Analysis of Networks via the Sparse β-Model*

Mingli Chen † Kengo Kato ‡ Chenlei Leng §

Abstract

Data in the form of networks are increasingly available in a variety of areas, yet statistical models allowing for parameter estimates with desirable statistical properties for sparse networks remain scarce. To address this, we propose the Sparse β-Model (SβM), a new network model that interpolates the celebrated Erdős-Rényi model and the β-model that assigns one different parameter to each node. By a novel reparameterization of the β-model to distinguish global and local parameters, our SβM can drastically reduce the dimensionality of the β-model by requiring some of the local parameters to be zero. We derive the asymptotic distribution of the maximum likelihood estimator of the SβM when the support of the parameter vector is known. When the support is unknown, we formulate a penalized likelihood approach with the ℓ_0-penalty. Remarkably, we show via a monotonicity lemma that the seemingly combinatorial computational problem due to the ℓ_0-penalty can be overcome by assigning nonzero parameters to those nodes with the largest degrees. We further show that a β-min condition guarantees our method to identify the true model and provide excess risk bounds for the estimated parameters. The estimation procedure enjoys good finite sample properties as shown by simulation studies. The usefulness of the SβM is further illustrated via the analysis of a microfinance take-up example.

Key Words: β-min condition; β-model; ℓ_0-penalized likelihood; Erdős-Rényi model; Exponential random graph models; Power law; Sparse networks

1 Introduction

Complex datasets involving multiple units that interact with each other are best represented by networks where nodes correspond to units and edges to interactions. Thanks to the rapid development of measurement and information technology, data in the form of networks are becoming increasingly available in a wide variety of areas including science, health, economics, engineering, and sociology (Jackson 2010, Barabási 2016, De Paula 2017, Newman 2018). Observed networks tend to be sparse, namely having much fewer edges than the maximum possible numbers of links allowed, and exhibits various degrees of heterogeneity. One of the major goals of analysis of networks is to understand the generative mechanism of the interconnections among the nodes in such networks using statistical models. We refer to Goldenberg et al.

*First arXiv version: August 8, 2019. This version: August 12, 2019.

†Department of Economics, University of Warwick, Coventry, CV4 7AL, UK. Email: m.chen.3@warwick.ac.uk

‡Department of Statistics and Data Science, Cornell University, 1194 Comstock Hall, Ithaca, NY 14853. Email: kk976@cornell.edu

§Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK. Email: C.Leng@warwick.ac.uk

(2009) and Fienberg (2012) for reviews, Kolaczyk (2009) for a comprehensive treatment, and Kolaczyk (2017) for foundational issues and emerging challenges. The study of various statistical properties of a network model is usually conducted by allowing the number of nodes n to go to infinity.

The earliest, simplest and perhaps the most studied network model is the Erdős-Rényi model (Erdős & Rényi 1959, 1960, Gilbert 1959) where connections between pairs of nodes independently occur with the same probability p. The resulting distribution of the degree of any node is Poisson for large n if np equals a constant. Probabilistically, the simplicity of the Erdős-Rényi model has permitted the development of many insights on networks as a mathematical object such as the existence of giant components and phase transition. The Erdős-Rényi model is also attractive from a theoretical perspective as discussed in Section 2. In particular, the maximum likelihood estimator (MLE) of its parameter is consistent and asymptotically normal for both dense and sparse networks. By sparse, we mean that the number of edges of a network scales sub-quadratically with the number of nodes. Similar phenomena are discussed for a closely related model for directed networks in Krivitsky & Kolaczyk (2015) that study the fundamental issue of the effective sample size of a network model. Despite its theoretical attractiveness, however, the Erdős-Rényi model is not suitable for modeling real networks whose empirical degree distributions are often heavy-tailed because it tends to produce degree distributions similar to Poisson (Clauset et al. 2009, Newman 2018). For example, many real networks are found to be scale-free (Barabási & Bonabau 2003) with their empirical degree distribution following a *power law*, at least asymptotically, in that, for large values of k, the fraction of nodes in the network having k connections is proportional to $k^{-\tau}$ for some $\tau > 1$. Intuitively, a typical network often exhibits a certain level of degree heterogeneity, usually having few high degree "core" nodes with many edges and many low degree individuals with few links (Barabási & Bonabau 2003, Clauset et al. 2009, Newman 2018). We refer further to Caron & Fox (2017) for related discussions in statistics and a novel attempt in using exchangeable random measures to model sparse networks with power-law degree distributions.

Many statistical models have been developed to directly account for degree heterogeneity. Two prominent examples are the stochastic block model and the β-model. The former aims to capture degree heterogeneity by clustering nodes into communities with similar connection patterns (Holland et al. 1983, Wang & Wong 1987, Bickel & Chen 2009, Abbe 2018), while the latter explicitly models degree heterogeneity by using node-specific parameters (Britton et al. 2006, Chatterjee et al. 2011). The β-model is a generalization of the Erdős-Rényi model where the probability that two nodes are connected depends on the corresponding two node parameters. It is one of the simplest exponential random graph models (Robins et al. 2007) and a special case of the p_1 model (Holland & Leinhardt 1981). Britton et al. (2006) show that the β-model can essentially generate degree sequences following a power law. The recent work of Mukherjee et al. (2019) studies sharp thresholds for detecting sparse signals in the β-model from a hypothesis testing perspective.

Statistically, however, the β-model has a limitation when sparse networks are considered. Namely, until now, the MLE of the β-model parameters is known to be consistent and asymptotically normal only for relatively dense networks (Chatterjee et al. 2011, Yan & Xu 2013). We refer also to Rinaldo et al. (2013) and Karwa & Slavković (2016) for further results concerning the MLE for the β-model, and Yan et al. (2016) for similar results on the MLE of the parameters in the p_1 model. The gap between the need for modeling sparse networks that are commonly seen in practice and the theoretical guarantees of the β-model that are available for much denser networks thus necessitates the development of new models.

In this paper, we propose a new network model which we call the *Sparse β-Model* (abbreviated as SβM) that can capture node heterogeneity and at the same time allows parameter estimates with desirable statistical properties under sparse network regimes, thereby complementing the Erdős-Rényi and β-models. Specifically, the SβM is defined by a novel reparameterization of the β-model to distinguish parameters characterizing global and local sparsity of the network. Using a cardinality constraint on the local parameters, the SβM can effectively interpolate the Erdős-Rényi and β-models with a continuum of intermediate models while reducing the dimensionality of the latter. As will become clear soon, the word "sparse" in SβM refers to the sparsity of the parameters as often used in high-dimensional statistics in the sense that many parameters in the SβM are assumed irrelevant, and should not be confused with sparsity of the network.

We study several statistical properties of the SβM in the asymptotic setting where the number of nodes tends to infinity. We first show that, similarly to Britton et al. (2006), if the parameters in the SβM are randomly generated in a suitable way, then the empirical degree distribution converges in probability to a power law. Second, we study parameter estimation in the SβM. We derive the asymptotic distribution of the maximum likelihood estimator when the support of the parameter vector is known. Although this result should be considered as a theoretical benchmark, it leads to the following important properties of the SβM: 1) the MLE of the parameters in the SβM can achieve consistency and asymptotic normality under sparse network regimes, and 2) the SβM can also capture the heterogeneous sparsity patterns for the individual nodes. Next, we consider a more practically relevant case where the support is unknown and formulate a penalized likelihood approach with the ℓ_0-penalty. Remarkably, we show via a monotonicity lemma that the seemingly combinatorial computational problem due to the ℓ_0-penalty can be overcome by assigning nonzero parameters to those nodes with the largest degrees. We show further that a β-min condition guarantees our method to identify the true model with high probability and derive excess risk bounds for the estimated parameters. In particular, we show that the ℓ_0-penalized MLE is *persistent* in the sense of Greenshtein & Ritov (2004) for (dense and) sparse networks under mild regularity conditions. The simulation study confirms that the ℓ_0-penalized MLE with its sparsity level selected by Bayesian Information Criterion (BIC) works well in the finite sample, both in terms of model selection and parameter estimation.

Our development of the SβM is practically motivated by the microfinance take-up dataset of 43 rural Indian villages in Banerjee et al. (2013). A detailed description of this dataset can be found in Section 5. In Figure 1, we plotted a sub-network of the dataset corresponding to one of the villages with $n = 150$ nodes as well as their empirical degree distribution. The average degree is 7.21, the maximum degree is 32, and there are 10 nodes with no connections at all. From the left plot, we can see that there are few nodes with many edges and many peripheral nodes with few connections. The right plot presents the empirical degree distribution on the log-log scale. Fitting a linear regression model to the points with degrees greater than 4, we can see that the resulting red dashed line gives a reasonable approximation to the tail of the distribution. This indicates that the tail of the distribution may be captured by a power law (Barabási 2016). In contrast, assuming a Poisson distribution, we also plotted the fitted Poisson probabilities in the black dash dotted line. It is clearly seen that a Poisson model fails to provide a reasonable fit to the tail of the empirical distribution. Because the Erdős-Rényi model tends to generate an empirical degree distribution following a Poisson law, we conclude that it is not able to capture the spreadout of the observed degrees.

The network structure in Figure 1 depicts features in so-called core-periphery or leaders-followers networks commonly seen in financial economics due to the presence of one group of core nodes and another group of peripheral nodes, e.g. over-the-counter markets for financial

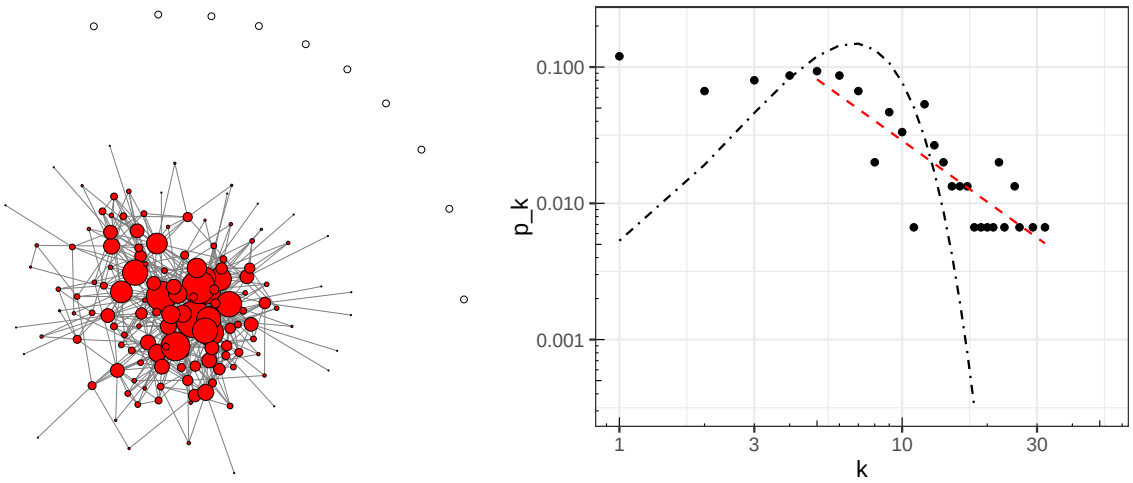

Figure 1: Left: The network of Village 14. The size of each node is proportional to its degree. Right: The solid dots are the degree distribution (frequency of degree denoted as p_k versus degree k) on the log-log scale for which a power law will follow a straight line. The red dashed line is the best linear regression fit to the points with degrees greater than 4. The fitted degree distribution assuming the frequencies follow a Poisson distribution is plotted as the black dash dotted line.

assets are dominated by a relatively small number of core intermediaries and a large number of peripheral customers. The core nodes are densely connected with each other and to the peripheral nodes, while the peripheral nodes are typically only connected to the core nodes but not to each other. This structure has important policy implications. For example, small shocks to those core/hub/leading players will affect the entire network (Acemoglu et al. 2012) because of their roles in facilitating diffusion (Banerjee et al. 2013). It is thus natural to associate those important core nodes with their individual parameters while leaving the less important peripheral nodes as background nodes without associated parameters. The SβM is a model for doing this.

The rest of the paper is organized as follows. In Section 2, we define the SβM, establish its connection to the Erdős-Rényi and β-models, discuss its properties, and show that the SβM can generate an empirical degree distribution converging in probability to a power law if the parameters are generated in a suitable way. We also derive some auxiliary asymptotic results for the Erdős-Rényi model. In Section 3, we consider estimation of the parameters in the SβM. We first consider the ideal situation that the support of the parameter vector is known and derive consistency and asymptotic normality results for the MLE. Next, we consider a more practically relevant situation where the support is unknown and formulate a penalized likelihood approach with the ℓ_0-penalty building on a monotonicity lemma, and derive some statistical properties of the estimator. In Section 4, we provide extensive simulation results. In Section 5, we analyze the microfinance take-up example. A summary and discussion on future research are given in Section 6. All the proofs are relegated to the Appendix.

1.1 Notation

Let $\mathbb{R}_+ = [0, \infty)$ denote the nonnegative real line. For a finite set F, let $|F|$ denote its cardinality. For a vector $\boldsymbol{\beta} \in \mathbb{R}^n$, let $S(\boldsymbol{\beta}) = \{i \in \{1, \ldots, n\} : \beta_i \neq 0\}$ denote the support of $\boldsymbol{\beta}$, and let $\|\boldsymbol{\beta}\|_0$ denote the number of nonzero elements of $\boldsymbol{\beta}$, i.e., $\|\boldsymbol{\beta}\|_0 = |S(\boldsymbol{\beta})|$. We use $\boldsymbol{\beta}_S$ to denote the subvector of $\boldsymbol{\beta}$ with indices in S and S^c as the complement of S. For two sequence of

positive numbers a_n and b_n, we write $a_n \sim b_n$ if $-\infty < \liminf_{n\to\infty} a_n/b_n \le \limsup_{n\to\infty} a_n/b_n < \infty$.

A network with n nodes is represented by a graph $G_n = G_n(V, E)$ where V is the set of nodes or vertices and E is the set of edges or links. Let $A = (A_{ij})_{i,j=1}^n$ be the adjacency matrix where $A_{ij} \in \{0, 1\}$ is an indicator whether nodes i and j are connected:

$$A_{ij} = \begin{cases} 1 & \text{if nodes } i \text{ and } j \text{ are connected} \\ 0 & \text{if nodes } i \text{ and } j \text{ are not connected} \end{cases}.$$

We focus on undirected graphs with no self loops, so that the adjacency matrix A is symmetric with zero diagonal entries. The degree of node i is defined by $d_i = \sum_{j=1}^n A_{ij} = \sum_{j\neq i} A_{ij}$, and the vector $\boldsymbol{d} = (d_1, \ldots, d_n)^T$ is called the degree sequence of G_n. The total number of edges is denoted by $d_+ = \sum_{i=1}^n d_i/2 = \sum_{1\le i<j\le n} A_{ij}$. Modeling a random network or graph is carried out by modeling the entries of A as random variables (Bollobás et al. 2007). Denote by $D_+ = E[d_+]$ the expected number of total edges, which is a function of n, typically a polynomial. We say that a (random) network is *dense* if $D_+ \sim n^2$ and that it is *sparse* if $D_+ \sim n^\kappa$ for some $\kappa \in (0, 2)$ (Bollobás & Riordan 2011). Apparently, the smaller κ is, the sparser the network is.

2 Sparse β-Model

We first review the Erdős-Rényi model and the β-model as a motivation to our SβM. The Erdős-Rényi model assumes that A_{ij}'s are generated as independent Bernoulli random variables with

$$P(A_{ij} = 1) = p = \frac{e^\mu}{1 + e^\mu},$$

where p and μ are parameters possibly dependent on n. Given the graph G_n, the MLE of p is

$$\hat{p} = \frac{1}{\binom{n}{2}} \sum_{1\le i<j\le n} A_{ij} = \frac{2d_+}{n(n-1)},$$

which is also known as the density of the network. The next proposition shows that the MLE $\hat{p}$ retains asymptotic normality even for sparse networks. That is, we assume that $p = p_n$ may tend to zero as $n \to \infty$ to accommodate sparse network regimes.

