldots, K$, $k < l$. If we define $\theta_w \doteq \theta_{11} = \cdots = \theta_{KK}$ and $\theta_b \doteq \theta_{12} = \cdots = \theta_{K-1,K}$, we have a simpler model with single within and between probabilities (Hofman and Wiggins, 2008).

We regard SBMs as logistic models and exploit this formulation to define a *degree-corrected* SBM by

$$A_{ij} \,|\, \sigma, \gamma, \eta \overset{\text{ind}}{\sim} \mathsf{Bern}\big(\text{logit}^{-1}(\gamma_{\sigma_i\sigma_j} + \eta_i + \eta_j)\big) \tag{2.2}$$

where, in logit scale, parameters $\gamma$ capture within and between community probabilities of association and node intercepts $\eta = (\eta_1, \ldots, \eta_n)$ capture the expected degrees of the nodes. To avoid redundancies, we only code $\gamma_{kl}$ for $k \leq l$. We note that without $\eta$, model (2.2) is equivalent to model (2.1) with $\gamma_{kl}$ =logit$(\theta_{kl})$. We also remark that we call the above model node-corrected, which is arguably more suitable for a broader generalized linear model formulation besides the logistic formulation; in Karrer and Newman's approach the observed $A_{ij}$ follow a Poisson distribution, and so $\eta$ is related to expected log degrees, and hence their degree-correction denomination (Karrer and Newman, 2011). Note that the failure in considering degree corrections may result in clustering the hubs into one community as shown in the left panel of Figure 2.1.

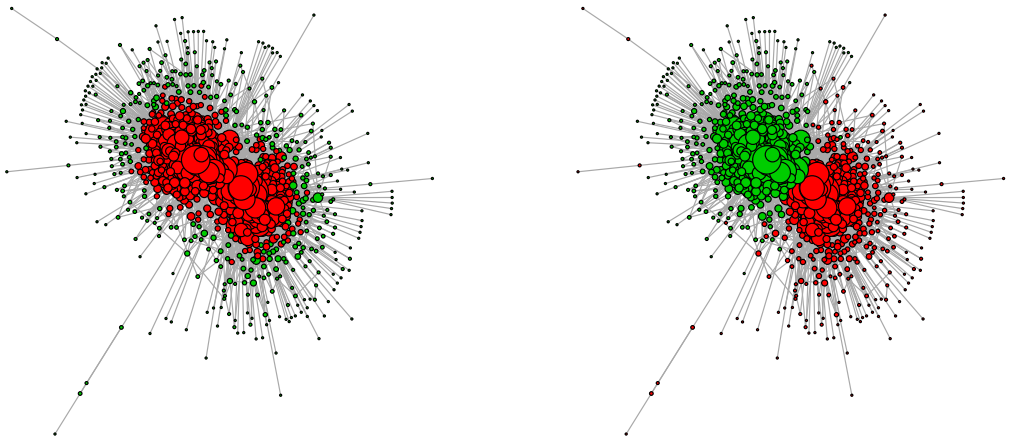

Figure 2.1: Political blogs network, illustrated in Section 2.6. Node sizes are proportional to degree; node colors (red/green) represent groups in non-degree-corrected estimator (left) and degree-corrected estimator (right).

### 2.2.1 Parameter Identifiability

In what follows, to simplify the notation we group $\beta = (\gamma, \eta)$ and define the design matrix $X$ associated to model (2.2) such that

$$A_{ij} \,|\, \sigma, \beta \overset{\text{iid}}{\sim} \mathsf{Bern}\big(\text{logit}^{-1}(x_{ij}(\sigma)^\top \beta)\big).$$

Note that we make explicit the dependence of each row $x_{ij}$ on the labels $\sigma$. Model (2.2) has then $\binom{K}{2} + K + n$ parameters, but the next result shows that only $\binom{K}{2} + n$ parameters are needed for the model to be identifiable if each community has at least two nodes (the proof is in Appendix 6.0.1.)

**Theorem 1.** *The design matrix $X$ associated with model* (2.2) *has the following properties:*

*(1) It has $K$ linearly dependent columns.*

*(2) It is full column-ranked if and only if each community has at least two nodes.*

Based on these two criteria, to attain an identifiable model we remove $K$ parameters from $\gamma$ and modify the prior on $\sigma$ to a constrained multinomial distribution,

$$\mathbb{P}(\sigma) \propto \prod_{k=1}^{K} I(N_k > 1) \prod_{i=1}^{n} \pi_k^{I(\sigma_i = k)},$$

where $I(\cdot)$ is the indicator function and $N_k = \sum_i I(\sigma_i = k)$ is the number of nodes in community $k$. There are still problems with label identifiability that we address by label remapping in the Section 2.3; for now, to allow for a straightforward remapping of community labels, we just set

$$\gamma_{11} = \cdots = \gamma_{KK} = 0 \tag{2.3}$$

to remove the redundant $\gamma$ parameters.

### 2.2.2 Hierarchical model for community detection

We attain a more flexible model by further setting a hyper-prior distribution on $\gamma = (\gamma_{12}, \ldots, \gamma_{K-1,K})$, $\eta$, and $\pi$,

$$\begin{aligned} \beta = (\gamma, \eta) &\sim I(\gamma \leq 0) \cdot N\Big(0, \tau^2 I_{n+\binom{K}{2}}\Big), \\ \pi &\sim \mathsf{Dir}(\alpha_1, \ldots, \alpha_K), \end{aligned} \tag{2.4}$$

where $\tau^2$ controls how informative the prior is. The prior on $\gamma$ and $\eta$ can be seen as a ridge regularization for the logistic regression in (2.2). The constraint $\gamma \leq 0$ in this SBM is essential to community detection since we should expect as many as or fewer edges between communities than within communities on average, and thus that the log-odds of between and within probabilities is non-positive. The conjugate prior on $\pi$ adds more flexibility to the model, and is important when identifying communities of varied sizes and alleviating resolution limit issues.

## 2.3 Label Identifiability

Since the likelihood in (2.2) only considers if individuals are in the same community or not, labels are not identifiable due to this stochastic equivalence. Moreover, if $\pi$ follows a strongly informative symmetric Dirichlet, $\alpha = W \cdot 1_K$ with $W$ large, then the marginal prior on $\sigma$ is approximately non-identifiable:

$$\mathbb{P}(\sigma) = \int \mathbb{P}(\sigma \,|\, \pi)\mathbb{P}(\pi)\mathrm{d}\pi = \frac{\prod_k \Gamma(N_k + W)/\Gamma(W)}{\Gamma(n + KW)/\Gamma(KW)} \approx \frac{\prod_k W^{N_k}}{(KW)^n} = \frac{1}{K^n}.$$

Since $\sigma_i$ are i.i.d. multinomial, then if $\pi$ is non-informative, $\pi = (1/K, \ldots, 1/K)$, the labels are not identifiable in the posterior $\mathbb{P}(\sigma|A)$ either. In fact, non-identifiability issues occur within a group of labels $\mathcal{I}$ whenever $\pi_i = \pi_j$ for all $i, j \in \mathcal{I}$. We need to address non-identifiability issues since here we discuss a non-informative $\pi$, a common modeling choice for simplicity.

A common approach in latent class models to fix label non-identifiability is to fix an arbitrary order in the parameters (Gelman et al., 2003, Chapter 18), e.g. $\gamma_{12} < \cdots < \gamma_{K-1,K}$. However, as Nowicki and Snijders (2001) pointed out, this solution can lead to imperfect identification of the classes if the parameters are close with high posterior probability; a major drawback then is that parameters and labels can be interpreted incorrectly. To address this problem, a label switching algorithm was proposed by Stephens (2000) in the context of MCMC sampling, but it is slow in practice. Another approach is to simply focus on permutation-invariant functions; in particular, when estimating $\sigma$, we can adopt a permutation-invariant loss, such as Binder's loss (Binder, 1978). We discuss such an approach in more detail in Section 2.5. Next, we propose an alternative, simpler procedure to remap labels and address non-identifiability.

### 2.3.1 Canonical Projection and Remapping Labels

Let $L \doteq \{1, \ldots, K\}$ and $\mathcal{L} = \{\sigma \in L^n : N_k(\sigma) > 1, k = 1, \ldots, K\}$ be the space of labels with positive prior probability. If $\rho$ is any *permutation* of the labels then $\mathbb{P}(\sigma|A) =$

$\mathbb{P}(\rho(\sigma)|A)$, where $(\rho(\sigma))_j = \rho(\sigma_j)$ for $j = 1, \ldots, n$. Non-identifiability here means that $\mathbb{P}(\cdot \mid A)$ is *invariant* under $\rho$, and that $\sigma$ and $\rho(\sigma)$ are $\mathbb{P}(\cdot|A)$-*equivalent*, which we denote by $\sigma \sim_P \rho(\sigma)$. Moreover, we can partition $\mathcal{L}$ according to $\sim_P$: if $S$ is one subspace defined by that partition by $\sim_P$, then any $\sigma \in S$ is such that $\sigma$ is not $\mathbb{P}(\cdot \mid A)$-equivalent to any other label configuration in $S$. To achieve label identifiability we anchor one such subspace as a *reference* space $Q$ and regard all other subspaces as permuted copies of $Q$.

Let $\mathsf{ind}(\sigma)$ be the vector with the first positions in $\sigma$ where each label appears, $\mathsf{ind}(\sigma)_k \doteq \min\{i : \sigma_i = k\}$, and further define $\mathsf{ord}(\sigma)$ as the vector with the order in which the labels appear in $\sigma$,

$$\mathsf{ord}(\sigma)_k = \sigma^{-1}\Big[\mathsf{ind}(\sigma)_{(k)}\Big], \quad k \in L. \tag{2.5}$$

Note that $\mathsf{ind}(\sigma)_{(k)}$ is the $k$-th position in the ordered vector $\mathsf{ind}(\sigma)$. As an example, if $\sigma = (2, 2, 3, 1, 3, 4, 2, 1)$ with $K = 4$ (and $n = 8$) then $\mathsf{ind}(\sigma) = (4, 1, 3, 6)$, ordered $\mathsf{ind}(\sigma)$ is $(1, 3, 4, 6)$ and so $\mathsf{ord}(\sigma) = (2, 3, 1, 4)$. To maintain identifiability we then simply constrain label assignments to the subset of $\mathcal{L}$ where $\mathsf{ord}(\cdot)$ is fixed. As a simple, natural choice, let us restrict assignments to $Q = \{\sigma : \mathsf{ord}(\sigma) = L\}$. Note that any $\sigma$ can be mapped to its *canonical* assignment by

$$\rho(\sigma) \doteq \mathsf{ord}(\sigma)^{-1}(\sigma). \tag{2.6}$$

Taking our previous example, $\sigma = (2, 2, 3, 1, 3, 4, 2, 1)$ would then be mapped to $\rho(\sigma) = (1, 1, 2, 3, 2, 4, 1, 3)$. The definitions of $\mathsf{ind}$ and $\mathsf{ord}$ can then be used to derive a procedure that *remaps* $\sigma$ to $\rho(\sigma)$; for completeness, we list an algorithm that implements such remap procedure in Appendix 6.0.2.

Our proposed reference set above is also described by $Q = \{\sigma \in \mathcal{L} : \sigma = \rho(\sigma)\}$, the *quotient* space of $\mathcal{L}$ with respect to $\mathsf{ord}$, $\mathcal{L}/\mathsf{ord}$: any pair of label configurations $\sigma_1$ and $\sigma_2$ such that $\rho(\sigma_1) = \rho(\sigma_2)$ are identified to a single label $\rho(\sigma_1)$ in $Q$. By constraining the labels to a reference quotient space we not only achieve identifiability, but also make the labels interpretable: label $j$ marks the $j$-th community to appear in the sequence of labels. As a consequence, we are not restricted to estimating permutation-invariant functions of

the labels, as in the approach of Nowicki and Snijders (2001), since now, for example, $\mathbb{P}(\sigma_i = j \,|\, A)$ is meaningful. As a particular application, we derive a direct estimator of $\sigma$ based on Hamming loss in Section 2.5; in the next section we discuss how the constraint to $Q$ is implemented in practice.

## 2.4 Posterior Sampling

To sample from the joint posterior on $\sigma$, $\beta$ and $\pi$, we use a Gibbs sampler (Geman and Geman, 1984; Robert and Casella, 1999) that iteratively alternates between sampling from

$$[\sigma \,|\, \gamma, \eta, \pi, A], \quad [\pi \,|\, \sigma, \gamma, \eta, A], \quad [\gamma, \eta \,|\, \sigma, \pi, A]$$

until convergence. Next, we discuss how we obtain each conditional distribution in closed form.

### 2.4.1 Sampling $\sigma$ and $\pi$

Let us start with the most relevant parameters: the labels $\sigma$. We can sample a candidate, unconstrained assignment for actor $i$, $\sigma_i$, conditional on all the other labels $\sigma_{[-i]}$, parameters $(\beta, \pi)$, and data $A$ from a multinomial with probabilities:

$$\begin{aligned}\mathbb{P}(\sigma_i \,|\, \sigma_{[-i]}, \beta, \pi, A) &\propto \pi_k \\ &\prod_{j \neq i} \Big(\text{logit}^{-1}(\gamma_{\sigma_i\sigma_j} + \eta_i + \eta_j)\Big)^{A_{ij}} \Big(1 - \text{logit}^{-1}(\gamma_{\sigma_i\sigma_j} + \eta_i + \eta_j)\Big)^{1-A_{ij}} \\ &= \pi_k \prod_{j \neq i} \frac{\exp\{A_{ij}(\gamma_{\sigma_i\sigma_j} + \eta_i + \eta_j)\}}{1 + \exp\{\gamma_{\sigma_i\sigma_j} + \eta_i + \eta_j\}}. \quad (2.7)\end{aligned}$$

To guarantee that parameters are identifiable, we reject the candidate $\sigma$ if $N_k \leq 1$ for any community $k$. Moreover, to keep the labels identifiable, we remap $\sigma$ using the routine in Section 2.3 and remap $\gamma$ accordingly.

As an example, consider the label samples obtained from running the Gibbs sampler on

the political blogs study in Section 2.6. In Figure 2.2, we plot a multidimensional scaling [MDS (Gower, 1966)] representation of the samples. We have $K = 2$ communities, and so $\mathcal{L}$ is partitioned into a reference quotient space in the right and a "mirrored" space in the left; any point in the mirrored space can be obtained by swapping labels 1 and 2 in the reference space and vice versa. The green arrow shows a valid sampling move $\sigma^{(t)} \to \sigma^{(t+1)}$ at iteration $t$ that does not require a remap, while the red arrow is an invalid move since it crosses spaces. The blue arrow remaps $\sigma^{(t+1)}$ to $\rho(\sigma^{(t+1)})$ in the reference space. The dashed green arrow summarizes both operations.

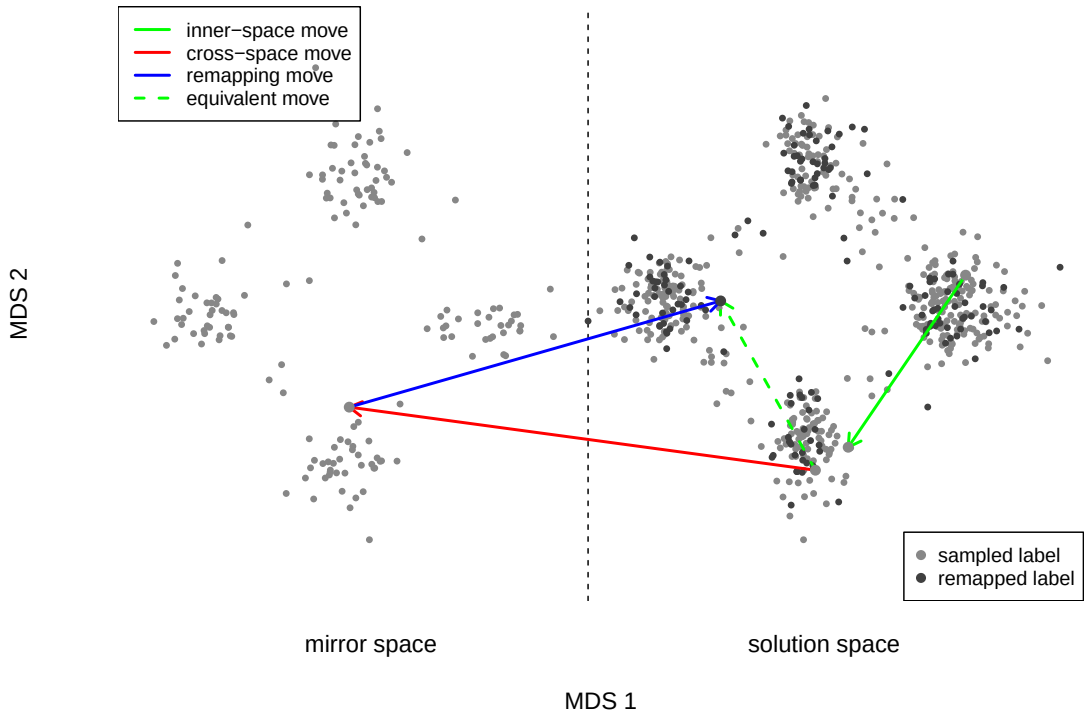


Figure 2.2: MDS representation of the two copies of the quotient space $\mathcal{L}/\mathsf{ord}$ using posterior samples for the political blogs example in Section 2.6. Arrows are described in text.

For the nuisance parameter $\pi$ we summon conjugacy to obtain

$$\pi \,|\, \sigma, \gamma, \eta, A \sim \mathsf{Dir}(\alpha + \mathbf{N}(\sigma)), \tag{2.8}$$

where $\mathbf{N}(\sigma) = (N_1, \ldots, N_K)$ and $N_k$ are community sizes.

### 2.4.2 Sampling $\gamma$ and $\eta$

Sampling $\beta$ conditional on $\sigma$, $\pi$, and data $A$ is more challenging since the logistic likelihood in (2.2) does not specify a closed form distribution. However, if we explore a data augmentation strategy by introducing latent variables $\omega = (\omega_{ij})_{i<j:i,j\in\{1,\ldots,n\}}$ from a Pólya-Gamma distribution, then the above conditional distribution of $\beta$ given $\omega$ is now available in closed form (Polson et al., 2012). More specifically, if $\omega_{ij} \,|\, \sigma, \beta \sim \mathsf{PG}(1, x_{ij}(\sigma)^\top \beta)$, then

$$\beta \,|\, \omega, \sigma, A \sim I[\gamma \leq 0] \cdot N(m, V)$$

where, with $\Omega = \mathrm{Diag}(\omega_{ij})$ and latent weighted responses $z_{ij} = (A_{ij} - 1/2)\omega_{ij}^{-1}$,

$$V = \left( X^\top \Omega X + \frac{1}{\tau^2} I_{n+\binom{K}{2}} \right)^{-1} \quad \text{and} \quad m = V X^\top \Omega z. \tag{2.9}$$

The assortativity constraint $\gamma \leq 0$ in the $\beta$ prior is clearly also present in the conditional posterior, and so we can use a simple rejection sampling step for the truncated normal: sample from unconstrained marginals $N(m, V)$ and accept only if $\gamma \leq 0$. However, since

$$\beta = \begin{bmatrix} \gamma \\ \eta \end{bmatrix} \,\Big|\, \omega, \sigma, A \sim N\left( m = \begin{bmatrix} m_\gamma \\ m_\eta \end{bmatrix}, V = \begin{bmatrix} V_\gamma & V_{\gamma\eta} \\ V_{\eta\gamma} & V_\eta \end{bmatrix} \right),$$

we can adopt a more efficient way of sampling $\beta$ by first sampling $\eta$ marginally,

$$\eta \,|\, \omega, \sigma, A \sim N(m_\eta, V_\eta), \tag{2.10}$$

and then sampling

$$\gamma \,|\, \eta, \omega, \sigma, A \sim I(\gamma \leq 0) \cdot N(m_\gamma + V_{\gamma\eta} V_\eta^{-1}(\eta - m_\eta), V_\gamma - V_{\gamma\eta} V_\eta^{-1} V_{\eta\gamma}) \tag{2.11}$$

from a truncated normal. In practice, we compute the Schur complement of $V_\eta$, $V_\gamma - V_{\gamma\eta} V_\eta^{-1} V_{\eta\gamma}$, using the SWEEP operator (Goodnight, 1979).

### 2.4.3 Gibbs sampler

To summarize, after setting initial parameters $\sigma$, $\beta$ and $\pi$ arbitrarily, we then iterate until convergence the following Gibbs sampling steps:

1. Sample $\sigma \,|\, \beta, \pi, A$: for each node $i$,
   (a) Sample $\sigma_i \,|\, \sigma_{[-i]}, \beta, A$ from a multinomial distribution as in (2.7). If $N_k(\sigma) < 2$ for some community $k$, reject and keep the previous value of $\sigma_i$.
   (b) Remap $\sigma$ using the procedure in Section 2.3.
2. Sample $\pi \,|\, \sigma, \beta, A$ from the Dirichlet distribution in (2.8).
3. Sample $\beta \,|\, \sigma, \pi, A$:
   (a) Sample $\omega \,|\, \sigma, \beta, \pi, A$: for each pair $i < j$, $\omega_{ij} \,|\, \sigma, \beta \sim \mathsf{PG}(1, x_{ij}(\sigma)^\top \beta)$.
   (b) Sample $\beta \,|\, \sigma, \pi, \omega, A$: compute $m$ and $V$ as in (2.9), sample $\eta$ marginally as in (2.10), and then sample $\gamma \,|\, \eta$ from a truncated multivariate normal distribution as in (2.11).

To speed up convergence and improve precision, we set the initial $\sigma$ to be an approximate posterior mode obtained from a greedy optimization version of the above routine, similar to a gradient cyclic descent method. The main changes are:

1. In Step 1.a we take $\sigma_i$ to be the mode of $\sigma_i \,|\, \sigma_{[-i]}, \beta, A$ (but we might still reject $\sigma_i$ if $N_k(\sigma) < 2$ for some $k$ and remap $\sigma$ in Step 1.b).

2. In Step 2, we take $\pi$ to be the mode of the Dirichlet distribution in (2.8).

3. Step 3 is substituted by a regularized iterative reweighted least-squares (IRLS) step, which is commonly used when fitting logistic regression models (McCullagh and Nelder, 1989). At the $t$-th iteration we define $\mu_{ij} = \text{logit}^{-1}(x_{ij}(\sigma)^\top \beta^{(t)})$ and $W = \text{Diag}(\mu_{ij}(1-\mu_{ij}))$ to obtain the update

$$V = \left(X^\top W X + \frac{1}{\tau^2} I_{n+\binom{K}{2}}\right)^{-1} \quad \text{and} \quad \beta^{(t+1)} = V X^\top W z^{(t)}$$

where $z^{(t)} = X\beta^{(t)} + W^{-1}(y-\mu)$ is now the "working response". To guarantee that the community constraints $\gamma \leq 0$ are met, we use an active-set method (Nocedal and Wright, 2006, Chapter 16).

