L-2 Regularized maximum likelihood for β-model in large and sparse networks

Meijia Shao

Department of Statistics, The Ohio State University, U.S.A.

Yu Zhang, Qiuping Wang

Department of Statistics, Central China Normal University, China

Yuan Zhang*

Department of Statistics, The Ohio State University, U.S.A.

*E-mail: yzhanghf@stat.osu.edu

Jing Luo

School of Mathematics and Statistics, South-Central Minzu University, China

Ting Yan

Department of Statistics, Central China Normal University, China

Summary. The β-model is a powerful tool for modeling network generation driven by degree heterogeneity. Its simple yet expressive nature particularly well-suits large and sparse networks, where many network models become infeasible due to computational challenge and observation scarcity. However, existing estimation algorithms for β-model do not scale up; and theoretical understandings remain limited to dense networks. This paper brings several significant improvements to the method and theory of β-model to address urgent needs of practical applications. Our contributions include: 1. method: we propose a new ℓ_2 penalized MLE scheme; we design a novel fast algorithm that can comfortably handle sparse networks of millions of nodes, much faster and more memory-parsimonious than all existing algorithms; 2. theory: we present new error bounds on β-models under much weaker assumptions than best known results in literature; we also establish new lower-bounds and new asymptotic normality results; under proper parameter sparsity assumptions, we show the first local rate-optimality result in ℓ_2 norm; distinct from existing literature, our results cover both small and large regularization scenarios and reveal their distinct asymptotic dependency structures; 3. application: we apply our method to large COVID-19 network data sets and discover meaningful results.

Keywords: Network analysis; β-model; Sparse networks; Big data; Regularization.

1. Introduction

1.1. The β-model: formulation, motivating data examples and previous work

The β-model (Chatterjee et al., 2011), is a heterogeneous exponential random graph model (ERGM) with the degree sequence as the exclusively sufficient statistic. It is a popular model for networks mainly driven by degree heterogeneity. Under this model, an undirected and binary network of n nodes, represented by its adjacency matrix $A = (A_{i,j})_{1\le\{i,j\}\le n} \in \{0,1\}^{n\times n}$, is generated by

$$\mathbb{P}(A_{i,j}=1) = \frac{e^{\beta_i^*+\beta_j^*}}{1+e^{\beta_i^*+\beta_j^*}}, \quad A_{i,j}=A_{j,i}, \quad 1\le i<j\le n, \tag{1}$$

where $\beta^* = (\beta_1^*,\ldots,\beta_n^*)$ denotes the vector of true model parameters, and edges $\{A_{i,j} : i<j\}$ are mutually independent. Set $A_{i,i}\equiv 0$ for all i. The negative log-likelihood function is

$$\mathcal{L}(\beta) = \sum_{1\le i<j\le n} \log\left(1+e^{\beta_i+\beta_j}\right) - \sum_{i=1}^n \beta_i d_i. \tag{2}$$

where $d_1, \ldots, d_n$ are observed node degrees: $d_i = \sum_{j \neq i} A_{i,j}$.

The β-model is a simple, yet expressive tool for describing degree heterogeneity, a feature of paramount importance in many networks (Babai et al., 1980; Fienberg, 2012). Here we briefly describe two motivating data sets. First, (Elmer et al., 2020) collected social networks between a group of Swiss students before and during COVID-19 lockdown. For privacy protection, they only released node degrees instead of adjacency matrix. The β-model can be fitted to this data while most other popular models which need adjacency matrix cannot be applied. The second example is a massive COVID-19 knowledge graph (Steenwinckel et al., 2020). It contains $n \approx 10^7$ non-isolated nodes with just 2.1×10^7 million edges. The β-model shows its unique advantages of high speed and memory parsimony in handling this data. We will present detailed analyses of these data sets in Section 6.

Since its birth, the β-model has attracted a lot of research interest. The early works Holland and Leinhardt (1981); Chatterjee et al. (2011); Park and Newman (2004); Rinaldo et al. (2013); Hillar and Wibisono (2013) studied basic model properties and established existence, consistency and asymptotic normality of the vanilla MLE. Chen and Olvera-Cravioto (2013); Yan et al. (2016, 2019); Stein and Leng (2021) extended the model for directed and bipartite networks. Karwa and Slavković (2016) studied differential privacy in β-model. Graham (2017); Su et al. (2018); Gao (2020); Yan et al. (2019); Stein and Leng (2020, 2021) incorporated nodal or edge-wise covariates. Wahlström et al. (2017) established the Cramer-Rao bound with repeated network observations. Mukherjee et al. (2018) studied a different variant of β-model for sparse network with a known sparsity parameter.

1.2. *Regularized β-models*

The most relevant to the topic of this paper are the recent pioneering works Chen et al. (2021) and Stein and Leng (2020). They introduced regularization into the β-model literature. The main idea is that a simplified parameter space would help address sparse networks. Chen et al. (2021) proposed the following sparse β-model (SβM). Suppose there exists a set $\mathcal{S} \subseteq [1:n]$, called *active set*, such that

$$\beta_i^* = \begin{cases} -\gamma/2 \log n + \mu^\dagger/2, & \text{if } i \notin \mathcal{S}, \\ (\alpha - \gamma/2) \log n + \mu^\dagger/2 + \beta_i^\dagger, & \text{if } i \in \mathcal{S}, \end{cases} \tag{3}$$

where α and γ are positive constants, and $\max\{|\mu^\dagger|, |\beta_i^\dagger|\} = o(\log n)$. They assume that γ and α are *known* and show that one can consistently estimate the $o(\log n)$ parts of the parameters $(\mu^\dagger, \beta_i^\dagger)$ using a constrained MLE with user-specified upper bound on the size of the estimated $\mathcal{S}$. Their approach can be viewed as an ℓ_0-regularized MLE. Stein and Leng (2020) proposed an ℓ_1-regularized MLE. The model assumption of Stein and Leng (2020) can be roughly understood as a relaxed version of (3), where we have α_i for each $i \in \mathcal{S}$ instead of a uniform α. Their method can be reformulated as

$$\arg\min_{\beta, \beta_0} \mathcal{L}_{\lambda;\beta_0}(\beta) = \mathcal{L}(\beta) + \lambda \big\| \beta - \beta_0 \cdot \mathbb{1} \big\|_1, \tag{4}$$

where $\mathbb{1}$ is an all-one vector and λ is a tuning parameter. Remark that our choice of λ is different from theirs. This is an ℓ_1-regularized MLE approach.

