Non-asymptotic model selection for models of network data with parameter vectors of increasing dimension

Sean Eli[1], Michael Schweinberger[1]

Abstract

Model selection for network data is an open area of research. Using the β-model as a convenient starting point, we propose a simple and non-asymptotic approach to model selection of β-models with and without constraints. Simulations indicate that the proposed model selection approach selects the data-generating model with high probability, in contrast to classical and extended Bayesian Information Criteria. We conclude with an application to the Enron email network, which has 181,831 connections among 36,692 employees.

Keywords: β-model, sparse β-model, random graph, Bayesian Information Criterion, Extended Bayesian Information Criterion

1. Introduction

The statistical analysis of network data can provide insight into the spread of infectious diseases and other network-mediated phenomena. Despite progress in the last decade, open problems abound: for example, there are many models for infectious diseases spreading by contact, but no model selection approach with theoretical guarantees is known. In the past decade, the β-model (Chatterjee et al., 2011) has emerged as a convenient starting point for answering statistical questions about network models with parameter vectors of increasing dimension (e.g., Yan and Xu, 2013, Rinaldo et al., 2013, Karwa and Slavković, 2016, Mukherjee et al., 2018, Chen et al., 2021). It assumes that nodes i and j are connected with probability $p_{i,j} = \text{logit}^{-1}(\beta_i + \beta_j)$, where $\beta_i \in \mathbb{R}$ and $\beta_j \in \mathbb{R}$ are the propensities of nodes i and j to form connections. The β-model can be used for detecting potential superspreaders, that is, population members who have many contacts and thus can infect many others during an epidemic (Schweinberger et al., 2021).

Despite interesting applications of the β-model and other network models, model selection is challenging. As a case in point, consider the Bayesian Information Criterion or BIC (Schwarz, 1978, Kass and Raftery, 1995), which penalizes two times the negative loglikelihood by $p \log n$, where p is the number of parameters and n is the number of observations. At a first glance, the BIC would appear to be a simple and natural approach to selecting network models, but it turns out to be problematic: the β-model of a random graph with N nodes has $p = N \to \infty$ parameters, but the BIC assumes that the number of parameters p is fixed (see, e.g., Schwarz, 1978), suggesting that the penalty imposed by the BIC may be inappropriate. Despite concerns about whether the BIC is appropriate, it is widely used in practice, because there are hardly any alternatives: for example, the popular R package ergm (Krivitsky et al., 2021, p. 25) reports the BIC with penalty $p \log \binom{N}{2}$, assuming a random graph with N nodes and $\binom{N}{2}$ possible connections is observed. However, simulation results in Section 3.1 suggest that the BIC with penalty $p \log \binom{N}{2}$ tends to select more restrictive models over less restrictive models, even when data are generated by less restrictive models. A possible reason is that the penalty $p \log \binom{N}{2}$ is too high: while the parameter vector $\beta \in \mathbb{R}^N$ is estimated from $\binom{N}{2}$ observations, the number of observations with a direct bearing on the propensity β_i of a given

*Corresponding Author
Email addresses: seaneli@rice.edu (Sean Eli), michael.schweinberger@missouri.edu (Michael Schweinberger)

node i are the $N-1$ possible connections of node i to the other $N-1$ nodes. As a consequence, it is an open question whether $N-1$, $\binom{N}{2}$, or some other function of N between $N-1$ and $\binom{N}{2}$ should be used to compute an appropriate penalty. A related concern is that some graphs are more sparse than others and the penalty of the BIC may have to be decreased in sparse-graph settings, but it is not obvious by how much. These considerations suggest that it is unclear how an appropriate penalty should be computed in practice, leaving aside the fact that the assumptions of the BIC are violated.

