Journal of Machine Learning Research 24 (2023) 1-36 Submitted 3/22; Revised 9/23; Published 10/23

Community models for networks observed through edge nominations

Tianxi Li TIANXILI@UMN.EDU
Department of Statistics
University of Virginia
Charlottesville, VA 22904, USA

Elizaveta Levina ELEVINA@UMICH.EDU

Ji Zhu JIZHU@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109, USA

Editor: Xiaotong Shen

Abstract

Communities are a common and widely studied structure in networks, typically assuming that the network is fully and correctly observed. In practice, network data are often collected by querying nodes about their connections. In some settings, all edges of a sampled node will be recorded, and in others, a node may be asked to name its connections. These sampling mechanisms introduce noise and bias, which can obscure the community structure and invalidate assumptions underlying standard community detection methods. We propose a general model for a class of network sampling mechanisms based on recording edges via querying nodes, designed to improve community detection for network data collected in this fashion. We model edge sampling probabilities as a function of both individual preferences and community parameters, and show community detection can be performed by spectral clustering under this general class of models. We also propose, as a special case of the general framework, a parametric model for directed networks we call the nomination stochastic block model, which allows for meaningful parameter interpretations and can be fitted by the method of moments. In this case, spectral clustering and the method of moments are computationally efficient and come with theoretical guarantees of consistency. We evaluate the proposed model in simulation studies on unweighted and weighted networks and under misspecified models. The method is applied to a faculty hiring dataset, discovering a meaningful hierarchy of communities among US business schools.

Keywords: Community detection; edge nomination; partial networks; spectral method; method of moment

1. Introduction

Networks have been widely used to describe relationships between individuals or interactions between units of complex systems in numerous fields (Newman, 2010). Community detection, the task of clustering nodes into groups with relatively homogeneous connection patterns, has been intensively studied since communities occur naturally in many real-world networks (Fortunato, 2010). Many statistical network models with communities have now

been proposed, from the simple stochastic block model (Holland et al., 1983) to more complex extensions with mixed membership (Airoldi et al., 2008) or temporal evolution (Xu and Hero, 2013; Matias and Miele, 2017). Such models can provide a rigorous statistical framework and theoretical performance guarantees (see, for example, Rohe et al. (2011); Zhao et al. (2012)), as well as lead to improved algorithms, e.g., Joseph and Yu (2016); Gao et al. (2017).

A practical difficulty in many empirical studies of networks arises from imperfect data collection. We loosely use the term "edge nomination" for any situation where edge information is obtained through a data collection mechanism which may not record the entire networks. This may occur in observational studies where interactions or connections are recorded from observations of the experimenters (Hass, 1991), and interactions not observed would be missing from the data. This can also include traditional surveys, since many social networks are constructed by asking subjects to name their friends or contacts (Harris, 2009). Sometimes these surveys limit how many friends one can name, and sometimes subjects may choose to name their friends selectively. Another example is internet crawlers that follow only a subset of the paths (Clauset et al., 2015; Ji and Jin, 2016). In all these situations, the missing edges may undermine the validity or efficiency of standard network analysis methods.

One important property that often arises in the aforementioned settings is that which edges are missing may depend on the properties of the individual node reporting them, which automatically invalidates all missing completely at random assumptions. We will use the term *nomination network* to refer to any situation where the missing edge pattern may depend on the node from which the edge information is collected.

Missing edges in networks can also be viewed as erroneous observations (a 0 instead of a 1). There has been a significant amount of work on denoising networks, which often considers both missing edges and falsely reported edges. Butts (2003) propose a Bayesian method to evaluate how reliable an observed network is. Following a similar set of model assumptions, Newman (2018a) propose a link prediction framework to recover underlying networks without specific structures. Newman (2018b) extends this work to a general framework to estimate networks under non-informative observational errors. Under the framework of exponential random graph models, Handcock and Gile (2010) study ways to handle general ignorable missing mechanisms. Related link prediction problems are studied in Zhao et al. (2017). However, Zhao et al. (2017) focus on the general model-free link prediction without specific structural assumptions and are not directly applicable to community detection problems. For networks with communities, Guimerà and Sales-Pardo (2009) propose a Bayesian model and inference method to detect both missing and spurious edges. Martin et al. (2016) take a similar modeling strategy but assume more flexible nonparametric error distributions. All these models for noisy networks assume the missing mechanism is independent of any network structure such as communities. In some situations, this assumption is reasonable, for instance, for recording errors. But for a network resulting from a survey, such an assumption is hard to justify. For example, in a high school survey of friendships, there may be individual differences in whether to name friends from their own"true" community. Here the missing mechanism potentially depends on both the community structure and individual node characteristics, requiring different models from those used for network denoising. Recently, Le and Levina (2017) considered a

scenario where the missing mechanism depends on the community labels of the node pairs in the context of jointly analyzing multiple networks sampled from the same probability model, but their method does not apply to a single network. Another challenge of modeling the missing edge mechanism in community detection problems lies in the computation. While likelihood-based approaches are widely used to modeling missing data, such methods are generally computationally infeasible for community detection problems. Variational inference can be used to approximate the likelihood in these setting, as studied in Tabouy et al. (2020). However, it is still far from being scalable for large scale networks.

In this paper, we introduce a general framework of modeling communities in networks collected from the edge nomination and collection procedure where the observed relations suffer from missingness. The framework can be used for both unweighted and weighted networks. We also propose a new directed network model we call the *nomination stochastic block model* (NSBM), a special case of the general framework which has interpretable model parameters. Under this model, we propose computationally efficient model fitting algorithms based on spectral clustering and the method of moments and show statistical consistency for both communities and estimated parameters.

2. Community models for networks with edge nominations

2.1 A general nomination framework based on the directed stochastic block model

The stochastic block model (SBM) (Holland et al., 1983) is one of the most widely used and well-understood models for communities in a network. It has been shown to recover communities in various settings successfully and can serve as a building block for more complicated models.

A network of n nodes can be represented by an $n \times n$ adjacency matrix A such that each entry $A_{ij} = \mathbf{1}(i \to j)$ is 1 if there is an edge from node i to node j and 0 otherwise. In particular, $A_{ii} = 1$ indicates a self-loop: $i \to i$. The standard SBM is defined for undirected networks, where $A_{ij} = A_{ji}$. While it is not easy to trace the start of its natural directed extension, the directed SBM is studied by Rohe et al. (2016), which in our context, reduces to the following model: given n nodes, a positive integer K and a $K \times K$ matrix of probabilities B, let $c_i \in \{1, \ldots, K\}$ be the community label of node i, and $\boldsymbol{c}$ be the vector of community labels. Here we treat $\boldsymbol{c}$ as fixed. Let $G_k = \{i : c_i = k\}$ be the set of nodes in community k and $n_k = |G_k|$. The entries of the adjacency matrix A are then generated independently from the Bernoulli distribution with $P(A_{ij} = 1) = B_{c_i c_j}$. The difference between the undirected and the directed models is that the directed model does not require B to be symmetric. Throughout the paper, we will call this directed version SBM by default.

Errors in recording network edges are common and can arise in a variety of ways. We focus on the situation when some connections are missing but all observed connections are true edges. This is different from the setting considered in Zhao et al. (2017); Newman (2018a), where falsely reported edges are also allowed, but it is a reasonable assumption in many applications (Zachary, 1977; Hass, 1991; Connor et al., 1992; Gleiser and Danon, 2003; Clauset et al., 2015; Ji and Jin, 2016). In particular, this is exactly the setting for the network of hiring relationships analyzed in Section 5. Let $\tilde{A}$ be the adjacency matrix

we observe, with potentially missing edges, where $\tilde{A}_{ij} = 1$ indicates node i reported that there is an edge from it to node j. The generating process for $\tilde{A}$ can be thought of as taking the original network A generated from the SBM and applying a binary nomination "mask" matrix $R \in \{0,1\}^{n\times n}$, so that the observed matrix is given by

$$\tilde{A} = A \circ R,$$

where $\circ$ is the element-wise Hadamard matrix product. Here $R_{ij} = 1$ indicates that node i revealed its connection to node j. We assume a nominated link is always a true link in A, while $\tilde{A}_{ij} = 0$ may result from either $A_{ij} = 0$ or $R_{ij} = 0$.

We base our model for R_{ij} on the following two considerations. By nature of the edge nomination process, the probability of the edge A_{ij} being reported by node i should depend on node i. It is natural to assume that it also depends on the closeness between the communities of i and j, which can be expressed through $B_{c_i c_j}$. We therefore propose the following general model for the observed network:

$$A_{ij} \overset{indep.}{\sim} \text{Bernoulli}(B_{c_i c_j}), \quad R_{ij} \overset{indep.}{\sim} \text{Bernoulli}(f_i(B_{c_i c_j})), \quad \tilde{A}_{ij} = A_{ij} \cdot R_{ij} \tag{1}$$

where $f_i : [0,1] \to [0,1]$ is the *nomination function* of node i. This general form includes some of the previously studied settings. For example, when $f_i \equiv \rho_i$ for all $i \in [n]$, every node randomly nominates each of its links with a fixed probability ρ_i, the setting studied in Butts (2003).

In most situations, we are interested in learning about properties of the network expressed in $\boldsymbol{c}, B$, or f_i's rather than predicting the latent status R_{ij}. We can integrate out R_{ij} and write the distribution of $\tilde{A}$ directly as

$$\mathbb{P}(\tilde{A}_{ij} = 1) = \tilde{P}_{ij} = B_{c_i c_j} f_i(B_{c_i c_j}) = F_i(B_{c_i c_j}) \tag{2}$$

where $F_i(x) = x f_i(x)$. The general model defined by (2) can be specialized to many different forms by specifying f_i. Model (2) is explicitly incorporating an informative missing mechanism through its dependence on the strength of the connection between nodes.

2.2 Community detection under the general edge nomination model

A general model like (2) allows for developing a general algorithm for solving problems of this type. Spectral clustering algorithms are among the most popular methods for estimating community labels due to their computational efficiency, ease of implementation, and excellent theoretical properties. Many versions of spectral clustering have been proposed and studied for community detection, but the general strategy is to use the eigenvectors of a matrix as input to a standard clustering algorithm, with the matrix chosen so that the population version of its eigenvectors reflects the true communities. While spectral clustering was originally proposed for undirected networks, it can be generalized to directed networks by using appropriate matrices as input (Chung, 2005; Zhou and Burges, 2007; Li and Zhang, 2010); refer to Malliaros and Vazirgiannis (2013) for a comprehensive review. Our goal in this paper is to show that a simple version of spectral clustering used in the seminal papers Rohe et al. (2011), Lei and Rinaldo (2014) and Rohe et al. (2016) can identify the communities, as long as it is applied to the correct spectral space. There may

be room for improvement on the details of the spectral clustering implementation (such as normalization, regularization, and so on), which is outside the scope of this work.

For a standard undirected SBM, both the row space and the column space contain community information. Under model (2), since each node uses an individual function F_i to nominate links, we would expect the node-specific nomination function to confound the community information in the row space of $\tilde{A}$. However, the *column* space of $\tilde{A}$ should still reflect communities since each node i applies the same function to all entries of the column j. This intuition is rigorously justified in Proposition 7 in Section B.2 in the Appendix. It suggests that the right singular vectors of $\tilde{A}$ can be used to recover communities, as long as $\tilde{A}$ concentrates around $\tilde{P}$. The column space is equivalent to the subspace of the right singular vectors. Therefore, the strategy of Lei and Rinaldo (2014) can be applied to the column space for community recovery. This procedure is fully described in the "Right singular vectors Spectral Clustering" (Right SC) algorithm below. Note that although we take the standard approach of applying spectral clustering to the adjacency matrix $\tilde{A}$, one could replace it with the Laplacian as in Rohe et al. (2011, 2016), or a regularized version of either matrix as in Amini et al. (2013); Joseph and Yu (2016); Le et al. (2017). We do not focus on investigating these options here, although we have obtained similar empirical results in experiments replacing the adjacency matrix with the Laplacian (results not shown for lack of space).