In this paper, we adopt a symbol system based on the original formulation of the β-model (1) and propose an ℓ_2-regularized MLE, as follows

$$\arg\min_{\beta} \mathcal{L}_\lambda(\beta) = \mathcal{L}(\beta) + \frac{\lambda}{2} \big\| \beta - \bar{\beta} \cdot \mathbb{1} \big\|_2^2, \tag{5}$$

where notice that we use $\bar{\beta} = n^{-1} \sum_{i=1}^n \beta_i$ instead of β_0, since the optimal β_0 equals $\bar{\beta}$. Our ℓ_2 penalty is a softer version of ℓ_0 (Chen et al., 2021) and ℓ_1 (Stein and Leng, 2020, 2021) penalties, thus is more flexible for modeling networks where β^* may not be ℓ_0 sparse. In this paper, we shall propose a fast algorithm for solving (5), establish accompanying theory, and conduct comprehensive simulation studies and data applications.

1.3. Our contributions

The main theme of our paper is an ℓ_2-regularized MLE. Therefore, we will focus more on comparing with Chen et al. (2021) and Stein and Leng (2020) in this part. Table 1 highlights the main differences and improvements of our work over Chen et al. (2021) and Stein and Leng (2020) and provides pointers to corresponding sections in this paper.

	Chen et al. (2021)	Stein and Leng (2020)	Our paper	Discussed in
Penalty	l_0	l_1	l_2	Section 1.2
Model	(3)	(3) with α_i for each i, $\mu^\dagger = \beta_i^\dagger = 0$	(1)	Section 1.2
Parameter constraint	*Known* γ and α; $\alpha > 0$	$\alpha_i > 0$	None	Section 1.2
MLE existence	Not always guaranteed	Yes	Yes	Section 2.1
Computation cost	$O(n^3)$ per iteration	$O(n^3)$ per iteration	$O(n \cdot \bar{d})$ + ($O(m^2)$ per iteration) *a*	Section 2.2
Need network sparsity *b*	$\gg n^{-1/2}$	$\gg n^{-1/6}$	$\gg n^{-1}$	Section 3.1
Finite-sample error bound	No	Yes	Yes	Section 3.1
Sparsistency	Yes	Yes	Yes	Section 3.1
Lower bound result	No	No	Yes	Section 3.2
Local rate-optimality	No	No	Yes	Section 3.2
Asymptotic normality	Fixed-dimensional	Fixed-dimensional	High-dimensional	Section 3.3
Data-driven tuning of λ	Yes	Yes	Yes	Section 4
Empirical scalability	$n \approx 10^3$	$n \approx 10^3$	$n \approx 10^7$	Sections 5 and 6

Table 1: Comparison table between Chen et al. (2021), Stein and Leng (2020) and our paper.

a: $\bar{d}$ is the average degree; m is the number of different unique degrees. See Section 2.2.
b: All method's abilities to handle sparse networks would depreciate as the degrees become inhomogeneous. We only present the best cases in the table.

Now, we summarize our main contributions.

I. *Weaker model assumptions and simple MLE existence guarantee.* Both Chen et al. (2021) and Stein and Leng (2020) assume that most nodes have a common (or nearly common) low expected degree, whereas a few nodes in the active set $\mathcal{S}$ have higher degrees. Chen et al. (2021) further assumes that those high-degree nodes also have similar expected degrees. Our paper makes no such assumptions. Our method and accompanying theory can also be applied to β-models that do not have such "parameter sparsity" structure, which Chen et al. (2021) and Stein and Leng (2020) are not provably valid for.

In the past, Rinaldo et al. (2013) devoted much effort into finding a complicated sufficient and necessary condition for MLE existence. The same treatment was inherited by Chen et al. (2021). We show that a simple "$\lambda > 0$" is sufficient to guarantee the existence of our ℓ_2-regularization MLE, similar to Stein and Leng (2020).

II. *Proposing a very fast new algorithm for large and sparse networks.* Computation has long been a challenge for β-models. To out best knowledge, all existing works resort to either some general optimizer Chen et al. (2021) or use GLM packages Yan et al. (2019); Stein and Leng (2020) for parameter estimation, which typically cannot scale up to networks with 10^4 nodes (see Table 1) and do not take advantage of network sparsity. In this paper, we propose a novel algorithms that fully exploits the structure of the β-model and particularly specialize for handling large and sparse networks. For example, our new algorithm only takes a few minutes on a personal computer to fit a β-model to the Steenwinckel et al. (2020) data that contains $O(10^7)$ nodes.

III. Theory is the most highlighted part of this paper. We believe our work represents a giant leap in the theoretical understandings of β-models. Our main theoretical contributions include:

(i). *Handling much sparser networks.* Existing β-model literature (Chatterjee et al., 2011; Yan and Xu, 2013; Yan et al., 2016) typically requires that $\rho_n = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} A_{i,j} \gtrsim$ Polynomial$(\log^{-1} n)$ for consistency, where ρ_n is the network density. Apparently, Chen et al. (2021) can handle network sparsity down to $\rho_n \gg n^{-1/2}$. However, their theory is built upon an unrealistic assumption of *knowing* the true values of γ and α in their model, while in view of Stein and Leng (2020), estimating unknown γ and α is the really hard part of the problem. Stein and Leng (2020) does not make such assumption, but they can only handle a network sparsity of $\rho_n \gg n^{-1/6}$. In sharp contrast, our method is guaranteed to be consistent for much sparser networks. For example, if all edges probabilities are on the same asymptotic order, we only need $\rho_n \gg n^{-1}$ to guarantee ℓ_∞ consistency. For more details, see Section 3.1.