An alternative is the extended BIC or EBIC, which is applicable when the number of parameters p increases with the number of observations n (Chen and Chen, 2008) and has been applied to generalized linear models (Chen and Chen, 2012). Despite the fact that the β-model is a generalized linear model for independent, though not identically distributed network data, the EBIC is as problematic as the BIC: the EBIC imposes a higher penalty than the BIC and is hence more conservative, as confirmed by simulation results in Section 3.1. A second, more recent alternative is the ℓ_0-penalized maximum likelihood approach to s-sparse β-models by Chen et al. (2021). While promising, the sparsity level s needs to be chosen. Chen et al. (2021) proposed choosing s based on the BIC with penalty $s\log\binom{N}{2}$, which raises the same questions we discussed above.

The above discussion underscores the challenge of specifying an appropriate penalty for model complexity in information-based approaches to model selection of network models. Using the β-model as a natural starting point, we propose a simple and non-asymptotic model selection criterion for β-models with and without constraints. The criterion is obtained by adding an appropriate adjustment term to the loglikelihood, which may be positive or negative, and depends on the pair of models under consideration. We demonstrate that the proposed model selection approach selects the data-generating model with high probability, in contrast to the BIC and EBIC. We conclude with an application to the Enron email network with 181,831 connections among 36,692 employees.

2. Main results

Let $D = (d_1, \dots, d_N)$ be the degrees of a β-model random graph with $N \geq 4$ nodes, and consider $\beta^{(1)} \neq \beta^{(2)} \in \mathbb{R}^N$. Define $\|\beta\|_\infty = \max_{1\leq i\leq N} |\beta_i|$, let K_N be the set of $i \in \{1,\dots,N\}$ such that $\beta_i^{(1)} \neq \beta_i^{(2)}$, and let $|K_N|$ be the cardinality of K_N. Denote by $\langle a, b\rangle$ the inner product of $a, b \in \mathbb{R}^N$, and define

$$f_{i;j}(D) = \left\langle D,\, \beta^{(i)}\right\rangle - \frac{1}{2}\left\langle E_{\beta^{(i)}} D,\, \beta^{(i)} - \beta^{(j)}\right\rangle, \;\; i \neq j \in \{1, 2\}\,. \tag{1}$$

Theorem 1. *The probabilities* $\mathrm{pr}_{\beta^{(1)}}\{f_{1;2}(D) > f_{2;1}(D)\}$ *and* $\mathrm{pr}_{\beta^{(2)}}\{f_{2;1}(D) > f_{1;2}(D)\}$ *are bounded below by*

$$1 - 2\exp\left(-\frac{\left\langle E_{\beta^{(1)}} D - E_{\beta^{(2)}} D,\, \beta^{(1)} - \beta^{(2)}\right\rangle^2}{2\sum_{1\leq i<j\leq N}\left|\beta_i^{(1)} - \beta_i^{(2)} + \beta_j^{(1)} - \beta_j^{(2)}\right|^2}\right). \tag{2}$$

A sufficient condition for the lower bound (2) *to be at least* $1 - 1/N$ *is*

$$\frac{\sqrt{|K_N|}\;\delta(\beta^{(1)}, \beta^{(2)})}{2\max\left\{||\beta^{(1)}||_\infty, ||\beta^{(2)}||_\infty\right\}(1 + \exp(2\max\left\{||\beta^{(1)}||_\infty, ||\beta^{(2)}||_\infty\right\}))^2} > \sqrt{\frac{128\log(2N)}{N-2}}. \tag{3}$$

The sufficient condition (3) of Theorem 1 resembles the beta-min conditions of high-dimensional variable selection (e.g., Bühlmann and van de Geer, 2011), which ensure that the weights of relevant and irrelevant variables are well-separated. In the same spirit, condition (3) ensures that $\beta^{(1)}$ and $\beta^{(2)}$ are well-separated, requiring the number $|K_N|$ of non-matching coordinates of $\beta^{(1)}$ and $\beta^{(2)}$ and the function $\delta(\beta^{(1)}, \beta^{(2)})$ quantifying the smallest difference between non-matching coordinates to be large enough. The function