Since we expect the posterior space to be multimodal, we adopt a strategy similar to Karrer and Newman (2011) and sample multiple starting points for $\sigma$ according to its prior distribution and then obtain approximate posterior modes for each simulation. We elect the best approximate mode over all simulations as the starting point for the Gibbs sampler, which is then run until convergence to more thoroughly explore the posterior space.

## 2.5 Posterior Inference

One common estimator for label assignment is the MAP estimator,

$$\widehat{\sigma}_M = \underset{\tilde{\sigma}\in\{1,\ldots,K\}^n}{\arg\min}\ \mathbb{E}_{\sigma\,|\,A}\big[I(\tilde{\sigma}\neq\sigma)\big] = \underset{\tilde{\sigma}\in\{1,\ldots,K\}^n}{\arg\max}\ \mathbb{P}(\sigma=\tilde{\sigma}\,|\,A),$$

which, albeit based on a zero-one loss function (Besag, 1986), has the advantage of being invariant to label permutations. However, given the flexibility in our model due to the hierarchical levels, the posterior space is often complex and so the MAP might fail to capture the variability and might focus on sharp peaks that gather a small amount of posterior mass around them.

Another estimator for label assignment arises from minimizing Binder's loss $B$ (Binder, 1978, 1981),

$$\widehat{\sigma}_B = \underset{\tilde{\sigma}\in\{1,\ldots,K\}^n}{\arg\min}\ \mathbb{E}_{\sigma\,|\,A}\big[B(\tilde{\sigma},\sigma)\big], \tag{2.12}$$

where

$$B(\tilde{\sigma},\sigma) = \sum_{i<j} I(\tilde{\sigma}_i \neq \tilde{\sigma}_j)I(\sigma_i = \sigma_j) + I(\tilde{\sigma}_i = \tilde{\sigma}_j)I(\sigma_i \neq \sigma_j).$$

The advantage of Binder's loss is that since it penalizes pairs of nodes it is invariant to label permutations—that is, $B(\tilde{\sigma},\sigma) = B(\tilde{\sigma},\phi(\sigma)) = B(\phi(\tilde{\sigma}),\sigma)$ for any permutation $\phi$. However, Lau and Green (2007) have shown that minimizing Binder's loss is equivalent to binary integer programming, which is an NP-hard problem. Moreover, as Fritsch and Ickstadt (2009) pointed out, even the approximated solution given by Lau and Green (2007) is only feasible when the dataset is of moderate size.

In contrast, when compared to MAP inference, centroid estimation (Carvalho and Lawrence, 2008) offers a better representative of the space since it arises from a loss function:

$$\widehat{\sigma}_H = \underset{\tilde{\sigma}\in\{1,\ldots,K\}^n}{\arg\min}\ \mathbb{E}_{\sigma\,|\,A}\big[H(\tilde{\sigma},\sigma)\big],$$

where $H$ is *Hamming* distance, $H(\tilde{\sigma},\sigma) = \sum_{i=1}^{n} I(\tilde{\sigma}_i \neq \sigma_i)$. The Hamming loss is more refined than the 0-1 loss of MAP estimation in the sense of offering more resolution. Take $\tilde{\sigma} = (1,1)$ and $\sigma = (1,2)$ as an example, $H(\tilde{\sigma},\sigma) = 1$ while $L_{\mathrm{MAP}}(\tilde{\sigma},\sigma) = 0$. The centroid estimator also identifies the median probability model, and thus is known to offer better predictive resolution than the MAP estimator (Barbieri and Berger, 2004). However, Hamming loss is only invariant to double label permutations but not to single label permutations, i.e., $H(\tilde{\sigma},\sigma) = H(\phi(\tilde{\sigma}),\phi(\sigma))$ but it is not necessarily true that $H(\tilde{\sigma},\sigma) = H(\phi(\tilde{\sigma}),\sigma)$ or $H(\tilde{\sigma},\sigma) = H(\tilde{\sigma},\phi(\sigma))$, and thus, in order for Hamming loss to be meaningful for estimation when the labels are non-identifiable we need to account for label aliasing. We then redefine the centroid estimator to depend on a specific permutation, for instance the

canonical permutation $\rho$ in (2.6),

$$\widehat{\sigma}_C = \rho\left( \underset{\tilde{\sigma}\in\{1,\ldots,K\}^n}{\arg\min}\ \mathbb{E}_{\sigma\,|\,A}\big[H(\tilde{\sigma},\rho(\sigma))\big] \right).$$

This remapped centroid estimator considers only one version of the posterior space, namely the reference quotient space $\mathcal{L}/\mathsf{ord}$ with $\mathsf{ord}$ in (2.5). The main advantage of this new estimator is to allow the following characterization (see Appendix 6.0.3 for the proof):

**Theorem 2.** *The centroid estimator $\widehat{\sigma}_C$ is a mapped* consensus *estimator: if $\mathbb{P}^*(\sigma\,|\,A)$ is the induced posterior probability of $\sigma \in \mathcal{L}\,/\,\mathsf{ord}$ and*

$$(\widehat{\sigma}^*)_i = \underset{k\in\{1,\ldots,K\}}{\arg\max}\ \mathbb{P}^*(\sigma_i = k\,|\,A)$$

*then $\widehat{\sigma}_C = \rho(\widehat{\sigma}^*)$.*

In practice, we estimate

$$\widehat{\mathbb{P}}^*(\sigma_i = k\,|\,A) \approx \frac{1}{N}\sum_{t=1}^{N} I(\sigma_i^{(t)} = k)$$

using $N$ realizations from the Gibbs sampler presented in Section 2.4 to define $\widehat{\sigma}_C$. Since we only need to elect, for each actor, the most likely label according to Theorem 2, obtaining the centroid estimator is much simpler computationally than MAP and Binder estimation. Note that due to the remap step when sampling $\sigma\,|\,\theta, A$, we are always constrained to the quotient space $\mathcal{L}/\mathsf{ord}$ and identifying label realizations under $\rho$, and thus really approximating $\mathbb{P}^*(\sigma\,|\,A)$.

### 2.5.1 Relating Binder and Centroid Estimators

We start by noting that if we define an extended *matched* map $M(\sigma) = \{I(\sigma_i = \sigma_j)\}_{1\le i<j\le n}$ that makes pairwise comparisons among labels in $\sigma$, then Binder and Hamming losses are related through $B(\tilde{\sigma},\sigma) = H(M(\tilde{\sigma}), M(\sigma))$ and so Binder's estimator in (2.12)

is also a centroid estimator in the extended matched space $M(\mathcal{L})$.

Back to the original space $\mathcal{L}$ of labels, we observe that, in practice, the Binder and centroid estimators are often close (in either loss). To explain these observations, we need the next result relating Binder and Hamming losses (the proof can be found in Appendix 6.0.4):

**Theorem 3.** *For any pair of label assignments $\tilde{\sigma}$ and $\sigma$, Binder loss is related to Hamming loss through*

$$B(\tilde{\sigma},\sigma) \leq H(\tilde{\sigma},\sigma)\Big(n - \frac{1}{2}H(\tilde{\sigma},\sigma)\Big). \tag{2.13}$$

*Moreover, if $K = 2$ then $B(\tilde{\sigma},\sigma) = H(\tilde{\sigma},\sigma)(n - H(\tilde{\sigma},\sigma))$.*

From (2.13) we see that Binder's loss can be approximately linearly bounded by Hamming loss when the Hamming distance between $\tilde{\sigma}$ and $\sigma$ is small. Thus, when the marginal posterior distribution on $\sigma$ has a compact cluster of label configurations with high posterior mass we expect this cluster to contain the centroid estimator and also, according to (2.13), Binder estimator since minimizing the posterior expected Hamming loss is approximately equivalent to minimizing the posterior expected Binder loss in this case. In the next section we run experiments on simulated datasets and observe that the two estimators are often close and show similar performance for simple networks (check, for instance, Figure 2.6.)

## 2.6 Experimental Results

In this section, we demonstrate the performance of the centroid estimator and compare it to Binder estimator under our model and to KN estimator (Karrer and Newman, 2011), MAP estimator, Fast-Greedy (FG) estimator (Clauset et al., 2004), Multi-Level (ML) estimator (Blondel et al., 2008), Walktrap (WT) estimator (Pons and Latapy, 2004) and Label Propagation (LP) estimator (Raghavan et al., 2007) through an empirical study. In the case studies we run repeated experiments on the same dataset and obtain the error rates of the estimators mentioned above when compared to known or *bona fide* ground truth references. To compare those estimators, we define a *q-error interval* as the interval with endpoints being the $q/2$ and $1 - q/2$ quantiles of the error rates.

Before discussing the experimental results, we present two illustrative examples next.

### 2.6.1 Illustrative Examples

Even though the models reviewed above are flexible enough to identify social block structure, they might fail to actually recognize communities defined in Chapter 1. We now show two simple examples to demonstrate how this happens, and compare our proposed solution to the results from applying Karrer and Newman's (KN) popular degree-corrected SBM (Karrer and Newman, 2011).

The first dataset is a synthetic network, denoted as the "spike" dataset, which we intentionally designed to show that degree correction is not sufficient to elicit communities. The network considered is split into $K = 2$ communities. The first community contains $2n_1$ nodes with $n_1$ of them being strongly connected as a complete graph $K_{n_1}$ (a "kernel") and having a one-to-one connection with the remaining $n_1$ nodes (a "crown"). The other community is formed in a similar way, but with a complete $K_{rn_1}$ kernel connected to a $rn_1$ crown, totaling $2rn_1$ nodes (see Figure 2.3). We add some between-community edges in such a way that each node from the complete graph $K_{n_1}$ in the first community is connected to $r$ nodes from the complete graph $K_{rn_1}$ in the second community.

We note that this network can still be characterized as having a community behavior since the edge density between communities is smaller than the density within communities. Moreover, due to the crowns, we also need to account for degree heterogeneity in each community. Let us then consider the case when $n_1 = 10$ and $r = 5$. Figure 2.3 compares the KN estimator and our estimator. The kernel-crown structure of both communities is not reflected in KN estimator; moreover, there are more edges between groups than within groups, which is not prescribed by community behavior.

We observe that degree correction is not enough to correctly capture the community structure in the synthetic network that we designed. However, similar results are also observed in some real-world datasets. Consider, for example, the "sampson" network reported by Sampson (1968) at time point $T_4$ among a group of 18 trainee monks at a New England

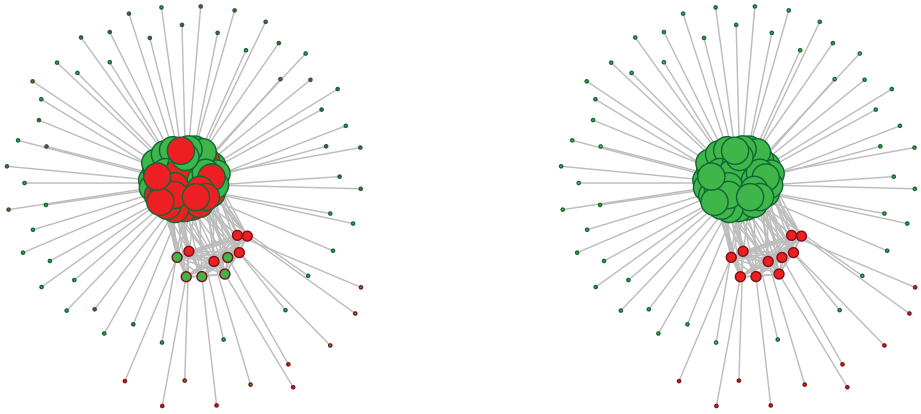

Figure 2.3: Spike network, $n_1 = 10$, $r = 5$. Node sizes are proportional to degree; node colors (red/green) represent groups in KN estimator (left) and our estimator (right). Node borders mark the reference.

monastery. Four types of relations—affection, esteem, influence, and sanctioning—between the monks are collected. In this network, each node represents a monk in the monastery, and two nodes are considered to be connected if they considered each other as being in at least one of the four relations when asked by Sampson. Sampson reported a partition of trainee monks into three communities ($K = 3$): Young Turks, Loyal Opposition and Outcasts. Figure 2.4 compares KN estimator to our estimator and shows a similar pattern where within group connections are sparser than between group connections according to the KN estimate; in particular, there are more edges between the red and green communities than within the green community.

### 2.6.2 Empirical Study

First, we evaluate our estimator on simulated datasets with known references. The networks are generated from a class of benchmark graphs that account for heterogeneities in node degree distributions and community sizes (Lancichinetti et al., 2008a). The model used in the simulation considers the following parameters: both degree distribution and the community sizes are assumed to follow power law distributions with exponents $a$ and $b$, respectively; each network consists of $n$ nodes and has average degree $\langle k \rangle$; and mixing

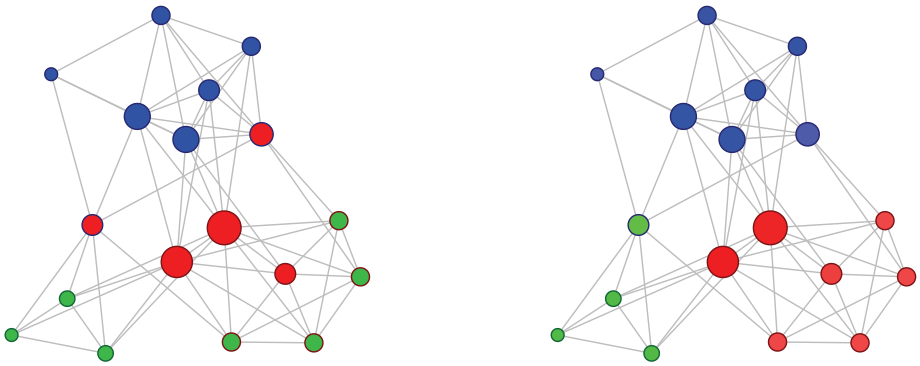

Figure 2.4: Sampson network at $T_4$, $n = 18$. Node sizes are proportional to degree; node colors mark KN estimator (left) and our estimator (right). Node borders mark the reference.

parameter $\mu$ represents the proportion of between-community edges.

We simulate 100 networks for each combination of $n = (100, 500)$, $a = (2, 3)$, $b = (1, 2)$, and $\mu = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6)$. Figure 2.5 shows one realization of the benchmark networks as an example. We compare the performance of the centroid, Binder, MAP, KN, FG, ML, WT and LP estimators in terms of the *normalized mutual information* (NMI) defined in (Danon et al., 2005). The NMI measures the similarity between two community labels $\sigma$ and $\tilde{\sigma}$:

$$\mathrm{NMI}(\sigma, \tilde{\sigma}) = \frac{2\mathrm{MI}(\sigma, \tilde{\sigma})}{\mathrm{H}(\sigma) + \mathrm{H}(\tilde{\sigma})}$$

where $\mathrm{MI}(\sigma, \tilde{\sigma})$ is the mutual information and $\mathrm{H}(\sigma)$ is the entropy of $\sigma$. The NMI is bounded below by 0, when two labels are independent, and above by 1, when two labels are identical.

The NMI of estimators is summarized in Figure 2.6. We observe from the figure that the centroid estimator performs comparably well as Binder, ML and WT estimators while slightly better than KN, MAP and FG estimators when the community structure is strong (mixing parameter $\mu \leq 0.3$). The centroid estimator outperforms KN, MAP, FG, ML and WT estimators to a large extent when the community structure is weak (mixing parameter $\mu \geq 0.4$). Of all estimators compared here, LP performs worst in terms of NMI while it

is the most computational efficient. Not surprisingly, all estimators perform worse as the mixing parameter $\mu$ increases (so that the communities are defined in a weaker sense) or the average degree $\langle k \rangle$ decreases (so that there are fewer edges). Similar results are found under other different combinations of $(a, b, \langle k \rangle)$, as shown in Figure **??** and Figure **??** in the Appendix.

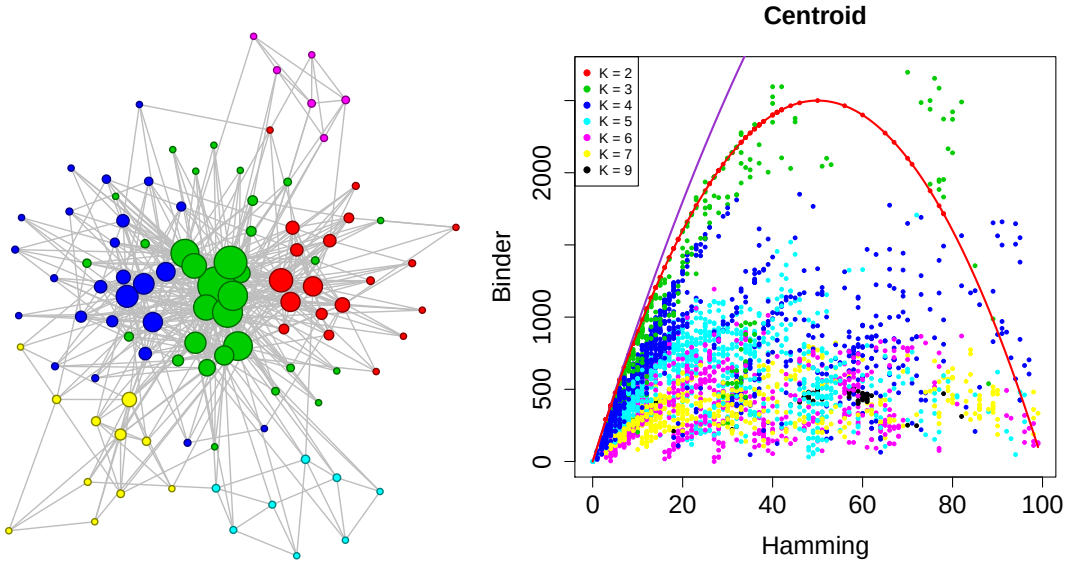


Figure 2.5: Left: one realization of the benchmark networks with $n = 100$ nodes, $a = 2$, $b = 1$, $\mu = 0.4$, and $\langle k \rangle = 10$. Right: Binder loss against Hamming loss over 50 graph realizations of such benchmark networks. Colors mark different values of $K$. Lines correspond to the upper bound in (2.13) for $K > 2$ and $K = 2$.

### 2.6.3 Case Study

Next, we evaluate our estimator for community detection on two real-world network datasets.

#### 2.6.3.1 Political blogs

The first case study is the political blogs network (Adamic and Glance, 2005), which is a medium real-world network containing over one thousand nodes. In this network, each node is a blog over the period of two months preceding the U.S. Presidential Election of

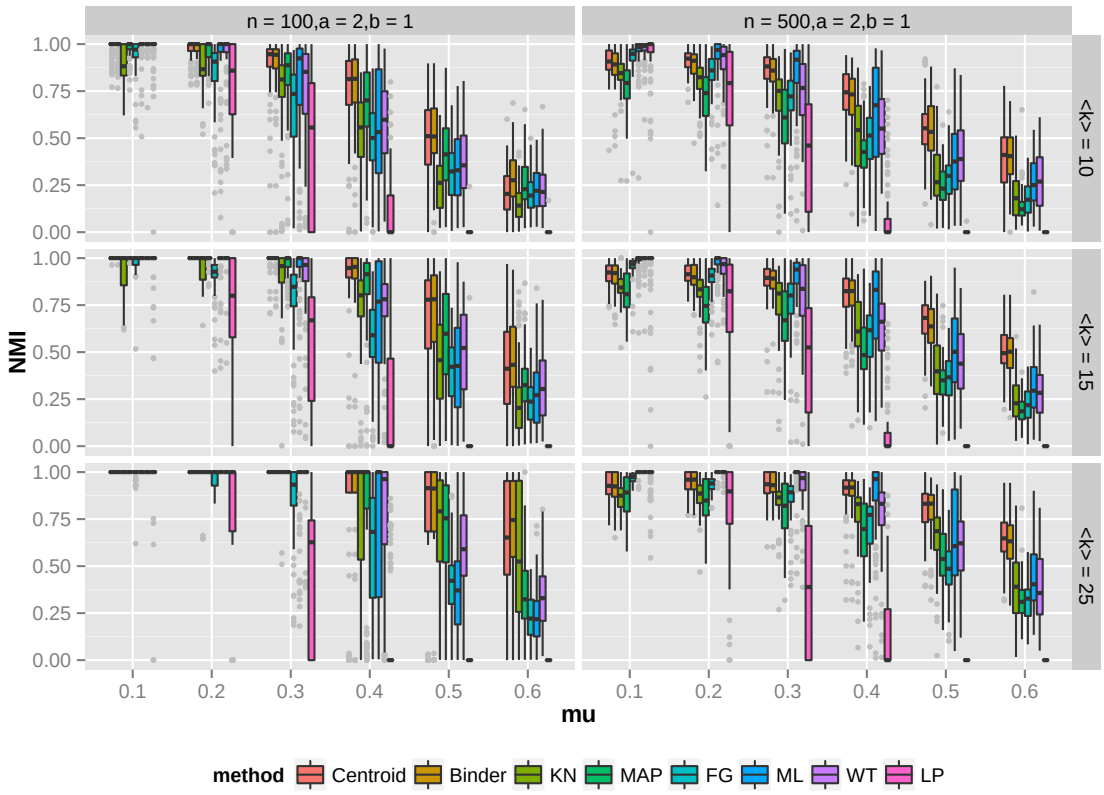


Figure 2.6: Benchmark networks of $n = 100$ and 500 nodes, with different combinations of exponents $a$, $b$ and average degrees $\langle k \rangle$. Each box plot corresponds to the NMI of the estimator over 100 graph realizations.

2004, and two nodes are considered to be connected if they referred to one another and there was overlap in the topics they discussed. Adamic and Glance label the blogs as either liberal or conservative and discover that a political blog rarely links to another blog of a different political ideology. It is expected that blogs in favor of the same political party are more likely to be linked and discussing the same topics than those in favor of different political parties, which corroborates a community behavior. This structure of the network leads to a polarization of those blogs into two communities, liberals and conservatives. We use this prior information on the political learning and pick $K = 2$, a choice also used by Karrer and Newman (2011).

The centroid estimator agrees well with the reference labeled by Adamic and Glance (2005), as depicted in the leftmost panel in Figure 2.7. We estimate each $\eta_i$ for node $i$ by its posterior mean using the converged samples and plot the estimated $\eta_i$ against the logit normalized degree of node $i$ in the middle panel. There is a positive linear relationship between $\eta_i$ and the logit of the normalized degrees, indicating that the expected degree, thus the probability of building an edge, is positively related to the observed degree of the node. If there is a community effect driven by political ideology, that is, if the network can be better explained by partitioning nodes into two different political communities, then $\gamma_{12}$ is expected to be significantly negative. The rightmost panel in Figure 2.7 shows the estimated posterior distribution of $\gamma_{12}$. An estimated 95% credible interval for $\gamma$ is $[-3.16, -2.99]$, which shows a clear deviation from 0 and thus indicates a strong community effect in the network.