Proposition 1. *Consider the Erdős-Rényi model. Assume that $n^\gamma p \to p^\dagger$ as $n \to \infty$ where $p^\dagger > 0$ is a fixed constant and $\gamma \in [0, 2)$. Then $n^{1+\gamma/2}(\hat{p}-p) \xrightarrow{d} N(0, \sigma^2_{p^\dagger})$, where $\sigma^2_{p^\dagger} = 2p^\dagger(1-p^\dagger)$ for $\gamma = 0$ and $\sigma^2_{p^\dagger} = 2p^\dagger$ for $\gamma \in (0, 2)$. If instead we assume $n^\gamma p = p^\dagger$, then the MLE of $p^\dagger$, denoted as $\hat{p}^\dagger = n^\gamma \hat{p}$, satisfies that $n^{1-\gamma/2}(\hat{p}^\dagger - p^\dagger) \xrightarrow{d} N(0, \sigma^2_{p^\dagger})$.*

The expected number of edges for the Erdős-Rényi model satisfies $D_+ \sim n^{2-\gamma}$ if $n^\gamma p \to p^\dagger$. The proposition shows that as long as $D_+ \to \infty$, which also allows for sparse networks, the MLE of p is asymptotically normal. If we assume further $n^\gamma p = p^\dagger$, then $p^\dagger$ as a non-degenerate constant can be consistently estimated with its MLE being asymptotically normal. In particular, for dense networks where $\gamma = 0$, $\hat{p}^\dagger$ is n-consistent. For sparse networks where $\gamma = 1$, $n\hat{p}^\dagger$ is $\sqrt{n}$-consistent. For a more general γ, the rate of convergence of $\hat{p}^\dagger$ is $n^{1-\gamma/2}$ and the asymptotic variance of $\hat{p}^\dagger$ is proportional to $n^{-2+\gamma}$. Thus $n^{2-\gamma}$ can be seen as the effective sample size for the size invariant parameter $p^\dagger$. The notion and importance of the effective

sample size of a network model has been discussed and highlighted by Krivitsky & Kolaczyk (2015) that study a closely related model for directed networks in the special case when $\gamma = 0$ or 1. We can also work with the parameter μ of the Erdős-Rényi model on the logit scale as follows.

Corollary 1. *Assume that $n^\gamma p \to p^\dagger$ as $n \to \infty$ where $p^\dagger > 0$ is a fixed constant and $\gamma \in [0, 2)$. Define $\mu^\dagger = \log[p^\dagger/(1-p^\dagger)]$ for $\gamma = 0$ and $\mu^\dagger = \log p^\dagger$ for $\gamma \in (0, 2)$. The MLE of $\mu = \log[p/(1-p)]$ is $\hat{\mu} = \log[\hat{p}/(1-\hat{p})]$ and we have $n^{1-\gamma/2}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2_{\mu^\dagger})$, where $\sigma^2_{\mu^\dagger} = 4 + 2e^{-\mu^\dagger} + 2e^{\mu^\dagger}$ if $\gamma = 0$ and $\sigma^2_{\mu^\dagger} = 2e^{-\mu^\dagger}$ if $\gamma \in (0, 2)$. In addition, we can expand μ as $\mu = -\gamma \log n + \mu^\dagger + o(1)$.*

Again the scaling factor $n^{2-\gamma}$ can be viewed as the effective sample size of the network model. From Proposition 1 and this corollary, the Erdős-Rényi model has a desirable statistical property that the MLE is asymptotically normal under a wide spectrum of sparsity levels of networks.

With a single parameter, however, the Erdős-Rényi model cannot capture the power law phenomenon often seen in practice. For example, when np converges to a constant, the degree distribution behaves similarly to a Poisson law for large n. An alternative model specifically designed for capturing degree heterogeneity is the β-model that assigns one parameter for each node (Chatterjee et al. 2011). In particular, this model assumes that A_{ij}'s are independent Bernoulli random variables with

$$P(A_{ij} = 1) = p_{ij} = \frac{e^{\beta_i + \beta_j}}{1 + e^{\beta_i + \beta_j}}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^T \in \mathbb{R}^n$ is an unknown parameter. In this model, β_i has a natural interpretation in that it measures the propensity of node i to have connections with other nodes. Namely, the larger β_i is, the more likely node i is connected to other nodes. The resulting log-likelihood under the β-model is easily seen as

$$\sum_{i=1}^{n} \beta_i d_i - \sum_{1 \le i < j \le n} \log(1 + e^{\beta_i + \beta_j})$$

and the degree sequence $\boldsymbol{d} = (d_1, \ldots, d_n)^T$ is thus a sufficient statistic. Because of this, the β-model offers a simple mechanism to describe the probabilistic variation of degree sequences, which serves as an important first step towards understanding the extent to which nodes participate in network connections. More importantly, the β-model has emerged in recent years as a theoretically tractable model amenable for statistical analysis. In particular, Chatterjee et al. (2011) prove the existence and consistency of the MLE of $\boldsymbol{\beta}$, while Yan & Xu (2013) show its asymptotic normality. Britton et al. (2006) show that, if β_i are randomly generated in a suitable way, the β-model can generate node degrees asymptotically following a power law and the empirical degree distribution converging in probability to the same power law. See Theorem 3 in Appendix B for more details.

Despite these attractive properties, the β-model has a limitation when sparse networks are considered. Up to now, the known sufficient condition for the MLE of the β-model to be consistent and asymptotically normal is $\max_{1 \le i \le n} |\beta_i| = o(\log \log n)$ (Chatterjee et al. 2011, Yan & Xu 2013), although this condition may not be the best possible. This condition implies that

$$\min_{1 \le i < j \le n} p_{ij} \gg \frac{e^{-C \log \log n}}{1 + e^{-C \log \log n}} \sim (\log n)^{-C}$$

for some positive constant C. Under this condition, the expected number of edges of the network should be of order at least $n^2(\log n)^{-C}$ and hence the network will be dense up to a logarithmic factor. Part of this requirement stems from the need to estimate n parameters, so we need a sufficient number of connections for each node to estimate all the β parameters well.

To conclude, the Erdős-Rényi model is simple enough to allow desirable asymptotic properties for the MLE under a variety of sparsity levels of the network but too under-parametrized to explain many notable features of the network. On the other hand, the over-parametrized β-model is more flexible at the expense of a minimal requirement for the density of the network. Motivated from these observations, we propose the Sparse β-Model (SβM) that retains their attractive properties. Specifically, the SβM assumes that A_{ij}'s are independent Bernoulli random variables with

$$P(A_{ij} = 1) = p_{ij} = \frac{e^{\mu+\beta_i+\beta_j}}{1 + e^{\mu+\beta_i+\beta_j}}, \tag{2}$$

where $\mu \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^n_+$ are both unknown parameters. To ensure identifiability, we require that the elements of $\boldsymbol{\beta}$ are nonnegative with at least one element equal to zero, i.e., $\min_{1\le i\le n} \beta_i = 0$. Hence $\|\boldsymbol{\beta}\|_0 \le n - 1$. A key assumption we make on the SβM is that $\boldsymbol{\beta}$ is sparse and we are mainly interested in the case where $\|\boldsymbol{\beta}\|_0 \ll n$.

In this model, $\mu \in \mathbb{R}$ can be understood as the intercept, a baseline term that may tend to $-\infty$ as $n \to \infty$, which allows various sparsity levels for the network similarly to the role of μ in the Erdős-Rényi model. Thus μ is the global parameter characterizing the sparsity of the entire network. On the other hand, $\boldsymbol{\beta} \in \mathbb{R}^n_+$ is a vector of node specific parameters. It can be understood that node i has no individual effect in forming connections if $\beta_i = 0$, and therefore β_i controls the local sparsity of the network around node i in addition to its baseline parameter μ. Such separate treatment of the global and local parameters corresponds to the roles that core and peripheral nodes play in a network. In the context of the microfinance example in Figure 1, this model allows us to differentially assign parameters only to certain nodes, e.g. those nodes that are considered "core". In Figure 2, three simulated examples with $n = 50$, 100 and 200 are presented to give a general idea of the networks generated from our model, where cores and peripherals are highly visible.

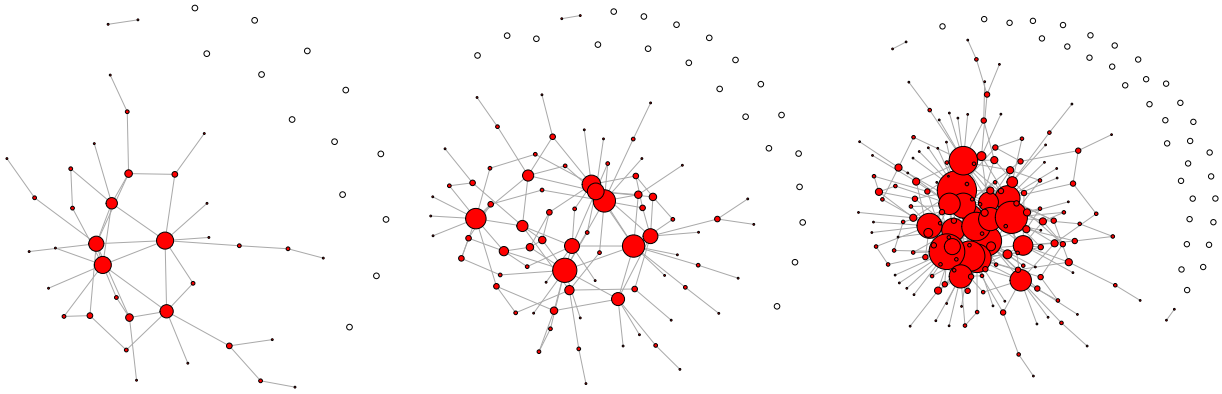

Figure 2: Some sample networks generated from the SβM. Left: $n = 50$, Middle: $n = 100$, Right: $n = 200$. The size of the vertex is proportional to its degree. The size of the support of $\boldsymbol{\beta}$ is set as $n/10$ with $\beta_i = \sqrt{\log n}$ or 0, while $\mu = -\log n$.

Without the sparsity assumption on $\boldsymbol{\beta}$, the SβM reduces to a reparametrized version of the β-model by shifting β_i in the latter by $\mu/2$. On the other extreme end when $\|\boldsymbol{\beta}\|_0 = 0$, the S$\beta$M reduces to the Erdős-Rényi model. Thus, the SβM interpolates the Erdős-Rényi and β-models. By allowing the sparsity level $\|\boldsymbol{\beta}\|_0$ to be much smaller than n, the SβM can drastically reduce

the number of parameters needed in the β-model, and, as will be discussed in Section 3, allow parameter estimators with desirable statistical properties under sparse network regimes.

We note that Mukherjee et al. (2019) consider a different reparameterization of the β-model to induce sparsity of resulting networks. Specifically, they consider the model

$$P(A_{ij} = 1) = \frac{\lambda}{n} \frac{e^{\beta_i + \beta_j}}{1 + e^{\beta_i + \beta_j}}.$$

The focus of Mukherjee et al. (2019) is different from ours and on testing the hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ against the alternative that $\boldsymbol{\beta}$ is nonzero but sparse. In addition, they do not consider estimation of the parameters when $\boldsymbol{\beta}$ is sparse, and assume that λ is a known constant, which should be contrasted with our SβM where both μ and $\boldsymbol{\beta}$ are unknown parameters.

In this paper, β parameters are mainly treated as fixed. To connect the SβM to power law, however, we invoke the following proposition by treating β_i's as random variables (Britton et al. 2006).

Proposition 2 (SβM and power law)**.** *Let $\{W_i\}_{i=1}^{\infty}$ be i.i.d. random variables supported in $[1, \infty)$ with $P(W_1 > w) \sim cw^{-\rho}$ as $w \to \infty$ for some $c > 0$ and $\rho \in (0, 1)$. For the SβM in (2), suppose that $\mu = -\rho^{-1} \log n$ and β_i's are generated as $\beta_i = \log W_i$. Then the limiting distribution of each node degree d_i as $n \to \infty$ is a power law with exponent $\tau = 2$, that is,*

$$p_k := \lim_{n \to \infty} P(d_i = k) \sim k^{-2}, \ k \to \infty.$$

In addition, for $N_k := |\{i \in \{1, \ldots, n\} : d_i = k\}|$, we have $N_k/n \xrightarrow{P} p_k$ as $n \to \infty$.

In the above proposition, we do not impose sparsity on $\boldsymbol{\beta}$, but it is possible to do so by assuming a mass at 1 to the distribution of W_1 since the assumption only requires the tail of the distribution of W_1 to behave like $cw^{-\rho}$. The proposition follows from the results of Britton et al. (2006), which are restated in Appendix B.

3 Parameter Estimation in SβM

In this section, we consider estimation of the parameters in the SβM. We will denote the true parameter value of $(\mu, \boldsymbol{\beta})$ by $(\mu_0, \boldsymbol{\beta}_0)$. We first discuss the case where the support of $\boldsymbol{\beta}_0$ is known. We consider the known support case for a theoretical purpose to study the properties of the SβM. Theorem 1 below reveals two important theoretical properties of the SβM: 1) the MLE of the parameters in the SβM can achieve consistency and asymptotic normality under sparse network regimes, and 2) the SβM can also capture the heterogeneous sparsity patterns for the individual nodes. Next, we consider a more practically relevant case where the support is unknown and study the ℓ_0-penalized MLE.

3.1 MLE with a known support

First, we consider the case where $S = S(\boldsymbol{\beta}_0)$, the support of $\boldsymbol{\beta}_0$, is known and study the asymptotic properties of the MLE for $(\mu_0, \boldsymbol{\beta}_{0S})$. The cardinality of the support $s_0 = |S| = \|\boldsymbol{\beta}_0\|_0$ may grow with the sample size n, i.e., $s_0 = s_{0n} \to \infty$. Similarly to Krivitsky et al. (2009) and Krivitsky & Kolaczyk (2015), we also introduce $\log n$ shifts to the parameters to accommodate sparsity of the network in the theoretical setup, and consider the statistical properties of the MLE of the scale-invariant parameters of the SβM.

Theorem 1 (Consistency and asymptotic normality of MLE with known support)**.** *Consider the reparameterization* $\mu = -\gamma \log n + \mu^\dagger$ *and* $\beta_i = \alpha \log n + \beta_i^\dagger$ *for all* $i \in S$ *for some* $\gamma \in [0, 2)$ *and* $\alpha \in [0, 1)$ *such that* $0 \le \gamma - \alpha < 1$*, and let* $[-M_1, M_1] \times [0, M_2]^{s_0}$ *be the parameter space for* $(\mu^\dagger, \boldsymbol{\beta}_S^\dagger)$ *where* (M_1, M_2) *are independent of* n*. Denote by* $(\mu_0^\dagger, \boldsymbol{\beta}_{0S}^\dagger)$ *the true parameter value for* $(\mu^\dagger, \boldsymbol{\beta}_S^\dagger)$*. Let* $(\hat{\mu}^\dagger, \hat{\boldsymbol{\beta}}_S^\dagger)$ *be an MLE of* $(\mu^\dagger, \boldsymbol{\beta}_S^\dagger)$*. Then:*

(i) If in addition $s_0 = o(n^{1-\alpha})$*, then the MLE* $(\hat{\mu}^\dagger, \hat{\boldsymbol{\beta}}_S^\dagger)$ *is uniformly consistent:* $\hat{\mu}^\dagger \xrightarrow{P} \mu_0^\dagger$ *and* $\max_{i \in S} |\hat{\beta}_i^\dagger - \beta_{0i}^\dagger| \xrightarrow{P} 0$.