Algorithm 1 (Right SC) *Given an adjacency matrix $\tilde{A}$ and the number of communities K:*

1. *Compute the rank K truncated SVD $\tilde{A}$, given by $\tilde{A} = \widehat{U}\widehat{D}\widehat{V}^T$.*
2. *Run the K-means clustering algorithm on rows of $\widehat{V}$ to assign each node to a community.*

2.3 The nomination stochastic block model (NSBM)

The general model (2) allows for a common algorithm of community detection. However, in many situations, in addition to community labels, one may also be interested in learning the nomination pattern F_i. Making the so far unspecified functions F_i both estimable and interpretable requires further modeling. Next, we introduce a specific nomination model under the framework of (2), which we believe achieves a good balance between generality and interpretability. In addition to the previously defined $\boldsymbol{c}$ and B, we introduce two new node-specific parameters, given by n-dimensional vectors $\boldsymbol{\lambda} = (\lambda_i)$ and $\boldsymbol{\theta} = (\theta_i)$. The proposed nomination stochastic block model (NSBM) assumes

$$f_i(B_{c_i c_j}) = \theta_i B_{c_i c_j}^{\lambda_i - 1}, i \in [n].$$

The parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ are easily interpretable. We can think of the parameter θ_i as measuring the overall propensity of node i to nominate links, and of λ_i as a measure of their preference for nominating links from their own or closely connected communities; both these factors may affect data collection in many applications. For example, suppose that $B_{kk} > B_{kj}, k \neq j$ so that the SBM is assortative. In this case, $\lambda_i > 1$ indicates that the node i tends to nominate connections from its own community while $\lambda_i < 1$ indicates a preference

for nominating connections from a different community. The marginal distribution of $\tilde{A}$ under the proposed NSBM is given by

$$\mathbb{P}(\tilde{A}_{ij} = 1) = \theta_i B_{c_i c_j}^{\lambda_i}. \tag{3}$$

In the current context of modeling Bernoulli probabilities, the model needs an implicit parameter constraint $\theta_i B_{c_i c_j}^{\lambda_i} \leq 1$ for all $i, j \in [n]$. This constraint usually does not explicitly impact later model estimation, as the key quantities (5)–(7) for the estimation algorithm automatically satisfy the constraint when the model itself is well-defined. So we will not focus on it in later discussions. This constraint can be dropped if we use the current model structure for a weighted network, as discussed in Section 2.5.

As with any model involving products of parameters, we need additional constraints for identifiability. We require that $\tilde{P}$ has no rows consisting entirely of zeros, and thus we require $\theta_i > 0$ for all i and that each row of B contains at least one positive entry. In addition, if $B_{kl} = 0$ for all $l \neq k$, then community k will not send edges to other communities and it will be impossible to identify λ_i's for nodes in community k. In addition, if $B_{kl} = B_{kk}$ for all l, then community k is not identifiable. We also need scaling constraints on B and $\boldsymbol{\lambda}$ to avoid invariance multiplying by a constant. Putting all these together leads to the following identifiability conditions.

Proposition 1 *The parameters of model* (3) *is identifiable if the following conditions hold:*

1. $B_{kk} = 1$ *for all* $k = 1, \ldots, K$.
2. *For each* k, *there exists at least one* $l \neq k$ *such that* $B_{kl} \neq B_{kk}$ *and* $B_{kl} \neq 0$.
3. $\theta_i > 0$ *for all* $i = 1, \ldots, n$.
4. $\frac{1}{n_k} \sum_{i \in G_k} \lambda_i = 1$ *for all* $k = 1, \ldots, K$, *where* $n_k = |G_k|$.

Compared with the general model (2), the NSBM offers the possibility of fitting an interpretable nomination mechanism model and learning each node's preference. The price we pay for interpretability, as usual, is less flexibility, since we now have parametric model assumptions. For example, the requirement $\theta_i > 0$ excludes the egocentric sampling mechanism (Tabouy et al., 2020; Li et al., 2023; Chan and Li, 2023), whereas the general model (2) includes it.

2.4 Parameter estimation under the NSBM

Given community labels $\boldsymbol{c}$, the other parameters in model (3) can be estimated by the method of moments under the identifiability constraints of Proposition 1. Specifically, if $B_{kl} > 0$, for any arbitrary $i \in G_k$ and $j \in G_l$, we have

$$\log(\tilde{P}_{ij}) = \mu_{il} = \log(\theta_i) + \lambda_i \log(B_{kl}). \tag{4}$$

Combining Proposition 1 and (4), we obtain the following identities:

$$\theta_i = \tilde{P}_{ij} \text{ for any } j \in G_k\,, \tag{5}$$

$$B_{kl} = \exp(-\frac{1}{n_k}\sum_{i\in G_k}(\mu_{ik}-\mu_{il}))\,, \tag{6}$$

$$\lambda_i = \frac{\mu_{ik}-\mu_{il}}{\sum_{j\in G_k}(\mu_{jk}-\mu_{jl})/n_k}\,,\ \text{if } B_{kk}\neq B_{kl}\,. \tag{7}$$

Therefore, we can use the method of moments to estimate μ_{il} by $\exp(\hat{\mu}_{il}) = \frac{1}{n_l}\sum_{j\in G_l}\tilde{A}_{ij}$ and plug it back to previous identities to estimate other parameters. This is summarized in the following algorithm.

Algorithm 2 (Parameter estimation for the NSBM by the method of moments) *Given the adjacency matrix $\tilde{A}$ and community labels $\boldsymbol{c}$, for $k = 1, 2, \cdots, K$ (obtained by, for example, right SC):*

1. *Set $T_{il} = \frac{\sum_{j\in G_l}\tilde{A}_{ij}}{n_l}$ for each $i \in G_k$ and $1 \le l \le K$.*
2. *Estimate θ_i for each $i \in G_k$ by $\hat{\theta}_i = T_{ik} \vee \frac{1}{n_k}$.*
3. *Find the set $\Psi_k = \{l : 1 \le l \le K, T_{il} > 0 \ \ \forall i \in G_k\}$. Set $\hat{B}_{kl} = 0$ for each $l \notin \Psi_k$.*
4. *(a) Define $Y_{il} = \log(T_{il} + \frac{1}{n_l})$ for each $i \in G_k$, where the $\frac{1}{n_l}$ is used to avoid overflow for the pathological case of $T_{il} = 0$ for some $i \in G_k$.*

   (b) Estimate λ_i for each $i \in G_k$ by

$$\hat{\lambda}_i = \frac{\sum_{l\in\Psi_k\setminus\{k\}}(Y_{ik}-Y_{il})}{\sum_{l\in\Psi_k\setminus\{k\}}\sum_{j\in G_k}(Y_{jk}-Y_{jl})/n_k} \tag{8}$$

   (c) Estimate B_{kl} for each $l \in \Psi_k \setminus \{k\}$ by

$$\hat{B}_{kl} = \exp(-\frac{1}{n_k}\sum_{i\in G_k}(Y_{ik}-Y_{il})). \tag{9}$$

2.5 Extensions

Our modeling strategy can be extended to handle other community models for the underlying true network, with appropriate modifications. We briefly discuss three possible extensions: weighted networks, undirected networks, and the degree-corrected SBM.

Networks with weighted edges are frequently encountered in practice. The NSBM can be applied directly to weighted networks: given community labels $\boldsymbol{c}$, assume each edge weight $\tilde{A}_{ij}$ is independently generated from a probability distribution satisfying

$$\mathbb{E}_\pi \tilde{A}_{ij} = \theta_i B^{\lambda_i}_{c_i c_j}. \tag{10}$$

The choice of weight distribution will depend on the problem at hand. For instance, the Poisson distribution is a popular choice for non-negative integer weights (Karrer and Newman,

2011). In this case, we can interpret A_{ij} as the number of interactions with node j coming from node i, and model it as generated from Poisson($B_{c_i c_j}$). Then $\tilde{A}_{ij}$ can be interpreted as the number of interactions chosen randomly by node i from Binomial($A_{ij}, \theta_i B_{c_i c_j}^{\lambda_i - 1}$) to report as their relationship with node j. Again, we are assuming that only true interactions are reported, so that $\tilde{A}_{ij} \leq A_{ij}$. Since the model is specified through the expectation of $\tilde{A}$, we can still apply the right spectral clustering and method of moments algorithms, and similar theoretical guarantees can be obtained as long as the generating distributions of the edge weights are not heavy-tailed. Since model (10) only specifies the mean structure of $\tilde{A}$, other constraints and modifications may be necessary if the edge distribution depends on other parameters.

Undirected networks have been the focus on most community detection work to date, and the standard SBM is a model for undirected network. Replacing the underlying network model in our setup with a standard SBM would result would not change the moment assumptions, and both community detection and model estimation can be done exactly the same way, though some details of the theoretical analysis will be slightly different.

Finally, a popular generalization of the SBM is the degree-corrected SBM (DCSBM) (Karrer and Newman, 2011; Rohe et al., 2016), introduced to account for degree heterogeneity frequently observed in real networks. If we assume that the underlying unweighted network follows the DCSBM model. That is,

$$\mathbb{E}A_{ij} = \psi_i \bar{\psi}_j B_{c_i c_j}.$$

With the same nomination model, the observed network now has

$$\mathbb{E}(\tilde{A}_{ij}) = \psi_i \bar{\psi}_j B_{c_i c_j} f_i(B_{c_i c_j}) = \theta_i \psi_i \bar{\psi}_j B_{c_i c_j}^{\lambda_i}.$$

In this case, only the product $\theta_i \psi_i$ is identifiable, so we can reparameterize and combine them into a new θ_i, fitting instead the model

$$\mathbb{E}(\tilde{A}_{ij}) = \theta_i \bar{\psi}_j B_{c_i c_j}^{\lambda_i}.$$

The logic of Section 2.2 still applies, and the right singular vectors contain the community information. The only difference is that the right singular vectors also reflect degree heterogeneity, so they should be normalized before clustering, using one of the normalization approaches developed for fitting DCSBM by spectral clustering (Jin, 2015; Lei and Rinaldo, 2014). Again, we can also handle weighted networks using the same strategy mentioned before.

3. Theoretical properties

Here we investigate asymptotic properties of community detection under the general model (2) and parameter estimation under the NSBM. We always assume that the B matrix is full-rank, and the number of communities K is known and fixed. In practice, K can be estimated by many data-driven methods such as the edge cross-validation of Li et al. (2020) by taking advantage of the property that the rank of $\tilde{P}$ equals the number of communities. We first introduce an additional assumption we need for theoretical developments, which is that none of the communities vanish relatively to the size of others as n grows.

Assumption A1 *Assume that* $n_{\min} := \min_k n_k \geq \kappa' n$ *for some constant* $\kappa' > 0$. *Also define* $n_{\max} = \max_k n_k$.

Theorem 2 (Consistency of the Right SC algorithm) *Assume the network* $\tilde{A}$ *is generated from model* (2). *Let* $\hat{\mathbf{c}}$ *be the output of the Right SC algorithm with a* $(1+\epsilon)$*-optimal solution,* $\sigma_K(\tilde{P})$ *be the* K*th largest singular value of* $\tilde{P}$, *and* $\|\tilde{P}\|_\infty = \max_{ij} \tilde{P}_{ij}$. *Assume A1 holds, and* $n\|\tilde{P}\|_\infty \geq C_0 \log n$ *for some constant* C_0. *If there exists a constant* C_1 *depending on* C_0, ϵ *and* κ', *such that*

$$\frac{Kn\|\tilde{P}\|_\infty}{\sigma_K(\tilde{P})^2} \leq \frac{1}{C_1}, \tag{11}$$

then with probability at least $1 - n^{-1}$, *there exists a permutation of labels* $\hat{\mathbf{c}}$, *such that*

$$\sum_k \frac{|G_k \setminus \hat{G}_k|}{n_k} \leq C_1 \frac{Kn\|\tilde{P}\|_\infty}{\sigma_K(\tilde{P})^2}.$$

Theorem 2 states that under the general model (2), the proportion of nodes misclustered by the RightSC algorithm is bounded above by the quantity $\frac{Kn\|\tilde{P}\|_\infty}{\sigma_K(\tilde{P})^2}$. Therefore, label estimation consistency is achieved when $\frac{Kn\|\tilde{P}\|_\infty}{\sigma_K(\tilde{P})^2} \to 0$. However, this bound depends on $\sigma_K(\tilde{P})$, a quantity without an obvious interpretation. To help build intuition about this result, we next present a more interpretable form of this result under a specific parameterization of the NSBM which satisfies the following two assumptions:

Assumption A2 (Simple scaling) *Assume* B *is a fixed matrix and* $\lambda_{\min} \leq \lambda_i \leq \lambda_{\max}, i \in [n]$ *for two constants* $\lambda_{\min}$ *and* $\lambda_{\max}$. *Furthermore, there exist a scalar sequence* ρ_n, *such that* $\theta_i = \rho_n \bar{\theta}_i$ *where* $\bar{\theta}_{\min} \leq \bar{\theta}_i \leq \bar{\theta}_{\max}$ *for two positive constants* $\bar{\theta}_{\min}$ *and* $\bar{\theta}_{\max}$.

Assumption A3 (Discretized parametrization) *Under A2, further assume* $\bar{\theta}_i$*'s are independently sampled from a fixed discrete distribution* g_θ *on* m_1 *different positive values and* λ_i*'s are independently sampled from a fixed discrete distribution* g_λ *with mean value 1 on* m_2 *different values and then rescaled to satisfy the identifiability constraints in Proposition 1.*

Combining A2 and A3 allows us to parameterize the edge density of the network by a single parameter depending on n, ρ_n, a widely used strategy in analyzing community detection algorithms (Lei and Rinaldo, 2014; Gao et al., 2017; Abbe, 2018). The discrete assumption A3, while not the most natural, has also been used for interpretation purposes in previous work (Zhao et al., 2012), and can be made less restrictive by choosing a large number of values the discrete distributions can take. It is not needed for the estimation theory that follow, only for the interpretation provided in Corollary 3.