$\delta(\beta^{(1)},\, \beta^{(2)})$ is defined as

$$\begin{aligned}
\delta(\beta^{(1)},\, \beta^{(2)}) &= \begin{cases} \delta_1(\beta^{(1)},\, \beta^{(2)}) & \text{if } |K_N| = 1 \\ \min\{\delta_1(\beta^{(1)},\, \beta^{(2)}),\ \delta_2(\beta^{(1)},\, \beta^{(2)})\} & \text{if } |K_N| \geq 2 \end{cases} \\
\delta_1(\beta^{(1)},\, \beta^{(2)}) &= \min_{i \,\in\, K_N}\ r(\beta_i^{(1)} - \beta_i^{(2)}) \\
\delta_2(\beta^{(1)},\, \beta^{(2)}) &= \min_{(i,j) \,\in\, K_N^2}\ \left\{ r\left(\beta_i^{(1)} - \beta_i^{(2)} + \beta_j^{(1)} - \beta_j^{(2)}\right) :\ \beta_i^{(1)} - \beta_i^{(2)} + \beta_j^{(1)} - \beta_j^{(2)} \neq 0 \right\},
\end{aligned}$$

where the map $r : \mathbb{R} \to \mathbb{R}$ defined by $r(x) = x\,(\exp(x) - 1)$ satisfies $r(x) > 0$ if and only if $x \neq 0$.

Theorem 1 provides a simple and non-asymptotic model selection approach based on the statistic $T_{1;2}(D) = f_{1;2}(D) - f_{2;1}(D)$. Suppose $\beta^{(1)} \neq \beta^{(2)} \in \mathbb{R}^N$ satisfy the beta-min condition (3). If $\beta^{(1)}$ is the data-generating vector, then $f_{1;2}(D) > f_{2;1}(D)$ with probability at least $1 - 1/N$. Likewise, if $\beta^{(2)}$ is the data-generating vector, then $f_{2;1}(D) > f_{1;2}(D)$ with probability at least $1 - 1/N$. Thus, if $\beta^{(1)}$ is chosen when $T_{1;2}(D) > 0$, and $\beta^{(2)}$ is chosen when when $T_{1;2}(D) < 0$, then the data-generating parameter vector is selected with probability at least $1 - 1/N$. Here, models are singleton subsets $\{\beta^{(1)}\}, \{\beta^{(2)}\} \subset \mathbb{R}^N$, and Theorem 1 bounds the probability that the data-generating parameter vector is selected.

In Section 3, we show that $T_{1;2}(D)$ can be used to select between more general models $B_1, B_2 \subset \mathbb{R}^N$. All we require is that B_1, B_2 are relatively closed, nonempty subsets of $\mathbb{R}^N$, so that constrained maximum likelihood estimates exist whenever unconstrained estimates exist (Theorem 5.7, Brown, 1986, p. 152). We then compute $T_{1;2}(D)$ by taking $\beta^{(1)}$ and $\beta^{(2)}$ to be maximum likelihood estimates constrained to B_1, B_2. Simulation results demonstrate that when the data-generating parameter vector belongs to $B_2 \setminus B_1$, then $T_{1;2}(D)$ selects the model B_2 with high probability, even when N is as small as 100. If the models are nested in the sense that $B_1 \subset B_2$ and the data-generating parameter vector belongs to B_1, then $T_{1;2}(D)$ selects the more restrictive model B_1 with high probability.

Computing $T_{1;2}(D)$ requires $O(N^2)$ operations, but the computing time can be reduced by imposing constraints on the parameters. For example, when parameters are identical within $K < N$ subpopulations of nodes, the computing time can be reduced to $O(K^2)$.