We further compare the centroid estimator with Binder estimator and KN estimator, as in the previous section. The estimated 90% error intervals for the centroid, Binder, and KN estimators are $[0.053, 0.054]$, $[0.053, 0.054]$, and $[0.045, 0.051]$, respectively. In general, the three estimators perform equally well while the KN estimator yields a slightly smaller error rate on average.

#### 2.6.3.2 Political books

Finally, we pick the political books dataset compiled by Krebs (2004). This is a network of 105 books on politics sold by the online bookseller Amazon around the time of the U.S. presidential election in 2004. Each node represents a book on politics, and an edge between two nodes is built if the two books are frequently copurchased by the same buyers. These books appear to form communities of copurchasing that align closely with political ideologies—liberal or conservative—except for a few books that were explicitly centrist. The books were labeled as "liberal", "neutral", or "conservative" by Newman (2006) based on a reading of the descriptions and reviews of the books posted on Amazon. We use the prior information that books are of three political opinions and obtain the centroid, Binder,

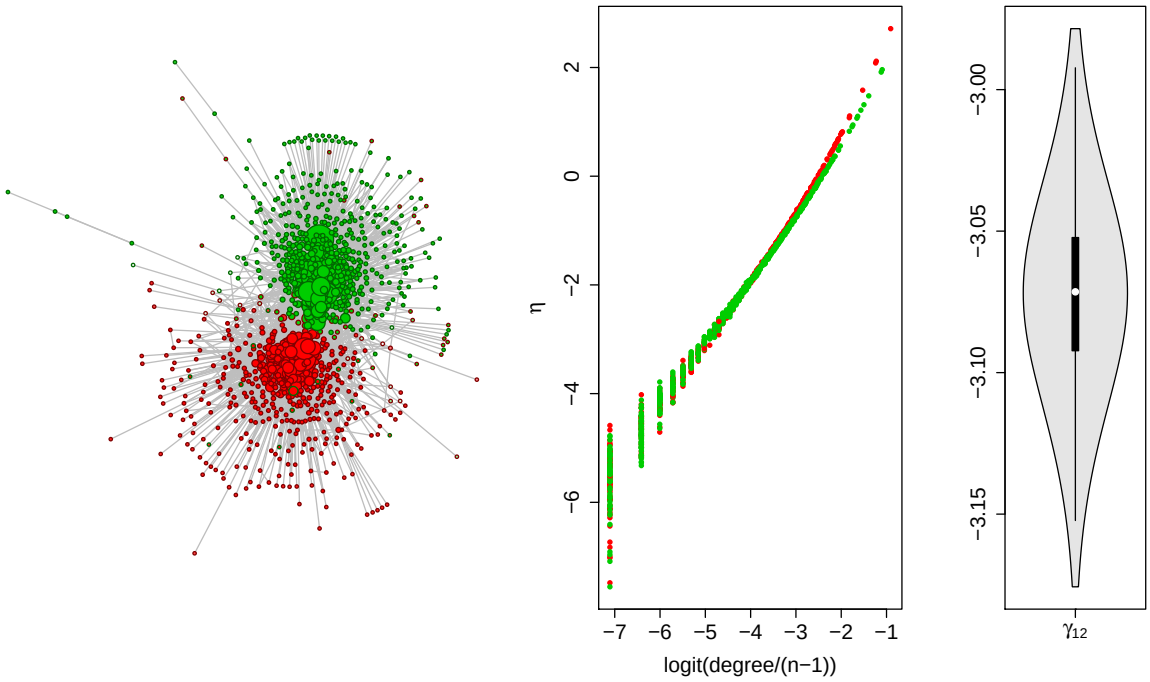


Figure 2.7: Political blogs network. Left: Node sizes are proportional to degree; node colors signal the centroid estimators (red/green). Node color intensities are proportional to $\widehat{\mathbb{P}}^*(\sigma_i \mid A)$ and node border colors mark the reference. Middle: $\eta_i$ on $\text{logit}(\text{degree}_i/(n-1))$ for each node $i$; color for each node $i$ represents $(\widehat{\sigma}_C)_i$. Right: estimated posterior distribution for $\gamma_{12}$.

and KN estimators under $K = 3$.

The left panel in Figure 2.8 shows the centroid estimator of the political books network. The communities corresponding to liberal (blue) and conservative (green) are clearly separated by the neutral (red) community. The estimated liberal and conservative communities agree with the reference well while there are a few nodes in or close to the estimated neutral community not matching the reference. The middle panel plots estimated $\eta_i$ against normalized degrees in logit scale. It is evident that the neutral (red) community has a different intercept for $\eta$, indicating that it is less connected. The right panel shows estimated marginal posterior distributions for $\gamma$. Not surprisingly, $\gamma_{23} < \gamma_{12}$ and $\gamma_{23} < \gamma_{13}$ with high posterior probability since communities 2 (green) and 3 (blue) are separated by community (red) and so do not share many edges.

We also use weakly-informative priors and run multiple chains as in the previous example. The estimated 90% error intervals for the centroid, Binder, and KN estimators are $[0.167, 0.175]$, $[0.167, 0.175]$, and $[0.171, 0.171]$, respectively. Most of the nodes not matching the reference are in or close to the neutral (red) community. The reason might be that those books appeal to buyers with different political opinions and thus are often copurchased with books in neighboring communities. The community labels (under the definition of "community" in this dissertation) of neutrals or nodes near neutrals do not reflect their stated political ideology in the descriptions or reviews.

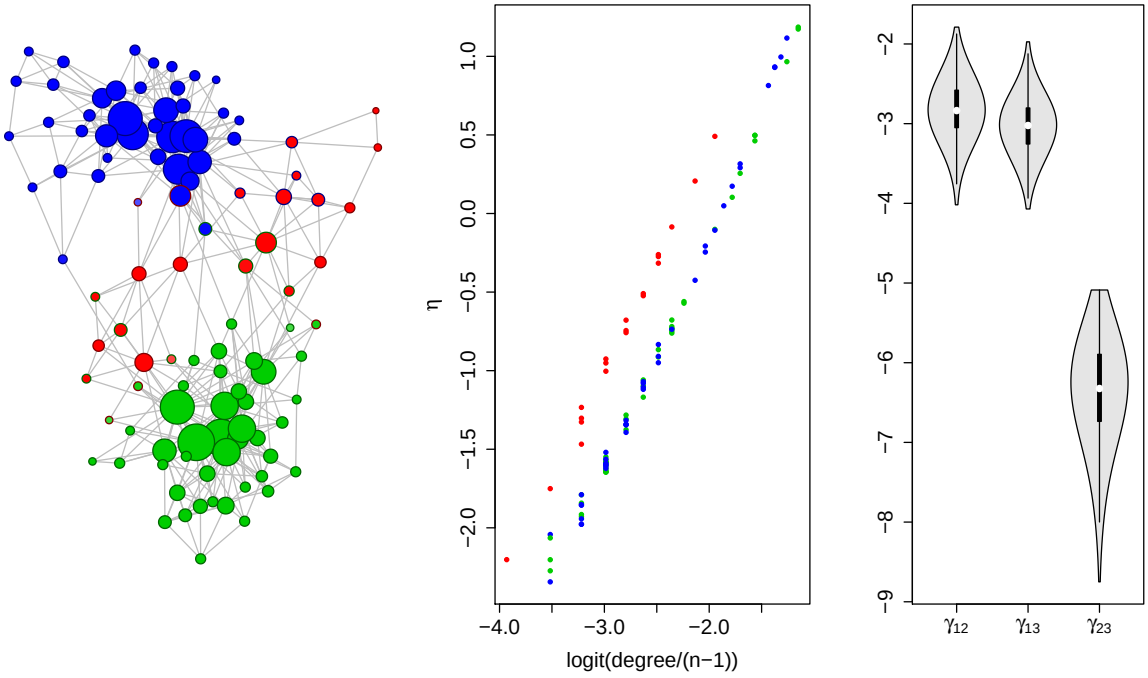


Figure 2.8: Political books network. Left: node sizes are proportional to degree; node colors signal the centroid estimators. Node color intensities are proportional to $\widehat{\mathbb{P}}^*(\sigma_i \mid A)$ and node borders mark the reference. Middle: $\eta_i$ on $\text{logit}(\text{degree}_i/(n-1))$ for each node $i$; color for each node $i$ represents $(\widehat{\sigma}_C)_i$. Right: estimated posterior distribution for $\gamma$.

## 2.7 Discussion

If the posterior space is multimodal then a single point estimator has difficulty in representing the space, and the centroid estimator is not immune to this problem. The

proposed estimation procedure can be extended to account for multiple modes by exploring *conditional* estimators on partitions of the space. While this can be done empirically by clustering posterior samples, we next pursue a more principled way of identifying partitions. As simple extensions to the proposed model, we incorporate parameters for node attributes and to generalize the formulation to account for count, categorical, and ordinal data in Chapter 3. Other directions for future work, albeit not related to community detection, include extending the remap procedure to other settings such as clustering and mixture model inference.

# Chapter 3

# Bayesian Group-Corrected SBM

## 3.1 Introduction

In this chapter, we generalize the degree-corrected SBM mentioned in Chapter 2 to make it suitable for a larger variety of applications. It is computationally prohibitive to fit the degree-corrected SBM on large networks ($\sim 10,000$) due to the sampling approach used in the inference. Besides, model (2.2) is based on a logistic regression, and thus limited to applications with binary data. We can only model the absence or presence of an interaction under model (2.2). However, strengths of connections can be of great significance and used to obtain more reliable inference results. We next present a simple example to demonstrate a case where modeling richer data improves inference.

The dataset we use here as an illustrative example is the well-known "karate club" network (Zachary, 1977), which is a social network of friendships between $n = 34$ members of a karate club at a U.S. university. In this network, nodes represent members of the karate club and weighted edges indicate how often two members interact outside club activities. This network is well-studied and is known to split into two communities ($K = 2$) due to a dispute over whether to raise the club fee. Figure 3.1 shows the community labels estimated using binary data (presence/absence of interactions) and count data (strengths of interactions), respectively. Node 10 prefers to be classified into the community with node 3 using binary data while into the community with node 34 using count data. A further look at node 10 shows that it is connected to only two nodes—3 and 34—which are "hubs" or "popular members" in the two communities. Binary data fails to tell the

Table 3.1: Karate network: weighted adjacency matrix

| Actor | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | . | 4 | 5 | 3 | 3 | 3 | 3 | 2 | 2 | . | 2 | 3 | 1 | 3 | . | . | . | 2 | . | 2 | . | 2 | . | . | . | . | . | . | . | . | . | 2 | . | . |
| 2 | 4 | . | 6 | 3 | . | . | . | 4 | . | . | . | . | . | 5 | . | . | . | 1 | . | 2 | . | 2 | . | . | . | . | . | . | . | . | 2 | . | . | . |
| 3 | 5 | 6 | . | 3 | . | . | . | 4 | 5 | 1 | . | . | . | 3 | . | . | . | . | . | . | . | . | . | . | . | . | . | 2 | 2 | . | . | . | 2 | . |
| 4 | 3 | 3 | 3 | . | . | . | . | 3 | . | . | . | . | 3 | 3 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 5 | 3 | . | . | . | . | . | 2 | . | . | . | 3 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 6 | 3 | . | . | . | . | . | 5 | . | . | . | 3 | . | . | . | . | . | 3 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 7 | 3 | . | . | . | 2 | 5 | . | . | . | . | . | . | . | . | . | . | 3 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 8 | 2 | 4 | 4 | 3 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 9 | 2 | . | 5 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 3 | . | 3 | 4 |
| 10 | . | . | 1 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 2 |
| 11 | 2 | . | . | . | 3 | 3 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 12 | 3 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 13 | 1 | . | . | 3 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 14 | 3 | 5 | 3 | 3 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 3 |
| 15 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 3 | 2 |
| 16 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 3 | 4 |
| 17 | . | . | . | . | . | 3 | 3 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 18 | 2 | 1 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 19 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 1 | 2 |
| 20 | 2 | 2 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 1 |
| 21 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 3 | 1 |
| 22 | 2 | 2 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 23 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 2 | 3 |
| 24 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 5 | . | 4 | . | 3 | . | . | 5 | 4 |
| 25 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 2 | . | 3 | . | . | . | 2 | . | . |
| 26 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 5 | 2 | . | . | . | . | . | . | 7 | . | . |
| 27 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 4 | . | . | . | 2 |
| 28 | . | . | 2 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 4 | 3 | . | . | . | . | . | . | . | . | 4 |
| 29 | . | . | 2 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 2 | . | 2 |
| 30 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 3 | . | . | 4 | . | . | . | . | . | 4 | 2 | . |
| 31 | . | 2 | . | . | . | . | . | . | 3 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 3 | 3 |
| 32 | 2 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 2 | 7 | . | . | 2 | . | . | . | 4 | 4 |
| 33 | . | . | 2 | . | . | . | . | . | 3 | . | . | . | . | . | 3 | 3 | . | . | 1 | . | 3 | . | 2 | 5 | . | . | . | . | . | 4 | 3 | 4 | . | 5 |
| 34 | . | . | . | . | . | . | . | . | 4 | 2 | . | . | . | 3 | 2 | 4 | . | . | 2 | 1 | 1 | . | 3 | 4 | . | . | 2 | 4 | 2 | 2 | 3 | 4 | 5 | . |

difference between these two interactions while count data in Table 3.1 shows that node 10 shares a stronger interaction with node 34, thus should be classified into the community with node 34.

To address these hurdles, I generalize the SBM by making changes as follows.

1. I adopt a broader family of generalized regression models to fit count, categorical and ordinal data. The broader family may include Poisson, Gaussian, zero-inflated Poisson and response factor model. A better resolution is expected to be achieved through fitting richer data.

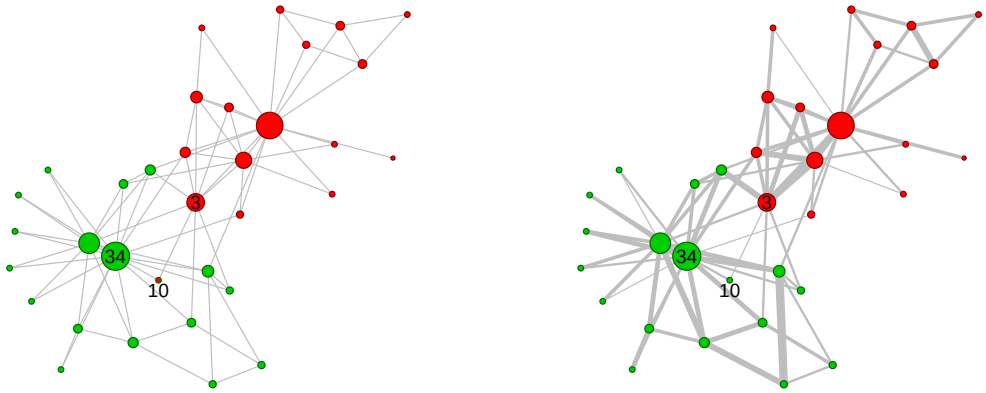


Figure 3.1: Karate network, $n = 34$. Node sizes are proportional to degree; node colors mark the estimator obtained using binary data (left) and that using count data (right). Node borders mark the reference given by Zachary. Edge widths indicate the weights on edges.

2. I model within-community interactions instead of between-community interactions to involve fewer parameters in a way that still captures the community behavior. By the definition of communities used in this dissertation, it is more likely to have an interaction within communities than between communities. Previously we required that $\binom{K}{2}$ log odds of between-community interactions to be non-positive while next we require $K$ log odds of within-community interactions be non-negative.

3. I make *group* corrections on groups of nodes instead of degree corrections on individual nodes and propose a simpler model where groups may correspond to the "popularity" or the core-periphery structure. Note that group corrections reduce to degree corrections when each node forms a "group" by itself.

4. I develop MAP estimators from an iterative optimization procedure, which is more computationally efficient compared with centroid estimators from Gibbs sampling.

The generalized SBM is discussed in Section 3.2 and Section 3.3. I then apply the model

to an amicus curiae network as an example of count data in Section 3.4. I also demonstrate its flexibility through fitting large-scale networks based on simulated benchmark networks and three real-world datasets with ground truth communities in Section 3.5. I conclude with a discussion in Section 3.6.

## 3.2 Group-Corrected Generalized SBM

Given a social network with $n$ individuals and observed interactions $A_{ij} \in \mathbb{N}$ between individuals $i$ and $j$, our goal is to identify $K$ clusters of nodes such that there are more interactions within clusters and fewer connections between clusters. This behavior is usually attributed to social assortativity—individuals with similar interests interact more intensively, "birds of a feather flock together"—and thus these clusters are called "communities". We can then see these social associations as a graph with $A$ as its adjacency matrix.

Following the approach in Chapter 2, we can consider a general Bayesian model to infer network assortativity parameters $\gamma$, degree correction terms $\eta$, and community labels $\sigma$ with

$$A_{ij} \,|\, \sigma, \eta, \gamma \sim \mathsf{F}\big[\mathrm{g}^{-1}(\gamma_{\sigma_i,\sigma_j} + \eta_i + \eta_j)\big].$$

Here, $\mathsf{F}$ represents a family of distributions where possible options are Bernoulli, Poisson and so on. g is the corresponding link function under $\mathsf{F}$. To guarantee identifiability, we set $\gamma_{ss} = 0$ for $s \in \{1, \ldots, K\}$, and to capture community behavior we set $\gamma_{rs} \leq 0$, $r \neq s$, w.p. 1 under the prior, that is, we expect the probability of an interaction between communities to be smaller than that within community.

This model has $K(K-1)/2 + n$ parameters, and is computationally infeasible for large $K$ and/or $n$. To alleviate this problem, we propose to amend the above model in two ways as follows.

1. We expect that, for large networks, individuals are exchangeable modulo certain criterion, for example "popularity" or centrality. Thus, we pool individuals into $L$ groups according to their criterion profiles, and assume that the degree distribution

within each criterion class is homogeneous. To this end, we assign a "group" label $Z_i$ to the $i$-th individual.

2. Previously, we set a different $\gamma_{rs} \leq 0$ w.p. 1 for each between-community interaction and $\gamma_{ss} = 0$ for all within-community interactions. We next reduce the number of $\gamma$ parameters by setting $\gamma_{rs} = 0$ for $r \neq s$, and requiring $\gamma_{ss} \geq 0$ w.p. 1 through the prior specification to keep the community behavior.

With these two changes, the model has $K + L$ parameters and is more computational amenable. The new likelihood is then

$$A_{ij} \,|\, \sigma, \gamma, \eta \overset{\text{ind}}{\sim} \mathsf{F}\left[\mathrm{g}^{-1}\left(\sum_{k=1}^{K} \gamma_k I(\sigma_i = \sigma_j = k) + \eta_{Z_i} + \eta_{Z_j}\right)\right]. \tag{3.1}$$

The prior distributions are

$$\begin{aligned} (\gamma, \eta) &\overset{\text{ind}}{\sim} \prod_k I(\gamma_k \geq 0)\mathsf{N}(0, \tau^2 I_{L+K}) \\ \sigma_i &\overset{\text{iid}}{\sim} \mathsf{MN}(1; \pi). \end{aligned} \tag{3.2}$$

Hyper-parameter $\tau^2$ can be chosen to be large to form a weakly informative prior. Similarly, $\pi$ informs about the expected size of the communities; for a flat prior we take $\pi = (1/K, \ldots, 1/K)$. In some networks, nodes with similar *centrality* properties are expected to behave similarly in building an interaction. An example of this is a network with *core-periphery* (CP) structure. CP structure is a common but informal notion in social network analysis which entails dense, cohesive *cores* and sparse, less-connected *peripheries* (Borgatti and Everett, 2000). It is worth mentioning that parameters $\eta$ indicate the local CP structure when nodes are grouped by centrality scores while parameters $\gamma$ capture the community structure. Both structures are important and irreplaceable by one another in the analysis of networks (Rombach et al., 2014). The differences between the CP structure and community structure can be found in Figure 3.2. The left and middle panels depict the adjacency of a network with global community structure and global

CP structure, respectively; the right panel shows the adjacency of a network with global community structure and local CP structure (or global CP structure and local community structure after reordering).

Additional covariates $y$ may also be included to model $A_{ij}$. An example is presented in Section 3.4.

$$\begin{aligned} A_{ij} \,|\, \sigma, \gamma, \eta, Y &\overset{\text{ind}}{\sim} \mathsf{F}\left[\mathrm{g}^{-1}\left(\sum_{k=1}^{K} \gamma_k I(\sigma_i = \sigma_j = k) + \eta_{Z_i} + \eta_{Z_j} + \sum_{s=1}^{S} \xi_s y_{is}[+/\cdot] y_{js}\right)\right], \\ (\gamma, \eta, \xi) &\overset{\text{ind}}{\sim} \prod_k I(\gamma_k \geq 0)\mathsf{N}(0, \tau^2 I_{L+K+S}). \end{aligned} \tag{3.3}$$

Figure 3.2: Simple illustrative networks with (i) global community structure; (ii)global core-periphery structure; (iii) global community structure and local core-periphery structure / global CP structure and global community structure

### 3.2.1 Parameter Identifiability

In what follows, to simplify the notation we set $\beta = (\gamma, \eta)$ and define the design matrix $X$ associated to model (3.1) such that

$$A_{ij} \,|\, \sigma, \beta \overset{\text{iid}}{\sim} \mathsf{Bern}\big(\text{logit}^{-1}(x_{ij}(\sigma)^\top \beta)\big).$$

Model (3.1) has then $K + L$ parameters, and the next result demonstrates the conditions required for the model to be identifiable (the proof is in Appendix 6.0.5).

**Theorem 4.** *The design matrix $X$ associated with model* (3.1) *is full column-ranked if each community has at least two nodes and*

*(1) $K > 2$; or*

*(2) none of the groups is completely covered by a community when $K = 2$.*

In practice, $K$ is chosen with some information criterion. The model is identifiable if the selected $K$ is greater than two. We only need to check if there is a group completely covered by one of the communities when $K = 2$ to guarantee model identifiability. It is worth pointing out that analyzing parameter identifiability is difficult for model (3.3) which includes covariates $y$ of various types and values in addition to community labels and group labels.

## 3.3 Model Inference

### 3.3.1 Initializing $Z$ and $\sigma$

Fitting community labels $\sigma$ and group labels $Z$ simultaneously is prohibitive due to combinatorial issues, especially when the groups and communities have overlaps. Moreover, it is difficult to determine the number of groups and the number of communities jointly. Hence, we adopt an optimization procedure with two passes: learning group labels $Z$ (as well as the number of groups) and then fitting labels $\sigma$. Heuristic methods are used to find good initials in our proposed procedure, as common in other SBM inference. Among the many ways to initialize $Z$, we settle with two main strategies stated below.