(ii) If in addition $|\mu_0^\dagger| < M_1$*,* $\eta \le \min_{i \in S} \beta_{0i}^\dagger \le \max_{i \in S} \beta_{0i}^\dagger \le M_2 - \eta$ *for some small constant* $0 < \eta < M_2$ *independent of* n*, and* $s_0 = o(n^{(1-\alpha)/2}/\sqrt{\log n})$*, then for any fixed subset* $F \subset S$*, we have*

$$\begin{pmatrix} n^{1-\gamma/2}(\hat{\mu}^\dagger - \mu_0^\dagger) \\ n^{1/2-(\gamma-\alpha)/2}(\hat{\beta}_i^\dagger - \beta_{0i}^\dagger)_{i \in F} \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, \Sigma_F),$$

where Σ_F *is the diagonal matrix with diagonal entries*

$$\begin{cases} 2e^{-\mu_0^\dagger} \text{ and } e^{-\mu_0^\dagger - \beta_{0i}^\dagger} \text{ for } i \in F & \text{if } \alpha < \gamma \\ 2e^{-\mu_0^\dagger} \text{ and } 2 + e^{-\mu_0^\dagger - \beta_{0i}^\dagger} + e^{\mu_0^\dagger + \beta_{0i}^\dagger} \text{ for } i \in F & \text{if } \gamma = \alpha \in (0, 1) \\ 4 + 2e^{-\mu_0^\dagger} + 2e^{\mu_0^\dagger} \text{ and } 2 + e^{-\mu_0^\dagger - \beta_{0i}^\dagger} + e^{\mu_0^\dagger + \beta_{0i}^\dagger} \text{ for } i \in F & \text{if } \gamma = \alpha = 0 \end{cases}.$$

Some comments on the theorem are in order. In what follows, we focus on the case where $\alpha < \gamma$ for simplicity of exposition. The expected number of edges of the SβM under the condition of the preceding theorem is

$$\begin{aligned} D_+ = E[d_+] &= \sum_{1 \le i < j \le n} p_{ij} \\ &= \binom{n - s_0}{2} \frac{n^{-\gamma} e^{\mu_0^\dagger}}{1 + n^{-\gamma} e^{\mu_0^\dagger}} + (n - s_0) \sum_{i \in S} \frac{n^{-(\gamma-\alpha)} e^{\mu_0^\dagger + \beta_{0i}^\dagger}}{1 + n^{-(\gamma-\alpha)} e^{\mu_0^\dagger + \beta_{0i}^\dagger}} + \sum_{\substack{i,j \in S \\ i<j}} \frac{n^{-(\gamma-2\alpha)} e^{\mu_0^\dagger + \beta_{0i}^\dagger + \beta_{0j}^\dagger}}{1 + n^{-(\gamma-2\alpha)} e^{\mu_0^\dagger + \beta_{0i}^\dagger + \beta_{0j}^\dagger}} \\ &= n^{2-\gamma} e^{\mu_0^\dagger}/2 + o(n^{2-\gamma}) \end{aligned}$$

provided that $s_0 = o(n^{1-\alpha})$ (see the proof of Theorem 1). Theorem 1 shows that the MLE of the parameters in the SβM (when the support is known) can achieve consistency and asymptotic normality for sparse networks, which should be contrasted with the β-model where the MLE is known to be consistent and asymptotically normal only for relatively dense networks.

In addition, the SβM can also capture the heterogeneous sparsity patterns for the individual nodes in the sense that

$$E[d_i] = \begin{cases} n^{1-(\gamma-\alpha)} e^{\mu_0^\dagger + \beta_{0i}^\dagger} + o(n^{1-(\gamma-\alpha)}) & \text{if } i \in S \\ n^{1-\gamma} e^{\mu_0^\dagger} + o(n^{1-\gamma}) & \text{if } i \notin S \end{cases}.$$

Intuitively speaking, γ (or the magnitude of μ) controls the global sparsity while α (or the magnitude of β_i) is controlling the local sparsity. Namely, as γ increases, all nodes will be less likely to be connected, while as α increases, the nodes in the support of $\boldsymbol{\beta}$ will be more likely to be connected. Theorem 1 shows how these global and local sparsity affect the effective sample sizes for μ (global parameter) and $\boldsymbol{\beta}_S$ (local parameter). The theorem implies that the effective sample size for the global parameter is $n^{2-\gamma}$ which is decreasing in γ similarly to the Erdős-Rényi graph (see the discussion after Corollary 1), while that for the local parameter is

$n^{1-\gamma+\alpha}$ which is decreasing in γ but increasing in α.

Finally, the proof of Theorem 1 is nontrivial since there are two types of parameters with different rates, one being common to the nodes and the other being node-specific, and the number of parameters $1 + s_0$ may diverge as n increases, which is reminiscent of the incidental parameter problem (Neyman & Scott 1948, Li et al. 2003, Hahn & Newey 2004). To prove the uniform consistency, we work with the concentrated negative log-likelihoods for $(\mu^\dagger, \boldsymbol{\beta}_S^\dagger)$ and show that they converge in probability to some nonstochastic functions uniformly in $i \in S$. To prove the asymptotic normality, we use iterative stochastic expansions to derive the uniform asymptotic linear representations for $(\hat{\mu}^\dagger, \hat{\boldsymbol{\beta}}_S^\dagger)$. See the proof in Appendix A.2 for the details.

3.2 ℓ_0-penalized MLE with an unknown support

In practice, the support of $\boldsymbol{\beta}_0$ is usually unknown. In this section, we consider and analyze the ℓ_0-norm constrained likelihood estimator for estimating the parameters of the model when this is the case. Writing the negative log-likelihood of the SβM as

$$\ell_n(\mu, \boldsymbol{\beta}) = -d_+\mu - \sum_{i=1}^n d_i\beta_i + \sum_{1\le i<j\le n} \log\left(1 + e^{\mu+\beta_i+\beta_j}\right),$$

we estimate the parameters as

$$(\hat{\mu}(s), \hat{\boldsymbol{\beta}}(s)) = \operatorname{argmin}_{\mu\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}_+^n} \ \ell_n(\mu, \boldsymbol{\beta}) \quad \text{subject to } \|\boldsymbol{\beta}\|_0 \le s, \tag{3}$$

where $s \in \{1, 2, \ldots, n-1\}$ is an integer-valued tuning parameter. We restrict s to be less than n so that the identifiability condition $\min_{1\le i\le n} \beta_i = 0$ is automatically satisfied. If there is a question of the existence of the global optimal solution in (3), we restrict the parameter space to be a (sufficiently large) compact rectangle.

The optimization problem (3) is a combinatorial problem that seems difficult to solve. For each s, a naive approach to compute the solution of (3) is to fit $\binom{n}{s}$ models, each assuming s out of n parameters in the SβM are nonzero, and then choose the model that gives the smallest negative log-likelihood. This strategy is used routinely in the so-called best subset selection for regression models, which is known for being unsuitable for datasets with a large number of parameters. Remarkably, the SβM has a property that at most only $n-1$ models need to be examined before the optimal choice is decided, making it attractive computationally. In particular, we have the following monotonicity lemma stating that the entries of $\hat{\boldsymbol{\beta}}(s) = (\hat{\beta}_1(s), \ldots, \hat{\beta}_n(s))^T$ are ordered according to those of the degree sequence $\boldsymbol{d} = (d_1, \ldots, d_n)^T$. Before presenting this lemma, we introduce the following notation to handle tied degrees. Let

$$d_{(1)} > d_{(2)} > \cdots > d_{(m)} \tag{4}$$

denote the distinctive values of d_i's. Denote by S_k the set of indices of those d_i's that equal to $d_{(k)}$ and by s_k its cardinality; that is, $S_k = \{i \in \{1, \ldots, n\} : d_i = d_{(k)}\}$ and $s_k = |S_k|$. By definition, $\sum_{k=1}^m s_k = n$. If no two degrees are tied, then $m = n$ and $s_k = 1$ for any $k = 1, \ldots, n$.

Lemma 1 (Monotonicity lemma)**.** *The estimate $\hat{\boldsymbol{\beta}}(s)$ in (3) has the following properties.*

(i) If $d_i < d_j$, then we have $\hat{\beta}_i(s) \le \hat{\beta}_j(s)$ for any $s < n$;

(ii) If $d_i = d_j$, then we have $\hat{\beta}_i(s) = \hat{\beta}_j(s)$ for any s such that $s = \sum_{k=1}^K s_k$ for some $K \le m-1$.

The proof of Lemma 1 and other proofs for Section 3.2 can be found in Appendix A.3. Lemma 1 implies that $\hat{\boldsymbol{\beta}}(s)$ as the constrained MLE of (3) has the same order as the degree sequence. That is, for the constrained optimization in (3) with a penalty parameter s, we just assign nonzero β to those nodes whose degrees are among the largest s nodes. More precisely, if $s = \sum_{k=1}^{K} s_k$ for some $K \leq m-1$, then $\hat{\beta}_i(s) \geq 0$ for $i \in \bigcup_{k=1}^{K} S_k$ and $\hat{\beta}_i(s) = 0$ for $i \in \bigcup_{K<k\leq m} S_k$. In other words, we can find *a priori* the support of $\hat{\boldsymbol{\beta}}(s)$ from the degree sequence and can compute $\hat{\boldsymbol{\beta}}(s)$ by solving the following optimization problem *without the ℓ_0-penalization*:

$$(\hat{\mu}(s), \hat{\boldsymbol{\beta}}(s)) = \operatorname{argmin}_{\mu\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^n_+} \ \ell_n(\mu, \boldsymbol{\beta}) \quad \text{subject to } \beta_i = 0 \text{ for } i \in \bigcup_{K<k\leq m} S_k.$$

This way, we can efficiently compute a solution path of $(\hat{\mu}(s), \hat{\boldsymbol{\beta}}(s))$ as a function of $s \in \{s_1, s_1+s_2, \ldots, \sum_{k=1}^{m-1} s_k\}$ without solving a computationally expensive combinatorial problem. We note that the set $\{1, \ldots, n-1\} \setminus \{s_1, s_1+s_2, \ldots, \sum_{k=1}^{m-1} s_k\}$ is excluded from consideration for the tuning parameter s, because otherwise the solution to the constrained optimization will not be unique.

The preceding lemma shows that there will be a sequence of supports

$$S_1, S_1 \cup S_2, \ldots, \bigcup_{k=1}^{m-1} S_k, \tag{5}$$

for $\hat{\boldsymbol{\beta}}(s)$ with $s \in \{s_1, s_1+s_2, \ldots, \sum_{k=1}^{m-1} s_k\}$. Next, we show that as long as the smallest nonzero element of $\boldsymbol{\beta}_0$ is above a certain threshold, with high probability the true support $S(\boldsymbol{\beta}_0)$ is included in the support sequence (5) constructed from the degree sequence $\boldsymbol{d}$.

Lemma 2. *Let $S = S(\boldsymbol{\beta}_0)$, and let $\tau \in (0,1)$ be given. Pick any $i \in S$ and $j \in S^c$. Suppose that*

$$\beta_{0i} > \log\left(1 + c_{n,\tau}(1+e^{\mu^-})(1+e^{2\overline{\beta}+\mu^+})\right), \tag{6}$$

where $c_{n,\tau} = \sqrt{(2/(n-2))\log(2/\tau)}$, $\overline{\beta} = \max_{1\leq k\leq n} \beta_{0k}$, $\mu^+ = \max\{\mu_0, 0\}$, and $\mu^- = \max\{-\mu_0, 0\}$. Then $d_i > d_j$ with probability at least $1-\tau$.

By the union bound, Lemma 2 immediately yields the following corollary.

Corollary 2 (β-min condition)**.** *Pick any $\tau \in (0,1)$. Suppose that the following β-min condition is satisfied:*

$$\min_{i\in S} \beta_{0i} > \log\left(1 + c_{n,\tau/n(n-1)}(1+e^{\mu^-})(1+e^{2\overline{\beta}+\mu^+})\right). \tag{7}$$

Then we have $\min_{i\in S} d_i > \max_{j\in S^c} d_j$ with probability at least $1-\tau$.

Corollary 2 specifies the minimum magnitude of the nonzero β's for the SβM to include the true support $S(\boldsymbol{\beta}_0)$ in the support sequence (5). For this reason, we call the condition in (7) the β-min condition. With this β-min condition, if we choose $s = |S(\boldsymbol{\beta}_0)|$, then we can identify the support of $\boldsymbol{\beta}_0$ by solving the optimization problem in (3) with a probability close to one. The issue of determining the sparsity level s will be discussed in Section 4.1. We note that $c_{n,\tau/n(n-1)} \sim \sqrt{(\log n)/n}$, and that the right hand side of (7) is of constant order as long as $e^{2\overline{\beta}+|\mu_0|} = O(\sqrt{n/\log n})$.

Finally, we evaluate the prediction risk for the estimator $(\hat{\mu}(s), \hat{\boldsymbol{\beta}}(s))$ for a given sparsity level s. Recall that the true value of $(\mu, \boldsymbol{\beta})$ is denoted by $(\mu_0, \boldsymbol{\beta}_0)$ with $s_0 = \|\boldsymbol{\beta}_0\|_0$. In general

s and s_0 may differ. Let $\mathcal{R}(\mu, \boldsymbol{\beta})$ be the risk of the parameter value $(\mu, \boldsymbol{\beta})$ which is defined by the expected normalized negative log-likelihood, i.e.,

$$\mathcal{R}(\mu, \boldsymbol{\beta}) = E[D_+^{-1} \ell_n(\mu, \boldsymbol{\beta})],$$

where we think of $D_+ = E[d_+]$ as the effective sample size. Normalization by D_+ is natural since the risk at the true parameter $\mathcal{R}(\mu_0, \boldsymbol{\beta}_0)$ is of constant order under sparse network scenarios; see the discussion after Theorem 2 (recall that in the linear regression case with squared loss function, the risk at the true parameter is the error variance, which is constant). For a given sparsity level s, consider the ℓ_0-constrained estimator $(\hat{\mu}(s), \hat{\boldsymbol{\beta}}(s))$ as in (3):

$$(\hat{\mu}(s), \hat{\boldsymbol{\beta}}(s)) = \operatorname{argmin}\{\ell_n(\mu, \boldsymbol{\beta}) : (\mu, \boldsymbol{\beta}) \in \Theta_s\},$$

where $\Theta_s = \{(\mu, \boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}_+^n : |\mu| \le M_1, \boldsymbol{\beta} \in [0, M_2]^n, \|\boldsymbol{\beta}\|_0 \le s\}$ and M_1, M_2 are given positive constants. We assume that M_1 and M_2 are sufficiently large and may increase with n, but suppress the dependence of the parameter space Θ_s on M_1 and M_2. In addition, both s_0 and s can depend on n. Following the empirical risk minimization literature (see, e.g., Greenshtein & Ritov 2004, Koltchinskii 2011), we will evaluate the performance of the estimator $(\hat{\mu}(s), \hat{\boldsymbol{\beta}}(s))$ by the (local) excess risk relative to the parameter space Θ_s

$$\mathcal{E}_s = \mathcal{R}(\hat{\mu}(s), \hat{\boldsymbol{\beta}}(s)) - \inf_{(\mu, \boldsymbol{\beta}) \in \Theta_s} \mathcal{R}(\mu, \boldsymbol{\beta}).$$

We note that the (global) excess risk relative to the true parameter $(\mu_0, \boldsymbol{\beta}_0)$ can also be bounded by the decomposition

$$\mathcal{R}(\hat{\mu}(s), \hat{\boldsymbol{\beta}}(s)) - \mathcal{R}(\mu_0, \boldsymbol{\beta}_0) = \left[\inf_{(\mu, \boldsymbol{\beta}) \in \Theta_s} \mathcal{R}(\mu, \boldsymbol{\beta}) - \mathcal{R}(\mu_0, \boldsymbol{\beta}_0) \right] + \mathcal{E}_s,$$

where the first term on the right hand side accounts for the deterministic bias. The following theorem derives high-probability upper bounds on the excess risk $\mathcal{E}_s$.