(ii). *Finite-sample upper bounds and provably sparsistent post-estimation variable selection.* Our theory provides finite-sample ℓ_∞ and ℓ_2 error bounds. The ℓ_∞ error bound enables a post-estimation variable selection algorithm (see Corollary 1) that allows us to consistently estimate $\mathcal{S}$ under the settings of either Chen et al. (2021) and Stein and Leng (2020). Our method not only computes much faster, but also requires much weaker assumption on the size of $\mathcal{S}$. Chen et al. (2021) assumes $|\mathcal{S}| \ll n$ and Stein and Leng (2020) assumes $|\mathcal{S}| \lesssim n^{1/2}$ in order to guarantee estimation consistency. We do not need such assumption and our theory, applied to Chen et al. (2021)'s model, only requires $\mathcal{S} \le Cn$ for any constant $C \in (0, 1/2)$.

(iii). *Finite-sample lower bound results and local rate-optimality.* Existing β-model literature presents very little understanding of lower bound results. We establish the first set of lower bound results for β-model with *one single* observation of the network. As a highlighted contribution, we find that our estimator is *locally rate-optimal* in ℓ_2 norm for estimating a β-model with true parameter β^* being "β-sparse" (in the sense of Chen et al. (2021)). This is the first estimation optimality result in β-model literature. We also established the local and non-local lower bounds in different ℓ_p norms, which are a little different from upper bounds, up to a logarithmic factor.

(iv). *High-dimensional asymptotic normality and characterization of estimator's behavior under heavy penalty.* All existing asymptotic normality results for β-models are fixed-dimensional. We present the first *high-dimensional* asymptotic normality result. Moreover, our theory reveals an interesting contrast in the asymptotic covariance structures under light and heavy penalties. When λ is small, the estimator's elements tend to be independent; and when λ is large, they become nearly perfectly dependent. This is verified by our simulation. Our paper provides the first characterization of the behavior of the regularized estimator with a large λ in the β-model literature. Our proof techniques in this part is original and very different from other β-model papers.

IV. Our other contributions include the follows. We develop a data-driven AIC-type criterion for automatically selecting the tuning parameter λ. Simulation results verify its effectiveness. In the discussion section, we also present empirical studies to gauge the feasibility of a popular idea on testing goodness-of-fit of network models applied to the β-model.

1.4. Notation

We inherit the asymptotic notion $O(\cdot)$, $o(\cdot)$, $\lesssim$ and $\asymp$ from standard calculus. Let $\mathbb{1} = (1, \ldots, 1)^T$. For any vector $u \in \mathbb{R}^{n\times 1}$ and matrix $U \in \mathbb{R}^{n\times n}$, define matrices $V(u)$ and $V(U)$ as follows: for all $1 \le \{i \ne j\} \le n$, define $\{V(u)\}_{i,j} = e^{u_i+u_j}/(1+e^{u_i+u_j})^2$ and $\{V(U)\}_{i,j} = e^{U_{i,j}}/(1+e^{U_{i,j}})$; and for all $1 \le i \le n$, define $V_{i,i} = \sum_{1\le j\le n, j\ne i} V_{i,j}$ for $V = V(u)$ or $V = V(U)$. For any vector x, inherit the standard notion of ℓ_p norms $\|x\|_p$ for $p = 0, 1, 2, \infty$. For any matrix $J \in \mathbb{R}^{n\times n}$, define $\|J\|_\infty = \sup_{x\ne 0} \|Jx\|_\infty/\|x\|_\infty = \max_{1\le i\le n} \sum_{j=1}^n |J_{i,j}|$. Also inherit Frobenius norm $\|\cdot\|_F$ and spectral norm $\|\cdot\|_{\text{op}}$ from standard matrix analysis. For two matrices $A_1, A_2 \in \mathbb{R}^{n\times n}$, write

$A_1 \geq A_2$ to denote element-wise comparison: $A_{1;i,j} \geq A_{2;i,j}$ for all $1 \leq \{i,j\} \leq n$. Finally, we import two concepts of "sparsity" from Chen et al. (2021): *network sparsity* refers to Chen et al. (2021) $\rho_n = \sum_{1\leq i<j\leq n} \mathbb{E}[A_{i,j}]/\binom{n}{2}$, and *$\beta$-sparsity* means that most β_i^*'s share a common value.

2. Our method

In this section, we present the parameter estimation and the fast algorithm for our model. Statistical inference would require quantitative theoretical study, therefore, we relegate them to Section 3.

2.1. Parameter estimation

Recall our method from (5). Denote the gradient of $\mathcal{L}_\lambda(\beta)$ by $F(\beta) = (F_1(\beta), \ldots, F_n(\beta))$. We have

$$F_i(\beta) = \frac{\partial \mathcal{L}_\lambda(\beta)}{\partial \beta_i} = \sum_{\substack{1\leq j\leq n \\ j\neq i}} \frac{e^{\beta_i+\beta_j}}{1+e^{\beta_i+\beta_j}} - d_i + \lambda(\beta_i - \bar{\beta}), \quad 1 \leq i \leq n. \tag{6}$$

We estimate the model parameters by solving the following equation set

$$F(\widehat{\beta}_\lambda) = 0. \tag{7}$$

There are two immediate questions regarding (7): the *existence* and the *uniqueness* of its solution. We first address the uniqueness of MLE, assuming its existence. It suffices to show the strict convexity of the ℓ_2-regularized likelihood $\mathcal{L}_\lambda(\beta)$. Denote the Jacobian matrix of $\mathcal{L}_\lambda(\beta)$ by $F'(\beta)$, we have

$$\{F'(\beta)\}_{i,j} = \frac{\partial F_i}{\partial \beta_j}(\beta) = \frac{\partial F_j}{\partial \beta_i}(\beta) = \frac{e^{\beta_i+\beta_j}}{(1+e^{\beta_i+\beta_j})^2} - \frac{\lambda}{n}, \qquad \text{for } 1 \leq \{i \neq j\} \leq n, \tag{8}$$

$$\{F'(\beta)\}_{i,i} = \frac{\partial F_i}{\partial \beta_i}(\beta) = \sum_{\substack{1\leq j\leq n \\ j\neq i}} \frac{e^{\beta_i+\beta_j}}{(1+e^{\beta_i+\beta_j})^2} + \frac{(n-1)\lambda}{n}, \qquad \text{for } 1 \leq i \leq n. \tag{9}$$