(i) *Probability of connecting*: We treat the connected network as an ergodic Markov chain and let $P = [p_{ij}]$ be the corresponding transition matrix where $p_{ij} = \frac{A_{ij}}{\sum_{\tilde{j}:(i,\tilde{j})\in E(G)} A_{i\tilde{j}}}$. We find the stationary distribution $\Pi$ such that $P\Pi = \Pi$ and expect that $\Pi_i$ roughly captures the probability of connecting with node $i$. We then perform hierarchical clustering on $\Pi$ and cut the hierarchical structure to form $L$ clusters for a range of $L$.

(ii) *Degree quantile*: In networks with degree assortativity, for example a network where hubs are more likely to be connected to hubs, quantiles of degrees are adopted to characterize initial groups. We define a *maximum* number $L_{\max}$ of popularity classes, ranking the observed degrees, and splitting them according to their quantiles. More precisely, if $d_i = \sum_{j \neq i} A_{ij}$ is the degree of $i$, we set $Z_i = j$ if $(j-1)/L_{\max} < \sum_v I(d_v < d_i)/n \leq j/L_{\max}$. Some degree strata might not have any nodes, and so the maximum popularity label $L$ might not be $L_{\max}$.

A heuristic approach based on the conductance is exploited to derive initial community labels. We use maximum spanning tree (MST) and cut the MST into $K$ parts by minimizing the "generalized" conductance for a range of $K$. The generalized conductance is approximated in MST by

$$\Phi = \max_k \Big\{ \frac{\sum_{i \in S_k, j \notin S_k, (i,j) \in T} P_{ij} \Pi_i}{\sum_{i \in S_k} \Pi_i} \Big\}$$

where $S_k$ are built by breaking edges in MST.

To estimate $\gamma$, $\eta$, $\sigma$ and $Z$ we explore a cyclic gradient descent method on the log posterior defined by (3.3) and (3.2) with three conditional steps:

$$[\gamma, \eta, \xi, \,|\, \sigma, Z, A] \qquad \text{and} \qquad [\sigma_i \,|\, \sigma_{[-i]}, \gamma, \eta, \xi, Z, A] \qquad \text{and} \qquad [Z_i \,|\, Z_{[-i]}, \gamma, \eta, \xi, \sigma, A],$$

where $\sigma_{[-i]}$ denotes all labels but the $i$-th one. The update step on $\sigma$ and $Z$ can get stuck in local maxima, and so we run this procedure from multiple starting points and select the fit with highest joint posterior probability. The following subsections explain these steps in detail.

### 3.3.2 Updating $\gamma$, $\eta$ and $\xi$

Conditional on community labels $\sigma$, we update $\gamma$, $\eta$ and $\xi$ using a ridge-regularized version of IRLS, an efficient and commonly used method when fitting generalized linear

models. In this case, we have a design matrix $X$ such that $A_{ij} \,|\, \sigma, \gamma, \eta, \xi \sim \mathsf{F}(\mu_{ij})$ with $\mu_{ij} = \mathrm{g}^{-1}(x_{ij}(\sigma)^\top \beta)$ and $\beta = (\gamma, \eta, \xi)$ according to (3.3). That is,

$$x_{ij}(\sigma)^\top \beta = \sum_{k=1}^{K} \gamma_k I(\sigma_i = \sigma_j = k) + \eta_{Z_i} + \eta_{Z_j} + \sum_s \xi_s y_{is}[+/\cdot] y_{js}.$$

Then, defining $W \doteq \mathrm{Diag}[\mathrm{Var}(\mu_{ij})]$, the update is $\beta^{(t+1)} = VX^\top W z^{(t)}$, with

$$V = \left( X^\top W X + \frac{1}{\tau^2} I_{K+L} \right)^{-1}$$

as the covariance and $z^{(t)} = X\beta^{(t)} + W^{-1}(A - \mu)$ as the "working response". In addition, to guarantee that $\gamma_{ss} \geq 0$ for every community $s$, we use an active-set method when updating $\beta$ (Nocedal and Wright, 2006). Instead of allocating the whole adjacency matrix $A$ and perform a GLM, we compute the sufficient statistics used in the update procedure which makes this step much more computational efficient and suitable for large-scale networks.

### 3.3.3 Updating $\sigma$ and $Z$

Now, given the updated values of $\beta$, we seek to update $\sigma$. A group update as in the previous step is however not possible, so we update each label $\sigma_i$ in turn, conditional on the remaining labels $\sigma_{[-i]}$ and model parameters $\beta$. From (3.3), we have that

$$\mathbb{P}(\sigma_i = k \,|\, \sigma_{[-i]}, \beta, A) \propto \pi_k \prod_{j \neq i} \frac{\exp\{A_{ij} x_{ij}^\top \beta\}}{1 + \exp\{x_{ij}^\top \beta\}}.$$

In practice, we do not compute the product above at each iteration but instead keep track of sufficient statistics when tentatively assigning $\sigma_i = k$ for $k = 1, \ldots, K$. We then pick $\sigma_i^{(t+1)}$ as the argument maximizer of $\mathbb{P}(\sigma_i \,|\, \sigma_{[-i]}^{(t)}, \beta^{(t)}, A)$ and update the sufficient statistics accordingly. We will not go into details of updating $Z$ since the procedure is very similar to that updating $\sigma$. The only difference is that we check for different identifiable constraints when updating $Z$.

| | |
|---|---|
| number of nodes | $3,512$ |
| number of edges | $54,134$ |
| edge density | 0.0088 |
| simple graph | TRUE |
| (mean, median, max, SD) degree | $(30.83, 19, 234, 30.75)$ |
| (mean, median, max, SD) adjacency | $(0.01, 0, 59, 0.15)$ |
| global clustering coefficients | 0.79 |

Table 3.2: Amicus curiae network: summary statistics of largest connected component

Table 3.3: Amicus curiae network: table summary of weighted adjacency matrix

| weight | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| number of edges | 6111182 | 46942 | 5389 | 1236 | 298 | 119 | 52 | 31 | 13 | 3 | 5 | 6 | 4 | 3 | 1 | 3 | 1 | 4 |
| weight | 21 | 22 | 28 | 30 | 31 | 34 | 35 | 36 | 37 | 38 | 40 | 41 | 44 | 45 | 47 | 49 | 50 | 59 |
| number of edges | 1 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 |

## 3.4 Case Study: Amicus Curiae Network

### 3.4.1 Background

In political science, an interest group refers to a formally organized association that seeks to influence public policy and other political outcomes (e.g., case decisions, contracts and appointments). Interest groups often make their opinion known before the U.S. Supreme Court through signing an *amicus curiae brief*. Interest groups usually work together by cosigning on amicus curiae briefs to achieve their goals at reduced cost and receive symbolic benefits (Hula, 1999; Box-Steffensmeier and Christenson, 2014). Such coalitions form an amicus curiae network, where nodes represent interest groups and edges indicate cosigner status. We obtain an amicus curiae network by considering interest groups active in the decade from 2000 to 2010 as nodes and the number of briefs cosigned by interest groups as weighted edges. The amicus curiae network measures not only interest groups coalitions, but also intensities of coalitions. The amicus curiae network consists of a highly connected component and other small isolated clusters of nodes. Of interest is the largest connected component with $n = 3,512$ nodes and $e = 54,134$ edges.

### 3.4.2 Zero-Inflated Poisson SBM

Poisson regression models are commonly used in modeling counts on edges. However, a strong assumption for Poisson regression is that the event's conditional mean and variance are equal. In practice, the phenomenon that counts having greater variance than the mean is often observed. And such phenomenon is described as overdispersion, which indicates that Poisson regression is not adequate. The descriptive analysis shown in Table 3.2 and Table 3.3 suggests the presence of excess zeros and large variance in the amicus curiae network, leading to overdispersion if fitted with Poisson. We address the problem by adopting a zero-inflated Poisson (ZIP) model, a simple mixture model for count data with excess zeros (Lambert, 1992). The model is a combination of a Poisson distribution and a degenerate distribution at zero.

Hence, we propose a two-state model: (i) a latent level binary network $B_{ij}$ modeling the linkage between interest groups $i$ and $j$; (ii) a main level weighted network $A_{ij}$ modeling strengths of connections. In addition, we consider other factors that help in explaining interest group coalitions. Among the many business characteristics of an interest group, industry is of significance since it is a measure of shared issue interests. Interest groups in the same industry are expected to cosign amicus curiae briefs more often because they share industrial demands to seek out mutual benefits via corporation (Box-Steffensmeier and Christenson, 2014). We measure industry by the associated U.S. Standard Industrial Classification (SIC) system available at `https://www.osha.gov/pls/imis/sic_manual.html`. We parse SIC codes in accordance with major divisions but do split one major division "Services" and its major groups such as "Membership Organization" to make sure the obtained industry groups are of moderate size. Sharing common membership and SIC codes may contribute to coalitions as well as frequencies of coalitions. Hence, communities and SIC codes are expected to play a role in modeling both latent and main network. The

proposed generalized SBM in (3.3) is adjusted as:

$$
\begin{aligned}
A_{ij} \,|\, B_{ij}, \sigma, \theta, \zeta, \rho, Y &\sim B_{ij} \cdot \mathsf{Poisson}\left( \sum_{k=1}^{K} \theta_k I(\sigma_i = \sigma_j = k) + \zeta_{Z_i} + \zeta_{Z_j} + \sum_{s=1}^{S} \rho_s y_{is} y_{js} \right), \\
B_{ij} \,|\, \sigma, \gamma, \eta, \xi, Y &\sim \mathsf{Bern}\left[ \mathrm{logit}^{-1}\left( \sum_{k=1}^{K} \gamma_k I(\sigma_i = \sigma_j = k) + \eta_{Z_i} + \eta_{Z_j} + \sum_{s=1}^{S} \xi_s y_{is} y_{js} \right) \right].
\end{aligned} \tag{3.4}
$$

Notation $y_{is}$ is the indicator that interest group $i$ belongs to industry $s$. Note that we need $S$ 0-1 indicators since one interest group may belong to more than one industry. There is expected to be a "boost" effect in both the probability of a coalition and the intensity of the coalition if two interest groups are in the same industry. The prior distributions are

$$
\begin{aligned}
(\theta, \zeta, \rho) &\overset{\text{ind}}{\sim} \prod_k I(\theta_k \geq 0) N(0, \nu^2 I_{L+K+S}) \\
(\gamma, \eta, \xi) &\overset{\text{ind}}{\sim} \prod_k I(\gamma_k \geq 0) N(0, \tau^2 I_{L+K+S}) \\
\sigma_i &\overset{\text{iid}}{\sim} \mathsf{MN}(1; \pi).
\end{aligned} \tag{3.5}
$$

Hyper-parameters $\nu^2$ and $\tau^2$ can take on large values to form weakly informative priors. Parameters $\eta, \zeta$ are expected to capture the coreness of the latent and main network, respectively. Parameters $\gamma$ and $\theta$ capture the community structure of the latent and main network, respectively. The inference on ZIP models is similar to that on one-state models as described in Section 3.3.

### 3.4.3 Results

Figure 3.3 shows the relation between industry and interest group coalitions. The higher darkness of diagonal terms relative to that of off-diagonal terms indicates that interest groups in the same industry tend to cosign more frequently in this amicus curiae network. Given this result, we always take industry into account when performing inference procedures later. We also observe that the edge densities within agriculture industry is

Table 3.4: Amicus curiae network: parsed industries

| | | | |
|---|---|---|---|
| 1 | A_Agriculture | 12 | I_EduServices |
| 2 | B_Mining | 13 | I_SocialServices |
| 3 | C_Construction | 14 | I_MbspBusiness |
| 4 | D_Manufacturing | 15 | I_MbspProfessional |
| 5 | E_Transportation | 16 | I_MbspLabor |
| 6 | F_WholesaleTrade | 17 | I_MbspCivic |
| 7 | G_RetailTrade | 18 | I_MbspPolitical |
| 8 | H_Finance | 19 | I_MbspReligious |
| 9 | I_GeneralServices | 20 | I_MbspOther |
| 10 | I_HealthServices | 21 | I_OtherServices |
| 11 | I_LegalServices | 22 | J_PublicAdmin |

very high and the sum of its between-industry densities is low, indicating that agricultural organizations work very closely with each other while rarely form coalitions with interest groups in other industries.

We partition interest groups into $L$ classes according to the connecting probabilities. To determine the number of classes, we perform a GGLM on the latent network with industry incorporated and select $L$ based on *posterior predictive loss* (PPL). PPL penalizes the departure from "fit" and "smoothness", and is commonly used in Bayesian analysis.

$$PPL_k = \frac{k}{k+1}G + P, \quad G = \sum_{l=1}^{n}(\mu_l - y_{l,\text{obs}}), \quad P = \sum_{l=1}^{n}\sigma_l^2,$$

where $\mu_l = \mathrm{E}[Y_{l,\text{rep}}|y]$ and $\sigma_l^2 = \mathrm{Var}[Y_{l,\text{rep}}|y]$.

Based on Figure 3.4, we choose $L = 6$. Similarly, the number of communities $K$ is chosen to be three. The left panels of Figure 3.5 depict the inferred communities and groups. The middle panels of Figure 3.5 show estimated degrees against degrees for the main and latent network, respectively. The right panels of Figure 3.5 demonstrate the heat maps of the expected adjacency matrix versus the observed adjacency matrix for the main and latent network, respectively. We observe from the heat maps that there is a local core-periphery structure besides community structure. Our model successfully captures the pattern of the network in general while tends to underestimate the strengths of a few connections. Some other factors are needed to fully explain the strong strengths of

Table 3.5: Amicus curiae: community & industry

| | Community 1 | Community 2 | Community 3 |
|---|---|---|---|
| G 6 | conservative(economic), agriculture | liberal(civil rights) | conservative(moral) |
| G 5 | conservative(more general economic), corporate interest | liberal(less issue-focused) | conservative(moral) |
| G 4 | conservative(media, oil, computer, finance) | liberal | conservative(moral) |
| G 3 | conservative | liberal | NA |
| G 2 | conservative | mostly liberal | NA |
| G 1 | conservative | mostly liberal | NA |

these coalitions, such as the issue area of an brief. Figure 3.6 presents the estimated 95% credible intervals for industry and community coefficients in the latent and main network. All industries and communities play an significant role in forming a coalition while only half is still significant in determining strengths of coalitions.

We are mostly interested in the inferred community labels. Table 3.5 is a list of what interest groups in the same community and group share in common. Communities appear to explain political ideologies to some extent. Interest groups with the highest coreness in the first community are conservative in the economic aspect. All such interest groups are agricultural organizations seeking anti-regulation from the government. Interest groups in the third community are mostly medical organizations, which are also conservative but in the moral aspect. Interest groups with the highest coreness in the second community are powerful liberal unions related to civil rights.

## 3.5 Application to Large-Scale Networks

In this section, we evaluate the performance of our proposed MAP estimator on both simulated benchmark networks and large-scale real-world networks. Similarly as in Section 2.6 we compare it to KN, FG, ML, WT and LP through an empirical study in terms of the NMI. Both benchmark networks and the large-scale real-world networks we use here involve binary data only, thus we apply the generalized SBM with $\mathsf{F} = \mathsf{Bern}$ and $\mathrm{g} = \mathrm{logit}$.

### 3.5.1 Empirical Study

Our empirical study generates networks with ground truth from a popular benchmark suite that accounts for heterogeneities in node degree distributions and community sizes (Lancichinetti et al., 2008b). The model generating networks considers the following parameters: assumed to follow power law distributions with exponents 2 and 1, respectively; the network consists of $n = (500, 1000)$ nodes and has average degree $\langle k \rangle = (10, 15, 25)$. Mixing parameter $\mu = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6)$ captures the proportion of between-community edges. The ratio between $n$ and the maximum degree controls the size of the communities relative to that of the network. We highlight two community structures formed by relatively large communities with ratio = 2, and relatively small communities with ratio = 10. The parameter "ratio" marks the difference between these benchmark networks and that in Chapter 2, and is introduced to control the two types of community structures.

We generate 100 networks for each combination of the parameters mentioned above and assume that the number of group classes is $\lfloor n/10 \rfloor$. The NMI of other estimators (KN, FG, ML, WT, LP) and our proposed MAP estimator is summarized in Figure **??**. We can conclude from the figure that the MAP estimator outperforms the other estimators on average in terms of the NMI, especially when the network is formed by relatively large communities. Not surprisingly, all estimators perform worse as the mixing parameter $\mu$ increases (so that the communities are defined in a weaker sense) or the average degree $\langle k \rangle$ decreases. Besides, our community detection procedure outperforms Karrer and Newman's significantly in computational time as shown in Figure **??**. Our MAP estimator is as computational efficient as FG, ML, WT estimators. While LP estimator beats all other estimators in terms of running time, it leads to the lowest NMI on average.

It is also worth pointing out that the NMI is less powerful in comparing communities as $K$ increases, as shown in column "ratio = 10" that the NMI is unusually high (above 0.75). Consider a simple illustration where

$$\begin{array}{lcccccccc} \sigma\text{:} & 1 & 1 & 2 & 2 & \dots & \dots & K & K \\ \tilde{\sigma}\text{:} & 1 & 2 & \dots & K & 1 & 2 & \dots & K. \end{array}$$

$\mathrm{NMI}(\sigma, \tilde{\sigma}) = 1 - \log 2 / \log K$ approaches 1 as $K$ increases while the actual labels might be quite different. Similar loss of power happens to some other measures, such as Binder's loss and adjusted Rand Index.

### 3.5.2 Case Study

Next, we evaluate our estimator for community detection on a collection of large-scale real-world network datasets with ground-truth communities (Yang and Leskovec, 2012). We consider three networks: an online social network *Youtube*, where nodes represent the users of Youtube, edges indicate the friendship formed by the users and ground-truth communities are defined by the user-defined interest groups; a co-authorship network *DBLP* where nodes represent authors published in a comprehensive list of research papers in computer science, edges indicate co-authorship in at least one paper and ground-truth communities are defined by Publication venues; a product co-purchasing network *Amazon*, where nodes represent products sold on Amazon website, edges indicate frequently co-purchase and ground-truth communities are defined by product categories provided by Amazon. All datasets are publicly available at `http://snap.stanford.edu/data/`.

An interesting phenomenon we discovered when processing the ground truth communities is that the difference in community sizes is huge, ranging from two to hundreds of thousands. A large proportion of the communities are negligible given the existence of the top largest communities. How to effectively learn the significant portions of networks and shrink the number of communities is fundamental to community detection on large-scale networks. First, we order the communities by their sizes $|C_{(i)}|, i = 1, \dots, K$ and consider the cumulative size of the largest $k$ communities $|\cup_{i \le k} C_{(i)}|$ (increasing trend) as well as the sequence $\{|C_{(i)}|, i = 1, \dots, k\}$ (decreasing trend) for some $k \le K$. We then pick the number of communities $k^\star$ that appears after the elbow of the decreasing trend while maintains most of the nodes. Figure 3.9 shows an example on the DBLP co-authorship network

where $k^\star = 1,000$ is a good choice. Then we consider the induced sub-graph generated by the nodes in the largest $k^\star$ communities and further shrink the number of communities by merging communities that are closely connected by conducting a hierarchical clustering. We regard each connected component in a group as a separate ground-truth community and provide an analysis on the largest connected component.

For each network, we choose a set of different numbers of popularity and carry out our proposed MAP estimation procedure. Figure 3.10 visualizes the edge densities within communities and between communities based on the MAP estimator. It is evident that the edge densities are greater within communities than between communities. Since KN estimator used for comparison in Section 3.5.1 is infeasible on real-world large networks within reasonable amount of time, we only make comparisons with some fast algorithms for community detection. The top panels in Figures 3.11, 3.12, and 3.13 plot the estimated $\eta_i$ (box plots) and the average normalized degrees in logit scale (points) against the popularity class; it is clear that $\eta$ is closely related to the degree of nodes. The remaining panels in Figures 3.11, 3.12, and 3.13 show a comparison among FG, LP, ML, WT estimators, our MAP estimator and randomly generated labels in terms of NMI. We conclude that our MAP estimator performs comparably well as FG, LP, ML, WT estimators on real-world large-scale networks while outperforms the random labels.

## 3.6 Discussion

The number of communities $K$ is first fixed and then selected under certain model selection choice in the approach proposed in this chapter. Possible extension of this work may include a procedure that efficiently models $K$ jointly with other parameters. Other directions for future work, albeit not related to community detection, include proposing more powerful measures for comparing communities, especially when the number of communities is large relative to the network.

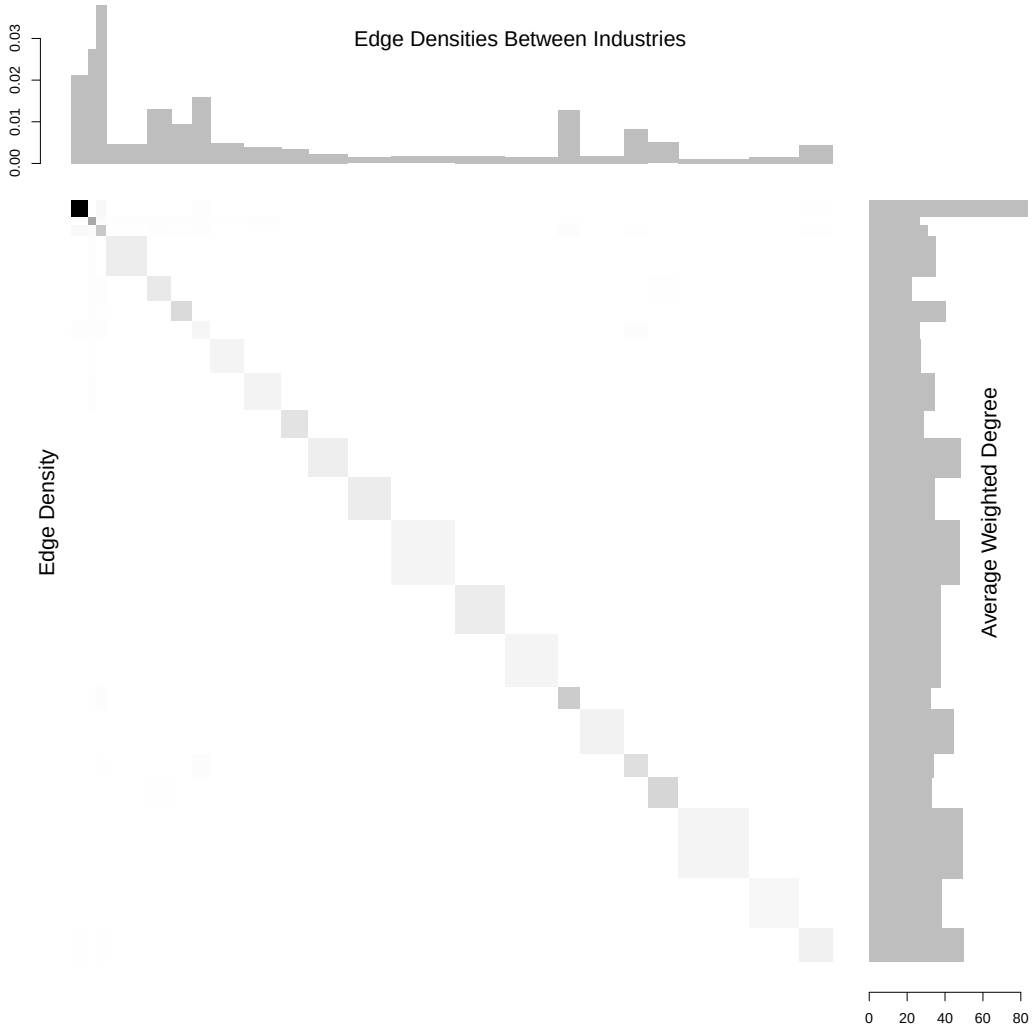


Figure 3.3: Diagonal/off-diagonal terms of the heat map represent edge densities within/between industries. The darker the color, the larger the edge densities. The number of nodes in each industry is proportional to the area of the corresponding square along the diagonal. The top bar graph shows the between-industry edge densities. The right-rail bar graph indicates the average weighted degree of nodes in each industry.