Theorem 2 (Excess risk bound)**.** *For any given $\tau \in (0, 1)$, we have*

$$\begin{aligned} \mathcal{E}_s \le \frac{2}{D_+} \Bigg[& M_1 \left\{ \sqrt{2 \operatorname{Var}(d_+) \log(4/\tau)} + (\log(4/\tau))/3 \right\} \\ & + M_2 s \left\{ \sqrt{2 \max_{1 \le i \le n} \operatorname{Var}(d_i) \log(4n/\tau)} + (\log(4n/\tau))/3 \right\} \Bigg] \end{aligned} \tag{8}$$

with probability at least $1 - \tau$. In particular, if $\mu_0 = -\gamma \log n + O(1)$ and $\beta_{0i} = \alpha \log n + O(1)$ uniformly in $i \in S(\boldsymbol{\beta}_0)$ for some $\gamma \in [0, 2)$ and $\alpha \in [0, 1)$ with $0 \le \gamma - \alpha < 1$, and $s_0 = o(n^{1-\alpha})$, then we have $D_+ \sim n^{2-\gamma}$ and

$$\mathcal{E}_s = O_P\left(\frac{M_1}{n^{1-\gamma/2}} + \frac{M_2 s \sqrt{\log n}}{n^{3/2-(\gamma+\alpha)/2}} \right). \tag{9}$$

In the latter setting of Theorem 2, it is not difficult to see that $E[\ell_n(\mu_0, \boldsymbol{\beta}_0)] \sim n^{2-\gamma}$ so that the risk at $(\mu_0, \boldsymbol{\beta}_0)$ normalized by D_+ is of constant order $\mathcal{R}(\mu_0, \boldsymbol{\beta}_0) \sim 1$. In addition, if e.g. $M_1 \sim \log n$ and $M_2 \sim \log n$, then the bound (9) becomes

$$\mathcal{E}_s = O_P\left(\frac{\log n}{n^{1-\gamma/2}} + \frac{s (\log n)^{3/2}}{n^{3/2-(\gamma+\alpha)/2}} \right).$$

Hence, the estimator $(\hat{\mu}(s), \hat{\boldsymbol{\beta}}(s))$ is *persistent* in the sense of Greenshtein & Ritov (2004), i.e., $\mathcal{E}_s \xrightarrow{P} 0$, as long as

$$s = o(n^{3/2-(\gamma+\alpha)/2}/(\log n)^{3/2}), \tag{10}$$

and provided that the true sparsity level satisfies $s_0 = o(n^{1-\alpha})$. Condition (10) is automatically satisfied if $\gamma + \alpha < 1$ since s is at most $n - 1$. In addition, the bound can achieve the near parametric rate $(\log n)/n^{1-\gamma/2}$ with respect to the effective sample size $D_+ \sim n^{2-\gamma}$ as long as $s = o(n^{(1-\alpha)/2}/\sqrt{\log n})$.

4 Simulation Study

4.1 Selection of sparsity level

In practice, we have to choose the sparsity level s for the ℓ_0-penalized MLE to work. In this simulation study, we will examine the following version of BIC

$$\mathsf{BIC}(s) = -2\ell_n(\hat{\mu}(s), \hat{\boldsymbol{\beta}}(s)) + s \log\left(n(n-1)/2\right). \tag{11}$$

We choose s that minimizes the BIC:

$$\hat{s} = \operatorname{argmin} \left\{ \mathsf{BIC}(s) : s \in \left\{ s_1, s_1 + s_2, \ldots, \sum_{k=1}^{m-1} s_k \right\} \right\}.$$

See Section 3.2 for the notation. The final estimator is then given by $(\hat{\mu}(\hat{s}), \hat{\boldsymbol{\beta}}(\hat{s}))$. We shall study the performance of the BIC via numerical simulations.

The BIC defined in (11) uses $n(n-1)/2$ as the sample size. In view of our previous discussion on the effective sample size, it would be natural to use D_+ or its unbiased estimate d_+ in place of $n(n-1)/2$ by defining a different BIC:

$$\mathsf{BIC}^*(s) = -2\ell_n(\hat{\mu}(s), \hat{\boldsymbol{\beta}}(s)) + s \log(d_+). \tag{12}$$

Preliminary simulation results suggest that, however, the performance of the BIC in (12) is similar or slightly worse than the one in (11) in most cases in terms of model selection and parameter estimation. Hence we only report the simulation results using (11).

4.2 Simulation results

In this simulation study, we consider the following configurations of $(\mu_0, \boldsymbol{\beta}_0)$:

(i) $\mu_0 = -\log\log n$, and $\beta_{0i} = \log\log n$, $\sqrt{\log n}$, or $\log n$ for $i \in S(\boldsymbol{\beta}_0)$;

(ii) $\mu_0 = -\sqrt{\log n}$, and $\beta_{0i} = \log\log n$, $\sqrt{\log n}$, or $\log n$ for $i \in S(\boldsymbol{\beta}_0)$;

(iii) $\mu_0 = -\log n$, and $\beta_{0i} = \log\log n$, $\sqrt{\log n}$, or $\log n$ for $i \in S(\boldsymbol{\beta}_0)$;

where $n = 50$, 100, 200 or 400. The sparsity level of $\boldsymbol{\beta}_0$ is either $s_0 = |S(\boldsymbol{\beta}_0)| = 2$, $\lfloor\sqrt{n/2}\rfloor$, $\lfloor\sqrt{n}\rfloor$, or $\lfloor 2\sqrt{n}\rfloor$, where $\lfloor a\rfloor$ denotes the largest integer smaller than a. Since the indices of the nonzero elements of $\boldsymbol{\beta}_0$ do not matter for our estimation procedure, we simply choose the first s_0 elements of $\boldsymbol{\beta}_0$ to be nonzero. The number of Monte Carlo repetitions is 1000 for each case of

simulation. To speed up our estimation procedure, in this simulation study, we restricted the maximum number of sparsity levels s examined to be 40.

These configurations are chosen to reflect various degrees of sparsity for the overall network globally and for individual nodes locally. Recall that an induced subgraph of a graph is another graph formed from a subset of the vertices of the graph and all of the edges connecting pairs of vertices in that subset. If $\mu_0 = -\log n$, then the subgraph induced by those nodes with zero β parameters will form a sparse Erdős-Rényi graph with $D_+ \sim n - s_0$. If $\mu_0 = -\log\log n$, then this subgraph is almost dense in that $D_+ \sim (n-s_0)^2/\log(n-s_0)$. If $\mu_0 = -\sqrt{\log n}$, then the induced subgraph lies somewhere between these two cases. The specification $\beta_{0i} = \log n$ is guided by the reparameterization in Theorem 1. By specifying $\beta_{0i} = \log\log n$ or $\sqrt{\log n}$, we want to consider those local parameters that are much smaller than $\log n$.

Figure 3 reports the frequencies when the support of $\boldsymbol{\beta}_0$ is correctly identified for different settings (Figure 6 in Appendix C provides the results on the average number of nonzero β selected based on our procedure in comparison to s_0). Figure 4 reports the simulation results on the ℓ_1-norm of $\hat{\boldsymbol{\beta}}(\hat{s}) - \boldsymbol{\beta}_0$ (Figure 7 in Appendix C reports the simulation results on $|\hat{\mu}(\hat{s}) - \mu_0|$).

Figure 3 shows that, in general, our estimation procedure works well in terms of model selection. In most cases, as n increases, BIC tends to correctly identify the support of $\boldsymbol{\beta}_0$. The only exception is the case when $\mu_0 = -\log n$ and β_{0i} is of lower magnitude as compared to μ_0, especially for the case when $\beta_{0i} = \log\log n$. In that case, the model selection results are worse than the other cases when the magnitude of μ_0 is smaller than that of β_{0i}. This is partly because the smaller the local parameter β_{0i} is, the harder it is to distinguish it from noise. When μ_0 is less negative, i.e., when the network is globally denser, BIC has better model selection results at all levels of β_{0i}. Figure 3 also shows that, given μ_0, the larger the magnitude of β_{0i} and the larger heterogeneity (more neighbors for nonzero β's) are, the better model selection results are. Figure 4 shows that the the estimation accuracy of $\hat{\boldsymbol{\beta}}(\hat{s})$ generally improves as n increases while it worsens as s_0 increases, reflecting the difficulty of estimating more parameters. Additional simulation results concerning the difference between $\hat{s}$ and s_0, and the estimation accuracy of $\hat{\mu}(\hat{s})$, found in Appendix C, support our findings.

From these simulation results, we may conclude that our estimation procedure works well for a wide variety of networks with varying degrees of local and global sparsity.

5 Data Analysis

In this section, we analyze the microfinance take-up example in Banerjee et al. (2013) to illustrate the usefulness of our model and estimation procedure. Banerjee et al. (2013) investigated the role of social networks, especially the role of those pre-identified as "leaders" (e.g., teachers, shopkeepers, savings group leaders), on households microfinancing decisions, and modelled the microfinancing decisions using a logit model.

Data. In 2006, data were collected for 75 rural villages in Karnataka, a state in southern India. A census of households was conducted, and a subset of individuals was asked detailed questions about the relationships they had with others in the village. This information was used to create network graphs for each village.

The social network data were collected along 12 dimensions in terms of whether individuals borrowed money from, gave advice to, helped with a decision, borrowed kerosene or rice from, lent kerosene or rice to, lent money to, obtained medical advice from, engaged socially with,

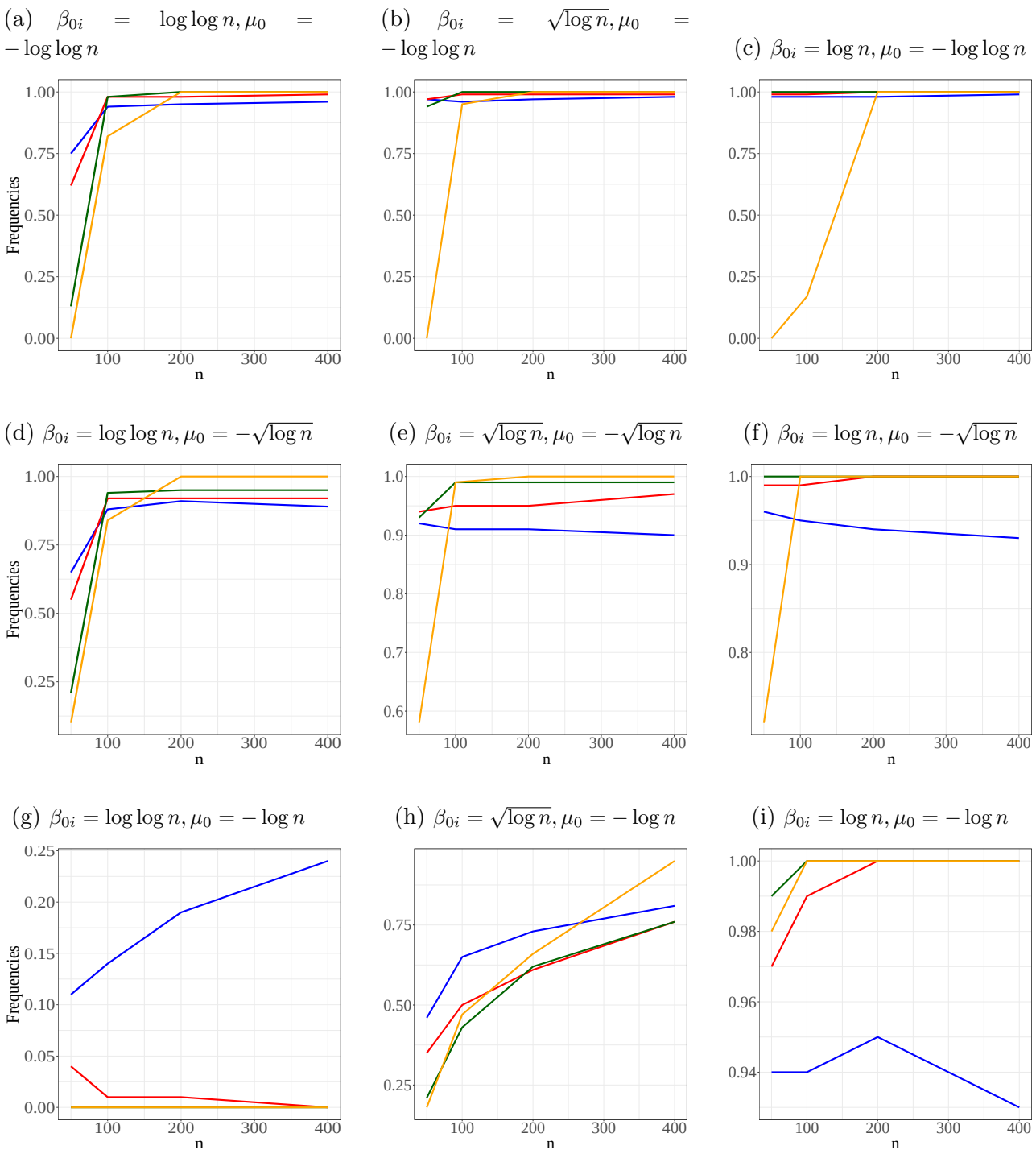

Figure 3: Simulation results on frequencies of the true support selected by BIC. ■ $s_0 = 2$, ■ $s_0 = \lfloor\sqrt{n/2}\rfloor$, ■ $s_0 = \lfloor\sqrt{n}\rfloor$, ■ $s_0 = \lfloor 2\sqrt{n}\rfloor$.

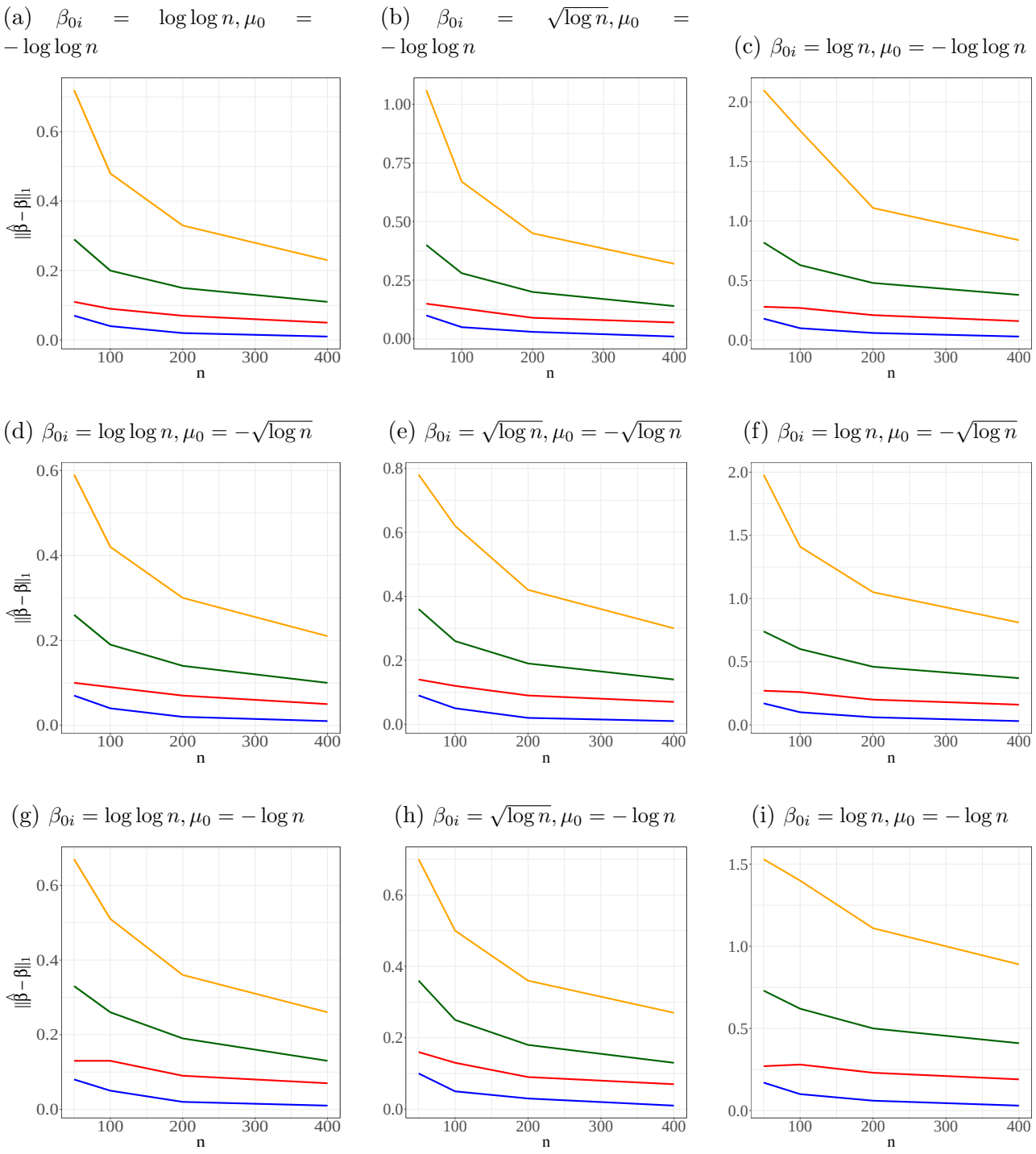

Figure 4: Simulation results on the ℓ_1-norm of $\hat{\boldsymbol{\beta}}(\hat{s}) - \boldsymbol{\beta}_0$ with $\hat{s}$ selected by BIC. $s_0 = 2$, $s_0 = \lfloor\sqrt{n/2}\rfloor$, $s_0 = \lfloor\sqrt{n}\rfloor$, $s_0 = \lfloor 2\sqrt{n}\rfloor$.

were related to, went to temple with, invited to one's home, or visited another's home. A relationship between households exists if any household member indicated a relationship with members from the other household. It should be noted that no relation exists between any two households in different villages due to the nature of data collection.