3. Numerical examples.

3.1. Simulations

We use simulations to shed light on the finite-N behavior of the proposed model selection approach, when models are closed subsets $B_1, B_2 \subset \mathbb{R}^N$. A practical and theoretical issue is that maximum likelihood estimates may not exist (Rinaldo et al., 2013): for example, many real-world networks have isolated nodes with degree 0, in which case maximum likelihood estimates do not exist, unless B_1, B_2 are compact (Theorem 5.7, Brown, 1986, p. 152). To alleviate existence issues, we consider β-models with block structure, in the sense that most nodes have low degrees and few nodes have high degrees. We assume that low-degree nodes have identical propensities to form connections, while high-degree nodes may have varying propensities. This simple approach to pooling strength reduces existence issues, and has the additional benefit of reducing computation time. In addition to having statistical and computational advantages, such scenarios are realistic in applications to network data, such as the spread of infectious diseases. In those applications, high-degree nodes are of primary interest, because such nodes are important drivers of many network-mediated phenomena: for example, population members with many contacts can infect many others, and are therefore potential superspreaders.

For various values of N and $|K_N|$, we consider the following models $B_1, B_2 \subset \mathbb{R}^N$: Model B_1 consists of parameter vectors for which the first $N - |K_N|$ coordinates are identical and the last $|K_N|$ coordinates are identical, whereas Model B_2 only requires the first $N - |K_N|$ coordinates to be identical. We specify $\beta^{(1)} \in B_1$, which consists of $|K_N|$ high propensities and $N - |K_N|$ low propensities, and obtain $\beta^{(2)} \in B_2$ from $\beta^{(1)}$ by adding independent Gaussian noise to the $|K_N|$ high propensities. We generate 500 β-model graphs from $\beta^{(1)}$ and compare $\beta^{(1)}$ and $\beta^{(2)}$ using $T_{1;2}(D)$. Next, using $\beta^{(1)}$ as the data-generating parameter vector, we compare the models B_1 and B_2 by using $T_{1;2}(D)$, replacing $\beta^{(1)}$ and $\beta^{(2)}$ by constrained maximum likelihood estimates. We repeat the process by drawing graphs from $\beta^{(2)}$ and selecting models in oracle and non-oracle scenarios. The estimated success probabilities of selecting the data-generating model with the proposed model selection approach, in both oracle and non-oracle scenarios, are presented in the O β and NO β columns in Table 1. As a reference point, we presented the estimated success probabilities of the BIC and EBIC in the non-oracle scenarios, which impose penalties $p \log \binom{N}{2}$ and $p \log \binom{N}{2} + 2\, p \log N$ on twice the negative loglikelihood. Here, $p \leq N$ is the number of unrestricted parameters of the model being considered, and N is the total number of parameters of the β-model. When data are generated from $\beta^{(2)}$, belonging to the less restrictive model B_2, the proposed model selection approach works well in both oracle and non-oracle scenarios: the data-generating model is selected with high probability, when N is as small as 100. However, when data are generated from $\beta^{(1)}$, belonging to the more restrictive model $B_1 \subset B_2$, there is a performance gap between the oracle and non-oracle results: while the oracle results are excellent when $N \geq 100$, the non-oracle results require $N \geq 500$. By contrast, in every situation, the BIC and EBIC select the more restrictive model B_1 regardless of whether data were generated from $\beta^{(1)}$ or $\beta^{(2)}$. We review possible reasons in Section 1.

3.2. Enron email network: forward model selection

We apply the proposed model selection approach to the Enron email network (Leskovec et al., 2008). The network consists of 181,831 connections among 36,692 employees, where a connection corresponds to an exchange of emails. We partition the employees into two subsets: S_{low} consists of employees with degrees less than the 95% quantile, and S_{high} consists of the 1,820 remaining employees. We perform a form of forward model selection. Starting with Model 0, under which all employees have the same propensity to be in email contact with others, we consider moving to Model 1, under which employees in S_{low} have identical propensities and employees in S_{high} have identical propensities. If Model 1 is selected over Model 0, we consider moving to Model 2, which only postulates that employees in S_{low} have identical propensities. We estimate the models with constrained maximum likelihood estimates $\beta^{(0)}, \beta^{(1)}$, and $\beta^{(2)}$, and compute $T_{1;0}(D) = 27.40 \times 10^4$ and $T_{2;1}(D) = -98.14 \times 10^4$. Thus, Model 1 is selected over Model 0, and Model 1 is selected over Model 2. These conclusions are not sensitive to the 95% threshold: Using 99%, 99.5%, or 99.9% quantiles as threshold leads to similar conclusions.