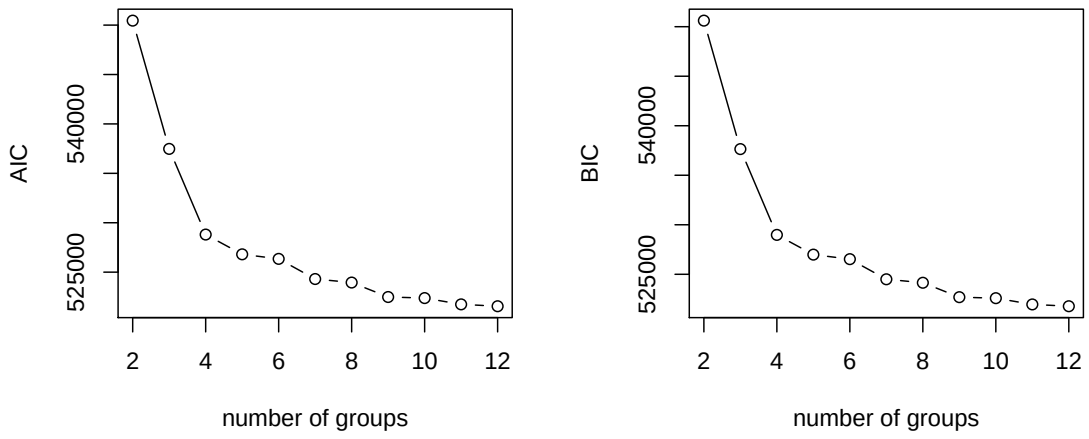


Figure 3.4: Edge PPL and degree PPL under the logistic regression fitted against industry and coreness classes

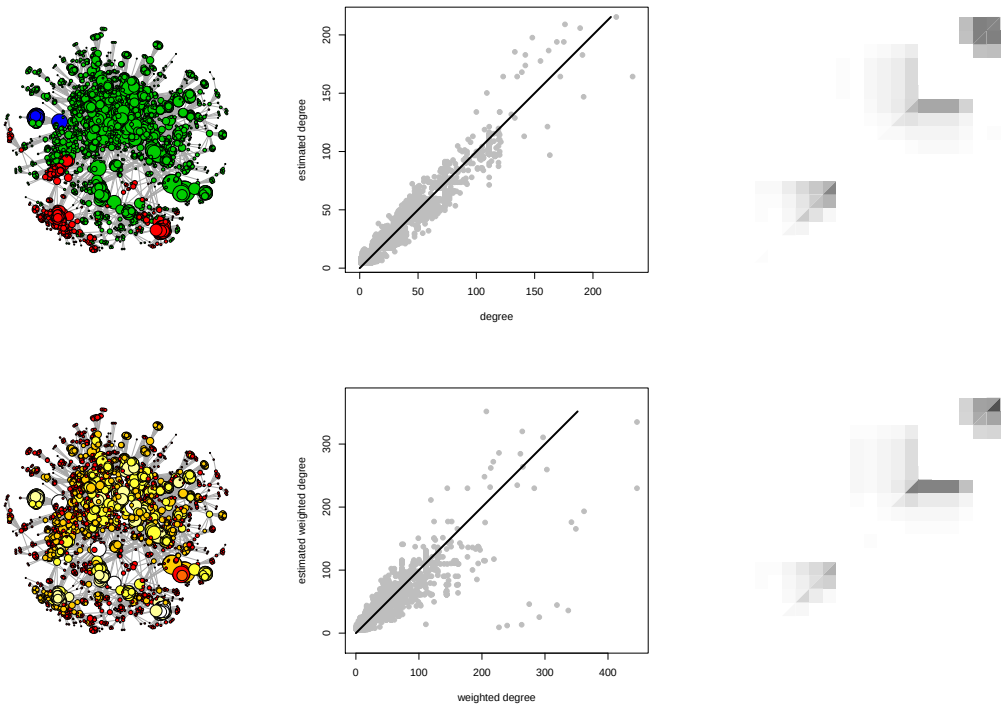


Figure 3.5: Left: (top) MAP estimate where node colors represent inferred communities; (bottom) MAP estimate where node colors indicate groups, coloring red to white from less powerful to powerful. Middle: (top) degree v.s. estimated degree plot; (bottom) weighted degree v.s. estimated weighted degree plot. Right: (top) estimated weighted adjacency matrix v.s. weighted adjacency matrix; (bottom) estimated adjacency matrix v.s. adjacency matrix, where nodes are order by key value pair (community, group).

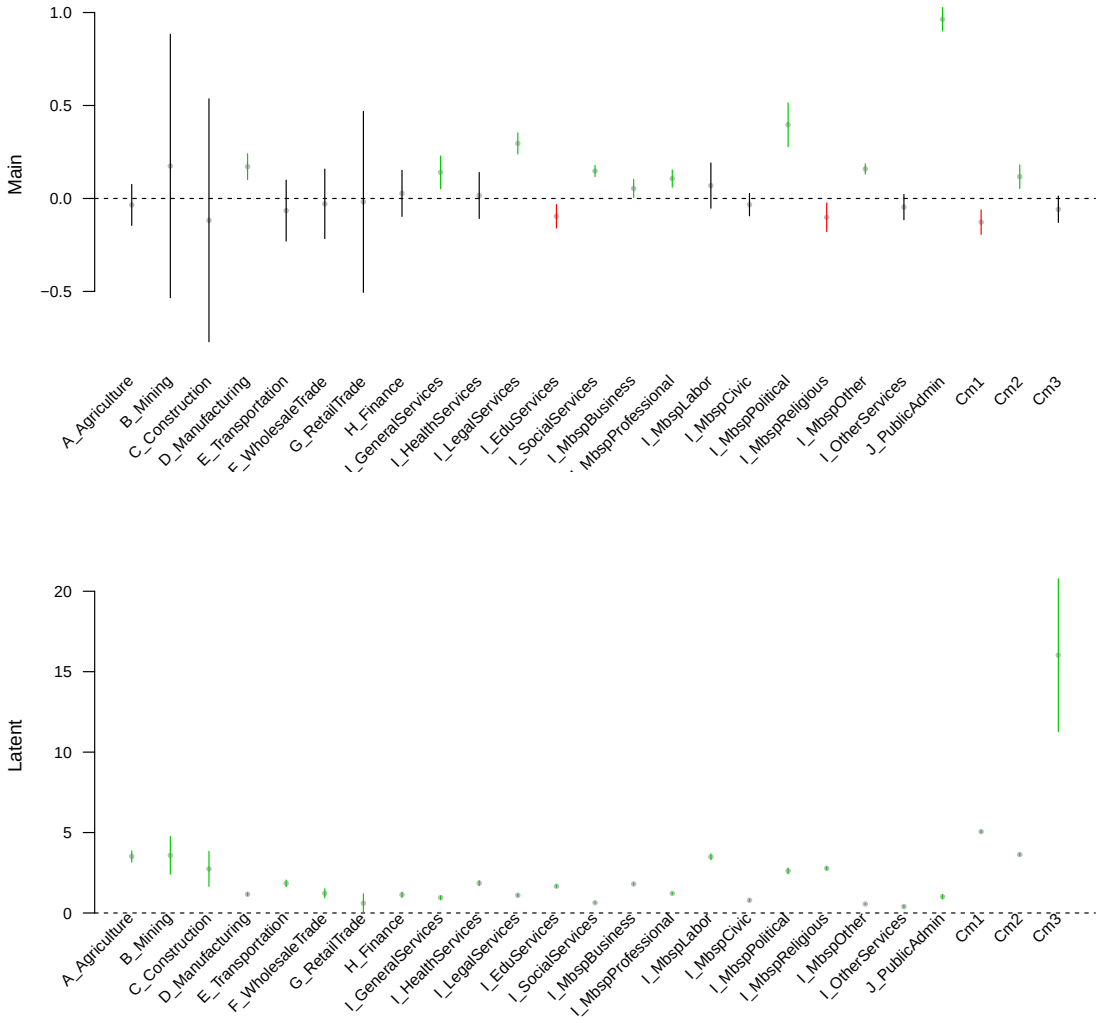


Figure 3.6: 95% credible interval of industry and community coefficients in both the main level and the latent level.

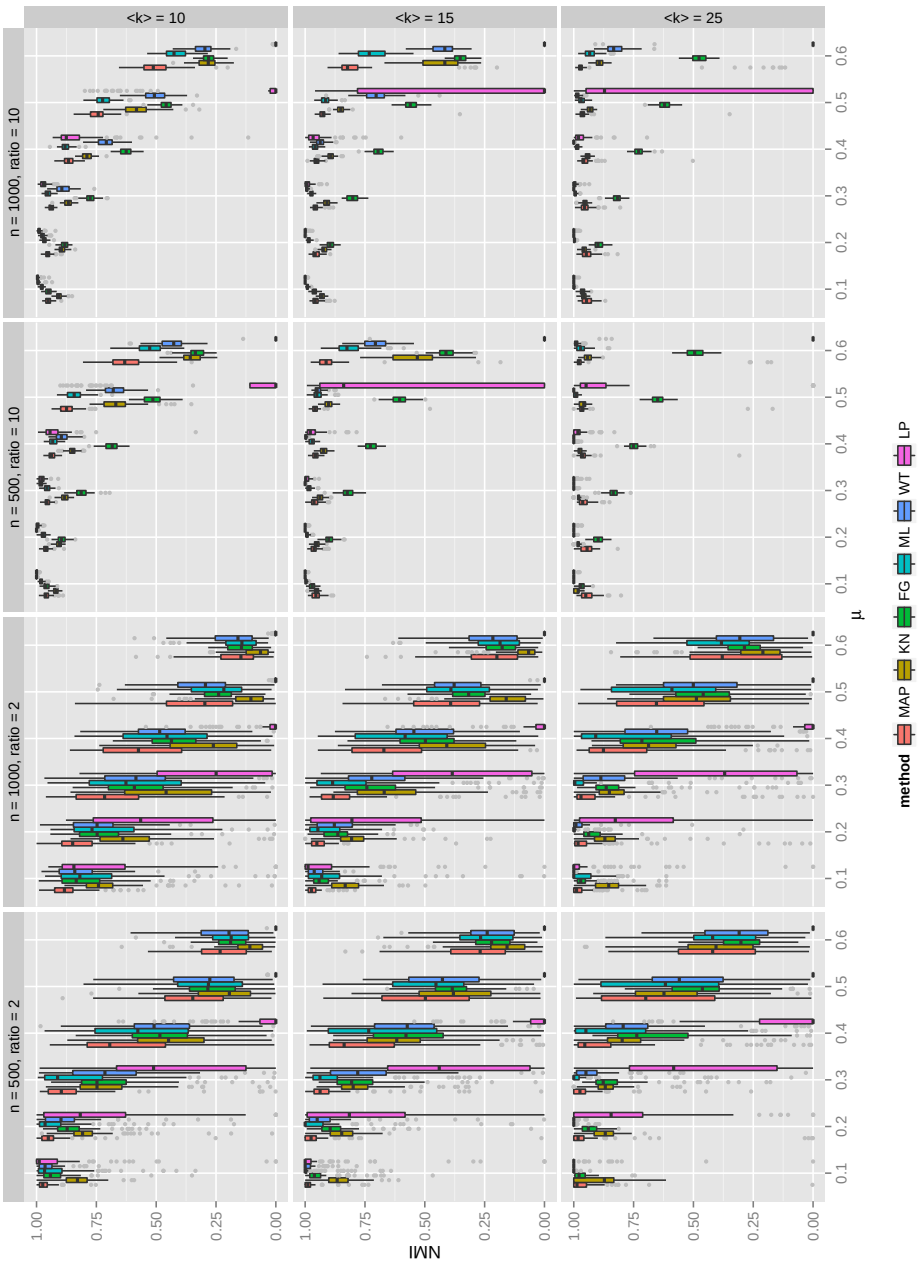


Figure 3.7: Comparison of our proposed MAP estimator and KN, FG, ML, WT and LP estimators in terms of the NMI. Benchmark networks of $n = 500$ and 1000 nodes, with different combinations of the average degree $\langle k \rangle$ and the ratio controlling the relative size of communities are used for comparison. Each box plot corresponds to the NMI of the estimator over 100 graph realizations.

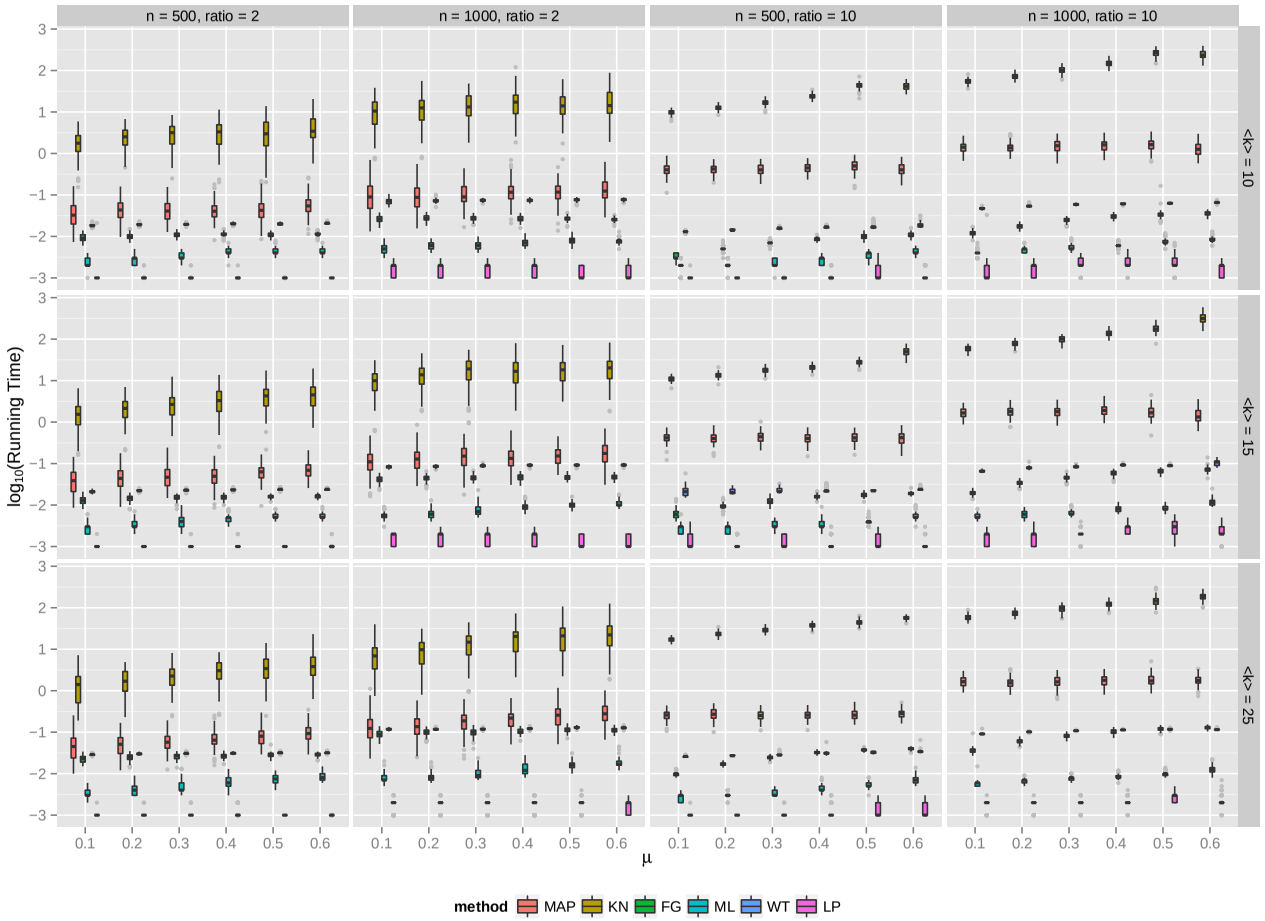


Figure 3.8: Comparison of our proposed MAP estimator and KN, FG, ML, WT and LP estimators in terms of the running time. Benchmark networks of $n = 500$ and 1000 nodes, with different combinations of the average degree $\langle k \rangle$ and the ratio controlling the relative size of communities. Each box plot corresponds to the $\log_{10}$(runtime) of the estimator over 100 graph realizations.

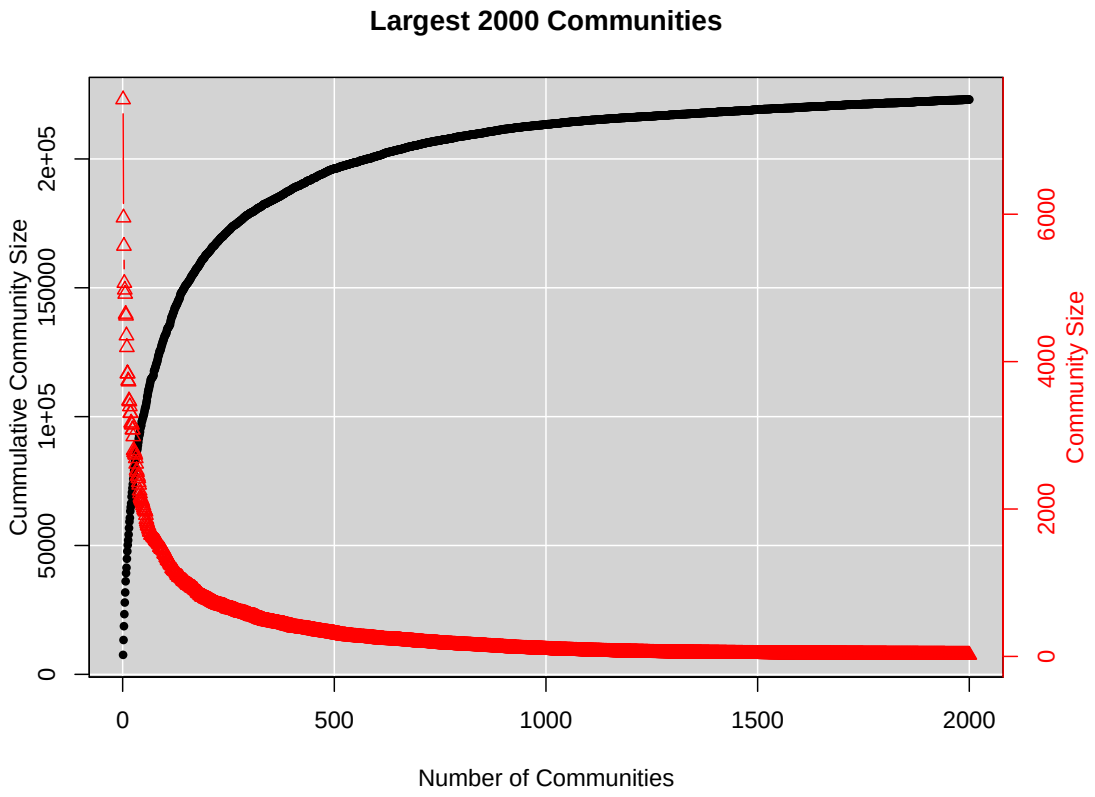


Figure 3.9: Co-authorship network *DBLP* with $k = 2,000$. Red: the decreasing sequence $\{|C_{(i)}|, i = 1, \ldots, k\}$; black: the cumulative size of the largest $k$ communities.

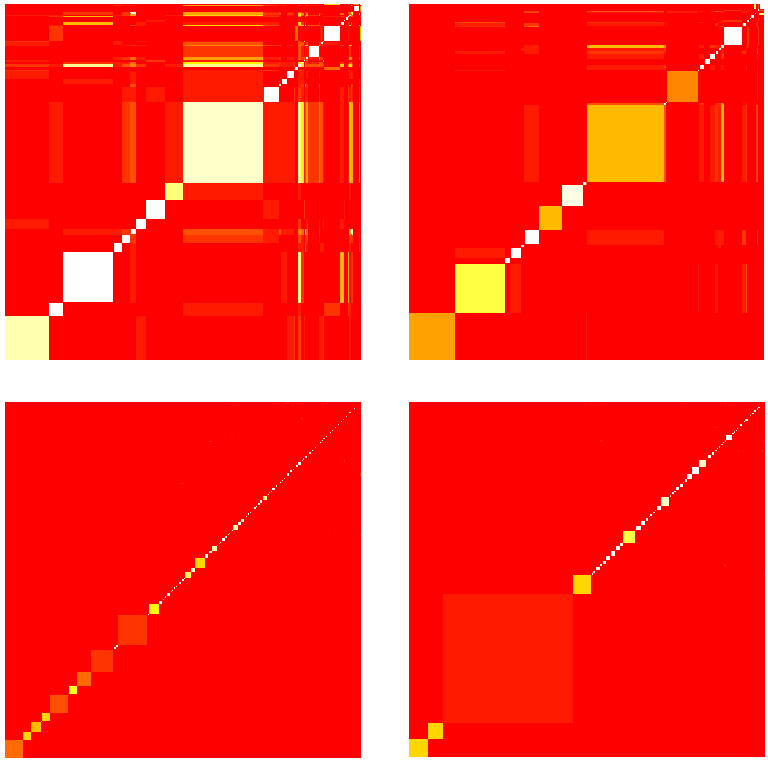

Figure 3.10: The community-by-community heat maps showing the edge densities within communities and between communities based on the MAP estimator. The left two plots correspond to the Youtube network under the smallest and largest $L$. The right two plots correspond to the DBLP and Amazon networks under the largest $L$. Red indicates low edge densities while white indicates high edge densities.