For narration convenience, we define the projection matrices for the subspace spanned by $\mathbb{1}$:

$$\mathcal{P} = n^{-1}\mathbb{1}\mathbb{1}^T, \quad \mathcal{P}_\perp = I - n^{-1}\mathbb{1}\mathbb{1}^T. \tag{10}$$

Using the notion $V(\beta)$ defined in Section 1.4, we rewrite the Jacobian matrix as

$$F'(\beta) = V(\beta) + \lambda I - \frac{\lambda}{n}\mathbb{1}\mathbb{1}^T = V(\beta) + \lambda\mathcal{P}_\perp. \tag{11}$$

By Theorem 1.1 of Hillar et al. (2012), $V(\beta)$ is globally strictly positive definite; also, $\mathcal{P}_\perp$ is semi-positive definite. Therefore, $\mathcal{L}_\lambda(\beta)$ is strictly convex, thus the solution to (7) is unique.

Next, let us discuss existence. The MLE existence conditions for the vanilla β-model (1) and its ℓ_0-regularized version (Chen et al., 2021) in existing literature assume that the degree vector d is an interior point of B, where B is solved from $d = B \cdot \mathrm{vec}(A)$ for $\mathrm{vec}(A)$ being the vectorization of A, see Proposition 2.1 in Hillar and Wibisono (2013), Rinaldo et al. (2013) and Appendix B of Chen et al. (2021). Such conditions are complicated and infeasible to verify in practice. In sharp contrast, our proposed ℓ_2-regularization provides a much cleaner guarantee of MLE existence.

LEMMA 1. *For any $\lambda > 0$ and any n, there exists a finite $\widehat{\beta}_\lambda \in \mathbb{R}^n$ such that $F(\widehat{\beta}_\lambda) = 0$.*

The MLE existence is thus guaranteed by Lemma 1 and the global strict convexity of $\mathcal{L}_\lambda(\beta)$ that we showed earlier. For the rest of this paper, we may sometimes set "$\lambda = 0$" for formula succinctness – readers may understand it as "setting λ to be a small positive value", in view of Lemma 1.

In the next subsection, we present our proposed fast algorithm to numerically solve (7).

2.2. *Fast algorithm via dimensionality reduction by degree-indexing*

The majority of existing β-model works (Yan et al., 2015, 2016; Chen et al., 2021; Stein and Leng, 2021) numerically estimate the parameters using generic GLM or optimization packages such as `glmnet`, which typically cost $O(n^3)$ computation per iteration (see Section "Cost of Computation" in Amazon H2O (2021) and Equation (4.26) in Section 4.4.1 of Hastie et al. (2009)). Consequently, they cannot scale above $O(10^3)$ nodes. Generic packages do not exploit the structure of the large and sparse but patterned design matrix under the β-model. Indeed, one can alternatively solve (7) by gradient descent or Newton's method, but the $O(n^2)$ per-iteration cost is still expensive.

Here, we present a novel algorithm that takes full advantage of (i) the structure of the β-model's likelihood function; and (ii) the widely-observed sparsity of large networks. The idea of our method stems from the following monotonicity lemma.

LEMMA 2. *The MLE $\widehat{\beta}_\lambda$ (solution to (7)) satisfies that $\widehat{\beta}_{\lambda;i} = \widehat{\beta}_{\lambda;j}$ if and only if $d_i = d_j$, for any $1 \le i < j \le n$, where $\widehat{\beta}_{\lambda;i}$ denotes the ith element of $\widehat{\beta}_\lambda$.*

We clarify that the monotonicity phenomenon was first discovered not by us, but by Hillar and Wibisono (2013) (Proposition 2.4). The proof of Lemma 2 is a close variant of its counterparts in earlier literature Hillar and Wibisono (2013); Chen et al. (2021). However, Hillar and Wibisono (2013) did not connect their monotonicity lemma to computation; and Chen et al. (2021) used it exclusively for variable selection. In contrast, we are the first to realize that Lemma 2 can greatly reduce the dimensionality of parameter estimation, which leads to a giant leap in both speed and memory efficiency.

We now describe our algorithm. Let $d_{(1)} < d_{(2)} < \cdots < d_{(m)}$ be the sorted unique values of observed degrees, where $m = |\text{Unique}(\{d_1, \ldots, d_n\})|$. For each $k \in \{1, \ldots, m\}$, let $\mathcal{D}_k := \{i_1^{(k)}, \ldots, i_{n_k}^{(k)}\} \subseteq \{1, \ldots, n\}$ collect all those nodes whose degrees equal $d_{(k)}$, namely, $d_{i_1^{(k)}} = \cdots = d_{i_{n_k}^{(k)}} = d_{(k)}$; and $d_j \neq d_{(k)}$, for all $j \notin \mathcal{D}_k$. Define $n_k := |\mathcal{D}_k|$. Then due to Lemma 2, in the MLE, all β_j's with $j \in \mathcal{D}_k$ should be set to a common value, denoted by δ_k. The objective function $\mathcal{L}_\lambda(\beta)$ can be "re-parameterized"† into a function of degree-indexed parameters $\delta := (\delta_1, \ldots, \delta_m)$, called $\widetilde{\mathcal{L}}_\lambda(\delta)$:

$$\widetilde{\mathcal{L}}_\lambda(\delta) := \mathcal{L}_\lambda(\beta) = \sum_{1 \le k < \ell \le m} n_k n_\ell \log\left(1 + e^{\delta_k + \delta_\ell}\right) + \sum_{1 \le k \le m} \frac{n_k(n_k - 1)}{2} \log\left(1 + e^{2\delta_k}\right) - \sum_{k=1}^m n_k d_{(k)} \delta_k + \frac{\lambda}{2} \sum_{k=1}^m n_k (\delta_k - \widetilde{\delta})^2, \tag{12}$$