In 2007, a microfinancing institution, Bharatha Swamukti Samsthe (BSS), began operations in these villages, and collected data on the households who participated in the microfinancing program. For the 43 villages that BSS entered, the total number of households is $n = 9598$; the average number of households for each village is 223 with a standard deviation of 56; and the average take-up rate for BSS is 18%, with a cross-village standard deviation of 8%.

For the data used in our paper, we considered the network in which two households are linked if and only if any of the 12 dimensions of social contact occurred between them. The adjacency matrix of this network is block diagonal as there exists no link between any two households in different villages.

Method. Following Banerjee et al. (2013), we study how social importance, in a network sense, will affect microfinance take-up decision. Building on the SβM, we identified "leaders" as those households whose β parameters are estimated as nonzero.

Since households in different villages are not connected, we allowed village-dependent μ parameters to capture the individual village effects in fitting the SβM. More precisely, for each village, we first fitted the SβM to the observed network of the village by choosing s via BIC. Having obtained the parameter estimates denoted as $\hat{\boldsymbol{\beta}}_m$ and $\hat{\mu}_m$ for village m, we used $\hat{\beta}^\star_{mi} = \hat{\beta}_{mi} + \hat{\mu}_m/2$ as a covariate for household i in village m to model the probability that this household participated in the microfinancing program

$$\text{Parti}_{mi} = \Lambda(c + \theta \cdot \hat{\beta}^\star_{mi}), \tag{13}$$

where Λ is the logistic function such that $\Lambda(a) = \log(a)/\log(1-a)$ for $a \in (0,1)$, and $c \in \mathbb{R}$ and $\theta \in \mathbb{R}$ are two unknown parameters. The role of these estimated β's will be referred to as β-centrality hereafter. As an alternative measure of leadership, we also examined the use of an indicator variable "Leader", defined as $\text{Leader}_{mi} := 1\{\hat{\beta}_{mi} > 0\}$. Below we suppress the dependence of $\hat{\boldsymbol{\beta}}$ and $\hat{\mu}$ on m for simplicity.

For comparison, degree centrality and eigenvector centrality, two widely used measures of the influence of a node, were also investigated. In the context of the data analysis, the degree centrality of household i is d_i, the number of links that this household has. This is a measure of how well-connected a household is in the network. In graph theory, eigenvector centrality is a recursively defined notion of importance by associating high scores to those nodes that are connected to high-scoring nodes. Mathematically, the eigenvector centrality of the ith household is the ith element of $\boldsymbol{x}$, where $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ is the nonnegative eigenvector associated with the largest eigenvalue of the adjacency matrix, normalized to have Euclidean norm n. Considering various combinations of these measures of influence, we examine the following models:

$$(1)\ \text{Parti}_i = \Lambda(c + \theta \cdot d_i); (2)\ \text{Parti}_i = \Lambda(c + \theta \cdot x_i); (3)\ \text{Parti}_i = \Lambda(c + \theta \cdot \hat{\beta}^\star_i);$$

$$(4)\ \text{Parti}_i = \Lambda(c + \theta \cdot (1\{\hat{\beta}_i > 0\} + \hat{\mu}/2));$$

$$(5)\ \text{Parti}_i = \Lambda(c + \theta_1 \cdot d_i + \theta_2 \cdot \hat{\beta}^\star_i); (6)\ \text{Parti}_i = \Lambda(c + \theta_1 \cdot d_i + \theta_2 \cdot (1\{\hat{\beta}_i > 0\} + \hat{\mu}/2));$$

$$(7)\ \text{Parti}_i = \Lambda(c + \theta_1 \cdot x_i + \theta_2 \cdot \hat{\beta}^\star_i); (8)\ \text{Parti}_i = \Lambda(c + \theta_1 \cdot x_i + \theta_2 \cdot (1\{\hat{\beta}_i > 0\} + \hat{\mu}/2));$$

where $c \in \mathbb{R}, \theta \in \mathbb{R}, \theta_1 \in \mathbb{R}$ and $\theta_2 \in \mathbb{R}$ are unknown parameters. Note that in these models, $\hat{\mu}$ can not be absorbed into c because it is a parameter dependent on the village m. In examining these models, we wanted to assess the effects of different centralities in models (1)–(4), and to compare their relative merits when competing with each other in models (5)–(8). Finally the parameters in models (1)–(8) were estimated via the method of maximum likelihood for a logistic regression model.

Results. Using BIC, the SβM gave a fit with an average 26% of the households having nonzero β parameter. To assess how the model fits the data graphically, in Figure 5, we plotted the empirical distribution of the degrees of the observed network (black solid points) and the degree distribution after fitting the SβM (red open dots). The latter was obtained by averaging the empirical degree distributions of 100 randomly generated networks from the SβM with the estimated parameters. For reference, we also included the degree distribution of the Erdős-Rényi model fit. It can be seen that the empirical degree distribution of the data in the upper tail follows roughly a straight line, suggesting that a power law may be appropriate. However, the huge discrepancy between the empirical distribution of the data degrees and the Erdős-Rényi model fit (black dash dotted line) implies that the Erdős-Rényi model does not fit the data. In contrast, the SβM fit tracks the empirical distribution of the data very closely in the upper tail, thus providing a much better fit to capture the heavy upper tail of the empirical degree distribution. Interestingly, it can also be seen that the SβM fit yields a Poisson curve for those points with small degrees and a different pattern for those with larger degrees. This pattern can be loosely understood as the result of assigning nonzero β parameters to the nodes with large degrees. In this sense, the SβM fit mimics a mixture of the Erdős-Rényi and β-models which echos our point made previously that the SβM interpolates these two. We remark further that the presence of many isolated nodes and many nodes with a small number of links prevented fitting the β-model to the network.

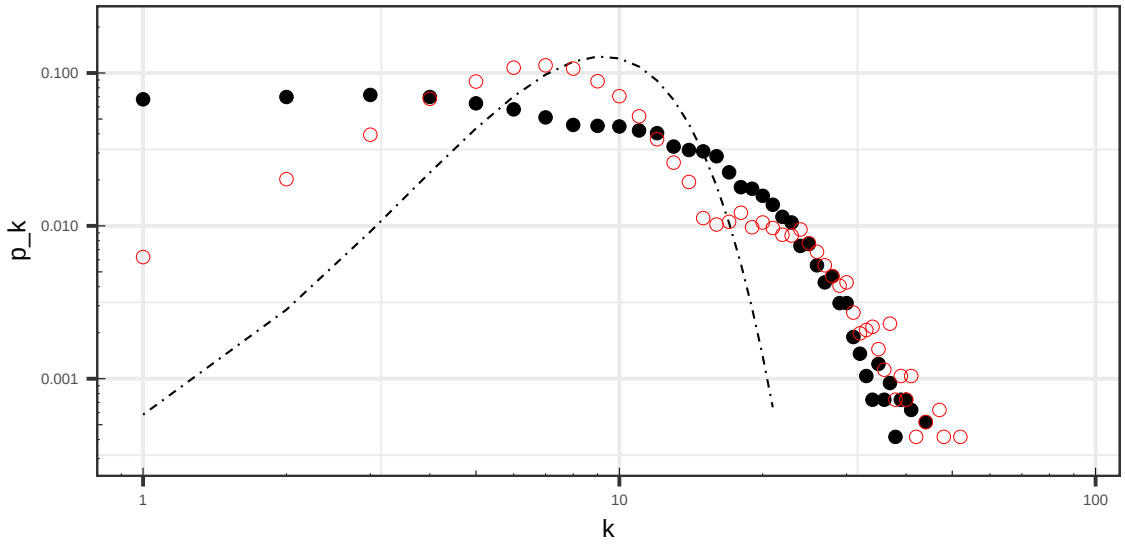

Figure 5: The black solid dots represent the degree distribution (frequency of degree denoted as p_k versus degree k) on the log-log scale using observed data from 43 villages. The red open dots correspond to the averaged empirical degree distributions of 100 randomly generated networks from the SβM with the estimated parameters. The fitted degree distribution assuming the frequencies follow a Poisson distribution is plotted as the black dash dotted line.

Table 1 provides the effects of degree centrality and eigenvector centrality on microfinance

take-up, along with the effects of β-centrality and how being a "Leader" can influence take-up. From the results on models (1)–(4), we can see that the effect on microfinance participation is much higher when β is larger (or when the node is identified as "Leader") for the SβM. On the other hand, although the effect of degree centrality is also statistically significant, the magnitude is much smaller when compared with eigenvector centrality or β-centrality or being a "Leader".

Table 1: Effect of Different Network Statistics on Take-Up. Standard errors are in parentheses.

	Dependent variable: take-up							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Degree	0.010^{***}				-0.001	-0.004		
	(0.003)				(0.005)	(0.005)		
Eigenvector		0.575^{***}					0.442^{**}	0.239
		(0.131)					(0.193)	(0.178)
Beta			0.198^{***}		0.212^{**}		0.071	
			(0.052)		(0.084)		(0.076)	
Leader				0.316^{***}		0.366^{***}		0.239^{***}
				(0.063)		(0.088)		(0.085)

Note: $^{*}p < 0.1;^{**}p < 0.05;^{***}p < 0.01$

Table 1 also provides effects of degree centrality or eigenvector centrality on microfinance take-up when controlling for β-centrality or being a "Leader". The regression results show that, after controlling for β-centrality or being a "Leader", the effects of degree centrality or eigenvector centrality are smaller, with the effect of degree centrality also becoming not statistically significant. Overall, we find the magnitude of β is significantly related to eventual microfinance participation. In particular, whether the household plays a leader role in the village is even more significantly related to eventual microfinance participation. Additional results can be found in Appendix D where a probit link and an identity link in (13) were used. All the additional results are consistent with the conclusions made from Table 1.

In our analysis, we did not distinguish causal or correlational effects in social networks. We note that other factors such as exogenous variation in the injection points could be useful for causal effects analysis. As the main objective of the current analysis is to provide insights on the role of social importance on program participation through the use of the SβM by defining new centrality measures such as β-centrality, we leave further investigation on dissecting causal and correlational effects or results from a structural economics model to future study.

6 Conclusion

We have proposed the Sparse β-Model (SβM) as a new generative model that can explicitly characterize global and local sparsity. We have shown that conventional asymptotic results including consistency and asymptotic normality results of its MLE are readily available for a wide variety of networks that are dense or sparse, when the support of the parameter is

known. When it is unknown, we have developed an ℓ_0-norm penalized likelihood approach for estimating the parameters and their support. We overcome the seemingly combinatorial nature of the optimization algorithm for computing the penalized estimator by fitting at maximum $n-1$ nested models with their support read from the degree sequence, thanks to a novel monotonicity lemma used to develop the solution path. A sufficient condition on the signal strength which is referred to as the β-min condition guarantees that, with high probability, the SβM chooses the correct model along its solution path. Therefore, the SβM represents a new class of models that are computationally fast, theoretically tractable, and intuitively attractive.

This paper has focused on the undirected simple graphs. It will be interesting to study the SβM for directed networks where, for each node, incoming and outgoing parameters are used for capturing the directional effect (Holland & Leinhardt 1981). We can adopt a similar strategy assuming that these parameters are sparse possibly after a reparametrization as developed for the SβM. It is not difficult to see, however, that the monotonicity lemma no longer holds. Therefore, when the support of these parameters is unknown, the ℓ_0-penalty based estimator is no longer computationally feasible. In view of this, we may develop ℓ_1-norm penalized likelihood estimation, immediately connecting this methodology to the vast literature on penalized likelihood methods for binary regression. Another future direction for research is to include covariate information at the nodal or link level. Progress has been made in this vein by extending the β-model (Graham 2017) and its generalization to directed networks (Yan et al. 2019). At the moment, however, these generalizations only work for relatively dense networks. The methodology proposed in this paper can be studied in this wider context. We note again that, where the support of the parameters is unknown, the ℓ_0-penalty based estimation is no longer feasible. We recommend using an ℓ_1-norm based penalization method. The results for these future directions will be reported elsewhere.

References

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

Acemoglu, D., Carvalho, V. M., Ozdaglar, A. & Tahbaz-Salehi, A. (2012), 'The network origins of aggregate fluctuations', *Econometrica* **80**, 1977–2016.

Banerjee, A., Chandrasekhar, A. G., Duflo, E. & Jackson, M. O. (2013), 'The diffusion of microfinance', *Science* **341**, 1236498.

Barabási, A. (2016), *Network Science*, Cambridge University Press.

Barabási, A. L. & Bonabau, E. (2003), 'Scale-free networks', *Scientific American* **50**, 50–59.

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

Bollobás, B., Janson, S. & Riordan, O. (2007), 'The phase transition in inhomogeneous random graphs', *Random Structures and Algorithms* **31**, 3–122.

Bollobás, B. & Riordan, O. (2011), 'Sparse graphs: Metrics and random models', *Random Structures and Algorithms* **39**, 1–38.

Boucheron, S., Lugosi, G. & Massart, P. (2013), *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press.

Britton, T., Deijfen, M. & Martin-Löf, A. (2006), 'Generating simple random graphs with prescribed degree distribution', *Journal of Statistical Physics* **124**, 1377–1397.

Caron, F. & Fox, E. (2017), 'Sparse graphs using exchangeable random measures (with discussion)', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 1295–1366.

Chatterjee, S., Diaconis, P. & Sly, A. (2011), 'Random graphs with a given degree sequence', *Annals of Applied Probability* **21**, 1400–1435.

Clauset, A., Shalizi, C. R. & Newman, M. E. (2009), 'Power-law distributions in empirical data', *SIAM review* **51**(4), 661–703.

De Paula, A. (2017), Econometrics of network models, *in* 'Advances in Economics and Econometrics: Theory and Applications, Eleventh World Congress', Cambridge University Press, pp. 268–323.

Erdős, P. & Rényi, A. (1959), 'On random graphs I', *Publ. Math. Debrecen* **6**, 290–297.

Erdős, P. & Rényi, A. (1960), 'On the evolution of random graphs', *Publ. Math. Inst. Hung. Acad. Sci* **5**, 17–60.

Fienberg, S. E. (2012), 'A brief history of statistical models for network analysis and open challenges.', *Journal of Computational and Graphical Statistics* **21**, 825–839.

Gilbert, E. G. (1959), 'Random graphs', *Annals of Mathematical Statistics* **30**, 1141–1144.

Goldenberg, A., Zheng, A. X., Feinberg, S. E. & Airoldi, E. M. (2009), 'A survey of statistical network models', *Foundations and Trends in Machine Learning* **2**, 129–233.

Graham, B. S. (2017), 'An econometric model of network formation with degree heterogeneity', *Econometrica* **85**, 1033–1063.

Greenshtein, E. & Ritov, Y. (2004), 'Persistence in high-dimensional linear predictor selection and the virtue of overparametrization', *Bernoulli* **10**, 971–988.

Hahn, J. & Newey, W. K. (2004), 'Jackknife and analytical bias reduction for nonlinear panel models', *Econometrica* **72**, 1295–1319.

Holland, P. W., Laskey, K. & Leinhardt, S. (1983), 'Stochastic blockmodels:First steps', *Social Networks* **5**, 109–137.

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

Jackson, M. O. (2010), *Social and economic networks*, Princeton university press.

Karwa, V. & Slavković, A. (2016), 'Inference using noisy degrees: Differentially private β-model and synthetic graphs', *Annals of Statistics* **44**, 87–112.