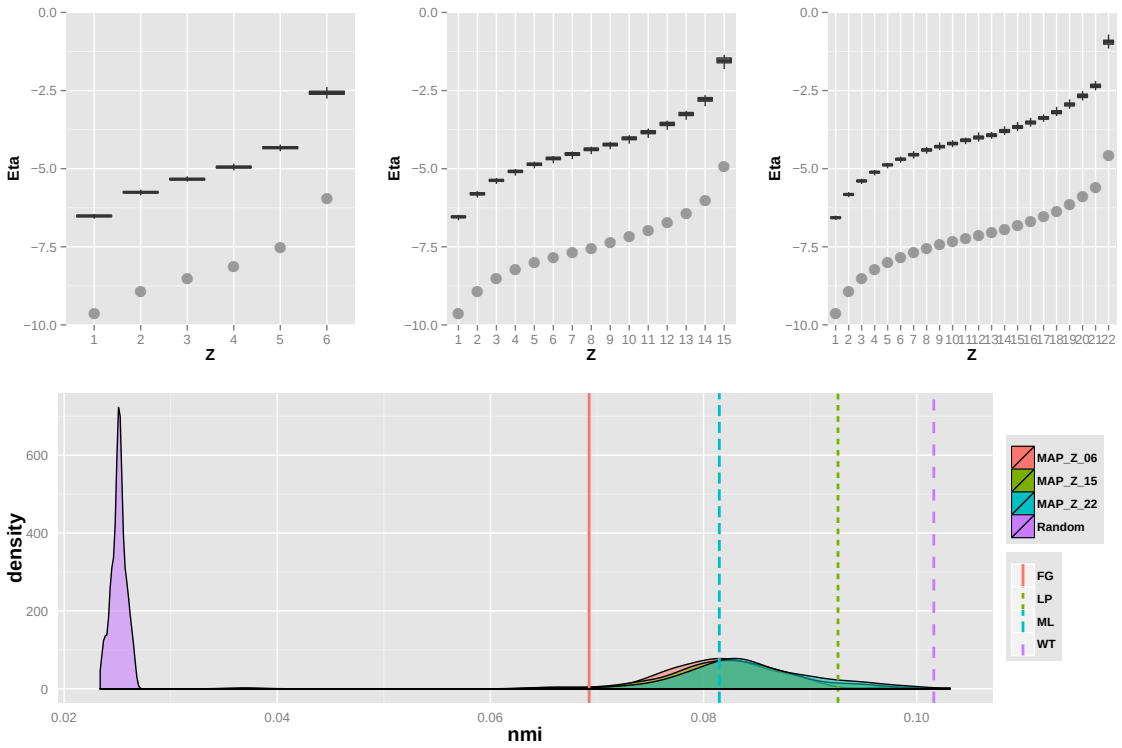


Figure 3.11: Online social network *Youtube*. Top: $\eta_i$ (box plots) and $\text{logit}(\overline{\text{degree}_i}/(n-1))$ (points) for each popularity class $i$. Bottom: the NMI of randomly generated labels, MAP estimates under different number of popularity classes, FG, LP, ML and WT estimates relative to the ground truth.

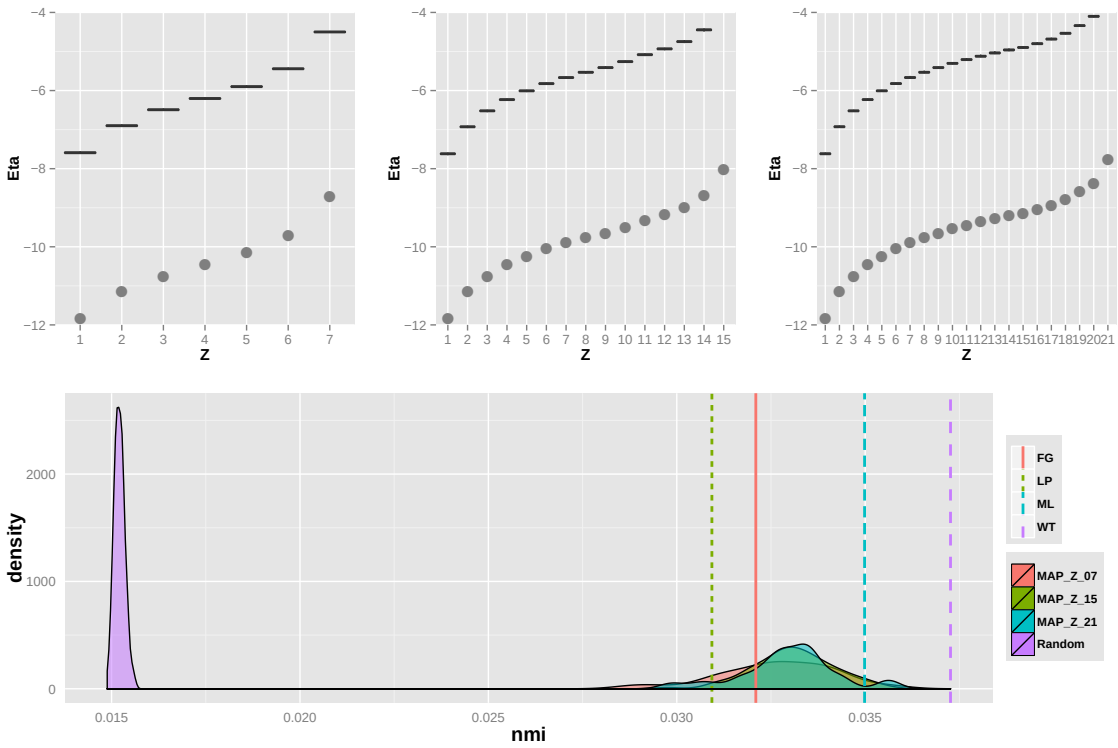


Figure 3.12: Co-authorship network *DBLP*. Top: $\eta_i$ (box plots) and $\text{logit}(\overline{\text{degree}_i}/(n-1))$ (points) for each popularity class $i$. Bottom: the NMI of randomly generated labels, MAP estimates under different number of popularity classes, FG, LP, ML and WT estimates relative to the ground truth.

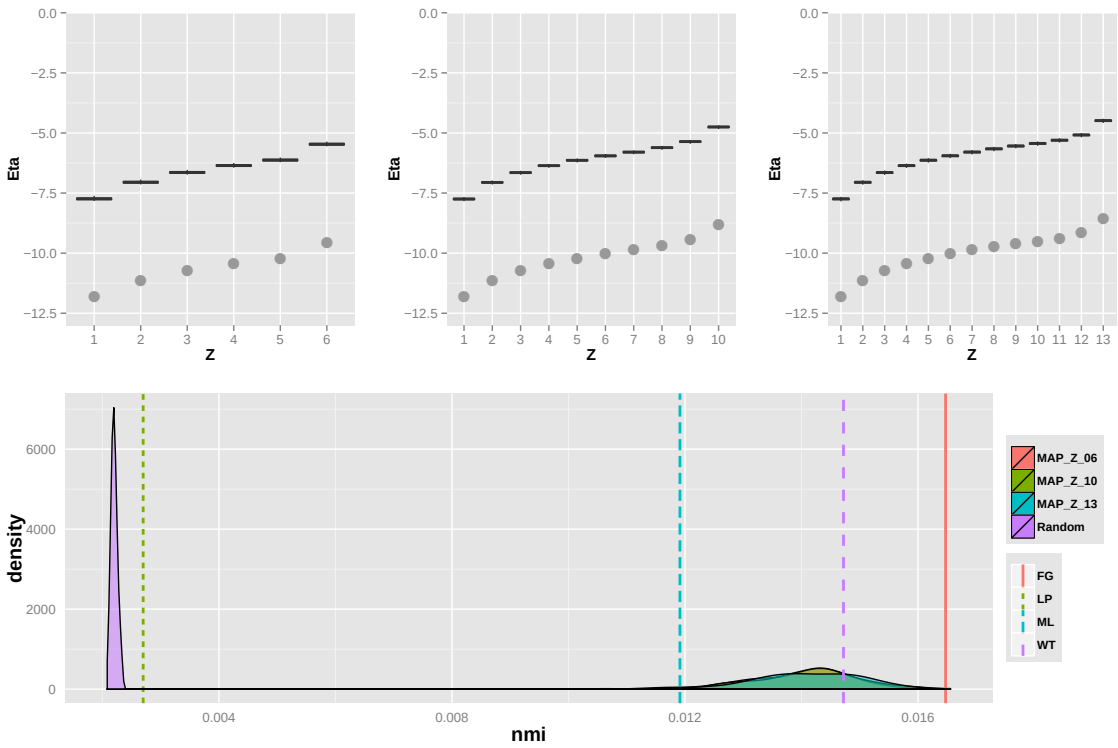


Figure 3.13: Product co-purchasing network *Amazon*. Top: $\eta_i$ (box plots) and $\text{logit}(\overline{\text{degree}_i}/(n-1))$ (points) for each popularity class $i$. Bottom: the NMI of randomly generated labels, MAP estimates under different number of popularity classes, FG, LP, ML and WT estimates relative to the ground truth.

# Chapter 4

# Ridge-Regularized Covariance Selection

In this chapter, we focus on covariance selection (introduced in Section 1.2), where networks describing interactions between nodes are unknown and to be inferred rather than given as data. Our motivation to covariance selection is driven by a dataset in periodontitis study. Periodontitis is the inflammation of tissues surrounding the teeth, and is caused by specific bacteria. These bacteria form a latent network, in which an edge indicates that the two connected bacteria share some common biological functions. Moreover, these bacteria often explore symbiotic relations, and are thus expected to be found in communities. We observe the ribonucleosomal expression level of each bacterium as data. Our goal is to identify the latent network as well as detect the bacteria communities based on the observed expression levels of each bacterium.

Our main contribution in this project is to jointly estimate concentration matrices and latent networks while taking community structure into account. To this end, we propose a Bayesian approach with a hierarchical prior with two levels in Section 4.1:

1. We develop a Bayesian ridge-regularized covariance selection that specifies a spike-and-slab prior on each off-diagonal entry of the concentration matrix. With this approach, we are able to obtain a positive-definite estimate of the concentration matrix and determine the underlying network simultaneously. We relate covariance selection and variable selection for Gaussian graphical models through an efficient algorithm in Section 4.2.

2. We offer a Bayesian approach for community detection that explicitly characterizes

community behavior and a MAP estimator to efficiently conduct inference in Section 4.2.

Results from a simulation study comparing our ridge-regularized covariance selection to other methods are reported in Section 4.3. We show that our proposed method is efficient and as reliable as other commonly used methods. A real-world meta-genomic dataset of complex microbial biofilms is used to demonstrate the covariance selection as well as community detection in Section 4.4. Finally, we offer some concluding remarks and directions for extension in Section 4.5.

## 4.1 Model Framework

We develop a hierarchical model to (i) perform covariance selection on a latent network of associations between individuals and (ii) identify the set of communities to which these individuals belong. We start by assuming that the data $X = (X_1, \ldots, X_n)$ for each sample follows

$$X_i \,|\, \mu, C \overset{\text{iid}}{\sim} \mathsf{N}(\mu, C^{-1}), \quad i = 1, \ldots, n, \tag{4.1}$$

where each $X_i$ is $p$-dimensional. The mean $\mu$ can have more structure than a single vector, as we will see in Section 4.4. We set an non-informative prior on $\mu$, $\mathbb{P}(\mu) \propto 1$.

Equation (4.1) implicitly defines a Gaussian graphical model on a undirected graph with $p$ nodes and adjacency matrix $A$. Recall that in a Gaussian graphical model, node $i$ and node $j$ are conditionally independent ($C_{ij} = 0$) if and only if there is no edge between them ($A_{ij} = 0$). To select which off-diagonal entries in $C$ are zero we adopt a spike-and-slab prior (George and McCulloch, 1993; Ishwaran and Rao, 2005) with $A$ as indicators:

$$C_{ij} \,|\, A_{ij} \overset{\text{ind}}{\sim} \mathsf{N}\big(0, \rho^2 A_{ij} + \rho^2 \nu_0 (1 - A_{ij})\big), \quad i, j = 1, \ldots, p, i < j, \tag{4.2}$$

where $\rho^2$ is chosen to be large (the "slab") while $\nu_0$ is small (the "spike".) For the diagonal

entries we set

$$C_{ii} \,|\, \lambda \overset{\text{ind}}{\sim} \mathsf{Exp}(\lambda/2), \quad i = 1, \ldots, p, \tag{4.3}$$

for computational convenience. In addition, we settle on a non-informative prior for $\lambda$, $\mathbb{P}(\lambda) \propto 1$.

Finally, to model the adjacency matrix $A$, we adopt a degree-corrected stochastic block-model which specifies that the probability of an edge between node $i$ and node $j$ depends on their labels $(\sigma_i, \sigma_j)$ and their expected degrees, and that $\sigma$ follows a product multinomial distribution as in Chapter 2:

$$\begin{aligned} A_{ij} \,|\, \sigma, \gamma, \eta &\overset{\text{ind}}{\sim} \mathsf{Bern}\big(\text{logit}^{-1}(\gamma_{\sigma_i\sigma_j} + \eta_i + \eta_j)\big), \quad i, j = 1, \ldots, p, i < j, \\ \sigma_i &\overset{\text{ind}}{\sim} \mathsf{MN}(1; \pi), \quad i = 1, \ldots, p. \end{aligned} \tag{4.4}$$

Hyper-parameters $\gamma$ capture within and between community probabilities of association (in logit scale) and node intercepts $\eta = (\eta_1, \ldots, \eta_p)$ capture the expected degrees of the nodes. A more realistic model is attained by further setting a hyper-prior distribution on $\gamma$ and $\eta$,

$$(\gamma, \eta) \sim I(\gamma \leq 0) \cdot \mathsf{N}\Big(0, \tau^2 I\Big), \tag{4.5}$$

where $\tau^2$ controls how informative the prior is. The constraint $\gamma \leq 0$ in this SBM is essential to community detection since we should expect as many as or fewer edges between communities than within communities on average, and thus that the log-odds of between and within probabilities is non-positive. The parameter identifiability of model (4.4) is described in the Section 3.2. The model is identifiable if each community has at least two nodes and the number of communities is greater than two.

To summarize, in the likelihood, we adopt a Gaussian graphical model; in the next level, we select the covariance structure in $C^{-1}$ with a spike-and-slab prior; and finally, we capture community behavior in the components of $X$ via a SBM on $A$.

## 4.2 Inference

To develop the MAP estimator for $C$, $A$, $\sigma$, $\gamma$ and $\eta$, we follow a cyclic gradient descent approach where each parameter is obtained by optimizing

$$[C, A \,|\, \sigma, \gamma, \eta, \mu, \lambda, X], \qquad [\sigma \,|\, \gamma, \eta, C, A, \mu, \lambda, X], \qquad [\gamma, \eta \,|\, \sigma, C, A, \mu, \lambda, X],$$
$$[\mu \,|\, \sigma, C, A, \sigma, \gamma, \eta, \lambda, X], \qquad [\lambda \,|\, \sigma, C, A, \sigma, \gamma, \eta, \mu, X]$$

in turn. While we have a step using $\mu$, in general we have $\widehat{\mu} = \sum_{i=1}^{n} X_i/n$ and so we often consider $X_i \,|\, C \overset{\text{iid}}{\sim} \mathsf{N}(0, C^{-1})$ by pre-centering $X$. Similarly, the MAP estimator for $\lambda$ is straightforward: $\widehat{\lambda} = 2/\sum_{i=1}^{p} C_{ii}$.

Now, we want to find a concentration matrix $C$ and latent network $A$ that maximize $\log \mathbb{P}(C, A|\sigma, \gamma, \eta, X)$, or equivalently,

$$\log \mathbb{P}(C, A, X|\sigma, \gamma, \eta) = \frac{n}{2} \log |C| - \frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^\top C (X_i - \mu)$$
$$- \frac{1}{2\rho^2} \sum_{1 \le i < j \le p} \frac{C_{ij}^2}{A_{ij} + \nu_0(1 - A_{ij})} - \frac{\lambda}{2} \sum_{i=1}^{p} C_{ii} + \sum_{1 \le i < j \le p} A_{ij}(\gamma_{\sigma_i \sigma_j} + \eta_i + \eta_j). \quad (4.6)$$

Note that the prior on $C$, $A$ and $\beta$ can be seen as a ridge regularization in the following optimization aiming to obtain a symmetric positive definite matrix $C$,

$$\max_{C \succ 0} \frac{n}{2} \log |C| - \frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^\top C (X_i - \mu) \quad (4.7)$$

To find the conditional MAP estimator for $C$ and $A$, we focus on each of their rows (or columns) at a time. For the $i$-th row and column, we consider the log-likelihood as a function of $C_{i,\cdot}$, that is,

$$\log \mathbb{P}(C_{i,\cdot}, X, A) = \frac{n}{2} \log |C_{ii} - C_{i,-i} C_{-i,-i}^{-1} C_{-i,i}| - \frac{1}{2}(S_{ii} C_{ii} + 2 S_{i,-i} C_{-i,i})$$
$$- \frac{1}{2\rho^2} \sum_{j \neq i} \frac{C_{ij}^2}{A_{ij} + \nu_0(1 - A_{ij})} - \frac{\lambda}{2} C_{ii},$$

up to terms that do not involve $C_{i,\cdot}$. Here $S = \sum_{i=1}^{n}(X_i - \widehat{\mu})(X_i - \widehat{\mu})^\top$ is a sufficient statistic. Then, if $V_i = \mathrm{Diag}_{j\neq i}\Big\{1/[\rho^2 A_{ij} + \rho^2\nu_0(1 - A_{ij})]\Big\}$, $C_{i,-i}$ is the $i$-th row with $i$-th column removed and $C^{-1}_{-i,-i}$ is the sub-matrix of $C^{-1}$ with $i$-th row and column removed, the ridge-regularized estimator for $C_{i,\cdot}$ is given by

$$\begin{aligned} \widehat{C}_{i,-i} &= -\big[(S_{ii} + \lambda)C^{-1}_{-i,-i} + V_i\big]^{-1} S_{i,-i}, \\ \widehat{C}_{i,i} &= \frac{n}{S_{ii} + \lambda} + \widehat{C}_{i,-i} C^{-1}_{-i,-i} \widehat{C}^\top_{i,-i}. \end{aligned} \tag{4.8}$$

We note that $C$ is kept positive definite along the whole procedure. Since $C \succ 0$, $C_{-i,-i} \succ 0$ for any $i$; after the update $C_{i,\cdot}$, $C$ is still positive definite since

$$C_{ii} - C_{i,-i}C^{-1}_{-i,-i}C_{-i,i} = \frac{n}{S_{ii} + \lambda} > 0,$$

as noted in Yuan (2008). We can obtain the estimate by solving (4.8) for each row and iterating until convergence. Thus, if the initial value of the concentration matrix is symmetric and positive definite, then the estimate based on (4.8) is also symmetric and positive definite throughout the iterative procedure.

This series of updates is conditional on $A$, as seen in Equation (4.6). However, we further propose an approach where the adjacency matrix is estimated along with the concentration matrix in the covariance selection procedure, that is, we update $C$ and $A$ jointly. Moreover, to avoid the risk of getting stuck if we update the whole row $A_{i,\cdot}$ each time, we propose to update at most one entry in $A$ at each iteration. That is, we move to $A^{(t+1)}_{i,\cdot}$ where the candidate move set for $A_{i\cdot}$ is $\{A_{i,\cdot} : H(A_{i,\cdot}, A^{(t)}_{i,\cdot}) \leq 1\}$ and $H(\cdot,\cdot)$ is the Hamming distance. In practice, we adopt the SWEEP operator (Goodnight, 1979) and Cholesky up/down-dates to make the iterative algorithm more efficient.

$$W = \left[\begin{array}{c|c} C_{ii} - C_{i,-i}C^{-1}_{-i,-i}C_{-i,i} & -C_{i,-i}W_{-i,-i} \\ \hline W_{-i,-i}C_{-i,i} & C^{-1}_{-i,-i} \end{array}\right] \xleftrightarrow{\text{SWEEP } i\text{th row}} \text{SWEEP } i\text{th row } C^{-1}$$

where $W_{11}$ remains constant in the $i$-th iteration according to (4.2).

Once we are given the adjacency matrix $A$ representing relationships between "actors" $i$ and $j$ in a network, we adopt a Bayesian degree-corrected SBM given in (4.4) to detect the community structure in the network, that is, to find a conditional MAP estimator for $[\sigma|\gamma, \eta, C, A, X]$ and $[\gamma, \eta|\sigma, C, A, X]$. First, we take $\sigma_i$ to be the mode of $\sigma_i|\sigma_{[-i]}, \beta, A$. Next, a regularized IRLS is carried out. To guarantee that the community constraints $\gamma \leq 0$ are met, we use an active-set method (Nocedal and Wright, 2006). More details can be found in Chapter 2.

To summarize, the Bayesian ridge-regularized graph estimate is obtained by iterating until convergence the steps in Algorithm 1.

**Algorithm 1** Bayesian ridge-regularized covariance selection

Set initial $C$, $A$; obtain initial $\lambda$, $\sigma$, $\eta$ and $\gamma$ based on $C$, $A$.
**repeat**
  **for** $i = 1, \ldots, p$ **do**
    Set $\text{lhood}_{\max} = -\infty$, $W = \text{SWEEP}(C^{-1}, i)$
    **for** $j \neq i$ **do**
      $\left\{ \begin{array}{lll} A_{ij} = 0 & \Longrightarrow & \widehat{C}^{(0)}_{i,\cdot};\ \text{compute lhood}_0 \\ A_{ij} = 1 & \Longrightarrow & \widehat{C}^{(1)}_{i,\cdot};\ \text{compute lhood}_1 \end{array} \right\}, \text{lhood} = \max_{k\in\{0,1\}} \text{lhood}_k$
      **if** $\text{lhood} > \text{lhood}_{\max}$
        $\text{lhood}_{\max} = \text{lhood}, j^\star = j, k^\star = \arg\max_{k\in\{0,1\}} \text{lhood}_k, \widehat{C}_{i,\cdot} = \widehat{C}^{(k^\star)}_{i,\cdot}$
      **end if**
    **end for**
    Update $A_{ij^\star} = A_{j^\star i} \leftarrow k^\star$
    Update $W_{-i,i} \leftarrow W_{-i,-i}\widehat{C}_{-i,i}$, $W_{i,-i} \leftarrow -W^\top_{-i,i}, W_{i,i} = \frac{n}{S_{ii}+\lambda}$
    Update $C^{-1} \leftarrow \text{SWEEP}(W, i)$
  **end for**
  Update $\lambda$
  Update $\sigma$, $\eta$ and $\gamma$ from the community detection procedure in Section 4.2
**until** the change in the log-likelihood is within certain tolerance

## 4.3 Experimental Results

In this section, we evaluate the performance of our proposed Bayesian ridge-regularized estimator in identifying latent networks. For comparison, we also estimate the concentra-

tion using sample estimates and a lasso-regularized estimator (Yuan, 2008). There has not been a method that jointly estimate the concentration matrices and detect the community structure in literature. The graph lasso estimator is the most similar approach as ours, but it fails to take the community structure into account when estimating the adjacency matrix. Hence, the following comparison among methods are made in terms of recovering concentration matrices. We expect that accounting for community structure when inferring adjacency matrices and concentration matrices yields more reliable estimation.