Corollary 3 *Assume the network is generated from the NSBM* (3). *Let* $\hat{\mathbf{c}}$ *be the clustering labels found by the Right SC algorithm with* $(1+\epsilon)$*-optimal solution. If assumptions of Proposition 1, A1, A2 and A3 hold, and* $n\rho_n \geq C_0 \log n$ *for some constant* C_0, *then for*

sufficiently large n, with probability at least $1-2n^{-1}$, there exists a permutation of labels $\hat{\mathbf{c}}$, such that

$$\sum_k \frac{|G_k \setminus \hat{G}_k|}{n_k} \leq C' \frac{1}{n\rho_n} \tag{12}$$

for some constant C' depending on $C_0, \kappa', \epsilon, \eta, K$ and the distributions of $\bar{\theta}_i$'s and λ_i's.

The Corollary 3 states that as long as the expected average degree of the network $n\rho_n$ grows at least in the order of $\log n$, the mis-clustered proportion is bounded by the order of $1/n\rho_n$.

Next, we show that $B, \boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ can be estimated consistently by Algorithm 2, under the NSBM (3) with parameterization A2. We assume known community labels for simplicity, since strong consistency can be established for a similar algorithm more amendable to theory, analogous to Vu (2018) and Lei and Zhu (2017). This algorithm and the resulting strong consistency of estimated community labels can be found in the supplementary material (Section A). In practice, we will always use the more accurate Right SC algorithm.

Theorem 4 *Assume the network is generated from the NSBM* (3)*. Let $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\lambda}}$ and $\hat{B}$ be the estimators for $\boldsymbol{\theta}, \boldsymbol{\lambda}$ and B, respectively, obtained by Algorithm 2. Assume conditions of Proposition 1, A1 and A2 hold. Then there exists constants c_1, c_2, c_3, depending on κ' in A1, $B, \bar{\theta}_{\min}, \bar{\theta}_{\max}, \lambda_{\min}, \lambda_{\max}$ in A2, and K, such that if $\rho_n \geq c_1 \frac{\log^4 n}{n}$ for sufficiently large n, we have*

$$\max_i \left|\hat{\theta}_i - \theta_i\right| \leq c_1 \frac{\log n}{\sqrt{n}}, \quad \max_i \left|\hat{\theta}_i - \theta_i\right| / \theta_i \leq \frac{1}{\log n}$$
$$\max_{k,l} |\hat{B}_{kl} - B_{kl}| \leq c_2 \max\left(\frac{\log^2 n}{\sqrt{n}}, \frac{1}{n\rho_n}\right)$$
$$\max_i |\hat{\lambda}_i - \lambda_i| \leq c_3 \max\left(\frac{\log^2 n}{\sqrt{n}}, \frac{\log n}{n\rho_n}\right)$$

with probability at least $1 - n^{-1}$.

Theorem 4 shows that the estimators are consistent when the average degree is on the order of $\log^4 n$.

4. Numerical results on synthetic networks

In this section, we demonstrate the proposed methods using simulation examples. We first show the importance of clustering based on the correct spectral information and that the NSBM cannot be approximated well by a few standard community models. We will illustrate this on both unweighted networks and weighted networks with Poisson-distributed edge weights. In Section 4.2, we evaluate our method under model specification.

4.1 Evaluation on networks from the NSBM

For this set of experiments, networks are generated from the NSBM as follows: $n = 1200$ nodes are randomly assigned to $K = 3$ communities with equal probability. The matrix

B has all diagonal elements equal to 1 and all off-diagonal elements equal to β. The parameters λ_i's are generated independently with $\log(\lambda)$ sampled uniformly from the interval $(-t, t)$, and then rescaled to satisfy the constraint $\sum_{\boldsymbol{c}_i=k} \lambda_i = n_k$ for each k. Each θ_i is independently set either to c or $0.05c$, with probability 0.5 each, with the value of c chosen so that the resulting average degree of the network is 50.

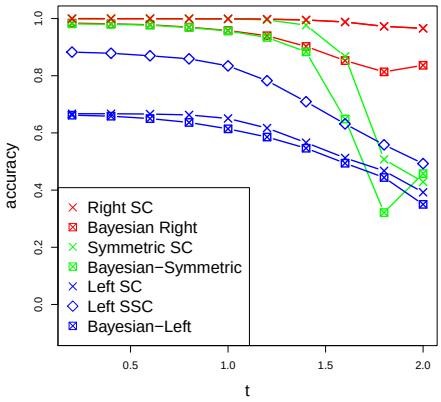
We evaluate several spectral clustering algorithms for community detection, showing importance of correctly identifying the informative part of the eigenstructure. For a directed network, one alternative to the Right SC algorithm is to cluster the left singular vectors ("Left SC"). However, under the NSBM Left SC will clearly fail, since it does not account for node heterogeneity. Therefore we instead consider the left *spherical* SC ("Left SSC"), which first normalizes each row of the matrix of left singular vectors before applying K-means clustering. We also tested the spherical version of Right SC (Right SSC), omitted here since its results are very similar to Right SC; this is as expected, as there is no need to normalize the right singular vectors to recover the community structure. The Right and Left SSC are using the spherical SC of (Lei and Rinaldo, 2014) on the right and left singular vectors, respectively, and if applied together, are essentially the co-clustering algorithm of Rohe et al. (2016). Our main purpose in comparing these versions of spectral clustering is to emphasize the importance of choosing the right singular vectors.

Another common approach to community detection in directed networks is to convert $\tilde{A}$ to a symmetric matrix and then apply an algorithm for an undirected network. This is typically accomplished by connecting two nodes in the undirected network if there is an edge in either direction. Applying SC and SSC to the symmetrized network gives two more options, "Symmetric SC" and "Symmetric SSC", but they again give similar results, and thus, we omit the spherical version. This strategy is equivalent to treating the network as generated from the SBM or the degree-corrected SBM (Karrer and Newman, 2011), respectively.

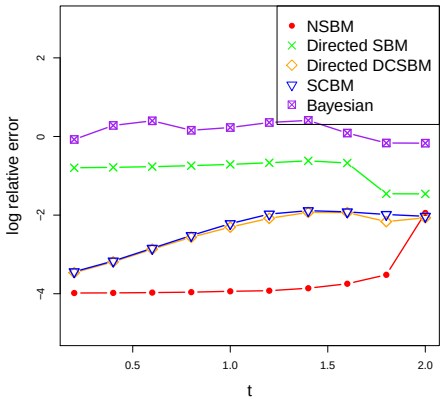
Lastly, we investigate other models for edge nomination. Specifically, we include the Bayesian method from Butts (2003), which includes the model of Newman (2018a) as a special case, assuming that the nomination process does not depend on the individual or the connections strength. This model is representative of the current literature on modeling missing links. The model, however, is not designed for community detection. Therefore, we take the posterior mean of the probability matrix as input for spectral clustering, which again results in "Left", "Symmetric" and "Right" versions. The posterior inference is implemented in R package `sna` (Butts, 2020) and we refer to the three versions as "Bayesian-Left", "Bayesian-Symmetric", and "Bayesian-Right". This method is computationally expensive and took a very long time to run; it cannot be applied to large networks.

We evaluate the community detection performance by using the cluster accuracy, defined to be mis-clustered proportion, under the best permutation within the K labels of the estimated clusters. In addition to community detection accuracy, we also evaluated the estimation error of $\tilde{P}$. Though using the NSBM is expected to give the best results, we are interested in how closely they can be approximated by using simpler models. Therefore, we compared results of Algorithm 2 to three other computationally feasible network models with communities: the directed SBM and its degree-corrected version, and the stochastic co-clustering block model (SCBM) of Rohe et al. (2016). Lastly, the Bayesian method of Butts (2003) also gives full posterior network distribution, and we take the posterior mean of

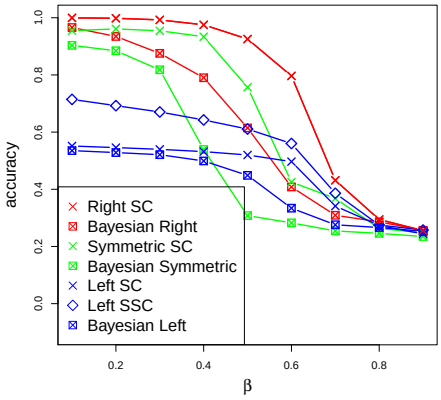
P together with the edge flipping probability to construct an estimate of $\tilde{P}$. The estimation accuracy is measured by the relative error $\|\tilde{P} - \hat{P}\|_F^2 / \|\tilde{P}\|_F^2$ averaged over 100 replications, where $\hat{P}$ is the estimated probability matrix.



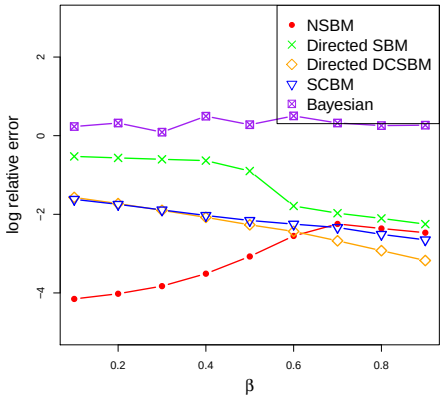
(a) Community detection accuracy.



(b) Estimation error of $\tilde{P}$ (log scale).



(c) Community detection accuracy.



(d) Estimation error of $\tilde{P}$ (log scale).

Figure 1: Community detection accuracy and probability matrix estimation error when the network is generated from the unweighted NSBM with $\beta = 0.2$ and varying t (Figure 1a-1b), and with $t = 1.5$ and varying β (Figure 1c-1d). The results are average values over 100 replications.

We start from varying t from 0.2 to 2 while keeping $\beta = 0.2$ fixed. The results are shown in Figure 1 (a-b). For community detection, all methods based on the right singular vectors are better than their counterparts using the other types of spectral structures. In particular, when t is small, the probability of nomination does not depend on the connection strength that much, and thus spectral clustering based on the standard SBM (or DCSBM) still works. As t increases and the nomination mechanism becomes more heterogeneous across the nodes, symmetric clustering methods fail. The Left SSC is even worse since it relies entirely on the senders' information, where the community structure is masked by heterogeneity of nominations. The Bayesian method is not effective in removing the

impact of the edge nomination effects. For estimating the probability matrix, the NSBM unsurprisingly works the best because it uses the correct model. More importantly, even for small values of t where symmetric methods perform ok on community detection, none of the other methods come close on estimating the probability matrix. Moreover, the estimation algorithm for the NSBM remains stable for most of the t in the range, only beginning to degrade when the clustering accuracy drops.

Next, we compare different methods while varying the signal-to-noise ratio, to see whether the effects of edge nomination become negligible when communities are well separated. Specifically, we vary β from 0.1 to 0.9 and fix $t = 1.5$. The corresponding results are shown in Figure 1 (c-d). The Right SC retains a consistent advantage over the entire range of β, though all methods fail to give informative clustering for $\beta \geq 0.7$. For model estimation, our method is better than the other for $\beta \leq 0.6$. For even larger β, because of the fading performance of clustering and the higher model complexity, the model estimation error becomes higher than the simpler SBM and SCBM. Notably, the directed SBM and SCBM slightly improve on estimation as β increases. This is because the two methods, while the two methods give poor community detection results, the model estimation is mainly based on averaging edges within detected blocks. As β grows, the probabilities of the true underlying model become more homogeneous, thus the estimation errors of the SCBM and the directed SBM becomes smaller.

We proceed to generate networks from a Poisson distribution with the same NSBM structure. The Bayesian method from the previous section is no longer applicable, but the other methods can still be used. We use the same two performance metrics and show that our method can be applied to weighted networks without any changes and has similar advantages over its competitors to what we observed in the previous section.

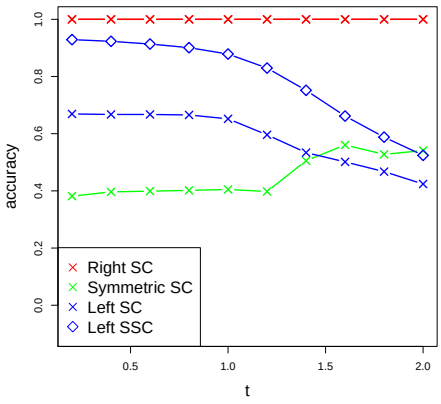
The only difference from the settings described above is that the value of c is set so that the average row sum of $\tilde{A}$ is 250, which roughly gives an average of 50 nonzero entries in each row of $\tilde{A}$, matching the degree of the unweighted networks. The results are shown in Figures 2. Our method retains the advantages it had on unweighted networks. Community detection becomes easier in this case, because Poisson distribution is more informative compared with Bernoulli. The Right SC remains accurate for the range of t from 0 to 2. The method is also very stable for most values of β, only deteriorating around $\beta > 0.8$.