4. Appendix.

Throughout, $\beta^* \in \{\beta^{(1)}, \beta^{(2)}\}$ is the data-generating parameter vector and $Y_{i,j} \in \{0,1\}$ are edge indicators. We define $p^k_{i,j} = E_{\beta^{(k)}}\, Y_{i,j} = \text{logit}^{-1}(\beta^{(k)}_i + \beta^{(k)}_j)$, and we suppress the arguments of the functions δ, δ_1 and δ_2 defined in Section 2.

The following auxiliary results are needed to prove Theorem 1.

Proposition 2. *For any $t > 0$, let $P_t = \text{pr}_{\beta^*}\left\{\left|\langle D,\, \beta^{(1)} - \beta^{(2)}\rangle - \langle E_{\beta^*} D,\, \beta^{(1)} - \beta^{(2)}\rangle\right| \geq t\right\}$ and let $M = \|\beta^{(1)} - \beta^{(2)}\|_\infty$. Then*

$$P_t \leq 2 \exp\left(-\frac{2\, t^2}{\sum_{1 \leq i < j \leq N} \left|\beta^{(1)}_i - \beta^{(2)}_i + \beta^{(1)}_j - \beta^{(2)}_j\right|^2}\right) \leq 2 \exp\left(-\frac{t^2}{N\, |K_N|\, M^2}\right). \tag{4}$$

Proof. Since $\langle D,\, \beta^{(1)} - \beta^{(2)}\rangle = \sum_{1 \leq i < j \leq N}\left(\beta^{(1)}_i - \beta^{(2)}_i + \beta^{(1)}_j - \beta^{(2)}_j\right) Y_{i,j}$ and the edge indicators $Y_{i,j} \in \{0,1\}$ are independent, the first inequality of (4) follows from Hoeffding's inequality. The denominator of the ratio in the first exponent in (4) equals

$$\sum_{i<j:\; i,j \,\in\, K_N} \left|\beta^{(1)}_i - \beta^{(2)}_i + \beta^{(1)}_j - \beta^{(2)}_j\right|^2 + \sum_{i \in K_N} \sum_{j \notin K_N} \left|\beta^{(1)}_i - \beta^{(2)}_i\right|^2.$$

Table 1: Simulation results: performance of the proposed model selection procedure in oracle and non-oracle scenarios, compared with BIC and EBIC

N	$\lvert K_N\rvert$	$\beta_i^{(1)}$	δ	O $\beta^{(1)}$	NO $\beta^{(1)}$	(E)BIC $\beta^{(1)}$	O $\beta^{(2)}$	NO $\beta^{(2)}$	(E)BIC $\beta^{(2)}$
100	30	(-0.5, 0)	1e-4	100	4	100	100	100	0
100	20	(-0.5, 0)	4e-6	100	3	100	100	100	0
100	10	(-0.5, 0)	9e-5	100	2	100	100	100	0
100	30	(-0.75, -0.25)	8e-7	100	4	100	100	100	0
100	20	(-0.75, -0.25)	2e-5	100	3	100	100	100	0
100	10	(-0.75, -0.25)	5e-7	100	1	100	100	100	0
250	40	(-1, -0.5)	1e-4	100	36	100	100	100	0
250	30	(-1, -0.5)	4e-6	100	21	100	100	100	0
250	20	(-1, -0.5)	3e-4	100	9	100	100	100	0
250	40	(-1.25, -0.75)	5e-7	100	19	100	100	100	0
250	30	(-1.25, -0.75)	2e-8	100	9	100	100	100	0
250	20	(-1.25, -0.75)	1e-4	100	6	100	100	100	0
500	50	(-1.5, -0.5)	3e-7	100	94	100	100	100	0
500	40	(-1.5, -0.5)	8e-7	100	70	100	100	100	0
500	20	(-1.5, -0.5)	2e-5	100	18	100	100	100	0
500	50	(-1.75, -0.75)	2e-8	100	65	100	100	100	0
500	40	(-1.75, -0.75)	3e-5	100	35	100	100	100	0
500	20	(-1.75, -0.75)	4e-6	100	9	100	100	100	0
1000	75	(-2, -1)	3e-7	100	100	100	100	100	0
1000	50	(-2, -1)	5e-7	100	90	100	100	100	0
1000	40	(-2, -1)	2e-8	100	68	100	100	100	0
1000	75	(-2.5, -1.5)	5e-7	100	63	100	100	100	0
1000	50	(-2.5, -1.5)	2e-9	100	23	100	100	100	0
1000	40	(-2.5, -1.5)	1e-7	100	13	100	100	100	0