Our simulation study generates networks from a popular benchmark suite due to Fortunato Lancichinetti et al. (2008b) that accounts for heterogeneities in node degree distributions and community sizes. The model used in the simulation considers the following parameters: both degree distribution and the community sizes are assumed to follow power law distributions with exponents $a = 2$ and $b = 1$, respectively; each network consists of $p = 50$ nodes and has average degree $\langle k \rangle = 10$. Mixing parameter $\mu$ captures the proportion of between-community edges. We highlight two community behaviors: gregarious, with $\mu = 0.1$ , or non-assortative, with $\mu = 0.4$.

We further generate concentration matrices based on the networks as ground truth according to Equation (4.2) with fixed $\rho^2 = 100$ and $\nu_0 = 10^{-6}$ for simplicity. The value $\rho^2 = 100$ is large enough to distinguish the differences in the concentration matrix when edges in the latent network are present or absent. The data $X = \{X_1, \ldots, X_n\}$ for $n = (10, 25, 50, 100, 200)$ is generated as in Equation (4.1). We estimate concentration matrices based on $X$ by sample concentration, our approach with $A$ known as well as unknown (latent), and Lasso estimates with different tuning parameters ranging from 0.001 to 10. The comparison in terms of the log relative Frobenius norm of estimated concentration matrices, $\log\left(\|\widehat{C} - C\|_F / \|C\|_F\right)$, is shown in Fig. 4.1. Our proposed approach outperforms the sample and Lasso estimates in terms of the log relative Frobenius norm. In addition, the error we made in estimating latent networks is mainly due to false negatives (failing to detect an edge when there is one), especially when we have fewer observations.

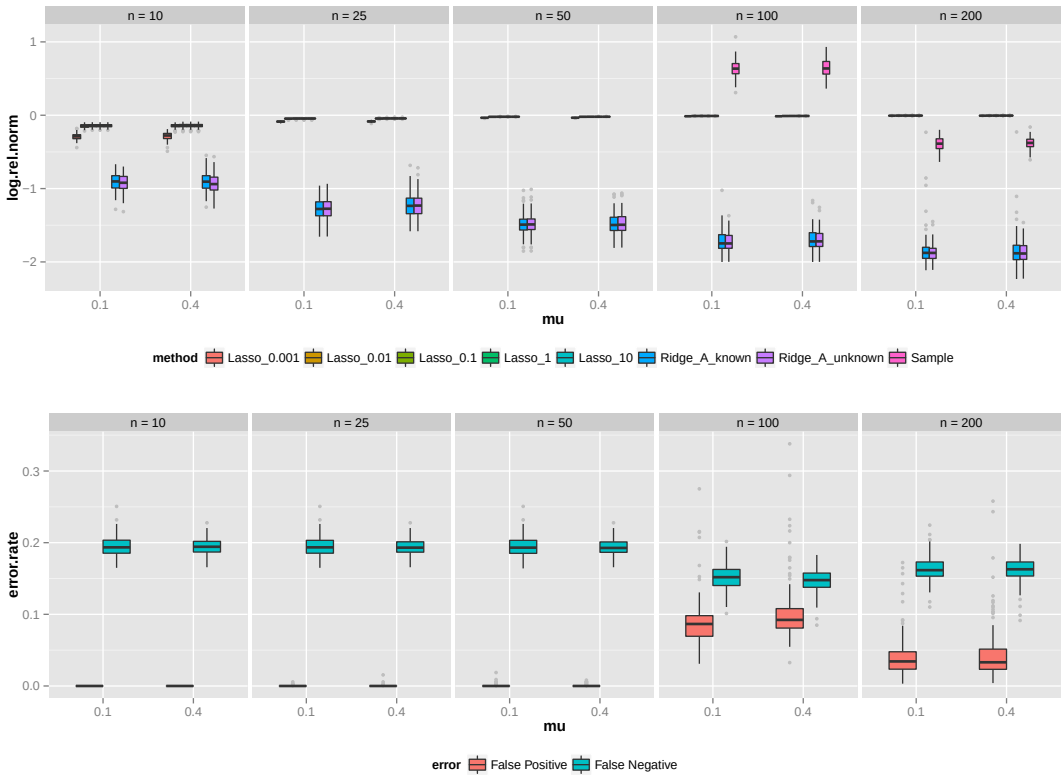


Figure 4.1: (Top) The log relative Frobenius norm of estimated concentration matrices under different approaches. The sample estimates when $n < p$ have relatively large norms are not shown to maintain a short scale. (Bottom) The false positive and negative rates of estimated adjacency matrices under our proposed model.

## 4.4 Case Study

In this case study, we take a dataset that measures 16S ribonucleosomal expression levels using the Human Oral Microbe Identification Microarray (HOMIM) for 276 bacteria and contrasts 90 sites in healthy individuals to 514 sites in patients with varying degrees of periodontitis (Duran-Pinedo et al., 2011). We assume, as before, that individual samples are independent, but now we exploit a decomposable mean model where the mean response for bacteria $i$ and sample $j$ is given by:

$$\mu_{ijc} = \theta_{ic} + \phi_j, \quad i = 1, \ldots, p, \quad j = 1, \ldots, n, \tag{4.9}$$

where $c$ is the condition, either healthy or diseased. Parameters $\theta_{ic}$ capture the expression effect of each bacteria per condition, while parameters $\phi_j$ represent the baseline expression level per sample and are considered nuisance.

After running our proposed procedure for $K = 2, \ldots, 10$ communities, we select $K = 3$ based on BIC. The two first panels in Figure 4.2 depict the inferred networks and communities. As can be seen, the "diseased" bacterial community is more connected and has a stronger community effect. To compare the joint effect of expression via $\theta$ and connectivity, we compute alpha-centralities (Bonacich and Lloyd, 2001) using $\widehat{\theta}_{\cdot c}$ as weights. The rightmost panel in Figure 4.2 contrasts alpha-centrality between the two conditions; for comparison, we mark bacterial species according to Socransky et al. (1998) complexes. Interestingly, bacteria from the red complexes—usually associated to the most severe forms of periodontitis—tend to have higher alpha-centralities in the diseased sample group relative to the healthy group.

## 4.5 Discussion

In this chapter, we developed a Bayesian ridge-regularized covariance selection model that incorporates community behavior through a latent network. This class of models has

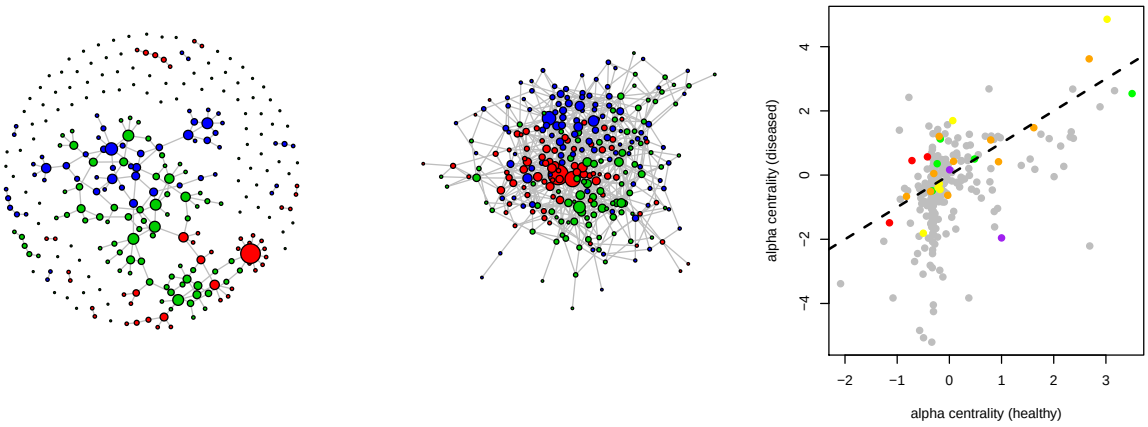


Figure 4.2: Inferred networks for healthy (left) and diseased (right) samples. Colors mark inferred communities. In the right, alpha-centrality with $\alpha = 0.5$ with $\theta$ estimates as exterior weights; colors mark Socransky complex classification Socransky et al. (1998).

many practical applications in social sciences and systems biology. Good results based on our simulation study indicate that the proposed approach is a serious contender for covariance selection when compared to Lasso-based estimators. Moreover, as the case study shows, our estimator reliably captures biological assortativity in bacterial communities, and is able to classify bacteria with respect to their different responses in expression and connectivity under two scenarios. Moreover, since most of the bacteria in dental biofilms are not cultivable, the proposed model gives insight into which partnerships are needed for these bacteria under different conditions. Possible extensions of this work may include developing dynamic models to fit time series data.

# Chapter 5

# Conclusions

In Chapter 2, I have proposed a Bayesian model based on degree-corrected SBMs that is tailored for community detection. More specifically, our model is flexible due to its hierarchical structure and aims to capture the notion of gregarious community behavior by requiring, through prior specification, that the probability of within-community associations to be no smaller than the probability of between-community associations. Moreover, I argue that the model is a better representative of assortatively mixing networks with binary data coding the associations instead of frequency counts, since I model binary observations using a suitable logistic regression with parameters for within and between-community probabilities of association. I devise a Gibbs sampler to obtain posterior samples and exploit a latent variable formulation to yield closed-form conditionals.

I formally address label identifiability by restricting label configurations to a canonical reference subspace, and propose a remap procedure to implement this constraint in practice. As a consequence, labels are interpretable and we are able to estimate any function of the labels as opposed to previous approaches that were restricted to permutation-invariant functions. In particular, I propose a novel remapped centroid estimator to infer community assignments. I contend that while the model can arguably represent the data well, the posterior space can be complex and a bad estimator can spoil the analysis; it is then imperative to adopt an estimator that arises from a principled and refined loss function and thus better summarizes the posterior space. Our proposed remapped centroid estimator is more similar to a posterior mean. Hence, it tends to situate itself in regions of high concentration of posterior mass in the meanwhile of considering the whole posterior

distribution in the space of remapped label assignments. From a practical point of view, I show that the proposed estimator performs as well as the Binder, KN, MAP, FG, ML, LP and WT estimators when the community structure is strong while performs better than other estimators mentioned when the community structure is defined in a weaker sense.

In Chapter 3, I extend the basic SBM to incorporate parameters for group attributes and generalize the formulation to account for more node and edge attributes. The model aims to capture the gregarious community behavior by requiring that the probability of within-community associations to be no smaller than that of between-community associations. What's more, I take the degree heterogeneity into consideration and make a group correction that reflects degree assortativity. The method has connection to some of the existing methods, but is expected to be more efficient and suitable for large-scale networks. I develop a graph generalized linear model (GGLM) procedure tailored for graphs to make the inference more computational efficient. GGLM implicitly computes sufficient statistics rather than generating responses and design matrices. I have presented an application to an amicus curiae network of count data. I have demonstrated the proposed MAP estimator on simulated benchmark networks as well as real-world networks with ground truth communities and shown that the MAP estimator outperforms KN, FG, ML, WT and WT estimators in terms of NMI on average.

In Chapter 4, I focus on the problem where networks are considered as latent and to be inferred. I relate identifying latent networks to estimating concentration matrices, and thus focus on covariance selection. To this end, a Bayesian approach that jointly estimate concentration matrices and identifying networks has been presented. I develop a Bayesian approach with a hierarchical prior with two levels: (i) a ridge-regularized covariance selection that specifies a spike-and-slab prior; (ii) a MAP estimator that explicitly characterizes community behavior. I compare our ridge-regularized covariance selection to other commonly used methods on simulated benchmark networks and have shown that including block structures when estimating concentration/adjacency matrices improves the inference results. Moreover, I have demonstrated the approach on a real-world meta-genomic dataset

of complex microbial biofilms. The proposed estimator reliably captures biological assortativity in bacterial communities, and is able to classify bacteria with respect to their different responses in expression and connectivity under healthy and diseased scenarios.

# Chapter 6

# Appendix

### 6.0.1 Proof of Theorem 1

For the proof we first note that we can split each row $x_{ij}$ in the design matrix of (2.2) according to $\gamma$ and $\eta$ entries, $x_{ij} \doteq [b_{ij} \quad c_{ij}]$, where

$$\begin{aligned} b_{ij,kl} &= I[\min(\sigma_i, \sigma_j) = k, \max(\sigma_i, \sigma_j) = l], \qquad k, l = 1, \ldots, K, k \leq l, \\ c_{ij,v} &= I(i = v) + I(j = v), \qquad v = 1, \ldots, n, \end{aligned} \tag{6.1}$$

that is, $b_{ij}$ identifies the pair of communities at the endpoints of $(i, j)$ for $\gamma$ and $c_{ij}$ marks each node-correction from $\eta$.

*Proof of (a).* Let us pick an arbitrary community $k$ and a pair $(i, j)$. There are then three ways to classify $(i, j)$: (i) it is outside of community $k$; (ii) one of its endpoints is in community $k$; or (iii) it is inside community $k$. If we now define $d_{ij,k} = \sum_{v:\sigma_v=k} c_{ij,v}$ then $(i, j)$ is classified exactly according to $d_{ij,k}$: $d_{ij,k} = 0, 1$, or 2 if $(i, j)$ is in cases (i), (ii), or (iii), respectively. Thus, it follows that

$$2b_{ij,kk} + \sum_{l \neq k} b_{ij,kl} = \sum_{v:\sigma_v=k} c_{ij,v},$$

for each $k = 1, \ldots, K$, and so $X$ has $K$ constraints in its columns. □

*Proof of (b).* Note that $X$ is full column-ranked if and only if $X^\top X$ is invertible, so we just need to show that $X^\top X$ is invertible if $N_k \geq 2$ for $k = 1, \ldots, K$. Let $B =$

$[b_{ij,12}, \ldots, b_{ij,K-1K}]_{i<j}$ and $C = [c_{ij,1}, \ldots, c_{ij,n}]_{i<j}$. Then $X = [B, C]$ and

$$X^\top X = \begin{bmatrix} B^\top B & B^\top C \\ C^\top B & C^\top C \end{bmatrix}.$$

Thus, $X^\top X$ is invertible if and only if both $B^\top B$ and the Schur complement of $C^\top C$, $\Delta \doteq C^\top [I - B(B^\top B)^{-1} B^\top] C$ are invertible. First,

$$B^\top B = \mathrm{Diag}\left( \sum_{i<j} I[\sigma_i = k, \sigma_j = l \text{ or } \sigma_i = l, \sigma_j = k] \right) = \mathrm{Diag}(N_k N_l),$$

and so, for this diagonal matrix to be invertible we need $N_k \neq 0$ for $k = 1, \ldots, K$.

As for the Schur complement $\Delta$, we have that

$$\Delta_{ii} = n - 1 - \sum_{k \neq i} \frac{\sum_{l \neq i} I[\sigma_i \neq \sigma_k = \sigma_l]}{N_{\sigma_i} N_{\sigma_k}},$$

and, for $i < j$,

$$\Delta_{ij} = 1 - \sum_{k \neq i} \frac{\sum_{l \neq j} I[\sigma_i = \sigma_j \neq \sigma_k = \sigma_l \text{ or } \sigma_i = \sigma_l \neq \sigma_k = \sigma_j]}{N_{\sigma_i} N_{\sigma_k}}.$$

But if $\sigma_i \neq \sigma_j$,

$$\Delta_{ij} = 1 - \sum_{k \neq i} \frac{\sum_{l \neq j} I[\sigma_i = \sigma_l \neq \sigma_k = \sigma_j]}{N_{\sigma_i} N_{\sigma_k}} = 0,$$

and otherwise, if $\sigma_i = \sigma_j$,

$$\Delta_{ij} = 1 - \sum_{k \neq i} \frac{\sum_{l \neq i} I[\sigma_i \neq \sigma_k = \sigma_l]}{N_{\sigma_i} N_{\sigma_k}}, \tag{6.2}$$

and so $\Delta_{ii} - \Delta_{ij} = n - 2$. Thus, after some row and column operations, $\Delta$ can be written

as a block diagonal matrix where each block of size $N_k$ has the form:

$$\begin{bmatrix} p & q & \dots & q \\ q & p & \dots & q \\ \vdots & & \ddots & \vdots \\ q & q & \dots & p \end{bmatrix}$$

with $q = \Delta_{ij}$ in (6.2) and $p = n - 2 + q$. The determinant of the block diagonal matrix is nonzero if and only if $n \neq 2$ and $N_k \neq 1$. Moreover, the determinant of $X^\top X$ is the same as that of the block diagonal matrix since one can be obtained from the other through row and column operations. Thus, the conditions $N_k \neq 0$ from $B^\top B$ and now $N_k \neq 1$ can be summarized into $N_k \geq 2$. □

### 6.0.2 Remap Algorithm

Algorithm 2 lists a routine that finds the canonical map $\rho$ based on the canonical order in $\sigma$ as in Equation (2.6) and remaps $\sigma$ in-place.

**Algorithm 2** Remapping labels in $\sigma$ to $\rho(\sigma)$.

```
assigned ← {}
ρ ← {}
n ← 0 {number of different labels in σ}
for i = 1, ..., |σ| do {obtain ρ ≐ ord(σ)^{-1}}
  if not assigned(σ(i)) then {first appearance?}
    assigned(σ(i)) ← true {mark σ(i)}
    n ← n + 1
    ρ(σ(i)) ← n
  end if
end for
for i = 1, ..., |σ| do {remap σ}
  σ(i) ← ρ(σ(i))
end for
return σ
```

### 6.0.3 Proof of Theorem 2

*Proof.* It is sufficient to find the pre-map estimator

$$\widehat{\sigma}^* \doteq \operatorname*{arg\,min}_{\tilde{\sigma}\in\{1,\dots,K\}^n} \mathbb{E}_{\sigma\,|\,A}\big[H(\tilde{\sigma},\rho(\sigma))\big]$$

since, by definition, $\widehat{\sigma}_C = \rho(\widehat{\sigma}^*)$.

Denoting $\Sigma = \{1,\dots,K\}^n$ and $\Sigma^* = \Sigma\,/\,\mathsf{ord}$, we have that

$$\begin{aligned}\mathbb{E}_{\sigma\,|\,A}\big[H(\tilde{\sigma},\rho(\sigma))\big] &= \sum_{\sigma\in\Sigma} H(\tilde{\sigma},\rho(\sigma))\mathbb{P}(\sigma\,|\,A)\\ &= \sum_{\sigma\in\Sigma^*}\sum_{\sigma^*:\rho(\sigma^*)=\sigma} H(\tilde{\sigma},\sigma)\mathbb{P}(\sigma^*\,|\,A).\end{aligned}$$

Since $\mathbb{P}(\sigma^*\,|\,A) = \mathbb{P}(\sigma\,|\,A)$ follows from the lack of identifiability we further obtain

$$\mathbb{E}_{\sigma\,|\,A}\big[H(\tilde{\sigma},\rho(\sigma))\big] = \sum_{\sigma\in\Sigma^*} n(\sigma)H(\tilde{\sigma},\sigma)\mathbb{P}(\sigma\,|\,A),$$

where $n(\sigma) = |\{\sigma^* : \rho(\sigma^*) = \sigma\}| = K!/(K-k(\sigma))!$ is the number of assignments that are identified to $\sigma$ through $\mathsf{ord}$, and $k(\sigma)$ is the number of different labels in $\sigma$. We can then define $\mathbb{P}^*(\sigma\,|\,A) \doteq n(\sigma)\mathbb{P}(\sigma\,|\,A)$ as the induced measure in the quotient space $\Sigma^*$ to thus have

$$\begin{aligned}\mathbb{E}_{\sigma\,|\,A}\big[H(\tilde{\sigma},\rho(\sigma))\big] = \sum_{\sigma\in\Sigma^*} H(\tilde{\sigma},\sigma)\mathbb{P}^*(\sigma\,|\,A) &= \sum_{\sigma\in\Sigma^*}\sum_{i=1}^{n} I(\tilde{\sigma}_i\neq\sigma_i)\mathbb{P}^*(\sigma\,|\,A)\\ &= n - \sum_{i=1}^{n}\sum_{\sigma\in\Sigma^*} I(\tilde{\sigma}_i=\sigma_i)\mathbb{P}^*(\sigma\,|\,A) = n - \sum_{i=1}^{n}\mathbb{P}^*(\sigma_i=\tilde{\sigma}_i\,|\,A).\end{aligned}$$

But then

$$\operatorname*{arg\,min}_{\tilde{\sigma}\in\{1,\dots,K\}^n} \mathbb{E}_{\sigma\,|\,A}\big[H(\tilde{\sigma},\rho(\sigma))\big] = \operatorname*{arg\,max}_{\tilde{\sigma}\in\{1,\dots,K\}^n} \sum_{i=1}^{n}\mathbb{P}^*(\sigma_i=\tilde{\sigma}_i\,|\,A)$$

and so

$$(\widehat{\sigma}^*)_i = \underset{k\in\{1,\ldots,K\}}{\arg\max}\ \mathbb{P}^*(\sigma_i = k \,|\, A),$$

that is, $\widehat{\sigma}^*$ is a consensus estimator, as desired. □

### 6.0.4 Proof of Theorem 3

*Proof.* To compare $\tilde{\sigma}$ and $\sigma$ let us define $n_{ij} \doteq \sum_{k,l} I(\sigma_k = i, \tilde{\sigma}_l = j)$, the number of nodes that belong to community $i$ in $\sigma$ and to community $j$ in $\tilde{\sigma}$. Then, $B(\tilde{\sigma},\sigma) = \sum_i \sum_{j<k}(n_{ij}n_{ik} + n_{ji}n_{ki})$, $H(\tilde{\sigma},\sigma) = \sum_{i\neq j} n_{ij}$, and $n = \sum_{i,j} n_{ij}$.