4.2 Evaluation under model misspecification

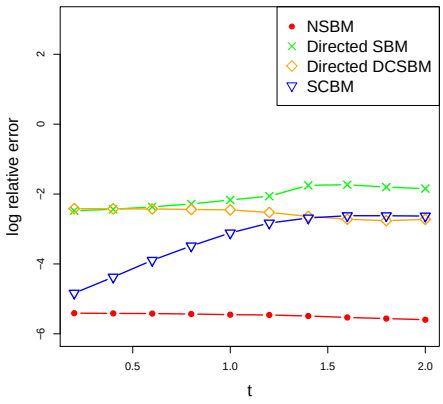
In practice, we never know the true model and may reasonably assume it is more complex than what we assume; however, we can also hope that our method will have some robustness to model misspecification. To investigate the degree of this robustness, we consider alternatives for both main components of our model: (1) the underlying network model, and (2) the edge nomination mechanism.

Alternative true network model. Instead of the SBM, we generate the true network from the random dot product graph model (Young and Scheinerman, 2007) with $n = 1200$. We first generate latent vectors $x_i \in \mathbb{R}^4, i = 1, \cdots, n$ from a Gaussian mixture model with $K = 3$ components $N(\mu_k, \sigma^2 I)$, and a uniform prior on components. The centers μ_k's are chosen so that $\mu_k^T \mu_{k'} / \mu_k^T \mu_k \approx 0.3$ for $k' \neq k$, to match the previous SBM setting. This gives the average distance between the K centroids $m_\mu \approx 1$. We use σ / m_μ as an intuitive

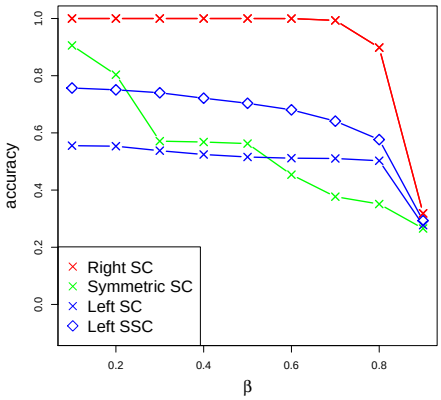
measure of separation between components and therefore of the difficulty of the problem. To guarantee positive edge probabilities, we truncate all x_i's to the positive quadrant. The probability matrix P is then given by $P_{ij} = x_i^T x_j$, rescaled so that the expected average degree is equal to 40.



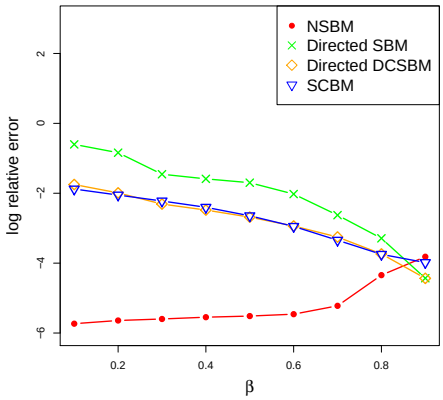
(a) Community detection accuracy.



(b) Estimation error of $\tilde{P}$ (log scale).



(c) Community detection accuracy.

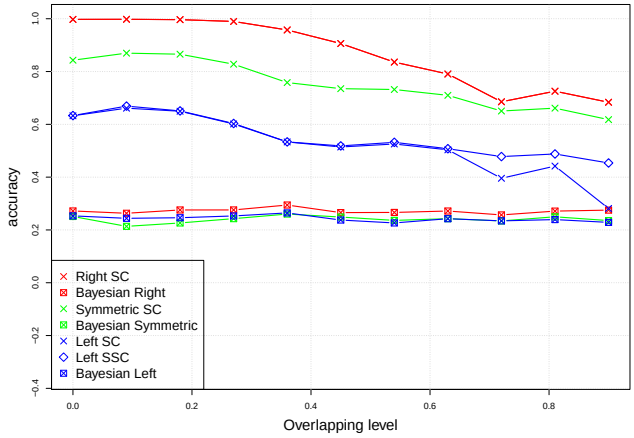


(d) Estimation error of $\tilde{P}$ (log scale).

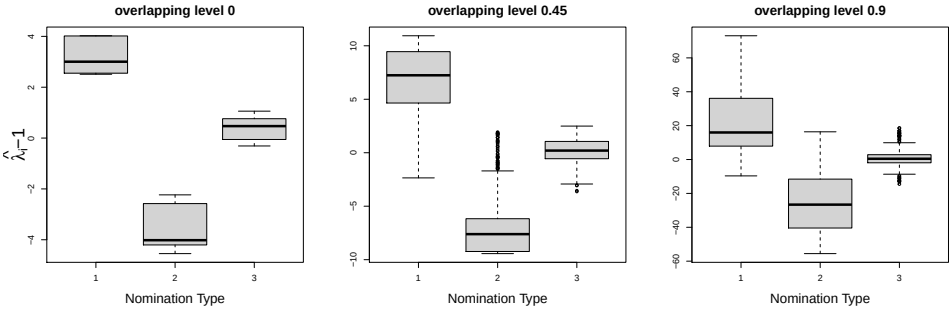
Figure 2: Community detection accuracy and probability matrix estimation error when the network is generated from NSBM with Poisson edge weights, with $\beta = 0.2$ and varying t (Figure 2a-2b), and with $t = 1.5$ and varying β (Figure 2c-2d). The results are average values over 100 replications.

Alternative edge nomination process. Here we consider a common survey scenario of limiting nominations per node. Specifically, we assume each node is allowed to nominate at most 15 connections. If the node has no more than 15 connections, all of them are nominated. If the node has more than 15, it will randomly choose which connections to nominate according to three types of preferences. Type 1: nominate as many nodes from its own community if possible. If there are no more than 15 in its own community, a Type 1 node will nominate all of them and fill in any remaining slots with nodes randomly chosen from the rest of the network; and if there are more than 15 in its own community, it will

randomly choose 15 of them to nominate. Type 2 is the opposite of Type 1: it always nominates connections to other communities if possible, choosing randomly among them if there are more than 15, and filling in the remaining spots with randomly chosen connections within its own community. Finally, Type 3 node nominates 15 randomly chosen connections without regard to communities. The three nomination types are randomly assigned to the 1200 nodes in equal proportions, independently of the community memberships.



(a) Community detection accuracy.



(b) The average $\hat{\lambda}_i$, centered at 1.

Figure 3: Results under misspecified model as a function of community overlap parameter σ/m_μ, averaged across 100 replications. Standard deviations over these 100 replications (not shown) are less than 1% of the mean values. 3a: Accuracy of community detection. 3b: Boxplots of average estimated $\hat{\lambda}_i$'s (centered at 1) for nodes of three different nomination types.

The estimated $\hat{\lambda}_i$'s are intuitively interpretable, especially when communities are well separated. Recall that, under the NSBM, $\lambda_i > 1$ indicates that node i tends to nominate edges from its own community (under assortativity) and $\lambda_i < 1$ indicates that the node tends to nominate edges from other communities. Since the model is misspecified, we cannot interpret the exact values of these parameters, but qualitatively we would expect that Type 1 nodes will have estimated $\hat{\lambda}$ s greater than 1, Type 2 less than 1, and Type 3 will be close to 1. Figure 3b shows boxplots of $\hat{\lambda}_i$'s for each of the three nomination node types, averaged over 100 replications. These values correctly distinguish between three different

types, although as the overlap between communities grows, the separation between types becomes less clear. The variance of estimated $\hat{\lambda}_i$'s also grows for larger overlaps, since we increase overlap by increasing σ^2 in the Gaussian mixture model, resulting in more heterogeneous connection probabilities.

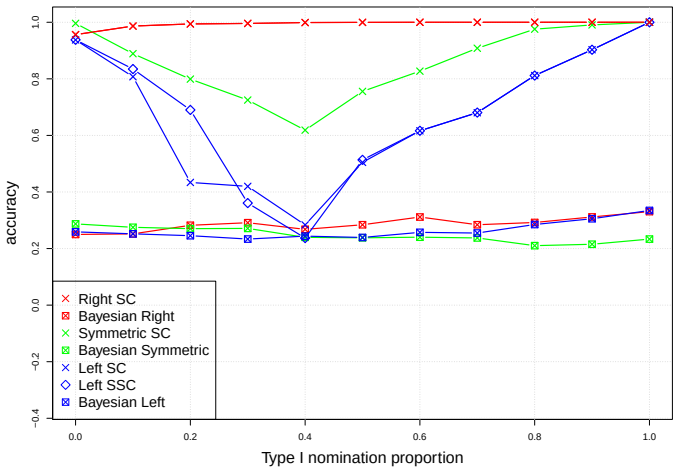


Figure 4: Community detection accuracy on networks generated from the mis-specified model with only Type 1 and Type 2 nominations, with varying proportion of Type 1 nodes.

When the overlap parameter $\sigma/m_\mu = 0$, the random dot product graph model reduces to the SBM; the larger σ/m_μ, the more blurred are the community boundaries. Further, the edge nomination process induces a complicated dependency structure between the edges in the observed network. We are not aware of any method that would fit such a model directly, but we expect that our model still provides a reasonable approximation, and the λ parameters NSBM fits may reflect the three types of nomination preferences. Community detection accuracy for σ/m_μ ranging from 0 to 0.9 under the alternative edge nomination procedure for all methods previously considered is shown in Figure 3a. The Right SC remains effective even if both the underlying network model and the nomination mechanism change, and community detection remains nearly perfect up to around $\sigma/m_\mu \approx 0.3$. More importantly, the relative ranking of different methods remains the same: while community detection accuracy drops as the overlap between mixture components increases, the Right SC remains the most accurate method available.

To investigate the impact of edge nomination types, we consider the special case of Type 1 and Type 2 nominations only, since the random nominations of Type 3 do not really offer much insight into the effect of the nomination mechanism. In this experiment, we vary the proportion of Type 1 nominations from 0 to 1, set $\sigma = 0$ to focus on the effect of type, and report community detection accuracy for all methods in Figure 4. In this setting, the Right SC delivers accurate community detection for any mixture of Types 1 and 2, whereas both Left SC and symmetric SC are effective if only one nomination type is present, but quickly lose power when there is more than one nomination type. Intuitively, this is expected

because if there is only one nomination type, the model is still approximately an SBM for which the left singular vectors are informative. With different nomination types, the left singular vectors no longer capture the community information correctly. The Bayesian clustering methods work poorly in this special case as well.

In summary, even when the nomination mechanism assumed by the NSBM is not correct, the estimated λ_i's still offer meaningful information. The results in Figure 3 and Figure 4 demonstrate the potential of our method as an approximation to more complex edge nomination models.

5. Business school faculty hiring network analysis

Here we apply the proposed NSBM to a faculty hiring network between US business schools. The data were collected by Clauset et al. (2015) via web crawling and contains data on 7856 faculty members from 112 business schools, recording the total number of PhDs from one institution hired by another institution. This gives us a weighted, directed hiring network. To reduce noise from very small schools, we removed institutions with either receiver or sender degree of 3 or less, resulting in 87 institutions remaining. The edge weights have a heavy-tailed distribution, as shown in Figure 6 of the supplementary material Section C: weight values of 0, 1, and 2 account for 90% of all edges, while the maximum weight is 60. As discussed in Section 2.5, our method can handle weighted edges but given we use the method of moments, but heavy tails present a problem. Given that most edge weights are 2 or less, we simply truncate all edge weights greater than 2 down to 2. The supplementary material in Section C shows the result is similar if we truncate at 3 instead of 2.

Table 1: Communities of business schools found by NSBM, their average and median rankings from US News 2012 and π-ranking of Clauset et al. (2015). Top 12 institutions according to π-ranking are listed for each community (sorted according to decreasing π-ranking).

	size	USNews (avg./med.)	π-ranking (avg./med.)	Top 12 Institutions
1	12	7.7/8	8.3/8	Stanford, MIT, Harvard, UC Berkeley, U Chicago, Cornell, U Michigan, Columbia, Yale, U Penn., NYU, Duke
2	12	29.8/32.5	17.7/17.5	U Rochester, Northwestern, Carnegie Mellon, U Wisconsin Madison, UCLA, U Minnesota-Twin Cities, UIUC, Purdue, U Florida, UT Austin, U Washington
3	19	53.1/54	45/45	Ohio State, UNC Chapel Hill, U Pittsburgh, Penn. State, Indiana U., Michigan State, Georgia Tech, U Arizona, SUNY Buffalo, Texas A&M, U Georgia, Arizona State
4	44	63.7/63	61.4/61.5	Washington U St. Louis, U Maryland College Park, U Colorado Boulder, UC Irvine, U Utah, U Oregon, U Southern California, UT Dallas, U Virginia, Boston U., UMass Amherst, Emory

Our goal is to investigate communities of institutions and patterns of hiring between these communities. We can think of the true unobserved edges as job offers extended, while the observed edges are offers accepted. In this context, it is safe to assume we do not observe any false edges. Additional missing edges are possible, however, even with accepted

offers – for example, PhD information not listed on the webpage of the faculty member, or faculty who have moved since first hired. Under the NSBM model, all these causes of missingness are reflected in the individual parameters λ_i, which is much more flexible than assuming edges are missing uniformly at random. The NSBM framework also assumes the true edge weight (number of offers extended) only depends on the communities of the institutions involved. This is clearly not true for individual offers, and a simplification for aggregate hires as well, but as the results below show, it provides a reasonable fit to this data and leads to a natural and meaningful interpretation. To determine the number of communities, we applied the edge cross-validation method with average stability selection proposed by Li et al. (2020), which selected $K = 4$. We then fit the NSBM to the network with $K = 4$ communities. Table 1 shows the four communities and their average rankings from two sources (which are also used to order communities). The rankings are the US NEWS graduate school rankings from 2012 (included in the data set) and the π-ranking proposed by Clauset et al. (2015), which is designed to measure hiring advantage, with a higher-ranked institution expected to be more successful in competing for top candidates. We list up to 12 institutions (the first two communities only have 12 each, whereas the other two have 19 and 44) with the highest π-ranking in each community in Table 1. Overall, the communities show a clear ordering which matches both rankings well.