Kolaczyk, E. D. (2009), *Statistical Analysis of Network Data: Methods and Models*, Springer.

Kolaczyk, E. D. (2017), *Topics at the Frontier of Statistics and Network Analysis: (Re)Visiting the Foundations*, Cambridge University Press.

Koltchinskii, V. (2011), *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. École d'été de probabilités de Saint-Flour XXXVIII-2008*, Springer.

Krivitsky, P. N., Handcock, M. S. & Morris, M. (2009), 'Adjusting for network size and composition effects in exponential-family random graph models', *Stat. Methodol.* **8**, 319–339.

Krivitsky, P. N. & Kolaczyk, E. D. (2015), 'On the question of effective sample size in network modeling: An asymptotic inquiry', *Statistical Science* **30**, 184–198.

Li, H., Lindsay, B. G. & Waterman, R. P. (2003), 'Efficiency of projected score methods in rectangular array asymptotics', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**, 191–208.

Mukherjee, R., Mukherjee, S. & Sen, S. (2019), 'Detection thresholds for the β-model on sparse graphs', *Annals of Statistics* **46**, 1288–1317.

Newman, M. (2018), *Networks (2nd Edition)*, Oxford University Press.

Neyman, J. & Scott, E. L. (1948), 'Consistent estimates based on partially consistent observations', *Econometrica* **16**, 1–32.

Rinaldo, A., Petrović, S. & Fienberg, S. E. (2013), 'Maximum likelihood estimation in the β-model', *Annals of Statistics* **41**, 1085–1110.

Robins, G., Pattison, P., Kalish, Y. & Lusher, D. (2007), 'An introduction to exponential random graph models for social networks', *Social Networks* **29**, 173–191.

van der Hofstad, R. (2016), *Random Graphs and Complex Networks*, Cambridge University Press.

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge University Press.

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

Yan, T., Jiang, B., Fienberg, S. E. & Leng, C. (2019), 'Statistical inference in a directed network model with covariates', *Journal of the American Statistical Association* **114**, 857–868.

Yan, T., Leng, C. & Zhu, J. (2016), 'Asymptotics in directed exponential random graph models with an increasing bi-degree sequence', *Annals of Statistics* **44**, 31–57.

Yan, T. & Xu, J. (2013), 'A central limit theorem in the β-model for undirected random graphs with a diverging number of vertices', *Biometrika* **100**, 519–524.

Appendices

The supplementary material contains the main proofs and additional numerical results.

A Proofs

A.1 Proofs for Section 2

Proof of Proposition 1. Since $A_{ij}, 1 \leq i < j \leq n$ are i.i.d. Bernoulli random variables with

$$E[A_{ij}] = p \quad \text{and} \quad \text{Var}(A_{ij}) = p(1-p),$$

we have

$$E\left[\sum_{i<j} A_{ij}\right] = \binom{n}{2}p \quad \text{and} \quad \text{Var}\left(\sum_{i<j} A_{ij}\right) = \binom{n}{2}p(1-p) \coloneqq s_n^2.$$

We will prove that

$$\frac{\sum_{i<j} A_{ij} - \binom{n}{2}p}{\sqrt{\binom{n}{2}p(1-p)}} \xrightarrow{d} N(0,1) \tag{14}$$

whenever $n^2p(1-p) \to \infty$. To this end, it suffices to verify following Lindeberg condition:

$$\forall \epsilon > 0, \quad \frac{1}{s_n^2}\sum_{i<j} E[B_{ij}^2 I(|B_{ij}| \geq \epsilon s_n)] \to 0,$$

where $B_{ij} = A_{ij} - E[A_{ij}]$. Since $|B_{ij}| \leq 1$, the left hand side will be zero whenever $\epsilon s_n > 1$, which is immediate as $n^2p(1-p) \to \infty$ implies that $s_n \to \infty$. We can rewrite the left hand side of (14) as

$$n^{\gamma/2}\sqrt{\binom{n}{2}}\frac{\sum_{i<j} A_{ij}/\binom{n}{2} - p}{\sqrt{n^\gamma p(1-p)}} = n^{\gamma/2}\sqrt{\binom{n}{2}}\frac{\hat{p}-p}{\sqrt{n^\gamma p(1-p)}}.$$

So we conclude that

$$n^{1+\gamma/2}(\hat{p}-p) \xrightarrow{d} \begin{cases} N(0, 2p^\dagger(1-p^\dagger)) & \text{if } \gamma = 0 \\ N\left(0, 2p^\dagger\right) & \text{if } \gamma \in (0,2) \end{cases}.$$

This completes the proof. □

Proof Corollary 1. The result is a simple application of the delta method when $\gamma = 0$, and so we focus on the case where $\gamma \in (0,2)$. Observe that $\hat{\mu} = \log[n^{-\gamma}n^\gamma\hat{p}/(1-\hat{p})] = -\gamma\log n + \log n^\gamma\hat{p} - \log(1-\hat{p})$. We have

$$\hat{\mu} - \mu = (\log n^\gamma\hat{p} - \log n^\gamma p) - (\log(1-\hat{p}) - \log(1-p)).$$

By Proposition 1, we have $\sqrt{n^{2-\gamma}}(n^\gamma\hat{p} - n^\gamma p) \xrightarrow{d} N(0, 2p^\dagger)$, so that by the delta method (or the

Taylor expansion) we have

$$\sqrt{n^{2-\gamma}}(\log n^\gamma \hat{p} - \log n^\gamma p) = \frac{1}{p^\dagger}\sqrt{n^{2-\gamma}}(n^\gamma \hat{p} - n^\gamma p) + o(1) \xrightarrow{d} N(0, 2/p^\dagger) = N(0, 2e^{-\mu^\dagger}).$$

Likewise, we have

$$\log(1-\hat{p}) - \log(1-p) = (-1 + o_P(1))(\hat{p} - p) = o(n^{-1+\gamma/2}).$$

Conclude that $\sqrt{n^{2-\gamma}}(\hat{\mu} - \mu) \xrightarrow{d} N(0, 2e^{-\mu^\dagger})$. □

Proof of Proposition 2. The proposition directly follows from Theorem 3 in Appendix B and the discussion following the theorem. □

A.2 Proof of Theorem 1

The proof of Theorem 1 uses Bernstein's inequality. We state Bernstein's inequality for the reader's convenience. See Boucheron et al. (2013) Theorem 2.10.

Lemma 3 (Bernsten's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables with mean zero such that $|X_i| \leq b$ a.s. for all $i = 1, \ldots, n$. Then*

$$P\left(\left|\sum_{i=1}^n X_i\right| \geq \sqrt{2t\sum_{i=1}^n E[X_i^2] + bt/3}\right) \leq 2e^{-t}$$

for every $t > 0$.

Proof of Theorem 1. In this proof, we focus on the case where $\alpha < \gamma$. The proofs for the other cases are analogous. In addition, to simplify the notation, below we use s in place of s_0 as the cardinality of $S = S(\boldsymbol{\beta}_0)$. We may assume without loss of generality that $S = \{1, \ldots, s\}$. Then the likelihood function for $(\mu, \boldsymbol{\beta}_S)$ is

$$\prod_{1\leq i<j\leq s}\left(\frac{e^{\mu+\beta_i+\beta_j}}{1+e^{\mu+\beta_i+\beta_j}}\right)^{A_{ij}}\left(\frac{1}{1+e^{\mu+\beta_i+\beta_j}}\right)^{1-A_{ij}}$$
$$\times \prod_{\substack{1\leq i\leq s\\ s+1\leq j\leq n}}\left(\frac{e^{\mu+\beta_i}}{1+e^{\mu+\beta_i}}\right)^{A_{ij}}\left(\frac{1}{1+e^{\mu+\beta_i}}\right)^{1-A_{ij}} \times \prod_{s+1\leq i<j\leq n}\left(\frac{e^{\mu}}{1+e^{\mu}}\right)^{A_{ij}}\left(\frac{1}{1+e^{\mu}}\right)^{1-A_{ij}}.$$

The negative log-likelihood for $(\mu, \boldsymbol{\beta}_S)$ is

$$\begin{aligned}\ell_n(\mu, \boldsymbol{\beta}_S) &= -\mu \underbrace{\sum_{1\le i<j\le n} A_{ij}}_{=d_+} - \underbrace{\sum_{1\le i<j\le s} (\beta_i+\beta_j)A_{ij}}_{=\sum_{i=1}^s \beta_i \sum_{\substack{1\le j\le s\\ j\neq i}} A_{ij}} - \sum_{i=1}^s \beta_i \sum_{j=s+1}^n A_{ij} + \binom{n-s}{2}\log(1+e^{\mu}) \\ &\quad + (n-s)\sum_{i=1}^s \log(1+e^{\mu+\beta_i}) + \sum_{1\le i<j\le s}\log(1+e^{\mu+\beta_i+\beta_j}) \\ &= -\mu d_+ - \sum_{i=1}^s \beta_i d_i + \binom{n-s}{2}\log(1+e^{\mu}) + (n-s)\sum_{i=1}^s \log(1+e^{\mu+\beta_i}) \\ &\quad + \sum_{1\le i<j\le s}\log(1+e^{\mu+\beta_i+\beta_j}).\end{aligned}$$

Recall the reparameterization $\mu = -\gamma \log n + \mu^\dagger$ and $\beta_i = \alpha \log n + \beta_i^\dagger$.

Part (i). We first prove the uniform consistency of the MLE $(\hat{\mu}^\dagger, \hat{\boldsymbol{\beta}}_S^\dagger)$ in the sense that $\hat{\mu}^\dagger \xrightarrow{P} \mu_0^\dagger$ and $\max_{1\le i\le s}|\hat{\beta}_i^\dagger - \beta_{0i}^\dagger| \xrightarrow{P} 0$. Consider the concentrated negative log-likelihood for $\mu^\dagger$:

$$\begin{aligned}\ell_n^c(\mu^\dagger) &= -\mu^\dagger d_+ + \binom{n-s}{2}\log(1+n^{-\gamma}e^{\mu^\dagger}) + (n-s)\sum_{i=1}^s \log(1+n^{-(\gamma-\alpha)}e^{\mu^\dagger+\hat{\beta}_i^\dagger}) \\ &\quad + \sum_{1\le i<j\le s}\log(1+n^{-(\gamma-2\alpha)}e^{\mu^\dagger+\hat{\beta}_i^\dagger+\hat{\beta}_j^\dagger}),\end{aligned}$$

which is minimized at $\mu^\dagger = \hat{\mu}^\dagger$ on $[-M_1, M_1]$. Since $\hat{\beta}_i \in [0, M_2]$ for $i \in S$, we see that

$$\sup_{|\mu^\dagger|\le M_1}\left|\ell_n^c(\mu^\dagger) - \left(-\mu^\dagger d_+ + \binom{n-s}{2}\log(1+n^{-\gamma}e^{\mu^\dagger})\right)\right| \le O(sn^{1-(\gamma-\alpha)} + s^2 n^{-(\gamma-2\alpha)}) = o(n^{2-\gamma}).$$

In addition, we have

$$E[d_+] = \sum_{1\le i<j\le n} p_{ij} = \binom{n-s}{2}\frac{n^{-\gamma}e^{\mu_0^\dagger}}{1+n^{-\gamma}e^{\mu_0^\dagger}} + O(sn^{1-(\gamma-\alpha)} + s^2 n^{-(\gamma-2\alpha)}) = (n^{2-\gamma}/2)e^{\mu_0^\dagger} + o(n^{2-\gamma}) \quad \text{and}$$
$$\mathrm{Var}(d_+) = \sum_{i<j} p_{ij}(1-p_{ij}) = (n^{2-\gamma}/2)e^{\mu_0^\dagger} + o(n^{2-\gamma}),$$

so that $d_+ = (n^{2-\gamma}/2)e^{\mu_0^\dagger} + o_P(n^{2-\gamma})$. Conclude that

$$2n^{-2+\gamma}\ell_n^c(\mu^\dagger) \xrightarrow{P} -\mu^\dagger e^{\mu_0^\dagger} + e^{\mu^\dagger}$$

uniformly in $|\mu^\dagger| \le M_1$. The right hand side is uniquely minimized at $\mu^\dagger = \mu_0^\dagger$, and hence we have $\hat{\mu}^\dagger \xrightarrow{P} \mu_0^\dagger$ by Theorem 5.7 in van der Vaart (1998).

Next, consider the concentrated negative log-likelihood for $\beta_i^\dagger$:

$$\ell_n^c(\beta_i^\dagger) = -\beta_i^\dagger d_i + (n-s)\log(1+n^{-(\gamma-\alpha)}e^{\hat{\mu}^\dagger+\beta_i^\dagger}) + \sum_{\substack{1\le j\le s\\ j\neq i}}\log(1+n^{-(\gamma-2\alpha)}e^{\hat{\mu}^\dagger+\beta_i^\dagger+\hat{\beta}_j^\dagger}),$$

which is minimized at $\beta_i^\dagger = \hat{\beta}_i^\dagger$ on $[0, M_2]$. The last term on the right hand side is $O(sn^{-(\gamma-2\alpha)}) = o(n^{1-(\gamma-\alpha)})$ uniformly in $|\beta_i^\dagger| \le M_2$ and $1 \le i \le s$. We note that

$$E[d_i] = \sum_{j \neq i} p_{ij} = \sum_{j>s} p_{ij} + \underbrace{\sum_{\substack{1 \le j \le s \\ j \neq i}} p_{ij}}_{=O(sn^{-(\gamma-2\alpha)})} = n^{1-(\gamma-\alpha)} e^{\mu_0^\dagger + \beta_{0i}^\dagger} + o(n^{1-(\gamma-\alpha)}),$$

$$\mathrm{Var}(d_i) = \sum_{j \neq i} p_{ij}(1-p_{ij}) = n^{1-(\gamma-\alpha)} e^{\mu_0^\dagger + \beta_{0i}^\dagger} + o(n^{1-(\gamma-\alpha)}),$$

where the o terms are uniform in $1 \le i \le s$. Since $d_i = \sum_{j \neq i} A_{ij}$ is the sum independent random variables with $|A_{ij} - E[A_{ij}]| \le 1$, applying Bernstein's inequality (Lemma 3) to d_i, we have

$$P\left(|d_i - E[d_i]| > \sqrt{2t\,\mathrm{Var}(d_i)} + t/3\right) \le 2e^{-t}$$

for every $t > 0$. Choosing $t = 2\log(2n)$ and using the union bound, we have

$$\max_{1 \le i \le s} |d_i - E[d_i]| \le 2\sqrt{\max_{1 \le j \le s} \mathrm{Var}(d_j)\log(2n)} + 2(\log(2n))/3$$

with probability approaching one. Using the preceding evaluation of $\mathrm{Var}(d_i)$, we have

$$\max_{1 \le i \le s} |d_i - E[d_i]| = O_P(n^{1/2-(\gamma-\alpha)/2}\sqrt{\log n}),$$

which implies that $d_i = n^{1-(\gamma-\alpha)} e^{\mu_0^\dagger + \beta_{0i}^\dagger} + o_P(n^{1-(\gamma-\alpha)})$ uniformly in $1 \le i \le s$. Together with the consistency of $\hat{\mu}^\dagger$, we have

$$n^{-1+(\gamma-\alpha)} \ell_n^c(\beta_i^\dagger) \xrightarrow{P} -\beta_i^\dagger e^{\mu_0^\dagger + \beta_{0i}^\dagger} + e^{\mu_0^\dagger + \beta_i^\dagger} = e^{\mu_0^\dagger}(-\beta_i^\dagger e^{\beta_{0i}^\dagger} + e^{\beta_i^\dagger})$$

uniformly in $\beta_i^\dagger \in [0, M_2]$ and $1 \le i \le s$. Pick any fixed $\delta > 0$. It is not difficult to show that

$$\epsilon := \min_{1 \le i \le s} \min_{|\beta_i^\dagger - \beta_{0i}^\dagger| > \delta} e^{\mu_0^\dagger}\{-(\beta_i^\dagger - \beta_{0i}^\dagger)e^{\beta_{0i}^\dagger} + e^{\beta_i^\dagger} - e^{\beta_{0i}^\dagger}\}$$

is bounded away from zero (indeed $\epsilon \ge e^{\mu_0^\dagger}\delta^2$). Now, if $|\hat{\beta}_i^\dagger - \beta_{0i}^\dagger| > \delta$ for some $1 \le i \le s$, then

$$\begin{aligned} & n^{-1+(\gamma-\alpha)} \ell_n^c(\hat{\beta}_i^\dagger) - n^{-1+(\gamma-\alpha)} \ell_n^c(\beta_{0i}^\dagger) \\ & \ge \epsilon - \underbrace{2 \max_{1 \le j \le s} \sup_{|\beta_j^\dagger| \le M_2} \left| n^{-1+(\gamma-\alpha)} \ell_n^c(\beta_j^\dagger) - e^{\mu_0^\dagger}(-\beta_j^\dagger e^{\beta_{0j}^\dagger} + e^{\beta_j^\dagger}) \right|}_{=o_P(1)}, \end{aligned}$$

but by the definition of the MLE, the left hand side is nonpositive. Conclude that

$$P\left(\max_{1 \le i \le s} |\hat{\beta}_i^\dagger - \beta_{0i}^\dagger| > \delta\right) \le P(o_P(1) \ge \epsilon) = o(1).$$

This implies that $\max_{1 \le i \le s} |\hat{\beta}_i^\dagger - \beta_{0i}^\dagger| = o_P(1)$.