Column $\beta_i^{(1)}$ with entries $(a,\, b)$ indicates that low-degree nodes have propensities $\beta_i^{(1)} = a$ and high-degree nodes have propensities $\beta_i^{(1)} = b$. Under $\beta^{(2)}$, low-degree nodes have propensities $\beta_i^{(2)} = \beta_i^{(1)}$, and high-degree nodes have propensities $\beta_i^{(2)} = \beta_i^{(1)} + \epsilon_i$, where $\epsilon_i \sim N(1/2, 1)$. Column δ refers to $\delta(\beta^{(1)}, \beta^{(2)})$. In all rows, the bound (2) of Theorem 1 computed from $\beta^{(1)}$ and $\beta^{(2)}$ is greater than 0.999, so we expect near-perfect oracle results. Columns O $\beta^{(i)}$, NO $\beta^{(i)}$, and (E)BIC $\beta^{(i)}$ report the frequency that the data-generating model is selected by the proposed model selection procedure in the oracle scenario ($\beta^{(1)}$ and $\beta^{(2)}$ known), the non-oracle scenario (models B_1 and B_2 are estimated), and by the BIC and EBIC, when data are generated from $\beta^{(i)} \in \{\beta^{(1)},\, \beta^{(2)}\}$. Since the BIC is conservative and the EBIC imposes a higher penalty than the BIC, both criteria give the same results, which are presented in column (E)BIC.

This is bounded above by $4\,(\lvert K_N\rvert\,(\lvert K_N\rvert - 1)\,/\,2)\,M^2 + \lvert K_N\rvert\,(N - \lvert K_N\rvert)\,M^2 \leq 2\,N\,\lvert K_N\rvert\,M^2$ when $\lvert K_N\rvert \geq 2$, and by $(N-1)\,M^2 \leq 2\,N\,\lvert K_N\rvert\,M^2$ when $\lvert K_N\rvert = 1$. In both cases, the denominator is bounded above by $2\,N\,\lvert K_N\rvert\,M^2$, proving (4). □

Lemma 3. *Suppose $n \geq 2$, and let $x = (a_1, ..., a_n) \in \mathbb{R}^n$ with $a_i \neq 0$ for all i. The number of pairs $i < j$ of indices with $a_i + a_j = 0$ is at most $n^2/4$.*

Proof. Suppose the most common zero-sum pair of entries of x consists of a and $-a$. Replacing the other zero-sum pairs with a and $-a$ does not decrease the number of zero-sum pairs in x. Thus, we may assume x has k entries of 1 and $(n-k)$ entries of -1; taking $k = n/2$ gives a maximum of $n^2/4$. □

Proof of Theorem 1. First, we bound $R = \left\langle E_{\beta^{(1)}} D - E_{\beta^{(2)}} D,\, \beta^{(1)} - \beta^{(2)} \right\rangle$ below. A direct computation gives

$$R = \sum_{i<j:\, i,j \in K_N} \left(\beta_i^{(1)} - \beta_i^{(2)} + \beta_j^{(1)} - \beta_j^{(2)}\right) \, (p_{i,j}^1 - p_{i,j}^2) + \sum_{i \in K_N} \sum_{j \notin K_N} (\beta_i^{(1)} - \beta_i^{(2)}) \, (p_{i,j}^1 - p_{i,j}^2).$$