The community-level parameters of NSBM can be directly interpreted in terms of a hiring "hierarchy", which was reported by Clauset et al. (2015). Based on the weighted NSBM in Section 2.5, we define connection strength from community k to community l as the expectation of average edge weights from nodes $i \in G_k$ to nodes $j \in G_l$, $M_{kl} = \frac{1}{n_k n_l} \sum_{i \in G_k, j \in G_l} \theta_i B_{ij}^{\lambda_i}$. Table 2 shows estimated connection strengths for the business schools hiring network. It shows that Group 1 institutions tend to hire the most from their own group, and about half as many from Group 2. They are not very likely to hire from Groups 3 and 4. Group 2 institutions hire roughly equally from Groups 1 and 2, and a fraction from Group 3, but very few from Group 4. Group 3 institutions hire the most from Group 2, not Group 1. Group 4 hire more from Group 1 and Group 2 than Group 3. The estimated model parameters thus indicate a strong hierarchy in hiring relationships between the groups, which aligns closely with the rankings in Table 1.

Table 2: Estimated strengths of connections between business school communities.

	Group 1	Group 2	Group 3	Group 4
Group 1	1.94	0.98	0.24	0.09
Group 2	1.64	1.39	0.47	0.15
Group 3	0.96	1.40	1.03	0.39
Group 4	1.01	0.94	0.63	0.24

The NSBM also allows us to estimate hiring preferences of individual institutions, represented by parameters λ_i. Taking Group 1 as an example, Yale and Columbia show the strongest preference ($\hat{\lambda}_i = 1.35$ and 1.22, respectively) for hiring within their own group, while the University of Michigan and the University of Pennsylvania are the least stringent ($\hat{\lambda}_i = 0.75$ and 0.71, respectively). More details on the fitted $\hat{\lambda}_i$'s are available in the supplementary materials Section C.

Overall, the NSBM reveals a clear hierarchical structure in the hiring relationships between US business schools in this network, which matches both our expectations and the observations of Clauset et al. (2015).

6. Discussion

We have proposed a general framework to model a directed network with communities, with network data collected by asking nodes to report or nominate their connections, a common scenario in practice. A particular parametric form of the general model, the NSBM, allows for meaningful interpretation of the parameters and computationally efficient fitting algorithms. Other parameterizations can be set up within the general framework, perhaps for specific data collection procedures and/or applications. We show that the right singular vectors of the adjacency matrix can be used for community detection even under this general nomination mechanism, whereas the parameter estimation algorithm would naturally need to be derived for every parametric model separately. In all cases, the critical point is that when we do not observe the whole network, pretending that we do tends to lead to a drop in accuracy and loss of efficiency. We saw this empirically in both simulated networks and the business school faculty hiring network. We also demonstrate that even with a more complicated nomination procedure that introduces dependence edges, our clustering method can still effectively find the communities and the NSBM can still deliver insights about the nomination mechanism. One potentially fruitful direction for future work is modeling network structures other than communities and investigating how incomplete and heterogeneous link nominations can affect our estimation of different types of structures and what models can be used to account for and correct the nomination process.

Acknowledgments

T. Li is currently affiliated with the University of Minnesota, but the work was completed when he was at the University of Virginia. T. Li was supported by a NSF grant (DMS-2015298) and the 3Caverlier award from the University of Virginia. E. Levina were supported in part by an ONR grant (N000141612910) and NSF grants (DMS-1521551 and DMS-1916222). J. Zhu were supported in part by NSF grants (DMS-1407698 and DMS-1821243). The authors acknowledge Research Computing at the University of Virginia for providing computational resources and technical support that have contributed to the results reported within this publication.

References

Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008.

Arash A Amini, Aiyou Chen, Peter J Bickel, and Elizaveta Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41

(4):2097–2122, 2013.

Carter T Butts. Network inference, error, and informant (in) accuracy: a bayesian approach. *social networks*, 25(2):103–140, 2003.

Carter T. Butts. *sna: Tools for Social Network Analysis*, 2020. URL `https://CRAN.R-project.org/package=sna`. R package version 2.6.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Ga Ming Angus Chan and Tianxi Li. Fitting low-rank models on egocentrically sampled partial networks. In *International Conference on Artificial Intelligence and Statistics*, pages 10635–10649. PMLR, 2023.

Kehui Chen and Jing Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113(521):241–251, 2018.

Fan Chung. Laplacians and the cheeger inequality for directed graphs. *Annals of Combinatorics*, 9:1–19, 2005.

Aaron Clauset, Samuel Arbesman, and Daniel B Larremore. Systematic inequality and hierarchy in faculty hiring networks. *Science advances*, 1(1):e1400005, 2015.

Richard C Connor, Rachel A Smolker, and Andrew F Richards. Dolphin alliances and coalitions. *Coalitions and alliances in humans and other animals*, 414:443, 1992.

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

Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block models. *The Journal of Machine Learning Research*, 18(1):1980–2024, 2017.

Pablo M Gleiser and Leon Danon. Community structure in jazz. *Advances in complex systems*, 6(04):565–573, 2003.

Roger Guimerà and Marta Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106 (52):22073–22078, 2009.

Mark S Handcock and Krista J Gile. Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1):5, 2010.

Kathleen Mullan Harris. *The National Longitudinal Study of Adolescent to Adult Health (Add Health), Waves I & II, 1994-1996; Wave III, 2001-2002; Wave IV, 2007–009 [machine-readable data file and documentation].* Carolina Population Center, University of North Carolina at Chapel Hill, 2009.

Christine C Hass. Social status in female bighorn sheep (ovis canadensis): expression, development and reproductive correlates. *Journal of Zoology*, 225(3):509–523, 1991.

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

Pengsheng Ji and Jiashun Jin. Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10(4):1779–1812, 2016.

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

Antony Joseph and Bin Yu. Impact of regularization on spectral clustering. *The Annals of Statistics*, 44(4):1765–1791, 2016.

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

Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1+\varepsilon)$-approximation algorithm for geometric k-means clustering in any dimensions. In *Proceedings-Annual Symposium on Foundations of Computer Science*, pages 454–462. IEEE, 2004.

Can M Le and Elizaveta Levina. Estimating a network from multiple noisy realizations. *arXiv preprint arXiv:1710.04765*, 2017.

Can M Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 2017.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2014.

Jing Lei and Lingxue Zhu. Generic sample splitting for refined community recovery in degree corrected stochastic block models. *Statistica Sinica*, pages 1639–1659, 2017.

Lihua Lei. Unified $\ell_{2\to\infty}$ eigenspace perturbation theory for symmetric random matrices. *arXiv preprint arXiv:1909.04798*, 2019.

Tianxi Li, Elizaveta Levina, and Ji Zhu. Network cross-validation by edge sampling. *Biometrika*, 107(2):257–276, 2020.

Tianxi Li, Yun-Jhong Wu, Elizaveta Levina, and Ji Zhu. Link prediction for egocentrically sampled networks. *Journal of Computational and Graphical Statistics*, pages 1–24, 2023.

Yanhua Li and Zhi-Li Zhang. Random walks on digraphs, the generalized digraph laplacian and the degree of asymmetry. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 74–85. Springer, 2010.

Fragkiskos D Malliaros and Michalis Vazirgiannis. Clustering and community detection in directed networks: A survey. *Physics reports*, 533(4):95–142, 2013.

Travis Martin, Brian Ball, and Mark EJ Newman. Structural inference for uncertain networks. *Physical Review E*, 93(1):012306, 2016.

Catherine Matias and Vincent Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141, 2017.

Mark Newman. *Networks: an introduction*. Oxford university press, 2010.

MEJ Newman. Network structure from rich but noisy data. *Nature Physics*, page 1, 2018a.

MEJ Newman. Network reconstruction and error estimation with noisy network data. *arXiv preprint arXiv:1803.02427*, 2018b.

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

Karl Rohe, Tai Qin, and Bin Yu. Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences*, 113(45): 12679–12684, 2016.

Timothée Tabouy, Pierre Barbillon, and Julien Chiquet. Variational inference for stochastic block models from sampled data. *Journal of the American Statistical Association*, 115 (529):455–466, 2020.

Van Vu. A simple svd algorithm for finding hidden partitions. *Combinatorics, Probability and Computing*, 27(1):124–140, 2018.

Kevin S Xu and Alfred O Hero. Dynamic stochastic blockmodels: Statistical models for time-evolving networks. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 201–210. Springer, 2013.

Stephen J Young and Edward R Scheinerman. Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 138–149. Springer, 2007.

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

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

Yunpeng Zhao, Yun-Jhong Wu, Elizaveta Levina, and Ji Zhu. Link prediction for partially observed networks. *Journal of Computational and Graphical Statistics*, 26(3):725–733, 2017.

Dengyong Zhou and Christopher JC Burges. Spectral clustering and transductive learning with multiple views. In *Proceedings of the 24th international conference on Machine learning*, pages 1159–1166, 2007.

Appendix A. Spectral minimum spanning tree clustering for the nomination model

In this section, we introduce another spectral method, Spectral Minimum Spanning Tree Clustering (Right SMST), which also uses the right singular vectors. The clusters are obtained by cutting the minimum spanning tree between embedded nodes; an algorithm studied in Vu (2018) and Lei and Zhu (2017). This algorithm is much easier to analyze than K-means, and we show in Section 3 that it can achieve the exact recovery of community labels for all nodes. However, in practice, the Right SC is always faster, and more importantly, much better on sparse networks. Therefore, the SMST algorithm is primarily of theoretical interest.

Algorithm 3 (Right SMST) *Given an adjacency matrix $\tilde{A}$ and the number of communities K:*

1. *Compute the rank-K truncated singular value decomposition $\tilde{A} = \widehat{U}\widehat{D}\widehat{V}^T$.*
2. *Run minimum spanning tree algorithm of Vu (2018) on $\widehat{V}$:*
   (a) *Construct the undirected distance graph between n nodes based on the distance matrix, where the edge weight between i and j is the distance between $\widehat{V}_{i\cdot}$ and $\widehat{V}_{j\cdot}$, the i-th and j-th rows of the matrix $\widehat{V}$.*
   (b) *Find the minimum spanning tree of the distance graph.*
   (c) *Remove the $K-1$ edges with the highest weights from the minimum spanning tree .*
   (d) *Return the resulting connected components as clusters.*

Intuitively, it is not hard to see why the Right SMST may be inferior to Algorithm 1 in practice. Algorithm 3 is designed with the expectation that the between-cluster distances are always larger than within-cluster distances. This works when the signal is strong enough, but with weaker signal the minimum between-cluster distance and the maximum within-cluster distance can be unstable. In contrast, K-means looks at the average behavior of observations within the same cluster and thus can be a lot more stable.

Next, we introduce the consistency of the Right SMST (Algorithm 3). The strong consistency can be obtained by using the recently ℓ_∞ perturbation theory from Lei (2019).

Theorem 5 (Consistency of the Right SMST algorithm) *Assume the network $\tilde{A}$ is generated from the general model* (2)*. Let $\hat{\mathbf{c}}$ be the output of Algorithm 3, and $\|\tilde{P}\|_\infty = \max_{ij}\tilde{P}_{ij}$. Assume A1 holds, $n\|\tilde{P}\|_\infty \geq C_0 \log n$, and*

$$\sigma_K(\tilde{P}) \geq C_1 n\|\tilde{P}\|_\infty \tag{13}$$

for some constants $C_0, C_1 > 0$. If the following condition (14) *is true*

$$\sqrt{\frac{\log n}{n\|\tilde{P}\|_\infty}}\max\left(\sqrt{n}\|U\|_{2,\infty}, \sqrt{n}\|V\|_{2,\infty}\right) = o(1), \tag{14}$$

for sufficiently large n, then there exists a permutation $\Psi : [K] \to [K]$ of community labels such that

$$\Psi(\hat{\boldsymbol{c}}) = \boldsymbol{c}$$

with probability at least $1 - n^{-1}$.