Part (ii). Next, we will derive the limiting distribution of $(\hat{\mu}^\dagger - \mu_0^\dagger, \hat{\boldsymbol{\beta}}_F^\dagger - \boldsymbol{\beta}_{0F}^\dagger)$ for any fixed subset $F \subset S$. Since the true parameter vector $(\mu_0^\dagger, \boldsymbol{\beta}_{0S}^\dagger)$ is bounded away from the boundary of the parameter space, the MLE satisfies the first order condition with probability approaching

one by the uniform consistency. The first order condition is described as follows:

$$
\begin{aligned}
&-d_+ + \binom{n-s}{2}\frac{n^{-\gamma}e^{\mu^\dagger}}{1+n^{-\gamma}e^{\mu^\dagger}} + (n-s)\sum_{i=1}^{s}\frac{n^{-(\gamma-\alpha)}e^{\mu^\dagger+\beta_i^\dagger}}{1+n^{-(\gamma-\alpha)}e^{\mu^\dagger+\beta_i^\dagger}} + \sum_{1\le i<j\le s}\frac{n^{-(\gamma-2\alpha)}e^{\mu^\dagger+\beta_i^\dagger+\beta_j^\dagger}}{1+n^{-(\gamma-2\alpha)}e^{\mu^\dagger+\beta_i^\dagger+\beta_j^\dagger}} = 0,\\
&-d_i + (n-s)\frac{n^{-(\gamma-\alpha)}e^{\mu^\dagger+\beta_i^\dagger}}{1+n^{-(\gamma-\alpha)}e^{\mu^\dagger+\beta_i^\dagger}} + \sum_{\substack{1\le j\le s\\ j\neq i}}\frac{n^{-(\gamma-2\alpha)}e^{\mu^\dagger+\beta_i^\dagger+\beta_j^\dagger}}{1+n^{-(\gamma-2\alpha)}e^{\mu^\dagger+\beta_i^\dagger+\beta_j^\dagger}} = 0,\ i\in S.
\end{aligned}
\tag{15}
$$

The left hand sides are $-d_+ + E[d_+]$ and $-\boldsymbol{d}_S + E[\boldsymbol{d}_S]$ at $(\mu^\dagger, \boldsymbol{\beta}_S^\dagger) = (\mu_0^\dagger, \boldsymbol{\beta}_{0S}^\dagger)$, where $\boldsymbol{d}_S = (d_1, \ldots, d_s)^T$. We will derive the joint limiting distribution for $(d_+ - E[d_+], \boldsymbol{d}_F - E[\boldsymbol{d}_F])$. Decompose d_+ as

$$
d_+ = \sum_{s<i<j\le n} A_{ij} + \sum_{\substack{1\le i\le s\\ s<j\le n}} A_{ij} + \sum_{1\le i<j\le s} A_{ij}.
$$

The variance of the first term on the right hand side is $(n^{2-\gamma}/2)e^{\mu_0^\dagger} + o(n^{2-\gamma})$, while the variances of the last two terms are $o(n^{2-\gamma})$. Hence we have

$$
d_+ - E[d_+] = \sum_{s<i<j\le n}(A_{ij} - p_{ij}) + o_P(n^{1-\gamma/2}) \quad \text{and} \quad n^{-1+\gamma/2}\sum_{s<i<j\le n}(A_{ij}-p_{ij}) \xrightarrow{d} N(0, e^{\mu_0^\dagger}/2).
$$

On the other hand, for $i, j \in S$, we have

$$
\begin{aligned}
\mathrm{Var}(d_i) &= \sum_{k\neq i} p_{ik}(1-p_{ik}) = n^{1-(\gamma-\alpha)}e^{\mu_0^\dagger+\beta_{0i}^\dagger} + o(n^{1-(\gamma-\alpha)}),\\
\mathrm{Cov}(d_i, d_j) &= p_{ij}(1-p_{ij}) = o(n^{1-(\gamma-\alpha)}),
\end{aligned}
$$

so that we have

$$
n^{-1/2+(\gamma-\alpha)/2}(\boldsymbol{d}_F - E[\boldsymbol{d}_F]) \xrightarrow{d} N(\boldsymbol{0}, \Lambda_F),
$$

where $\Lambda_F = \mathrm{diag}\{e^{\mu_0^\dagger+\beta_{0i}^\dagger} : i \in F\}$. Since $\sum_{s+1\le i<j\le n} A_{ij}$ and $\boldsymbol{d}_F$ are independent, we conclude that

$$
\begin{pmatrix} n^{-1+\gamma/2}(d_+ - E[d_+]) \\ n^{-1/2+(\gamma-\alpha)/2}(\boldsymbol{d}_F - E[\boldsymbol{d}_F]) \end{pmatrix} \xrightarrow{d} N\left(\boldsymbol{0}, \begin{pmatrix} e^{\mu_0^\dagger}/2 & \boldsymbol{0} \\ \boldsymbol{0} & \Lambda_F \end{pmatrix}\right). \tag{16}
$$

Let $\hat{\delta} = \max_{1\le i\le s}|\hat{\beta}_i^\dagger - \beta_{0i}^\dagger|$. Applying the Taylor expansion to the first equation in (15), we have

$$
-d_+ + E[d_+] + (n^{2-\gamma}/2)(e^{\mu_0^\dagger} + o_P(1))(\hat{\mu}^\dagger - \mu_0^\dagger) + O_P(sn^{1-(\gamma-\alpha)}\hat{\delta}) = 0. \tag{17}
$$

In particular, this implies that

$$
\hat{\mu}^\dagger - \mu_0^\dagger = O_P(n^{-1+\gamma/2}) + O_P(sn^{-1+\alpha}\hat{\delta}).
$$

Likewise, applying the Taylor expansion to the second equation in (15), we have

$$
-d_i + E[d_i] + n^{1-(\gamma-\alpha)}(e^{\mu_0^\dagger+\beta_{0i}^\dagger} + o_P(1))\Big\{\hat{\beta}_i^\dagger - \beta_{0i}^\dagger + O_P(n^{-1+\gamma/2}) + \underbrace{O_P(sn^{-1+\alpha}\hat{\delta})}_{=o_P(\hat{\delta})}\Big\} = 0 \tag{18}
$$

uniformly in $1 \leq i \leq s$. Since $\max_{1\leq i\leq s} |d_i - E[d_i]| = O_P(n^{1/2-(\gamma-\alpha)/2}\sqrt{\log n})$, we have

$$\hat{\delta} = O_P(n^{-1/2+(\gamma-\alpha)/2}\sqrt{\log n}).$$

Plugging this evaluation into (18), we have

$$n^{1/2-(\gamma-\alpha)/2}(\hat{\beta}_i^\dagger - \beta_{0i}^\dagger) = n^{-1/2+(\gamma-\alpha)/2}e^{-\mu_0^\dagger-\beta_{0i}^\dagger}(d_i - E[d_i]) + \underbrace{O_P(sn^{-1+\alpha}\sqrt{\log n})}_{=o_P(1)} \tag{19}$$

uniformly in $1 \leq i \leq s$. Likewise, plugging the preceding evaluation of $\hat{\delta}$ into (17), we have

$$-d_+ + E[d_+] + (n^{2-\gamma}/2)(e^{\mu_0^\dagger} + o_P(1))(\hat{\mu}^\dagger - \mu_0^\dagger) + O_P(sn^{1/2-(\gamma-\alpha)/2}\sqrt{\log n}) = 0$$

and the last term on the left hand side is $o_P(n^{1-\gamma/2})$ under our assumption that $s = o(n^{(1-\alpha)/2}/\sqrt{\log n})$. Hence we have

$$n^{1-\gamma/2}(\hat{\mu}^\dagger - \mu_0^\dagger) = 2e^{-\mu_0^\dagger}n^{-1+\gamma/2}(d_+ - E[d_+]) + o_P(1). \tag{20}$$

The desired conclusion follows from combining the expansions (20) and (19) with (16). □

A.3 Proofs for Section 3.2

Proof of Lemma 1. In this proof, we omit the argument s and write $(\hat{\mu}(s), \hat{\boldsymbol{\beta}}(s)) = (\hat{\mu}, \hat{\boldsymbol{\beta}})$.

Part (i). Suppose on the contrary that there exist i and j such that $d_i < d_j$ but $\hat{\beta}_i > \hat{\beta}_j$. Define $\tilde{\boldsymbol{\beta}}$ by

$$\tilde{\beta}_k = \begin{cases} \hat{\beta}_k & \text{if } k \neq i, j \\ \hat{\beta}_j & \text{if } k = i \\ \hat{\beta}_i & \text{if } k = j \end{cases}.$$

Now, since $-(d_i\hat{\beta}_j + d_j\hat{\beta}_i) < -(d_i\hat{\beta}_i + d_j\hat{\beta}_j)$, we have $\ell_n(\hat{\mu}, \tilde{\boldsymbol{\beta}}) < \ell_n(\hat{\mu}, \hat{\boldsymbol{\beta}})$, which contradicts the fact that $(\hat{\mu}, \hat{\boldsymbol{\beta}})$ is an optimal solution to (3)

Part (ii). Suppose on the contrary that there exist i and j such that $d_i = d_j$ but $\hat{\beta}_i \neq \hat{\beta}_j$. Define $\tilde{\boldsymbol{\beta}}$ by

$$\tilde{\beta}_k = \begin{cases} \hat{\beta}_k & \text{if } k \neq i, j \\ (\hat{\beta}_i + \hat{\beta}_j)/2 & \text{if } k = i, j \end{cases}.$$

It is not difficult to see that $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^n_+$ and $\|\tilde{\boldsymbol{\beta}}\|_0 \leq s$ if $s = \sum_{k=1}^K s_k$ for some $K \leq m-1$. Since for any $k \neq i, j$,

$$2\log\left(1 + e^{\mu+(\beta_i+\beta_j)/2+\beta_k}\right) < \log\left(1 + e^{\mu+\beta_i+\beta_k}\right) + \log\left(1 + e^{\mu+\beta_j+\beta_k}\right),$$

we have $\ell_n(\hat{\mu}, \tilde{\boldsymbol{\beta}}) < \ell_n(\hat{\mu}, \hat{\boldsymbol{\beta}})$, which contradicts the fact that $(\hat{\mu}, \hat{\boldsymbol{\beta}})$ is an optimal solution to (3). □

The proof of Lemma 2 relies on Hoeffding's inequality; for the reader's convenience, we state it as the following lemma. For its proof, see, e.g., Theorem 2.8 in Boucheron et al. (2013).

Lemma 4 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables such that*

each X_i takes values in $[a_i, b_i]$ for some $-\infty < a_i < b_i < \infty$. Then

$$P\left(\sum_{i=1}^{n}(X_i - E[X_i]) > t\right) \leq \exp\left\{-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right\}$$

for every $t > 0$.

Proof of Lemma 2. For the sake of notational convenience, we use $(\mu, \boldsymbol{\beta})$ for $(\mu_0, \boldsymbol{\beta}_0)$. Recall that $i \in S$ and $j \in S^c$, i.e., $\beta_i \neq 0$ and $\beta_j = 0$. Observe that

$$d_i = \sum_{k \neq i} A_{ik} = \sum_{k \neq i,j} A_{ik} + A_{ij}, \quad d_j = \sum_{k \neq j} A_{jk} = \sum_{k \neq i,j} A_{jk} + A_{ij},$$

where

$$A_{ik} \sim Ber\left(\frac{e^{\mu+\beta_i+\beta_k}}{1+e^{\mu+\beta_i+\beta_k}}\right), \quad A_{jk} \sim Ber\left(\frac{e^{\mu+\beta_k}}{1+e^{\mu+\beta_k}}\right).$$

Then,

$$\begin{aligned} d_i - d_j &= \sum_{k \neq i,j}(A_{ik} - A_{jk}) = \sum_{k \neq i,j}(A_{ik} - E[A_{ik}]) - \sum_{k \neq i,j}(A_{jk} - E[A_{jk}]) + \sum_{k \neq i,j}(E[A_{ik}] - E[A_{jk}]) \\ &= \sum_{k \neq i,j}(A_{ik} - E[A_{ik}]) - \sum_{k \neq i,j}(A_{jk} - E[A_{jk}]) + \sum_{k \neq i,j}\left(\frac{e^{\mu+\beta_i+\beta_k}}{1+e^{\mu+\beta_i+\beta_k}} - \frac{e^{\mu+\beta_k}}{1+e^{\mu+\beta_k}}\right). \end{aligned}$$

Define

$$\varepsilon_{ij} = \min_{k \neq i,j}\left(\frac{e^{\mu+\beta_i+\beta_k}}{1+e^{\mu+\beta_i+\beta_k}} - \frac{e^{\mu+\beta_k}}{1+e^{\mu+\beta_k}}\right) > 0,$$

and observe that

$$d_i - d_j \geq \sum_{k \neq i,j}(A_{ik} - E[A_{ik}]) - \sum_{k \neq i,j}(A_{jk} - E[A_{jk}]) + (n-2)\varepsilon_{ij}.$$

Now, by Hoeffding's inequality (Lemma 4), for every $t > 0$,

$$P\left(\sum_{k \neq i,j}(A_{ik} - E[A_{ik}]) < -t\right) \leq e^{-2t^2/(n-2)},$$

and so with probability at least $1 - \tau/2$,

$$\sum_{k \neq i,j}(A_{ik} - E[A_{ik}]) \geq -\sqrt{\frac{(n-2)}{2}\log(2/\tau)}.$$

Likewise, with probability at least $1 - \tau/2$,

$$\sum_{k \neq i,j}(A_{jk} - E[A_{jk}]) \leq \sqrt{\frac{(n-2)}{2}\log(2/\tau)}.$$

Hence, with probability at least $1-\tau$,

$$d_i - d_j \geq (n-2)\Big\{ -\underbrace{\sqrt{\frac{2}{n-2}\log(2/\tau)}}_{=c_{n,\tau}} + \varepsilon_{ij} \Big\}.$$

Next, we establish a lower bound on ε_{ij}. Observe that

$$\frac{e^{\mu+\beta_i+\beta_k}}{1+e^{\mu+\beta_i+\beta_k}} - \frac{e^{\mu+\beta_k}}{1+e^{\mu+\beta_k}} = \frac{e^{\mu+\beta_k}}{1+e^{\mu+\beta_k}} \cdot \frac{e^{\beta_i}-1}{1+e^{\mu+\beta_i+\beta_k}} \geq \frac{e^{-\mu^-}}{1+e^{-\mu^-}} \cdot \frac{e^{\beta_i}-1}{1+e^{2\overline{\beta}+\mu^+}},$$

so that

$$\varepsilon_{ij} \geq \frac{1}{1+e^{\mu^-}} \cdot \frac{e^{\beta_i}-1}{1+e^{2\overline{\beta}+\mu^+}}.$$

The right hand side is larger than $c_{n,\tau}$ under Condition (6). This completes the proof. □