where $\widetilde{\delta} := n^{-1} \sum_{k=1}^m n_k \cdot \delta_k$ is the weighted average, conceptually analogous to $\bar{\beta}$ under the original parameterization. The gradient of $\widetilde{\mathcal{L}}_\lambda(\beta)$ is

$$G_k(\delta) := \frac{\partial \widetilde{\mathcal{L}}_\lambda(\beta)}{\partial \delta_k} = \sum_{\substack{1 \le \ell \le m \\ \ell \neq k}} n_k n_\ell \frac{e^{\delta_k + \delta_\ell}}{1 + e^{\delta_k + \delta_\ell}} + n_k(n_k - 1) \frac{e^{2\delta_k}}{1 + e^{2\delta_k}} - n_k d_{(k)} + n_k(\delta_k - \widetilde{\delta})\lambda, \tag{13}$$

and its Jacobian matrix, denoted by J, is

$$J_{k\ell}(\delta) = \frac{\partial^2 \widetilde{\mathcal{L}}_\lambda(\delta)}{\partial \delta_k \partial \delta_\ell} = n_k n_\ell \frac{e^{\delta_k + \delta_\ell}}{\left(1 + e^{\delta_k + \delta_\ell}\right)^2} - \frac{n_k n_\ell}{n} \lambda, \quad \text{for } 1 \le \{k \neq \ell\} \le m, \tag{14}$$

$$J_{kk}(\delta) = \frac{\partial^2 \widetilde{\mathcal{L}}_\lambda(\delta)}{\partial \delta_k^2} = \sum_{\substack{1 \le \ell \le m \\ \ell \neq k}} n_k n_\ell \frac{e^{\delta_k + \delta_\ell}}{\left(1 + e^{\delta_k + \delta_\ell}\right)^2} + 2n_k(n_k - 1) \frac{e^{2\delta_k}}{\left(1 + e^{2\delta_k}\right)^2} + n_k\left(1 - \frac{n_k}{n}\right)\lambda, \quad \text{for } 1 \le k \le m. \tag{15}$$

†We put quotation marks around *re-parameterized*, because the mapping between β_i's and δ_k's is data-dependent – this is different from the common definition of "reparameterization".

We do not need to separately study the existence and uniqueness of $\widehat{\delta}_\lambda := \arg\min_\delta \widetilde{\mathcal{L}}_\lambda(\delta)$, because (1) minimizing $\widetilde{\mathcal{L}}_\lambda(\delta)$ is the same as minimizing $\mathcal{L}_\lambda(\beta)$ under the additional constraint introduced by Lemma 2; and (2) the optimal solution $\widehat{\beta}_\lambda$ to the unconstrained problem $\min_\beta \mathcal{L}_\lambda(\beta)$ exists, is unique (recall Section 2.1), and satisfies that constraint. Therefore, the $\widehat{\delta}_\lambda$, which corresponds to $\widehat{\beta}_\lambda$, would also be the unique optimal solution to $\min_\delta \widetilde{\mathcal{L}}_\lambda(\delta)$.

Our new algorithm is particularly effective in handling *large* and *sparse* networks, where the total number of different node degrees m is much smaller than the network size n. Its computational complexity is bottle-necked by the first step: computing all node degrees, which would cost $O(\rho_n \cdot n^2)$ time, where recall that ρ_n represents network sparsity. Then each iteration in the gradient method would cost $O(m^2)$. All these are much cheaper than the $O(n^3)$ per-iteration cost in existing literature.

The data set Steenwinckel et al. (2020) provides a striking example of the scalability of our method. It contains $n \approx 1.3$ million nodes, but only $m = 459$ different degrees. Many nodes share common low degrees, see Table 2. Our algorithm makes it feasible to run a Newton's method on this network of seemingly prohibitive size.

Degree	1	2	3	4	5
Number of nodes with this degree	684003	132123	48126	24586	15189
Percentage of nodes, unit: %	52.45	10.13	3.69	0.19	0.12

Table 2: Frequencies of low degrees in the data set Steenwinckel et al. (2020), $n = 1304155$.

We conclude this subsection by clarifying that the "re-parameterization" (12)–(15) is employed only for computation, namely, finding the MLE. Throughout the theoretical analysis in the next Section, we would always stick to the MLE $\widehat{\beta}_\lambda$ under the original parameterization.

3. Theory and theory-based statistical inference

Throughout this section, we assume the true parameters satisfy $\beta^* \in \mathcal{S}_1 = [a_1^* \log n - M^*, a_2^* \log n + M^*]^n$ for some constants $a_1^* < a_2^*$ and $M^* > 0$. To better connect to existing β-model literature, we define the following shorthand.

Definition 1. *For the true parameter value β^*, define*

$$b_n = \max_{1\le i<j\le n} \frac{(1+e^{\beta_i^*+\beta_j^*})^2}{e^{\beta_i^*+\beta_j^*}}, \quad c_n = \min_{1\le i<j\le n} \frac{(1+e^{\beta_i^*+\beta_j^*})^2}{e^{\beta_i^*+\beta_j^*}},$$
$$\text{and} \quad q_n = \left\{ \max_{1\le i\le n} (n-1)^{-1} \sum_{\substack{1\le j\le n \\ j\neq i}} \frac{e^{\beta_i^*+\beta_j^*}}{(1+e^{\beta_i^*+\beta_j^*})^2} \right\}^{-1}. \tag{16}$$

By definition, b_n^{-1} and c_n^{-1} are minimum and maximum edge probabilities, and $(n-1)q_n^{-1}$ is the maximum expected degree. Clearly, $b_n^{-1} \le q_n^{-1} \le c_n^{-1}$. In this paper, we do not discuss very dense networks where edges are nearly all 1's, not only due to that most real-world networks are sparse, but also because we can model $\{1 - A_{i,j}\}_{1\le\{i,j\}\le n}$ as a sparse network. Formally:

Assumption 1. *Zero edges are not sparse, that is, $c_n^{-1} \le 1 - C_0$ for some constant $C_0 \in (0,1)$.*

To simplify the presentation of our main results, we make the following assumption for convenience.