Compared to Theorem 2 for the Right SC, Theorem 5 requires one additional condition (14) to achieve strong consistency. Condition (14) reduces to (13) only when $\max(\|U\|_{2,\infty}, \|V\|_{2,\infty}) = O(\frac{1}{\sqrt{n}})$, which also means that $\tilde{P}$ has perfect incoherence (Candès and Recht, 2009; Candès and Tao, 2010). Again, we give a simplified form of Theorem 5 in the special case of the NSBM with the parameterization assumed in A2.

Corollary 6 (Consistency of the Right SMST algorithm under NSBM) *Assume the network $\tilde{A}$ is generated from the NSBM* (3)*. Let $\hat{\boldsymbol{c}}$ be the output of Algorithm 3. If assumptions of Proposition 1, A1, and A2 hold, and*

$$n\rho_n / \log n \to \infty,$$

then for sufficiently large n, with probability at least $1 - n^{-1}$, there exists a permutation $\pi : [K] \to [K]$ of labels $\hat{\boldsymbol{c}}$ such that

$$\pi(\hat{\boldsymbol{c}}) = \boldsymbol{c}.$$

Appendix B. Proofs

B.1 Model identifiability

Proof [Proof of Proposition 1] We need to show that given the probability matrix $\tilde{P} = (\tilde{P}_{ij}) = (\theta_i B^{\lambda_i}_{c_i c_j})$ and the community labels $\boldsymbol{c}$, all parameters are uniquely determined under the current constraints. Without loss of generality, we focus on identifying the parameter for one arbitrary community k. For any $i \in G_k$, we have

$$\mu_{il} = \log(\theta_i B^{\lambda_i}_{kl}) = \log(\theta_i) + \lambda_i \log(B_{kl}) \tag{15}$$

where we treat $\log(0)$ as $-\infty$. It can be seen that $\log(\theta_i) = \mu_{ik}$ by setting $l = k$ under the constraint $B_{kk} = 1$.

Write $\boldsymbol{b} = (\log(B_{k1}), \cdots, \log(B_{k,K}))$. Notice that for any $1 \le l \le K$ such that $B_{kl} \neq 0$, we have

$$\mu_{ik} - \mu_{il} = \lambda_i (b_k - b_l). \tag{16}$$

The constraint on λ_i indicates that there exists at least one node i with non-zero λ_i. Since there exists at least one l such that $0 < B_{kl} \neq B_{kk}$, $b_k - b_l \neq 0$, we can locate one such node (denoted by i_0) and community (denoted by l_0) by identifying i and l corresponding to a non-zero $\mu_{ik} - \mu_{il}$. Given this l_0, we can uniquely determine the ratio between all non-zero λ_i's. The nodes with $\lambda_i = 0$ can be directly identified from $\mu_{ik} - \mu_{il_0} = 0$. Therefore, with the constraint $\sum_{i \in G_k} \lambda_k = n_k$, the identification of λ_i's is guaranteed.

Fixing the node i_0, b can be determined by (16) up to a shift. Since we constrain $b_k = B_{kk} = 1$, all the other entries of $B_{k\cdot}$ are also identifiable.

■

B.2 Community detection

Proposition 7 *Let $\tilde{P} = \tilde{U}\tilde{D}\tilde{V}^T$ be the SVD of $\tilde{P}$. There exists a matrix $X \in \mathbb{R}^{K \times K}$ such that*

$$\tilde{V} = ZX \tag{17}$$

where Z is the $n \times K$ community membership matrix, defined by $Z_{ik} = 1(c_i = k)$. In addition, $\|X_{k\cdot} - X_{l\cdot}\|_2 = \sqrt{n_k^{-1} + n_l^{-1}}$ for any $1 \le k \neq l \le K$.

Proof [Proof of Proposition 7] It is easy to check that $\tilde{P} = FZ^T$ where F is the matrix obtained by applying function F_i to each element of the ith row of the matrix ZB. Write $\Delta = diag(\sqrt{n_1}, \cdots, \sqrt{n_K})$. Assume that the SVD of $F\Delta$ is given by

$$F\Delta = UDV^T.$$

We have

$$\tilde{P} = UDV^T(Z\Delta^{-1})^T = UD(Z\Delta^{-1}V)^T.$$

Notice that $Z\Delta^{-1}$ is an orthonormal matrix and so is $Z\Delta^{-1}V$. Taking $X = \Delta^{-1}V$ gives the SVD of $\tilde{P}$, up to the standard invariances (sign-flipping and rotation within subspaces

corresponding to equal singular values).Note that for a full rank B, V is a $K \times K$ orthonormal matrix. The distance claim follows directly from the orthogonality of rows of V. ■

We will use the following three known results on spectral clustering.

Lemma 8 (Lemma 7 of Chen and Lei (2018)) *Let $M, \widehat{M}$ be two matrices of size $n \times n$ and $V, \widehat{V}$ be the $n \times K$ orthogonal matrices of top K right singular vectors of M and $\widehat{M}$. Then there exists a $K \times K$ orthogonal matrix Q such that*

$$\|\widehat{V}Q - V\|_F \leq \frac{2\sqrt{2K}\|\widehat{M} - M\|}{\sigma_K(M)}.$$

The orthogonal matrix Q makes no difference for subsequent developments and will be omitted.

Lemma 9 (Lemma 5.3 of Lei and Rinaldo (2014)) *Let $V, \widehat{V}$ be two $n \times K$ matrices with V having only K distinct rows, corresponding to K communities denoted by $\mathbf{c}$. Let $\hat{\mathbf{c}}$ be the output of a K-means clustering algorithm on $\widehat{V}$, with objective value no larger than $1+\epsilon$ of the global optimum (Kumar et al., 2004). Denote the community indices corresponding to $\mathbf{c}$ and $\hat{\mathbf{c}}$ by $\{G_k\}$ and $\{\hat{G}_k\}$. Define $S_k = \{i : i \in G_k, \hat{c}_i \neq k\}$. For any δ smaller than the minimum distance between any two distinct rows of V, if*

$$8(2+\epsilon)\|\widehat{V} - V\|_F^2 \leq n_{\min}\delta^2$$

where $n_{\min} = \min_k |G_k|$, then there exists a permutation of the K community labels in $\hat{\mathbf{c}}$, such that

$$\sum_{k=1}^{K} |S_k| \leq 8(2+\epsilon)\frac{\|\widehat{V} - V\|_F^2}{\delta^2}.$$

Another result we need is the concentration of a random (directed) graph adjacency matrix from Le et al. (2017). A similar result was also obtained by Lei and Rinaldo (2014).

Lemma 10 *Let A be the adjacency matrix of a random graph on n nodes with independent edges. Set $\mathbb{E}(A) = P = [p_{ij}]_{n\times n}$ and assume that $n \max_{ij} p_{ij} \leq d$ for $d \geq C_0 \log n$ and $C_0 > 0$. Then there exists a constant C depending on C_0 such that*

$$\|A - P\| \leq C\sqrt{d}$$

with probability at least $1 - n^{-1}$.

With these three lemmas, we are ready to prove Theorem 2.

Proof [Proof of Theorem 2]

Let $\tilde{V}^*$ be the matrix of right singular vectors for $\tilde{P}$ and let $\tilde{V}$ be the right singular vectors of $\tilde{A}$. The assumption $n\|\tilde{P}\|_\infty \geq C_0 \log n$ implies the condition for concentration of Lemma 10. From Lemma 10, we have

$$\|\tilde{V} - \tilde{V}^*\|_F \leq \frac{2\sqrt{2K}}{\sigma_K(\tilde{P})}\|\tilde{A} - \tilde{P}\| \leq \frac{2C\sqrt{2K}}{\sigma_K(\tilde{P})}\sqrt{n\|\tilde{P}\|_\infty}$$

with probability at least $1 - n^{-1}$.

To apply Lemma 9, note that from Proposition 7, the minimum distance between distinct rows in $\tilde{V}^*$ is at least $\sqrt{\frac{2}{n_{\max}}}$. Therefore, according to Lemma 9,

$$\begin{aligned}\sum_k \frac{|S_k|}{n_k} &\leq \frac{1}{n_{\min}} \sum_{k=1}^{K} |S_k| \leq \frac{1}{n_{\min}} 8(2+\epsilon) \frac{\|\tilde{V} - \tilde{V}^*\|_F^2}{\frac{2}{n_{\max}}} \\ &\leq 32C^2(2+\epsilon) \frac{n_{\max} K n \|\tilde{P}\|_\infty}{n_{\min} \sigma_K(\tilde{P})^2} \leq \frac{32C^2(2+\epsilon)}{\kappa'} \frac{Kn\|\tilde{P}\|_\infty}{\sigma_K(\tilde{P})^2}\end{aligned}$$

as long as the condition of Lemma 9 holds,

$$\frac{32C^2(2+\epsilon)}{\kappa'} \frac{Kn\|\tilde{P}\|_\infty}{\sigma_K(\tilde{P})^2} \leq 1\,,$$

which can be guaranteed by the assumptions of Theorem 2 when setting $C_1 = \frac{32C^2(2+\epsilon)}{\kappa'}$. This completes the proof.

■

Proof [Proof of Corollary 3] Let g_θ and g_λ be the distributions of $\bar{\theta}_i$ and λ_i, respectively. We have $\|\tilde{P}\|_\infty \leq \rho_n \gamma_1 \|B\|_\infty$ from A2, where both γ_1 and $\|B\|_\infty$ are constants. We now need a bound on $\sigma_K(\tilde{P})$.

From Proposition 7, it follows that $\sigma_K(\tilde{P})$ is the K-th singular value of $F = \rho_n M$ where

$$M = \begin{pmatrix} \bar{\theta}_1 B_{c_1,1}^{\lambda_1} & \bar{\theta}_1 B_{c_1,2}^{\lambda_1} & \cdots & \bar{\theta}_1 B_{c_1,K}^{\lambda_1} \\ \bar{\theta}_2 B_{c_2,1}^{\lambda_2} & \bar{\theta}_2 B_{c_2,2}^{\lambda_2} & \cdots & \bar{\theta}_2 B_{c_2,K}^{\lambda_2} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{\theta}_n B_{c_n,1}^{\lambda_n} & \bar{\theta}_n B_{c_n,2}^{\lambda_n} & \cdots & \bar{\theta}_n B_{c_n,K}^{\lambda_n} \end{pmatrix}$$

and $\Delta = diag(\sqrt{n_1}, \cdots, \sqrt{n_K})$. Under A2, there are at most $m_1 m_2 K$ distinct rows of M. Denote the matrix with these $m_1 m_2 K$ rows by $\tilde{M} \in \mathbb{R}^{(m_1 m_2 K) \times K}$, and write

$$F = \rho_n \tilde{Z} \tilde{M}\,, \tag{18}$$

where F is the same quantity in the proof of Proposition 7, $\tilde{Z} \in \mathbb{R}^{n \times (m_1 m_2 K)}$ with exactly one 1 in each row and zeros in the other positions. $\tilde{Z}$ gives the correspondence from each row of M to the rows of $\tilde{M}$. Let $\tilde{n}_k$ be the number of times that the kth row of $\tilde{M}$ appears in rows in M, and define $\tilde{\Delta} = diag(\sqrt{\tilde{n}_1}, \cdots, \sqrt{\tilde{n}_{m_1 m_2 K}})$. It is easy to check $\tilde{Z}\tilde{\Delta}^{-1}$ is an orthogonal matrix. Therefore,

$$\sigma_K(\tilde{P}) = \sigma_K(\rho_n \tilde{\Delta} \tilde{M} \Delta) \geq \lambda \rho_n \min_{i,j,k} \sqrt{\tilde{n}_{ijk}} \min_k \sqrt{n_k}\,, \tag{19}$$

where $\lambda = \sigma_K(\tilde{M})$.