Proof of Theorem 2. For the sake of notational simplicity, we will write $(\hat{\mu}(s), \hat{\boldsymbol{\beta}}(s)) = (\hat{\mu}, \hat{\boldsymbol{\beta}})$. We begin with noting that

$$\begin{aligned}
\mathcal{E}_s &\leq \mathcal{R}(\hat{\mu}, \hat{\boldsymbol{\beta}}) - \inf_{(\mu,\boldsymbol{\beta})\in\Theta_s} D_+^{-1}\ell_n(\mu,\boldsymbol{\beta}) + \sup_{(\mu,\boldsymbol{\beta})\in\Theta_s} |D_+^{-1}\ell_n(\mu,\boldsymbol{\beta}) - \mathcal{R}(\mu,\boldsymbol{\beta})| \\
&= \mathcal{R}(\hat{\mu}, \hat{\boldsymbol{\beta}}) - D_+^{-1}\ell_n(\hat{\mu}, \hat{\boldsymbol{\beta}}) + \sup_{(\mu,\boldsymbol{\beta})\in\Theta_s} |D_+^{-1}\ell_n(\mu,\boldsymbol{\beta}) - \mathcal{R}(\mu,\boldsymbol{\beta})| \\
&\leq 2 \sup_{(\mu,\boldsymbol{\beta})\in\Theta_s} |D_+^{-1}\ell_n(\mu,\boldsymbol{\beta}) - \mathcal{R}(\mu,\boldsymbol{\beta})|.
\end{aligned} \tag{21}$$

Next, observe that

$$\begin{aligned}
|D_+^{-1}\ell_n(\mu,\boldsymbol{\beta}) - \mathcal{R}(\mu,\boldsymbol{\beta})| &\leq D_+^{-1}|\mu|\,|d_+ - E[d_+]| + D_+^{-1}\left|\sum_{i=1}^n \beta_i(d_i - E[d_i])\right| \\
&\leq D_+^{-1}\left(M_2\,|d_+ - E[d_+]| + M_1 s \max_{1\leq i\leq n}|d_i - E[d_i]|\right)
\end{aligned} \tag{22}$$

for $(\mu,\boldsymbol{\beta}) \in \Theta_s$ where we have used the fact that $\sum_{i=1}^n \beta_i \leq M_1 s$. Now, using Bernstein's inequality (Lemma 3) and the union bound, we have

$$\max_{1\leq i\leq n}|d_i - E[d_i]| \leq \sqrt{2\max_{1\leq j\leq n}\mathrm{Var}(d_j)\log(4n/\tau)} + (\log(4n/\tau))/3 \tag{23}$$

with probability at least $1-\tau/2$. Likewise, by Bernstein's inequality, we have

$$|d_+ - E[d_+]| \leq \sqrt{2\,\mathrm{Var}(d_+)\log(4/\tau)} + (\log(4/\tau))/3 \tag{24}$$

with probability at least $1-\tau/2$. Combining (21)–(24), we obtain the bound (8).

Finally, if $\mu_0 = -\gamma\log n + O(1)$ and $\beta_{0i} = \alpha\log n + O(1)$ for $i \in S(\boldsymbol{\beta}_0)$, then from the proof of Theorem 1, we know that $D_+ \sim n^{2-\gamma}$, $\mathrm{Var}(d_+) \sim n^{2-\gamma}$, and $\max_{1\leq i\leq n}\mathrm{Var}(d_i) \sim n^{1-(\gamma-\alpha)}$. This leads to the second bound (9). □

B Beta-model and power law

In this appendix, we restate the results of Britton et al. (2006) that show that the β-model can generate node degrees asymptotically following a power law and the empirical degree distribution converging in probability to the same power law if β_i are randomly generated in a suitable way. We first recall the definition of a mixed Poisson distribution.

Definition 1 (Mixed Poisson distribution)**.** Let F be a distribution function supported in $\mathbb{R}_+$. A random variable X taking values in the nonnegative integers follows the *mixed Poisson distribution* with mixing distribution F if

$$P(X = k) = \int_{[0,\infty)} e^{-w} \frac{w^k}{k!} dF(w), \quad k = 0, 1, 2, \ldots.$$

If $W \sim F$, then we also say that X follows the mixed Poisson distribution with parameter W.

The next lemma shows that the tail behavior of a mixed Poisson distribution is determined solely by that of the mixing distribution.

Lemma 5. *Let F be a distribution function supported in $\mathbb{R}_+$ such that $c_1 x^{1-\tau} \leq 1 - F(x) \leq c_2 x^{1-\tau}$ for large x for some $0 < c_1 < c_2 < \infty$. Then there exist $0 < c_1' < c_2' < \infty$ such that the distribution function G of a mixed Poisson distribution with mixing distribution F satisfies $c_1' x^{1-\tau} \leq 1 - G(x) \leq c_2' x^{1-\tau}$ for large x.*

Proof. See van der Hofstad (2016) Exercise 6.12. □

The following results are taken from Theorem 3.2 and Proposition 3.1 in Britton et al. (2006). In the following, the variables $d_1, \ldots, d_n$ are indeed a triangular sequence and hence should be indexed by n, but this is suppressed for the notational convenience.

Theorem 3 (β-model and mixed Poisson distribution)**.** *Let $\{W_i\}_{i=1}^{\infty}$ be i.i.d. positive random variables with $P(W_1 > w) \sim cw^{-\rho}$ as $w \to \infty$ for some $c > 0$ and $\rho \in (0, 1)$. For the β-model in (1), suppose that $\beta_1, \ldots, \beta_n$ are generated as $\beta_i = \log W_i - (\log n)/(2\rho)$ for $i = 1, \ldots, n$ for each $n = 1, 2, \ldots$. Then:*

(i) The limiting distribution of each node degree d_i as $n \to \infty$ is the mixed Poisson distribution with parameter ϱW_1^{ρ}, where $\varrho = c \int_0^{\infty} (1 + x)^{-2} x^{-\rho} dx$.

(ii) For $N_k := |\{i \in \{1, \ldots, n\} : d_i = k\}|$, we have $N_k/n \xrightarrow{P} P(\varrho W_1^{\rho} = k)$ as $n \to \infty$, where ϱ appears in (i).

Combined with Lemma 5, we know that

$$\begin{aligned} P(d_i \geq y) &\approx P(\varrho W_1^{\rho} \geq y) \\ &= P(W_1 \geq (y/\varrho)^{1/\rho}) \\ &\sim c\varrho y^{-1}, \end{aligned}$$

so that the limiting distribution of d_i is a power law with exponent $\tau = 2$ in the sense that

$$\lim_{n \to \infty} P(d_i = k) \sim k^{-2}, \ k \to \infty.$$

Likewise, the empirical degree distribution converges in probability to the same power law.

C Additional Simulations

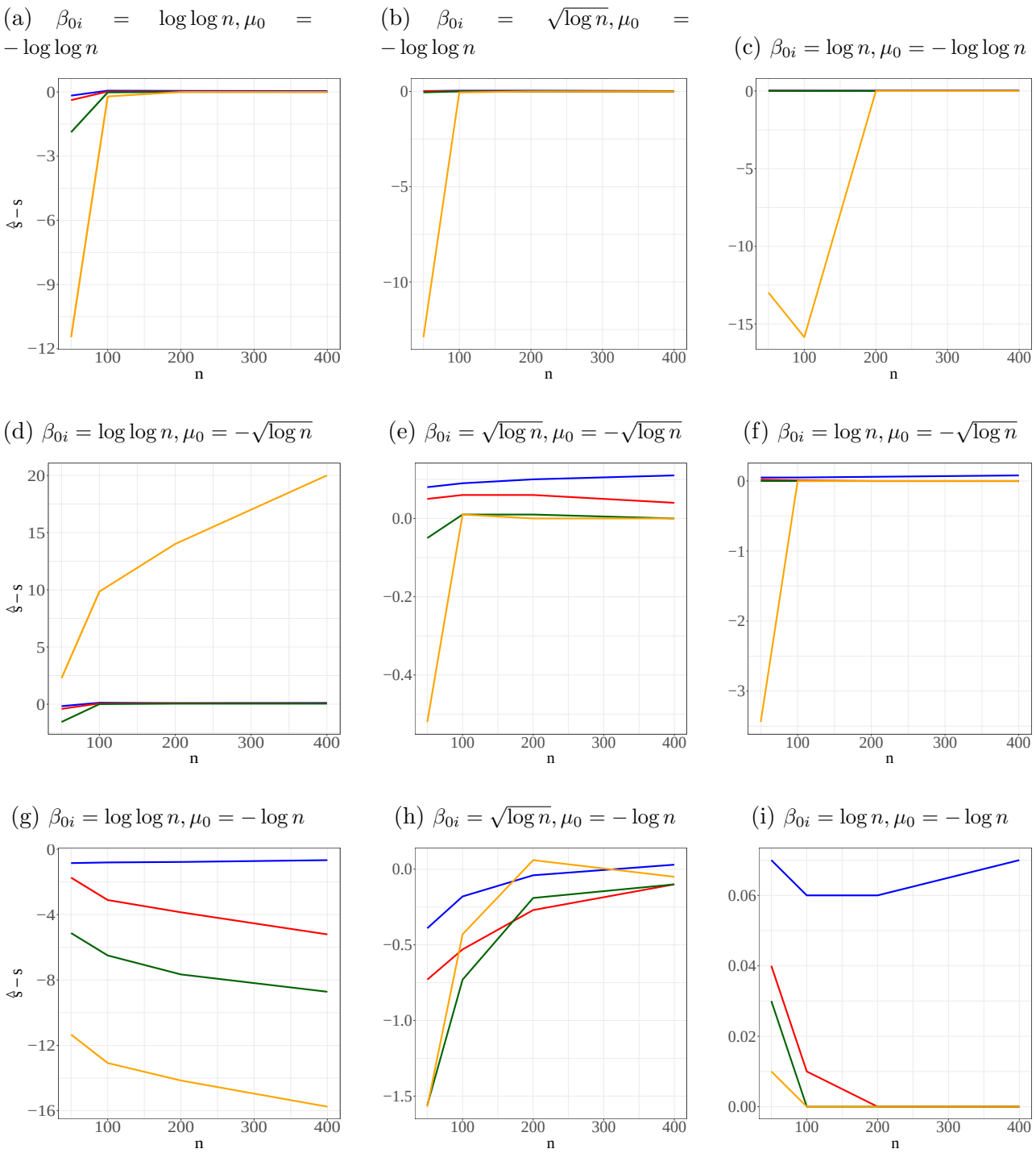

Figure 6: Simulation results on $\hat{s} - s_0$ with $\hat{s}$ selected by BIC. ■ $s_0 = 2$, ■ $s_0 = \lfloor\sqrt{n/2}\rfloor$, ■ $s_0 = \lfloor\sqrt{n}\rfloor$, ■ $s_0 = \lfloor 2\sqrt{n}\rfloor$.

D Additional Data Analysis Results

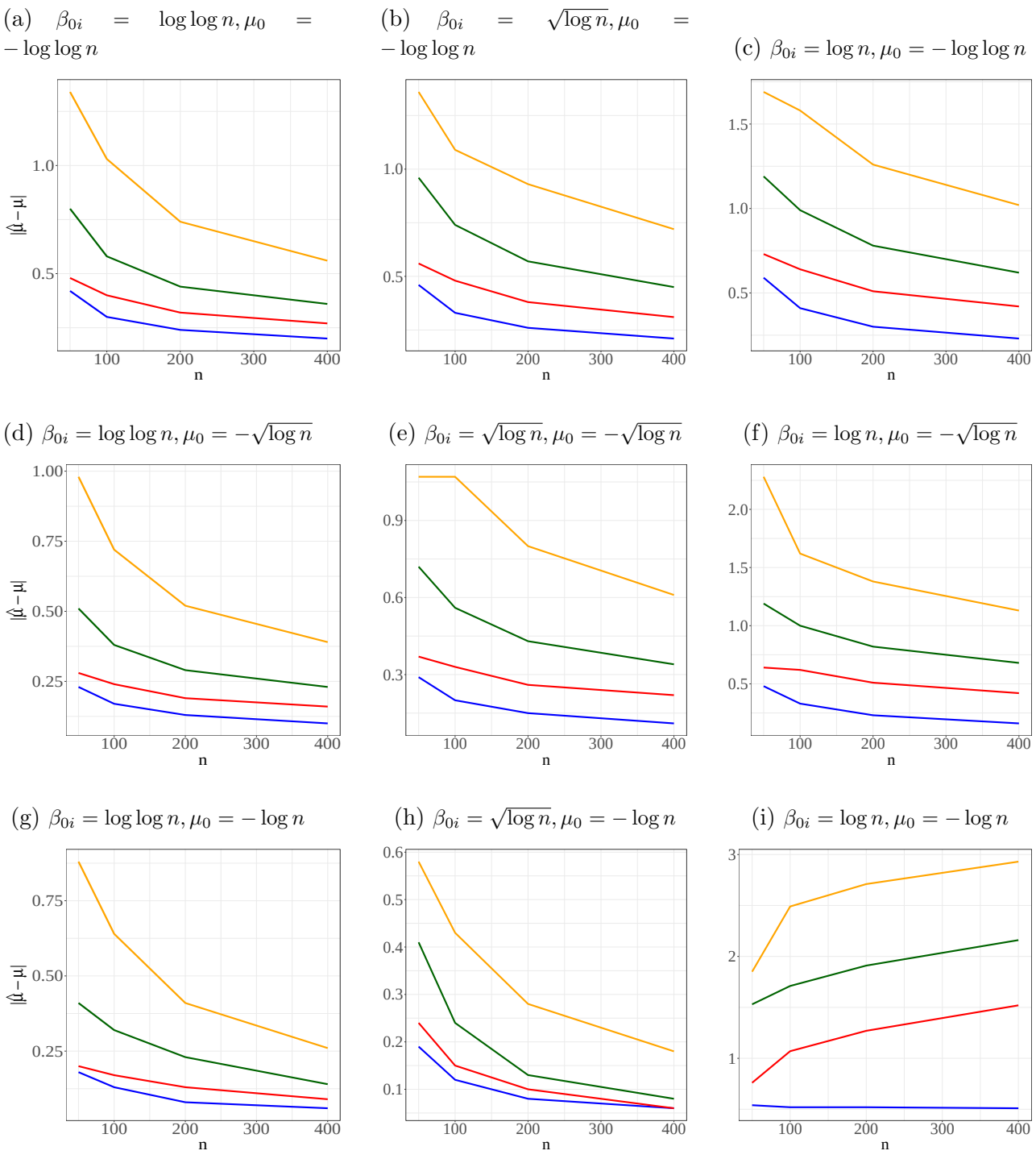

Figure 7: Simulation results on $|\hat{\mu}(\hat{s}) - \mu_0|$ with $\hat{s}$ selected by BIC. ■ $s_0 = 2$, ■ $s_0 = \lfloor\sqrt{n/2}\rfloor$, ■ $s_0 = \lfloor\sqrt{n}\rfloor$, ■ $s_0 = \lfloor 2\sqrt{n}\rfloor$.

Table 2: Effect of Different Network Statistics on Take-Up, Probit Link

	Dependent variable: take-up							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Degree	0.006***				−0.0004	−0.002		
	(0.002)				(0.003)	(0.003)		
Eigenvector		0.333***					0.255**	0.143
		(0.075)					(0.110)	(0.102)
Beta			0.114***		0.119**		0.042	
			(0.030)		(0.047)		(0.043)	
Leader				0.182***		0.208***		0.136***
				(0.036)		(0.050)		(0.049)

Note: $^{*}p < 0.1;^{**}p < 0.05;^{***}p < 0.01$

Table 3: Effect of Different Network Statistics on Take-Up, Identity Link (Linear Regression)

	Dependent variable: take-up							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Degree	0.001***				−0.0001	−0.001		
	(0.001)				(0.001)	(0.001)		
Eigenvector		0.090***					0.069**	0.039
		(0.020)					(0.029)	(0.027)
Beta			0.031***		0.032**		0.011	
			(0.008)		(0.013)		(0.012)	
Leader				0.049***		0.056***		0.037***
				(0.010)		(0.013)		(0.013)

Note: $^{*}p < 0.1;^{**}p < 0.05;^{***}p < 0.01$