For instance, if $K = 2$ then $H(\tilde{\sigma},\sigma) = n_{12} + n_{21}$ and

$$\begin{aligned} B(\tilde{\sigma},\sigma) &= (n_{11}n_{12} + n_{21}n_{22}) + (n_{11}n_{21} + n_{12}n_{22}) \\ &= (n_{12} + n_{21})(n_{11} + n_{22}) \\ &= H(\tilde{\sigma},\sigma)\big(n - H(\tilde{\sigma},\sigma)\big). \end{aligned}$$

More generally, for $K > 2$, we have:

$$\begin{aligned} nH(\tilde{\sigma},\sigma) &= \sum_{i\neq j} n_{ij} \sum_{i,j} n_{ij} = \sum_{i\neq j} n_{ij} \left( \sum_{i\neq j} n_{ij} + \sum_i n_{ii} \right) \\ &= \sum_{i\neq j} n_{ij} \sum_{i\neq j} n_{ij} + \sum_{i\neq j} n_{ij} \sum_i n_{ii} \\ &= \underbrace{\sum_{i\neq j} n_{ij}^2}_{A} + \underbrace{\sum_{\substack{i\neq j, k\neq l \\ k\neq i, j\neq l}} n_{ij}n_{kl}}_{B} + 2 \underbrace{\sum_{\substack{i\neq j, i\neq k \\ j<k}} (n_{ij}n_{ik} + n_{ji}n_{ki})}_{C} \\ &\quad + \underbrace{\sum_{\substack{i\neq j, i\neq k \\ j\neq k}} n_{ii}n_{jk}}_{D} + \underbrace{\sum_{i\neq j} (n_{ii}n_{ij} + n_{ii}n_{ji})}_{E}. \end{aligned}$$

Thus, $B(\tilde{\sigma},\sigma) = C + E$ and, in particular,

$$\begin{aligned} H^2(\tilde{\sigma},\sigma) &= \Big(\sum_{i\neq j} n_{ij}\Big)\Big(\sum_{i\neq j} n_{ij}\Big) \\ &= \sum_{i\neq j} n_{ij}^2 + \sum_{\substack{i\neq j,k\neq l \\ k\neq i,j\neq l}} n_{ij}n_{kl} + 2\sum_{\substack{i\neq j,i\neq k \\ j<k}} (n_{ij}n_{ik} + n_{ji}n_{ki}) \\ &= A + B + 2C. \end{aligned}$$

The bound $B(\tilde{\sigma},\sigma) \leq H(\tilde{\sigma},\sigma)(n - H(\tilde{\sigma},\sigma)/2)$ then follows from

$$nH(\tilde{\sigma},\sigma) - B(\tilde{\sigma},\sigma) - \frac{1}{2}H^2(\tilde{\sigma},\sigma) = \frac{1}{2}A + \frac{1}{2}B + D \geq 0$$

since $A, B$ and $D$ are all non-negative. □

### 6.0.5 Proof of Theorem 4

*Proof.* The proof of Theorem 4 is similar to that of Theorem 1. We first consider the case where each node forms a group. We can generalize the identifiability conditions found for this group partition with finest resolution to any group partitions, since any non-overlapping group partition is a linear combination of that with finest resolution. We split each row $x_{ij}$ in the design matrix of (3.3) according to $\gamma$ and $\eta$ entries, $x_{ij} \doteq [b_{ij} \quad c_{ij}]$, where

$$\begin{aligned} b_{ij,k} &= I(\sigma_i = \sigma_j = k) \qquad k = 1,\ldots,K, \\ c_{ij,v} &= I(i = v) + I(j = v), \qquad v = 1,\ldots,n, \end{aligned} \tag{6.3}$$

that is, $b_{ij}$ identifies the pair of nodes in the same community for $\gamma$ and $c_{ij}$ marks each group-correction (also node-correction since we are using the finest partition) from $\eta$.

Note that $X$ is full column-ranked if and only if $X^\top X$ is invertible, so we just need to show that $X^\top X$ is invertible if the conditions in Theorem 4 hold. Let $B = [b_{ij,1},\ldots,b_{ij,K}]_{i<j}$

and $C = [c_{ij,1}, \ldots, c_{ij,n}]_{i<j}$. Then $X = [B, C]$ and

$$X^\top X = \begin{bmatrix} B^\top B & B^\top C \\ C^\top B & C^\top C \end{bmatrix}.$$

Thus, $X^\top X$ is invertible if and only if both $B^\top B$ and the Schur complement of $C^\top C$, $\Delta \doteq C^\top [I - B(B^\top B)^{-1} B^\top] C$ are invertible. First,

$$B^\top B = \mathrm{Diag}\left( \sum_{i<j} I(\sigma_i = \sigma_j = k) \right) = \mathrm{Diag}\left( \binom{N_k}{2} \right),$$

and so, for this diagonal matrix to be invertible we need $N_k \geq 2$ for $k = 1, \ldots, K$.

Next, We consider the Schur complement $\Delta \doteq C^\top [I - B(B^\top B)^{-1} B^\top] C$. After rows and columns operations, the Schur complement $\Delta$ can be written in the form

$$\Delta = \begin{bmatrix} A_1 & \mathbf{1} & \ldots & \mathbf{1} \\ \mathbf{1} & A_2 & \ldots & \mathbf{1} \\ \vdots & & \ddots & \vdots \\ \mathbf{1} & \ldots & \mathbf{1} & A_K \end{bmatrix}$$

where

$$A_k = \begin{bmatrix} p & q & \ldots & q \\ q & p & \ldots & q \\ \vdots & & \ddots & \vdots \\ q & q & \ldots & p \end{bmatrix}_{N_k}, \quad q = \frac{2}{N_k} - 1, p - q = n - 2. \tag{6.4}$$

We consider the determinant of one block

$$|D| = \begin{vmatrix} A_1 & \mathbf{1} \\ \mathbf{1} & A_2 \end{vmatrix} = |A_1| \cdot |A_2 - \mathbf{1} A_1^{-1} \mathbf{1}^\top|$$

Simple linear algebra shows that matrix $A_2 - \mathbf{1} A_1^{-1} \mathbf{1}^\top$ is in the form of (6.4) with some $p$

and $q$ satisfying $p - q = n - 2$. In other words, $D$ shares the same formulation as $A_1$, with which we can compute the determinant of matrix including the next block

$$\begin{vmatrix} D & \mathbf{1} \\ \mathbf{1} & A_3 \end{vmatrix} = |D| \cdot |A_3 - \mathbf{1}D^{-1}\mathbf{1}^\top|.$$

Hence, the determinant of $\Delta$ can be computed consecutively using this property. And $|\Delta| \neq 0$ if and only if the first such block matrix $D$ is invertible, which requires $N_1 + N_2 \neq n$. In other words, $K > 2$.

Note that when $K = 2$, model (3.1) is not identifiable if each node is partitioned into a group by itself. However, the model is identifiable under other group partitions, as long as none of partitioned groups is completely covered in a community. The node correction is a special case when each group (node) is completely contained in the community of that node.

□

### 6.0.6 Derivation of $\mathbb{P}(\sigma_i = k \,|\, \sigma_{[-i]}, \beta, A)$

$$\begin{aligned}
\mathbb{P}(\sigma_i = k \,|\, \sigma_{[-i]}, \beta, A) &= \frac{\mathbb{P}(A \,|\, \sigma_i, \sigma_{[-i]}, \beta)\mathbb{P}(\sigma_i)\mathbb{P}(\sigma_{[-i]})\mathbb{P}(\beta)}{\sum_{\tilde{\sigma}_i} \mathbb{P}(A \,|\, \tilde{\sigma}_i, \sigma_{[-i]}, \beta)\mathbb{P}(\tilde{\sigma}_i)\mathbb{P}(\sigma_{[-i]})\mathbb{P}(\beta)} \\
&= \frac{\prod_{i \neq j} \mathbb{P}(A_{ij} \,|\, \sigma_i, \sigma_j, \beta)\mathbb{P}(\sigma_i)}{\sum_{\tilde{\sigma}_i} \prod_{i \neq j} \mathbb{P}(A \,|\, \tilde{\sigma}_i, \sigma_j, \beta)\mathbb{P}(\tilde{\sigma}_i)} \\
&\propto \prod_{i \neq j} \mathbb{P}(A_{ij} \,|\, \sigma_i, \sigma_j, \beta)\mathbb{P}(\sigma_i) \\
&= \pi_k \prod_{j \neq i} \frac{\exp\{A_{ij} x_{ij}^\top \beta\}}{1 + \exp\{x_{ij}^\top \beta\}}
\end{aligned}$$

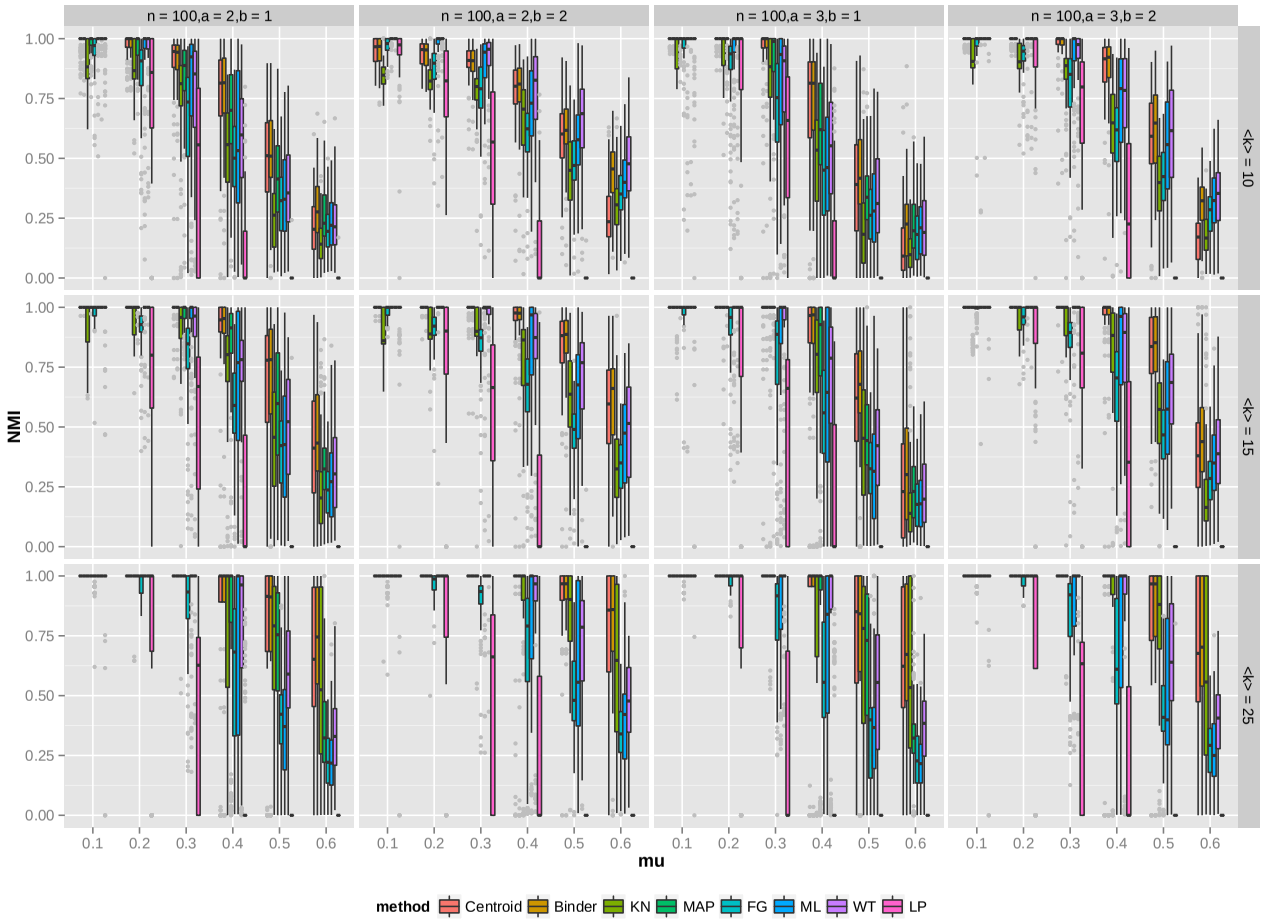


Figure 6.1: Benchmark networks of $n = 100$ nodes, with different combinations of the exponents $a \in \{2, 3\}$, $b \in \{1, 2\}$ and the average degree $\langle k \rangle \in \{10, 15, 25\}$. Each boxplot corresponds to the NMI of the estimator over 100 graph realizations.

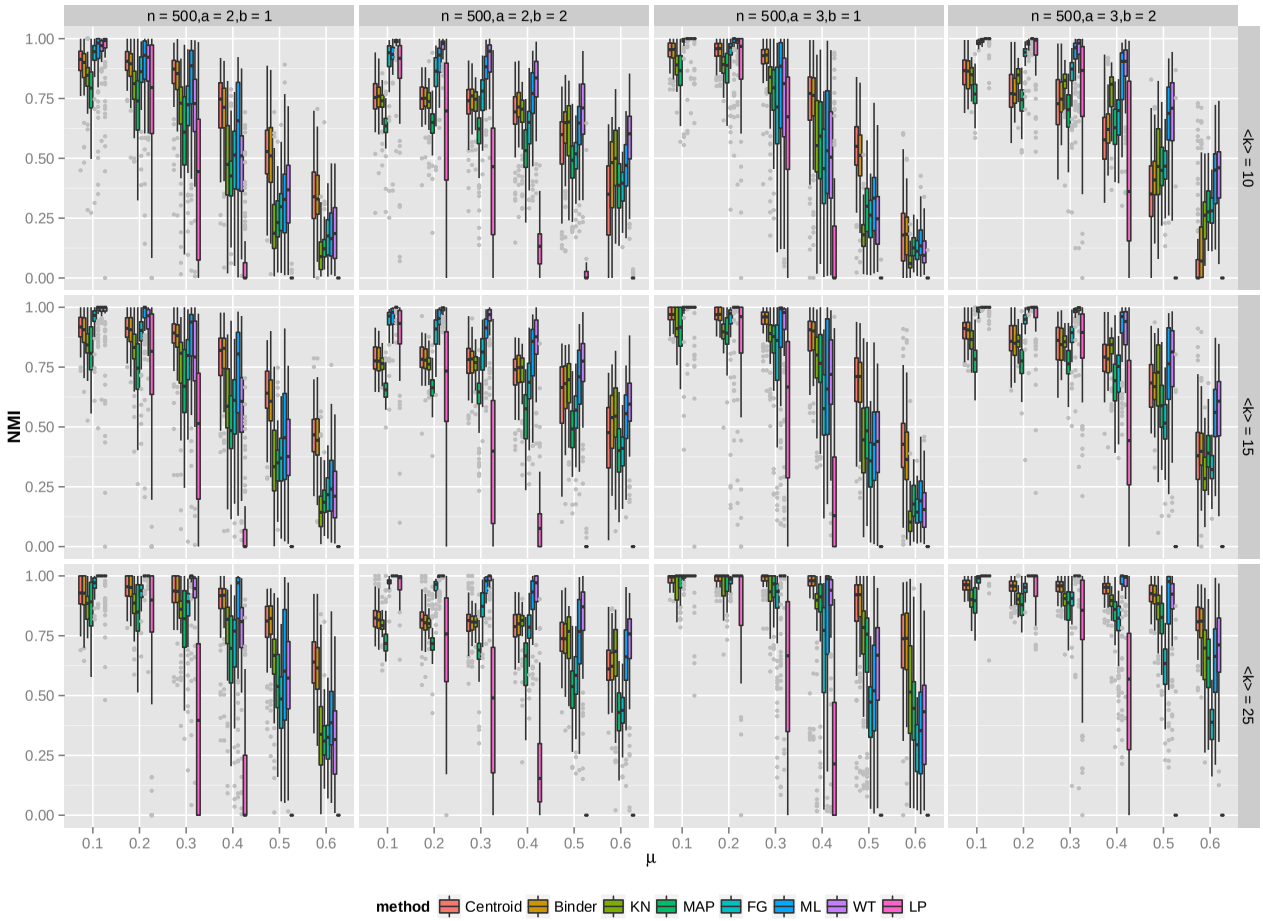


Figure 6.2: Benchmark networks of $n = 500$ nodes, with different combinations of the exponents $a \in \{2, 3\}$, $b \in \{1, 2\}$ and the average degree $\langle k \rangle \in \{10, 15, 25\}$. Each boxplot corresponds to the NMI of the estimator over 100 graph realizations.

# Bibliography


Adamic, L. 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.

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

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97.

Anderson, C., Wasserman, S., and Faust, K. (1992). Building stochastic blockmodels. *Social Networks*, 14(1):137–161.

Barbieri, M. and Berger, J. (2004). Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897.

Barnes, E. (1982). An algorithm for partitioning the nodes of a graph. *SIAM Journal on Algebraic Discrete Methods*, 3(4):541–550.

Batagelj, V., Mrvar, A., and Doreian, P. (2005). Generalized blockmodeling with pajek.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):259–302.

Bickel, P. 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., Choi, D., Chang, X., and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4):1922–1943.

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika*, 65(1):31–38.

Binder, D. A. (1981). Approximations to bayesian clustering rules. *Biometrika*, 68(1):275–285.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Bonacich, P. and Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23(3):191–201.

Borgatti, S. P. and Everett, M. G. (2000). Models of core/periphery structures. *Social networks*, 21(4):375–395.

Box-Steffensmeier, J. M. and Christenson, D. P. (2014). The evolution and formation of amicus curiae networks. *Social Networks*, 36(0):82 – 96. Special Issue on Political Networks.

Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2007). On finding graph clusterings with maximum modularity. In *Graph-Theoretic Concepts in Computer Science*, pages 121–132. Springer.

Carvalho, L. and Lawrence, C. (2008). Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proceedings of the National Academy of Sciences*, 105(9):3209–3214.

Celisse, A., Daudin, J.-J., and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899.

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

Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6):066111+.

Danon, L., Díaz-guilera, A., and Duch, J. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, page 09008.

Daudin, J. J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, pages 157–175.

Donath, E. and Hoffman, J. (1973). Lower bounds for the partitioning of graphs. *IBM J. Res. Dev.*, 17(5):420–425.

Drton, M. and Perlman, M. D. (2004). Model selection for gaussian concentration graphs. *Biometrika*, 91(3):591–602.

Duch, J. and Arenas, A. (2005). Community identification using extremal optimization. *Physical Review E*, 72:027104.

Duran-Pinedo, A. E., Paster, B., Teles, R., and Frias-Lopez, J. (2011). Correlation network analysis applied to complex biofilm communities. *PloS one*, 6(12):e28438.

Fienberg, S. E., Meyer, M. M., and Wasserman, S. S. (1985). Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80(389):51–67.

Fienberg, S. E. and Wasserman, S. (1981). An exponential family of probability distributions for directed graphs: Comment. *Journal of the American Statistical Association*, 76(373):54–57.

Fortunato, S. and Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41.

Fosdick, B. and Hoff, P. (2013). Testing and modeling dependencies between a network and nodal attributes. *arXiv:1306.4708v1*.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Fritsch, A. and Ickstadt, K. (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, 4(2):367–392.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. CRC press.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Goodnight, J. H. (1979). A tutorial on the sweep operator. *The American Statistician*, 33(3):149–158.

Gower, J. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338.

Hancock, T., Takigawa, I., and Mamitsuka, H. (2010). Mining metabolic pathways through gene expression. *Bioinformatics*, 26(17):2128âĂŞ2135.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2001). Maximum likelihood from incomplete data via the em algorithm. *The Elements of Statistical Learning*, pages 520–528.

Hoff, P., Raftery, A., and Handcock, M. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Hoff, P., Raftery, A., and Handcock, M. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, 100(469):286–295.

Hofman, J. and Wiggins, C. (2008). Bayesian approach to network modularity. *Physical Review Letters*, 100(25):258701.

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

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

Hula, K. (1999). *Lobbying Together: Interest Group Coalitions in Legislative Politics*. American Governance and Public Policy Series. Georgetown University Press.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *Annals of Statistics*, pages 730–773.

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

Kernighan, B. and Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. *Bell Sys. Tech. J.*, 49(2):291–308.

Kim, M. and Leskovec, J. (2011). Modeling social networks with node attributes using the multiplicative attribute graph model. *UAI*, 7AUAI Press:400–409.

Krebs, V. (2004). unpublished.

Krivitsky, P. N. (2012). Exponential-family random graph models for valued networks. *Electron. J. Statist.*, 6:1100–1128.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008a). Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(1):046110.

Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008b). Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(1):046110.

Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558.

Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press.

Lorrain, F. and White, H. C. (1971). Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology*, 1(1):49–80.

Mariadassou, M., Robin, S., and Vacher, C. (2010). Uncovering latent structure in valued graphs: A variational approach. *The Annals of Applied Statistics*, 4(2):715–742.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, volume 37. CRC Press.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462.

Newman, M. (2002). Assortative mixing in networks. *Phys. Rev. Lett.*, 89:208701.

Newman, M. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066133.

Newman, M. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

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

Newman, M. E. J. (2003). Mixing patterns in networks. *Phys. Rev. E*, (67).

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer-Verlag, 2nd edition.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087.

Parthasarathy, S., Ruan, Y., and Satuluri, V. (2011). Community discovery in social networks: Applications, methods and emerging trends. In *Social Network Data Analytics*, pages 79–113. Springer.

Polson, N. G., Scott, J. G., and Windle, J. (2012). Bayesian inference for logistic models using polya-gamma latent variables. *arXiv preprint arXiv:1205.0310*.

Pons, P. and Latapy, M. (2004). Computing communities in large networks using random walks. *J. of Graph Alg. and App. bf*, 10:284–293.

Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3).

Robert, C. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer New York.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential random graph ($p^*$) models for social networks. *Social networks*, 29(2):173–191.

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

Rombach, M., Porter, M., Fowler, J., and Mucha, P. (2014). Core-periphery structure in networks. *SIAM Journal on Applied Mathematics*, 74(1):167–190.

Sampson, S. F. (1968). *A novitiate in a period of change: an experimental and case study of social relationships*. PhD thesis, Cornell University, September.

Snijders, T. A. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100.

Socransky, S., Haffajee, A., Cugini, M., Smith, C., and Kent, R. (1998). Microbial complexes in subgingival plaque. *Journal of clinical periodontology*, 25(2):134–144.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society. Series B*, 62(4):795–809.

Tallberg, C. (2005). A bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology*, 29:1–23.

Vázquez, A. (2003). Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, 67(5):056104.

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

Vu, D. Q., Hunter, D. R., and Schweinberger, M. (2013). Model-based clustering of large networks. *Annals of Applied Statistics*, 7(2):1010–1039.

Wang, Y. J. and Wong, G. Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19.

Whittaker, J. (2009). *Graphical models in applied multivariate statistics*. Wiley Publishing.

Yang, J. and Leskovec, J. (2012). Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12, pages 3:1–3:8, New York, NY, USA. ACM.

Yuan, M. (2008). Efficient computation of âĎŞ1 regularized estimates in gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17(4):809–826.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473.

Zanghi, H., Picard, F., Miele, V., and Ambroise, C. (2010). Strategies for online inference of model-based clustering in large and growing networks. *The Annals of Applied Statistics*, 4(2):687–714.

# Curriculum Vitae

*Contact* Lijun Peng

Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, MA 02215, USA

*Education* **Zhejiang University**, B.S., Mathematics , September 2006 – June 2010.

**Boston University** PhD candidate, September 2004 – present. Thesis advisor: Luis Carvalho.

*Publications*

1. Lijun Peng and Luis Carvalho, *Group-Corrected Stochastic Block Models for Community Detection on Large-scale Networks.* 2014 NIPS Workshop, Networks: From Graphs to Rich Data.
2. Lijun Peng and Luis Carvalho, *Bayesian Ridge-Regularized Covariance Selection with Community Behavior in Latent Gaussian Graphical Models. Interdisciplinary Bayesian Statistics*, pp 207-216.
3. Lijun Peng and Luis Carvalho, *Bayesian Degree-corrected Stochastic Block Models for Community Detection.* arXiv:1309.4796.