Let $L_N = \max\left\{\|\beta^{(1)}\|_\infty,\, \|\beta^{(2)}\|_\infty\right\}$. For distinct $i, j \in K_N$,

$$\left(\beta_i^{(1)} - \beta_i^{(2)} + \beta_j^{(1)} - \beta_j^{(2)}\right) \, (p_{i,j}^1 - p_{i,j}^2) \geq \left(\beta_i^{(1)} + \beta_j^{(1)} - \beta_i^{(2)} - \beta_j^{(2)}\right) \, \frac{e^{\beta_i^{(1)} + \beta_j^{(1)} - \beta_i^{(2)} - \beta_j^{(2)}} - 1}{(1 + e^{2\,L_N})^2} \geq 0,$$

with equality if and only if $\beta_i^{(1)} - \beta_i^{(2)} + \beta_j^{(1)} - \beta_j^{(2)} = 0$. In addition, if $i \in K_N$ and $j \notin K_N$,

$$(\beta_i^{(1)} - \beta_i^{(2)}) \, (p_{i,j}^1 - p_{i,j}^2) \geq (\beta_i^{(1)} - \beta_i^{(2)}) \, \frac{e^{\beta_i^{(1)} - \beta_i^{(2)}} - 1}{(1 + e^{2\,L_N})^2} > 0.$$

We define the following sums

$$S_1(N) \quad = \sum_{i<j:\, i,j \in\, K_N} r\left(\beta_i^{(1)} - \beta_i^{(2)} + \beta_j^{(1)} - \beta_j^{(2)}\right), \qquad S_2(N) \quad = \sum_{i \in K_N} r(\beta_i^{(1)} - \beta_i^{(2)});$$

it follows that R is nonnegative and bounded below by

$$\frac{1}{(1 + e^{2\,L_N})^2} \left(S_1(N) + (N - |K_N|) \, S_2(N)\right). \tag{5}$$

By definition of δ_1, $S_2(N) \geq |K_N|\delta_1 > 0$. By convention, we take $S_1(N) = 0$ if $|K_N| = 1$. We use Lemma 3 to bound $S_1(N)$ from below, which provides useful lower bounds for R when $|K_N|$ is large. Observe that the (i,j)-th term of $S_1(N)$ is zero if and only if $\beta_i^{(1)} - \beta_i^{(2)} + \beta_j^{(1)} - \beta_j^{(2)} = 0$. Applying Lemma 3 to $\beta^{(1)} - \beta^{(2)} \in \mathbb{R}^N$ shows there are at least $|K_N\,|(|K_N| - 2)\,/\,4$ nonzero terms in $S_1(N)$. By definition of δ_2, $S_1(N) \geq \delta_2 |K_N|(|K_N| - 2)/4$. Thus, if $|K_N| \geq 2$, then (5) is greater than

$$\begin{aligned} \frac{1}{(1 + e^{2\,L_N})^2} \left(\frac{|K_N|(|K_N| - 2)}{4} \, \delta_2 + (N - |K_N|)|K_N|\delta_1\right) &\geq \frac{\delta}{4\,(1 + e^{2\,L_N})^2} |K_N| \, (4\,N - 3\,|K_N| - 2) \\ &\geq \frac{\delta}{4\,(1 + e^{2\,L_N})^2} |K_N| \, (N - 2)\,, \end{aligned}$$

and we have $R \geq \delta |K_N| \, (N - 2) \,/ (4 \, (1 + e^{2L_N})^2) > 0$. If $|K_N| = 1$, then (5) and thus R are greater than $\delta_1(N - 1)/((1 + e^{2L_N})^2) > 0$. Next, to obtain the desired probability bounds, notice

$$f_{1;2}(D) - f_{2;1}(D) = \left\langle D,\, \beta^{(1)} - \beta^{(2)} \right\rangle - \frac{1}{2} \left\langle E_{\beta^{(1)}} D + E_{\beta^{(2)}} D,\, \beta^{(1)} - \beta^{(2)} \right\rangle.$$