Assumption 2. *There exists at least* four *distinct nodes $i_1, i_2, j_1, j_2 \in [1:n]$, such that* $\max\left(|\beta_{i_1}^* - a_1^* \log n|, |\beta_{i_2}^* - a_1^* \log n|, |\beta_{j_1}^* - a_2^* \log n|, |\beta_{j_2}^* - a_2^* \log n|\right) = O(1)$.

The sole purpose of Assumption 2 is to reduce the number of symbols in theorems and ease reading. It implies that $b_n \asymp \widetilde{b}_n$, where $\widetilde{b}_n \asymp \sup_{\beta_1,\beta_2\in[a_1^*\log n+C_1, a_2^*\log n+C_2]} (1+e^{\beta_1+\beta_2})^2/e^{\beta_1+\beta_2}$ for any constants C_1, C_2, so that we will not need to set up a separate symbol $\widetilde{b}_n$ when presenting theoretical results, especially those places that may otherwise involve both b_n and $\widetilde{b}_n$. The same goes for c_n.

3.1. Consistency and finite-sample error bounds

Following the convention of the β-model literature, we first present an ℓ_∞ bound.

THEOREM 1. *Define*

$$\Gamma(n,\ \beta^*;\lambda) := \frac{\log^{1/2} n}{b_n^{-1}n} \cdot \frac{\lambda}{b_n^{-1}n+\lambda} + \frac{(q_n^{-1}n\log n)^{1/2} + \|\mathcal{P}_\perp\beta^*\|_\infty\lambda}{b_n^{-1}n+\lambda} \cdot \Big(1 + \frac{b_n}{q_n} \cdot \frac{\|\mathcal{P}_\perp\beta^*\|_\infty\lambda}{b_n^{-1}n+\lambda}\Big). \tag{17}$$

Assume that

$$(b_n/q_n) \cdot \Gamma(n,\beta^*;\lambda)\ \le 1/20. \tag{18}$$

Then our ℓ_2-regularized MLE $\widehat{\beta}_\lambda$ satisfies

$$\mathbb{P}\Big\{\|\widehat{\beta}_\lambda - \beta^*\|_\infty \le C_1\Gamma(n,\beta^*;\lambda)\Big\} \ge 1 - n^{-C_2} \tag{19}$$

for some constants $C_1, C_2 > 0$, where C_1 depends on C_2.

Theorem 1 holds for all choices of λ, as long as (18) is satisfied. To better parse it, we discuss two representative choices of λ. First, we set $\lambda = O(1)$. Recall that our Lemma 1 guarantees the existence of MLE for any $\lambda > 0$, so $\widehat{\beta}_\lambda$ is well-defined. Then the assumption (18) can be simplified to $(b_n^4/q_n^3)\cdot n^{-1}\log n \to 0$, and the error bound becomes $\Gamma(n,\beta^*;\lambda) \asymp \{(b_n^2/q_n)\cdot n^{-1}\log n\}^{1/2}$. To further simplify, we can consider the set up of Chen et al. (2021), where $\beta_i^* = (-\gamma\log n + \mu^\dagger)/2 + \mathbb{1}_{[i\in\mathcal{S}]}(\alpha\log n + \mathfrak{b}_i^\dagger)$, where $\mathcal{S}\subset[1:n]$ is the "active set" satisfying $|\mathcal{S}| \ll n^{1-\alpha}$, and $\vee_{i=1}^n|\mathfrak{b}_i^\dagger| \vee |\mu^\dagger| \ll \log n$.

Now we compare our Theorem 1 with $\lambda = O(1)$ to Chen et al. (2021) and Stein and Leng (2020) and elaborate the improvement. Under the setting of Chen et al. (2021), our assumption (18) becomes $\gamma + 3\alpha < 1$. Now $\rho_n \asymp n^{-\gamma}$ represents network sparsity, this says that when β_i^*'s are sufficiently homogeneous, our Theorem 1 can handle networks as sparse as $\rho_n \gg n^{-1}$ – this is a widely-recognized minimal sparsity assumption across many different topics in network literature Bickel and Chen (2009); Zhao et al. (2012); Zhang and Xia (2022).

In contrast, Theorem 1 and Lemma 2 in Chen et al. (2021) not only assume *knowing the true values of* γ and α‡ (in fact, γ is the bottle neck of estimation accuracy – compare the error rates of γ and that of the individualism parameters in Theorem 1 of Stein and Leng (2020)), but they also additionally assume that $\rho_n \gg n^{-1/2}$§. Stein and Leng (2020) makes a much stronger assumption that $\rho_n \gtrsim n^{-1/6}$¶. Therefore, our Theorem 1 significantly improves over these state-of-the-art results.

In general, our Theorem 1's assumption demands a compromise between network sparsity and degree heterogeneity. This is understandable, since the presence of high degree discrepancy makes Jacobian matrix $F'(\beta^*)$ ill-conditioned thus deteriorates the local convexity of the likelihood around β^*, making parameter estimation harder. In some special settings, such as Chen et al. (2021), where there are essentially only two kinds of nodes: hub nodes with $\beta_i \approx (\alpha-\gamma/2)\log n$ and peripheral nodes with $\beta_j \approx -\gamma/2\cdot\log n$, ad-hoc strategies such as estimating high- and low-degree nodes separately may effectively patch up this issue, but in general, high heterogeneity remains a major challenge for β-model studies.

Next, we briefly discuss the behavior of $\widehat{\beta}_\lambda$ when $\lambda\to\infty$. Theorem 1 of Stein and Leng (2020) shows an error bound that grows *quadratically* with λ in terms of the L_1 norm. In contrast, our error bound in terms of the L_∞ norm gives $\lim_{\lambda\to\infty}\Gamma(n,\beta^*;\lambda) = b_n\log^{1/2}/n + \|\mathcal{P}_\perp\beta^*\|_\infty\cdot\{1 + (b_n/q_n\cdot\|\mathcal{P}_\perp\beta^*\|_\infty)\}$ if $\lambda \gg n/b_n$. The first term approximately matches the error bound in fitting an Erdos-Renyi model, agreeing with the intuitions that large λ penalizes parameters to an Erdos-Renyi graph. The second error term is due to regularization.