By A1, A2, and Hoeffding's inequality, we have

$$\min_{i,j,k} \tilde{n}_{ijk} \geq C_2 n$$

with probability at least $1-\exp(-\gamma_2 n)$ for some constants γ_2, $C_2 > 0$ depending on κ', K and g_θ, g_λ. Under this event, we have

$$\sigma_K(\tilde{P}) \geq \sqrt{C_2 \kappa'} n \rho_n.$$

Finally, applying Theorem 2 directly gives

$$\sum_k \frac{|S_k|}{n_k} \leq C_1 \frac{Kn\|\tilde{P}\|_\infty}{\sigma_K(\tilde{P})^2} \leq \frac{C_1}{C_2 \kappa'} \frac{K}{n\rho_n}$$

with probability at least $1 - n^{-1} - e^{-\gamma_1 n} - e^{-\gamma_2 n} \geq 1 - 2n^{-1}$ for sufficiently large n. Setting $C' = \frac{C_1 K}{C_2 \kappa'}$ completes the proof. ■

Lemma 11 (Directed version of Corollary 3.6 in Lei (2019)) *Let $\tilde{A} \in \{0,1\}^{n\times n}$ be an adjacency matrix of a directed network with independent Bernoulli entries and the expectation $\tilde{P} \in [0,1]^{n\times n}$. Assume the rank of $\tilde{P}$ is K and K is fixed. Let $\tilde{A} = \hat{U}\hat{\Sigma}\hat{V}^T$ and $U\Sigma V^T$ be the rank K SVD of $\tilde{A}$ and $\tilde{P}$, respectively. If*

$$\Sigma_{KK} \geq C_0 n \|\tilde{P}\|_\infty,$$

and $n\|P\|_\infty \geq C_0 \log n$ for some constant $C_0 > 0$, then with probability at least $1 - n^{-1}$, we have

$$\max(\|\hat{U} - U\|_{2,\infty}, \|\hat{V} - V\|_{2,\infty}) \leq C\sqrt{\frac{\log n}{n\|\tilde{P}\|_\infty}} \max\left(\|U\|_{2,\infty}, \|V\|_{2,\infty}\right).$$

Proof [of Lemma 11] Let

$$\tilde{P}^s = \begin{bmatrix} 0 & \tilde{P} \\ \tilde{P}^T & 0 \end{bmatrix} \text{ and } \tilde{A}^s = \begin{bmatrix} 0 & \tilde{A} \\ \tilde{A}^T & 0 \end{bmatrix}.$$

Then $\tilde{A}^s$ is a symmetric matrix with independent Bernoulli entries in the upper triangular positions, drawn with probabilities $\tilde{P}^s$. The eigenvectors of $\tilde{P}^s$ are $[U^T, V^T]^T$. The result follows directly by applying Corollary 3.6 of Lei (2019) with the additional constraint on Σ_{KK}, which corresponds to formula (200) of Lei (2019). ■

Proof [of Theorem 5] By Proposition 7 and Assumption A1, the minimum spanning tree algorithm will perfectly recover communities if

$$\max_i \|\hat{V}_{i\cdot} - V_{i\cdot}\| < \frac{\sqrt{2}}{4\sqrt{(1-(K-1)\kappa')n}} \leq \frac{\sqrt{2}}{4\sqrt{n_{\max}}} \leq \frac{1}{4} \min_{c_i \neq c_j} \|V_{i\cdot} - V_{j\cdot}\|. \tag{20}$$

This is because (20) ensures that any between-community edge would have a higher weight than any within-community edge. In the minimum spanning tree, between any two communities, there is at most one edge connecting them, and in total there would be exactly $K-1$ between-community edges. Therefore, removing the $K-1$ edges with largest weights results in the correct community partition.

It remains to show (20). Lemma 11 gives

$$\max_i \|\hat{V}_{i\cdot} - V_{i\cdot}\| = \|\hat{V} - V\|_{2,\infty} \leq C\sqrt{\frac{\log n}{n\|\tilde{P}\|_\infty}} \max\left(\|U\|_{2,\infty}, \|V\|_{2,\infty}\right)$$

with probability at least $1 - n^{-1}$, which implies (20).

■

Proof [of Corollary 6] To apply Theorem 5, we just need to show the two conditions (13) and (14) hold. Equation (19) in the proof of Corollary 3 implies (13). As discussed after Corollary 6, to show (14) when $n\rho_n / \log n \to \infty$, it is sufficient to show that $\tilde{P}$ is perfectly incoherent, that is

$$\|U\|_{2,\infty} = O(1/\sqrt{n}) \quad \text{and} \quad \|V\|_{2,\infty} = O(1/\sqrt{n}).$$

From Proposition 7, we know that V has only K distinct rows and each unique row appears at least $n_{\min}$ times, from A1. Therefore, $\|V\|2, \infty = O(1/\sqrt{n})$.

The proof of Proposition 7 indicates that U consists of the left singular vectors of F, which is given by (18). Using the same notation, we have

$$F = \tilde{Z}\tilde{\Delta}^{-1}\Delta\tilde{M}\,,$$

where $\Delta\tilde{M}$ is an $m_1 m_2 K \times K$ matrix. Again, $\tilde{Z}\tilde{\Delta}^{-1}$ is an orthonormal matrix. Therefore, $U = \tilde{Z}\tilde{\Delta}^{-1}\tilde{U}$ where $\tilde{U}$ is the left singular vector of $\Delta\tilde{M}$. Hence U only has $m_1 m_2 K$ distinct rows. Since we assume m_1, m_2, and K to be fixed, we also know $\|U\|_{2,\infty} = O(1/\sqrt{n})$.

■

B.3 Proofs for parameter estimation under the NSBM

Proof [of Theorem 4] Without loss of generality, let us assume the first n_1 nodes are from community 1 and focus on estimating parameters in community 1. The same argument can be repeated for the other $K - 1$ communities. Note that consistency trivially holds for $B_{1l} = 0$, so for this proof we focus on the case $B_{1l} > 0$. For each $l \in [K]$ such that $B_{1l} > 0$, define

$$\tilde{P}_{il} = \theta_i B_{1l}^{\lambda_i}.$$

By Bernstein inequality, we have

$$\mathbb{P}\left(\left|\frac{\sum_{j\in G_l}\tilde{A}_{ij}}{n_l} - \tilde{P}_{il}\right| > t\right) \leq 2\exp\left(-\frac{n_l t^2/2}{\tilde{P}_{il} + t/3}\right). \tag{21}$$

To make the concentration nontrivial, we need to require at least $t \leq \tilde{P}_{il}$. Hence we have $\tilde{P}_{il} \geq t/3$, leading to

$$\mathbb{P}\left(\left|\frac{\sum_{j\in G_l}\tilde{A}_{ij}}{n_l} - \tilde{P}_{il}\right| > t\right) \leq 2\exp\left(-\frac{n_l t^2}{4\tilde{P}_{il}}\right) \leq 2\exp\left(-\frac{\kappa' n t^2}{4\tilde{P}_{il}}\right). \tag{22}$$

When $l = 1$, we have $\tilde{P}_{i1} = \theta_i$ and letting $t = \log n/\sqrt{n}$ in (22) gives

$$\mathbb{P}\left(\left|\hat{\theta}_i - \theta_i\right| > \frac{\log n}{\sqrt{n}}\right) \leq 2\exp\left(-\frac{\kappa' \log^2 n}{4}\right).$$

The first claimed result is obtained by taking the union of all i.

A useful special case is $t = \delta_n \tilde{P}_{il}$, for which (22) gives

$$\mathbb{P}(|T_{il} - \tilde{P}_{il}| > \delta_n \tilde{P}_{il}) \leq 2\exp\left(-\frac{\kappa'}{4} n \delta_n^2 \tilde{P}_{il}\right). \tag{23}$$

When $l = 1$, we have $\tilde{P}_{i1} = \theta_i$. If $\theta_i \geq \frac{8}{\kappa'}\frac{\log^4 n}{n}$, setting $t = \theta_i/\log n$ in (23) gives

$$\mathbb{P}\left(|\hat{\theta}_i - \theta_i|/\theta_i > \frac{1}{\log n}\right) \leq 2\exp\left(-\frac{\kappa' n \theta_i}{4\log^2 n}\right) \leq 2\exp(-2\log^2 n). \tag{24}$$

Therefore, when $\min_i \theta_i \geq \frac{8}{\kappa'}\frac{\log^4 n}{n}$, taking the union over all $i \in [n]$ gives

$$\mathbb{P}(\max_i |\hat{\theta}_i - \theta_i|/\theta_i > 1/\log n) \leq 2n\exp(-2\log^2 n) \leq \exp(-\log^2 n) \leq n^{-1}.$$

for sufficiently large n.

Recall that we only need to consider l such that $B_{1l>0}$. To control the estimation error of $\hat{B}_{1l}$, define

$$Z_{il} := \left(\log(T_{i1} + \frac{1}{n_1}) - \log(T_{il} + \frac{1}{n_l})\right)$$

and $\bar{Z}_{il} = \frac{1}{n_1}\sum_{i\in G_1} Z_{il}$. We have

$$\begin{aligned}
|\bar{Z}_{1l} - (-\log B_{1l})| &= |\bar{Z}_{1l} - \frac{1}{n_1}\sum_{i\in G_1}(\mu_{i1} - \mu_{il})| \\
&\leq |\bar{Z}_{1l} - \mathbb{E}\bar{Z}_{1l}| \\
&\quad + \frac{1}{n_1}\sum_{i\in G_1} |\mathbb{E}\log(T_{i1} + \frac{1}{n_1}) - \log(\mathbb{E}T_{i1} + \frac{1}{n_1})| \\
&\quad + \frac{1}{n_1}\sum_{i\in G_1} |\mathbb{E}\log(T_{il} + \frac{1}{n_l}) - \log(\mathbb{E}T_{il} + \frac{1}{n_l})| \\
&\quad + \frac{1}{n_1}\sum_{i\in G_1} |\log(\tilde{P}_{i1} + \frac{1}{n_1}) - \log(\tilde{P}_{i1})| + \frac{1}{n_1}\sum_{i\in G_1} |\log(\tilde{P}_{il} + \frac{1}{n_l}) - \log(\tilde{P}_{il})| \\
&\leq |\bar{Z}_{1l} - \mathbb{E}\bar{Z}_{1l}| \\
&\quad + \frac{1}{n_1}\sum_{i\in G_1} |\mathbb{E}\log(\sum_{j\in G_1} A_{ij} + 1) - \log(\mathbb{E}\sum_{j\in G_1} A_{ij} + 1)| \\
&\quad + \frac{1}{n_1}\sum_{i\in G_1} |\mathbb{E}\log(\sum_{j\in G_l} A_{ij} + 1) - \log(\mathbb{E}\sum_{j\in G_l} A_{ij} + 1)| \\
&\quad + \frac{1}{n_1}\sum_{i\in G_1} |\log(n_1\tilde{P}_{i1} + 1) - \log(n_1\tilde{P}_{i1})| + \frac{1}{n_1}\sum_{i\in G_1} |\log(n_l\tilde{P}_{il} + 1) - \log(n_l\tilde{P}_{il})|.
\end{aligned} \tag{25}$$

To control the first term in (25), note that Z_{il}'s are independent across different i's, and

$$-2\log n \leq -2\log n_1 \leq Z_{il} \leq 2\log 2.$$

By Hoeffding's inequality and A2, we have

$$\mathbb{P}\left(|\bar{Z}_{1l} - \mathbb{E}\bar{Z}_{1l}| > \frac{\log^2 n}{\sqrt{n}}\right) \leq 2\exp(-\frac{\kappa \log^2 n}{16}) \tag{26}$$

For the 2nd term in (25), because of the fact that $\sum_{j\in G_1} A_{ij}$ follows binomial distribution and Taylor expansion, we have

$$\frac{1}{n_1}\sum_{i\in G_1} |\mathbb{E}\log(\sum_{j\in G_1} A_{ij} + 1) - \log(\mathbb{E}\sum_{j\in G_1} A_{ij} + 1)| \leq c'\frac{1}{n\rho_n}.$$

The similar bound holds for the 3rd term. Finally, by Taylor expansion again, we have

$$\frac{1}{n_1}\sum_{i\in G_1} |\log(n_1\tilde{P}_{i1} + 1) - \log(n_1\tilde{P}_{i1})| + \frac{1}{n_1}\sum_{i\in G_1} |\log(n_l\tilde{P}_{il} + 1) - \log(n_l\tilde{P}_{il})| \leq \frac{c''}{n\rho_n}.$$

Combining the above results, we have

$$\begin{aligned}
&\mathbb{P}\left(|\log\hat{B}_{1l} - \log B_{1l}| > \tilde{c}\max(\frac{\log^2 n}{\sqrt{n}}, \frac{1}{n\rho_n})\right) \\
&= \mathbb{P}\left(|\frac{1}{n_1}\sum_{i\in G_1}(Y_{i1} - Y_{il}) - \frac{1}{n_1}\sum_{i\in G_1}(\mu_{i1} - \mu_{il})| > \tilde{c}\max(\frac{\log^2 n}{\sqrt{n}}, \frac{1}{n\rho_n})\right) \leq 2\exp(-\frac{\kappa\log^2 n}{16}).
\end{aligned} \tag{27}$$

Because the function $\exp(x)$ is convex, for any $x, y > 0$ we have

$$|\exp(x) - \exp(y)| \leq |x - y|\exp(\max(x, y)).$$

For sufficiently large n, under the event of (27),

$$|\hat{B}_{kl} - B_{kl}| \leq \tilde{c}'\max(\frac{\log^2 n}{\sqrt{n}}, \frac{1}{n\rho_n}).$$

Part 2 of the theorem comes directly from (27) after taking the union of at most K^2 events for community pairs with nonzero B_{kl}.