It follows that

$$\begin{aligned} \mathrm{pr}_{\beta^{(1)}} \left\{ f_{1;2}(D) \; < \; f_{2;1}(D) \right\} &= \mathrm{pr}_{\beta^{(1)}} \left\{ \left\langle D,\, \beta^{(1)} - \beta^{(2)} \right\rangle < \frac{1}{2} \left\langle E_{\beta^{(1)}} D + E_{\beta^{(2)}} D,\, \beta^{(1)} - \beta^{(2)} \right\rangle \right\} \\ &\leq \mathrm{pr}_{\beta^{(1)}} \left\{ \left| \left\langle D,\, \beta^{(1)} - \beta^{(2)} \right\rangle - \left\langle E_{\beta^{(1)}} D,\, \beta^{(1)} - \beta^{(2)} \right\rangle \right| > \frac{1}{2} R \right\}, \end{aligned} \tag{6}$$

and similarly

$$\mathrm{pr}_{\beta^{(2)}}\left\{f_{2;1}(D)\;<\;f_{1;2}(D)\right\} \leq \mathrm{pr}_{\beta^{(2)}}\left\{\left|\left\langle D,\,\beta^{(1)}-\beta^{(2)}\right\rangle - \left\langle E_{\beta^{(2)}}D,\,\beta^{(1)}-\beta^{(2)}\right\rangle\right| > \frac{1}{2}R\right\}. \tag{7}$$

Bound (2) follows from applying the sharper bound of (4) to (6) and (7). Applying the weaker bound of (4) to (6) and (7), and using $R \geq \delta\,|K_N|\,(N-2)\,/(4\,(1+e^{2L_N})^2)$ and $N \geq 4$, we obtain a weaker probability bound:

$$2\exp\left(-\frac{\delta^2\,|K_N|\,(N-2)^2}{64\,N\,M^2\,\left(1+e^{2\,L_N}\right)^4}\right) \leq 2\exp\left(-\frac{\delta^2\,|K_N|\,(N-2)}{128\,M^2\,\left(1+e^{2\,L_N}\right)^4}\right). \tag{8}$$

It is easily verified that the beta-min condition (3) implies that the bound (8) is less than $1/N$. □

5. Acknowledgements

This work was supported by the National Science Foundation, Division of Mathematical Sciences, Statistics (NSF award DMS-1812119).

References

Brown, L., 1986. Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory. Institute of Mathematical Statistics, Hayworth, CA, USA.

Bühlmann, P., van de Geer, S., 2011. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer-Verlag, New York.

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

Chen, J., Chen, Z., 2008. Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95, 759–771.

Chen, J., Chen, Z., 2012. Extended bic for small-n-large-p sparse glm. Statistica Sinica 22, 555–574.

Chen, M., Kato, K., Leng, C., 2021. Analysis of networks via the sparse β-model. Journal of the Royal Statistical Society, Series B (Statistical Methodology) To appear.

Karwa, V., Slavković, A.B., 2016. Inference using noisy degrees: Differentially private β-model and synthetic graphs. The Annals of Statistics 44, 87–112.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.

Krivitsky, P.N., Hunter, D.R., Morris, M., Klumb, C., 2021. ergm 4.0: New features and improvements ArXiv:2106.04997v1.

Leskovec, J., Lang, K., Dasgupta, A., Mahoney, M., 2008. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Internet Mathematics 6. doi:10.1080/15427951.2009.10129177.

Mukherjee, R., Mukherjee, S., Sen, S., 2018. Detection thresholds for the β-model on sparse graphs. The Annals of Statistics 46, 1288–1317.

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

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Schweinberger, M., Bomiriya, R.P., Babkin, S., 2021. A semiparametric Bayesian approach to epidemics, with application to the spread of the coronavirus MERS in South Korea in 2015. Journal of Nonparametric Statistics To appear.

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.