‡Notice that in their theoretical analysis, the MLE procedure does not optimize over γ and α.

§To see this, notice that the assumption Equation (7) in their Lemma 2 implies $\alpha+\gamma\le 1/2$.

¶For sparse networks where $\rho_n\to 0$ (ρ_n is called $\rho_{n,0}$ in their notation) their Equation (7) implies $K_n^{-1} \asymp e^{3r_{n,0}} \asymp \rho_{n,0}^3 = \rho_n^3$ when there is no excess risk. Then if $\rho_n \ll (\log n/n)^{1/6}$, the only network model that satisfies their Assumption 2 is Erdos-Renyi, which is uninteresting in the context of β-model studies.

Based on Theorem 1, when Assumption (18) is satisfied, our method can identify node i's whose true β_i^* is significantly different from the majority. In the spirit of Chen et al. (2021) and Chen et al. (2021), this can be called as "*sparsistency*" or variable selection consistency (Li et al., 2015). In practice, to carry out this post-estimation thresholding, we need to estimate q_n and b_n. To simplify narration, in this part of the paper, we are working under the β-model with the β-sparsity from Chen et al. (2021). The following corollary provides a theoretically guaranteed algorithm to carry out a post-estimation variable selection for our ℓ_2-regularized MLE.

COROLLARY 1 (SPARSISTENCY OF POST-ESTIMATION THRESHOLDING). *Consider the setting of Chen et al. (2021) in (3). Then under the assumptions of our Theorem 1 (with $\lambda = O(1)$) and the assumption from Chen et al. (2021) that $|\mathcal{S}| \ll n^{1-\alpha}$, when $\alpha > 0$, the following algorithm:*

- *Step 1: Run our ℓ_2-regularized MLE with a $\lambda = O(1)$ to produce $\widehat{\beta}_\lambda$;*
- *Step 2: Take the subset of nodes whose degrees are the middle ζ_0 proportion. Here ζ_0 can be any fixed number in $(0,1)$, such as $\zeta_0 = 50\%$. Let A_{ζ_0} denote the $(\zeta_0 n) \times (\zeta_0 n)$ submatrix induced by the selected nodes. Let $\bar{A}_{\zeta_0}$ denote $\binom{\zeta_0 n}{2}^{-1} \sum_{1 \le i < j \le n} (A_{\zeta_0})_{i,j}$. Set $\widehat{b}_n = \bar{A}_{\zeta_0}(1 - \bar{A}_{\zeta_0})$;*
- *Step 3: Output $\widehat{\mathcal{S}} = \{j : |\widehat{\beta}_{\lambda;j} - \log\{\bar{A}_{\zeta_0}/(1-\bar{A}_{\zeta_0})\}| > \{\widehat{q}_n^{-1} \log n/(\widehat{b}_n^{-2} n)\}^{1/2}\}$;*

outputs an estimator $\widehat{\mathcal{S}}$ that enjoys the sparsistency guarantee: $\mathbb{P}(\widehat{\mathcal{S}} = \mathcal{S}) \to 1$.

REMARK 1. *Now we compare our Corollary 1 to the counterparts in Chen et al. (2021) and Stein and Leng (2020). The main advantage of our approach compared to Chen et al. (2021) is computational speed. Chen et al. (2021) treats the estimation of $\mathcal{S}$ as a model selection problem; they therefore fit the model many times at different sparsity levels. In sharp contrast, our new theoretical results reveals a new and different understanding that in fact, the MLE with a* small *amount of regularization can achieve sparsistency, without going through model selection, thus saving a lot of computation. Notice that our argument here is not based on our fast degree-indexed estimation algorithm, because we expect our dimension reduction technique can also be applied to Chen et al. (2021), though it still needs some adaptation to carefully handle ties in observed node degrees. We notice that Stein and Leng (2020) did not discuss estimation of $\mathcal{S}$, possibly because they studied a more general β-model where the β_i^*'s for those $i \in \mathcal{S}$ can have substantive differences from each other. We believe that some separation conditions such as those we assumed in Corollary 1 ("$\alpha > 0$") and the "min-β condition" in Chen et al. (2021) would be necessary to discuss sparsistency in the setting of Stein and Leng (2020).*

Next we present our ℓ_2 error bound.

THEOREM 2. *Set $\lambda > 0$. There exists constants $C_1, C_2 > 0$, where C_1 depends on C_2, such that with probability at least $1 - n^{-C_1}$, we have*

(i) Set $\lambda = O(1)$. Suppose we perform a constrained MLE with $\widehat{\beta}_\lambda \in [a_1 \log n - M, a_2 \log n + M]^n$ for some $M > 0$ and a_1, a_2 satisfying $a_1 \le a_1^* \le a_2^* \le a_2$. Then we have

$$n^{-1/2} \cdot \|\widehat{\beta}_\lambda - \beta^*\|_2 \le C_2 \{(b_n')^2/q_n \cdot n^{-1} \log n\}^{1/2}, \tag{20}$$

where $b_n' := \max\{e^{2a_1}/(1+e^{2a_1})^2, e^{2a_2}/(1+e^{2a_2})^2\}$.

(ii) When $\lambda \gtrsim q_n^{-1} n$ with a large enough constant factor, we have

$$n^{-1/2} \cdot \|\widehat{\beta}_\lambda - \beta^*\|_2 \lesssim \frac{(q_n \cdot n \log n)^{1/2} + n^{-1/2} \|\mathcal{P}_\perp \beta^*\|_2 \lambda}{b_n^{-1} n + (b_n/c_n)^{-2} \lambda} + \left\{ \frac{(c_n/q_n)^2 \cdot (\log n/n^2) \lambda}{b_n^{-1} n + (b_n/c_n)^{-2} \lambda} \right\}^{1/2}$$
$$\overset{\lambda \to \infty}{\longrightarrow} (b_n/c_n)^2 \cdot n^{-1/2} \|\mathcal{P}_\perp \beta^*\|_2 + (b_n/q_n) \cdot n^{-1} \log^{1/2} n. \tag{21}$$

Part (i) of Theorem 2 has a similar flavor to the counterparts in Chen et al. (2021) and Stein and Leng (2020), where they optimized the model parameter β over a subset Θ_0 of the parameter space Θ. In Part (i) of our Theorem 2, one can understand Θ_0 to be $[a_1 \log n - M, a_2 \log n + M]^n$. Under the settings of Chen et al. (2021) and (exactly similarly) Stein and Leng (2020), we can accurately estimate b_n and q_n as described in Remark 1. Under our setting, if we additionally assume that only a small fraction of β_i^*'s are distinct from the common value in β^*, then one can accurately estimate the common parameter using the same method as in Remark 1. It will lead to consistent estimates for b_n and q_n. As a result, our error bound do not have the *excess risk* (like did in Chen et al. (2021) and Stein and Leng (2020)) caused by using a working (a_1, a_2) that does not satisfy the assumed relationship with a_1^* and a_2^*.