For Part 3, first note that because of the previous discussion, we only consider the settings when $T_{il} > 0$ for $B_{1l} > 0$. Therefore, we treat Ψ_1 as known. Let $\mu_{il} = \log(\tilde{P}_{il})$. For any $l \in \Psi_1$, define $b_l = \log(B_{1l})$ for $B_{1l} > 0$. We have

$$\mu_{il} - \mu_{i1} = \log\tilde{P}_{il} - \log\tilde{P}_{i1} = \lambda_i(b_l - b_1), i \in G_1.$$

Summing up across ψ_1,

$$\sum_{l\in\Psi_1}(\mu_{il} - \mu_{i1}) = \lambda_i\sum_{l\in\Psi_1}(b_l - b_1).$$

Under the identifiability constraint, we also have

$$\frac{1}{n_1}\sum_{i\in G_1}\sum_{l\in\Psi_1}(\mu_{il}-\mu_{i1})=\sum_{l\in\Psi_1}(b_l-b_1).$$

The two identities give

$$\frac{\sum_{l\in\Psi_1}(\mu_{il}-\mu_{i1})}{\frac{1}{n_1}\sum_{i\in G_1}\sum_{l\in\Psi_1}(\mu_{il}-\mu_{i1})}=\lambda_i.$$

To obtain an error bound for estimated parameters, we will separately bound the numerator and the denominator above. We now proceed to bound $Y_{il}-\mu_{il}$. A useful inequality is, for x, $y\geq 0$,

$$|\log(1+x)-\log(1+y)|\leq\frac{|x-y|}{1+\min(x,y)}.$$

Note that

$$|\log(\frac{\sum_{j\in G_l}\tilde{A}_{ij}}{n_l}+\frac{1}{n_l})-\log(\tilde{P}_{il}+\frac{1}{n_l})|=|\log(\sum_{j\in G_l}\tilde{A}_{ij}+1)-\log(n_l\tilde{P}_{il}+1)|\leq\frac{|\sum_{j\in G_l}\tilde{A}_{ij}-n_l\tilde{P}_{il}|}{1+\min(\sum_{j\in G_l}\tilde{A}_{ij},n_l\tilde{P}_{il})}.$$

In particular, under A2, there is a constant $\phi\leq\min_{ij}\bar{\theta}_i B^{\lambda_i}_{c_ic_j}$. Now let $\eta=2\max_{ij}B_{ij}$, $t=\frac{1}{\eta\log n}$. Using (21), we have

$$\mathbb{P}\left(|\sum_{j\in G_l}\tilde{A}_{ij}-n_l\tilde{P}_{il}|\geq\sqrt{n_l\tilde{P}_{il}}\log n\right)\leq 2\exp\left(-\frac{\log^2 n}{4}\right).$$

Suppose $\sqrt{n_l\tilde{P}_{il}\log n}<n_l\tilde{P}_{il}/2$, which holds for sufficiently large n under A2, as long as $n\rho_n\gg\log n$. So when n is sufficiently large, we have

$$\mathbb{P}\left(|Y_{il}-\log(\tilde{P}_{il}+\frac{1}{n_l})|\leq 2\frac{\log n}{\sqrt{\kappa n\rho_n\phi}}\right)\leq 2\exp\left(-\frac{\log^2 n}{8}\right). \tag{28}$$

From (25), we also know that $|\log(\tilde{P}_{il}+\frac{1}{n_l})-\mu_{il}|\leq c\frac{1}{n\rho_n}$. Therefore, we get

$$\mathbb{P}(|(Y_{i1}-Y_{il})-(\mu_{i1}-\mu_{il})|\leq 5\frac{\log n}{\sqrt{\kappa n\rho_n\phi}})\geq 1-4\exp\left(-\frac{\log^2 n}{8}\right) \tag{29}$$

for sufficiently large n. Applying (29) for all $l\in\Psi_1$ leads to

$$\mathbb{P}(|\sum_{l\in\Psi_1}(Y_{i1}-Y_{il})-\sum_{l\in\Psi_1}(\mu_{i1}-\mu_{il})|\leq 5K\frac{\log n}{\sqrt{\kappa n\rho_n\phi}})\leq 1-20\exp\left(-\frac{\log^2 n}{8}\right) \tag{30}$$

for sufficiently large n.

For the denominator, apply (27) across $l\in\Psi_1$, we have

$$\mathbb{P}\left(\left|\frac{\sum_{i\in G_1}\sum_{l\in\Psi_1}(Y_{i1}-Y_{il})}{n_1}-\frac{\sum_{i\in G_1}\sum_{l\in\Psi_1}(\mu_{i1}-\mu_{il})}{n_1})\right|>\tilde{c}''\max(\frac{\log^2 n}{\sqrt{n}},\frac{1}{n\rho_n})\right)\leq 2K\exp(-\frac{\kappa\log^2 n}{16}). \tag{31}$$

Another useful inequality we need is, for any x, y, x_0, y_0 such that $x \cdot x_0 > 0, y \cdot y_0 > 0$,

$$\left|\frac{x}{y} - \frac{x_0}{y_0}\right| \leq \sqrt{\frac{1}{\min(|y|,|y_0|)^2} + \frac{\max(|x|,|x_0|)^2}{\min(|y|,|y_0|)^4}} \left(|x - x_0| + |y - y_0|\right). \tag{32}$$

By A2, there are constants $\alpha, \beta > 0$ such that

$$|\lambda_i| < \alpha \quad \text{and} \quad 1/\beta < \min_k \sum_{l \in \Psi_k} (b_l - b_1) \leq \max_k \sum_{l \in \Psi_k} (b_l - b_1) < \beta.$$

Under the complement event of the union of (30) and (31) and assuming $\tilde{c}'' \max(\frac{\log^2 n}{\sqrt{n}}, \frac{1}{n\rho_n}) < \beta/2$, we apply (32) with

$$x_0 = \sum_{l \in \Psi_1} (\mu_{i1} - \mu_{il}), y_0 = \frac{\sum_{i \in G_1} \sum_{l \in \Psi_1} (\mu_{i1} - \mu_{il})}{n_1},$$

$$x = \sum_{l \in \Psi_1} (Y_{i1} - Y_{il}), y = \frac{\sum_{i \in G_1} \sum_{l \in \Psi_1} (Y_{i1} - Y_{il})}{n_1}.$$

This gives

$$\begin{aligned} |\hat{\lambda}_i - \lambda_i| &\leq \sqrt{\frac{1}{(|y_0|/2)^2} + \frac{(|x_0| + 5K\frac{\log n}{\sqrt{\kappa n \rho_n \phi}})^2}{(|y_0|/2)^4}} (\tilde{c}'' \max(\frac{\log^2 n}{\sqrt{n}}, \frac{1}{n\rho_n}) + 5K\frac{\log n}{\sqrt{\kappa n \rho_n \phi}}) \\ &\leq \tilde{c}''' \max(\frac{\log^2 n}{\sqrt{n}}, \frac{\log n}{n\rho_n}) \end{aligned}$$

with probability at least $1 - 2K\exp(-\frac{\kappa \log^2 n}{16}) - 4\exp\left(-\frac{\log^2 n}{8}\right)$. This completes the proof of Part 3. ■

Appendix C. Additional results about hiring network analysis

The network we use is shown in Figure 5, where the node size is proportional to the receiver degree, i.e., the number of institutions to which institution i has sent its graduates.

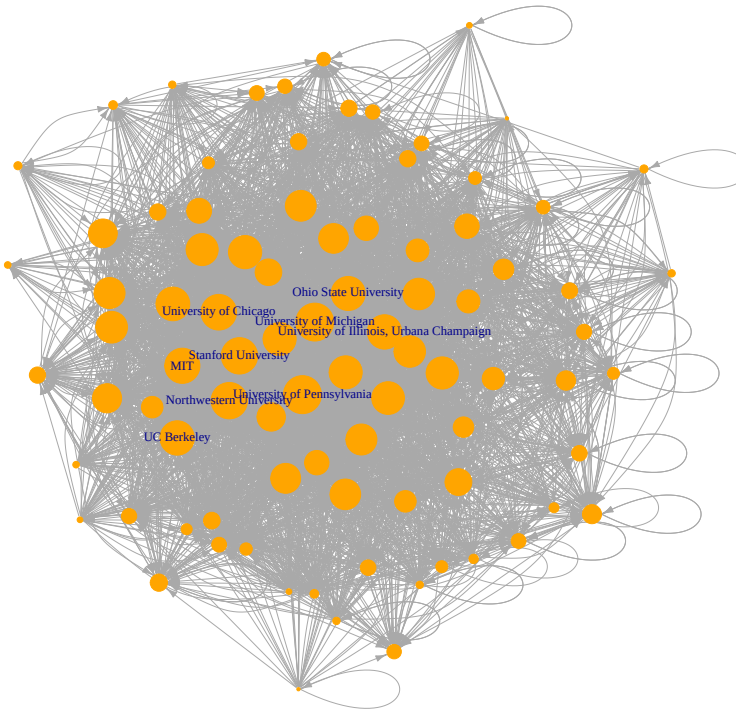


Figure 5: The hiring network between 87 U.S. business schools. An edge from i to j indicates that institution i has hired Ph.D. graduates from institution j. The node size is proportional to the number of incoming edges.

The edge weights in the hiring network have heavy tails, as shown in Figure 6. Our fitting strategy, the method of moments, relies on well-concentrated observations, and we truncated the edge weights at 2 for the main analysis in the paper, which affects 10% of edges. Table 4 compares those communities to the results obtained if we truncate the edge weights at 3 instead, changing the weights of 6% of edges. Only 7 out of 87 schools change communities as a result, and all but one of them change to an adjacent group. Overall, the results are stable relative to the choice of threshold, as one would expect from the proportions in Table 3.

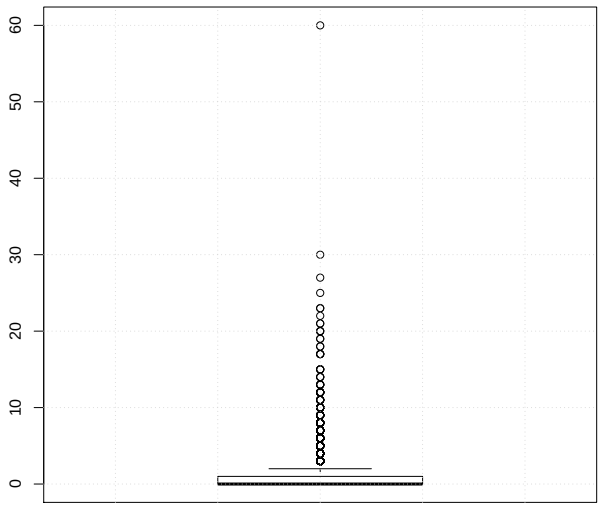


Figure 6: The boxplot of edge weights in the original hiring data.

Table 3: The distribution of edge weights in the original hiring data

edge weights	0	1	2	3	4	5 or more
proportion	64%	18%	8%	4%	3%	3%

Table 4: Confusion matrix of the community labels based on edge weights truncated at 2 and at 3.

		Truncated at 3			
		1	2	3	4
Truncated at 2	1	12	0	0	0
	2	3	8	0	1
	3	0	0	16	3
	4	0	0	0	44

Table 5 shows the estimated λ_i values for Group 1 schools.

For comparison, we briefly discuss community detection results on this network obtained by spectral clustering applied to the undirected version. The four communities are shown in Table 6, with their average and median ranking by US News and π-ranking, and 15 institutions with the highest π-ranking in each community. The first group is still higher-ranked even though it no longer includes universities like Yale, Columbia and Cornell. The other three groups, however, all have similar average rankings and no discernible interpretation that we could think of that might result in such groupings. The striking difference between

Table 5: Estimated λ_i's for Group 1 institutions.

Institution	$\hat{\lambda}_i$	USN ranking	π-ranking
Yale	1.35	10	11
Columbia	1.22	9	10
Cornell	1.18	16	7
Harvard	1.12	2	3
MIT	1.11	3	2
UC Berkeley	0.99	7	4
U of Chicago	0.91	5	6
Stanford	0.90	1	1
New York U	0.86	10	16
Duke	0.85	12	19
U Michigan	0.75	14	9
U Pennsylvania	0.71	3	12

the average rankings of groups from these two clustering results confirms the importance of accounting for the nomination mechanism and using the correct spectral information.

Table 6: Communities of business schools found by symmetric spectral clustering, their average and median rankings from US News 2012 and π-ranking of Clauset et al. (2015). Up to 15 institutions with the highest π-ranking are listed for each community.

	size	USN (avg./med.)	π-ranking (avg./med.)	Institutions
1	19	19.2/14	17.8/13	Stanford, MIT, Harvard, UC Berkeley, U Rochester, U Chicago, Northwestern, U Michigan, U Penn., Carnegie Mellon, NYU, U Minnesota Twin Cities, Duke, UNC Chapel Hill, U Washington St. Louis
2	20	55.1/56.5	44.6/42	Cornell, Columbia, U Wisconsin-Madison, UIUC, Ohio State, U Florida, U Pittsburgh, Penn State, Michigan State, SUNY Buffalo, U Mass Amherst, Syracuse, Tulane, U Connecticut, U Cincinnati
3	24	52.7/40	54/49	Yale, UCLA, U Washington, U Colorado Boulder, UC Irvine, U Utah, U Oregon, UT Dallas, U Virginia, Boston U, UC Davis, Vanderbilt, Claremont Graduate U, U Houston, Rice U
4	24	63.8/63	56/56.5	Purdue, U Iowa, UT Austin, Indiana U, Georgia Tech, U Arizona, Texas A&M, U Georgia, Arizona State, U South Carolina, Virginia Tech, Florida State, U Oklahoma, U Kansas, Louisiana State