Under this β-sparsity assumption as that in Chen et al. (2021) and Stein and Leng (2020), we can think that effectively $b_n' = b_n$. Part (i) of Theorem 2 requires significantly weaker assumptions compared to the $\lambda = O(1)$ case of Theorem 1. This is not surprising, since $n^{-1/2}\|u\|_2 \le \|u\|_\infty$. On the other hand, the result is also weaker than Theorem 1. This is particularly noticeable if we compare their Part (i)'s: Theorem 1's Part (i) leads to sparsistency (discussed in Remark 1); whereas Theorem 2's Part (i) does not. On the other side, both theorems agree on the intuition that some structural homogeneity assumptions might be necessary to achieve consistent parameter estimation in very sparse networks. This echoes the similar opinion implicitly expressed in the main theorems of Chen et al. (2021) and Stein and Leng (2020).

3.2. *Lower bounds*

Readers may naturally wonder how tight the error bounds in Section 3.1 are. Existing β-model literature contains little study in this regard. In this section, we provide local lower bounds in ℓ_p $(p = 0, 1, 2)$ norms. Here "local" means that our lower bounds consider estimators that search in a small ℓ_∞ neighborhood around the true β^*. We present local lower bounds with a "Cramer-Rao flavor", rather than conventional minimax lower bounds that consider a much larger parameter space, because the problem's difficulty critically depends on the value of β^*. This is very different from the classical problem of estimating a population mean from i.i.d. observations, in which, the difficulty of the problem does not depend on the value of the population mean.

Since throughout this paper, we are constantly most interested in studying β^*'s with β-sparsity (Chen et al., 2021), in this section, we study the following setting:

$$\beta_i^* = \begin{cases} -\gamma/2 \log n + O(1), & \text{if } i \notin \mathcal{S}, \\ (\alpha_i - \gamma/2) \log n + O(1), & \text{if } i \in \mathcal{S}. \end{cases}, \tag{22}$$

where $\alpha_i > 0$ for all $i \in \mathcal{S}$ and suppose $\alpha = \max_{i \in \mathcal{S}} \alpha_i > 0$ is a constant that satisfies all the conditions that Chen et al. (2021) imposed on their (common) α. Compared to (3), we made two changes in (22). First, like Stein and Leng (2020), we allow α to be different between nodes: in (22), they become α_i's. Second, for cleanness of presentation, we change the $o(\log n)$ part in the definition of β^* in (3) to $O(1)$. This is an illustrative special case that both Stein and Leng (2020) and Chen et al. (2021) carefully studied (see pages 11–12 of Stein and Leng (2020)).

Now, we formally state our main results in this subsection. Here, we focus on the ℓ_2 norm. For a better comparison, we will also state an upper bound that *matches* the lower bound.

THEOREM 3 (LOCAL LOWER BOUND AND MATCHING UPPER BOUND IN ℓ_2 NORM). *Consider the setting (22). Assume $\alpha = \max_i \alpha_i$ satisfies the assumptions $|\mathcal{S}| \ll n^{1-\alpha}$ from Chen et al. (2021); and $\min_{i \in \mathcal{S}} \alpha_i > \alpha_0$ for some constant $\alpha_0 > 0$. Let $\mathcal{B}(\|\cdot\|, r)(\beta')$ denote the closed ball in the Banach space equipped with $\|\cdot\|$, centered at β' with radius r. Then we have*

$$\inf_{\widehat{\beta}} \sup_{\beta_0 \in \mathcal{B}_{\|\cdot\|_\infty, M_\beta}(\beta^*)} n^{-1/2} \cdot \|\widehat{\beta} - \beta_0\|_2 \gtrsim (b_n^{-1} n)^{-1/2}, \tag{23}$$

where $M_\beta > 0$ is an arbitrary constant.

Under the same setting, we can obtain $\widehat{\beta}_\lambda$ *by our* ℓ_2*-regularized MLE constrained on* $[a_1 \log n - M, a_2 \log n + M]$ *with a large enough constant* $M > 0$ *and* (a_1, a_2) *evaluated by the method described in Remark 1 and* $\lambda = O(1)$*. Then* $\widehat{\beta}_\lambda$ *achieves a nearly matching upper bound*

$$n^{-1/2} \cdot \|\widehat{\beta}_\lambda - \beta^*\|_2 \lesssim (b_n^{-1} n)^{-1/2} \cdot \log^{1/2} n. \tag{24}$$

Theorem 3 shows that our estimator $\widehat{\beta}_\lambda$ is locally rate-optimal in ℓ_2 norm for estimating a "β-sparse" true β^* that conforms to (22). Parallel local lower bounds in other norms can be built by slightly varying the proof of Theorem 3. Specifically, (23) continues to hold if $n^{-1/2}\|\cdot\|_2$ is replaced by $n^{-1}\|\cdot\|_1$ or $\|\cdot\|_\infty$. But currently we are yet unable to establish matching upper bounds.

Theorem 3 is arguably among the first few of its kind in the β-literature. The most closely related results to our best knowledge are Wahlström et al. (2017) and Lee and Courtade (2020). Section III.A of Wahlström et al. (2017) presents a marginal Cramer-Rao bound of $(\sum_{j\neq i} \mathbb{E}[A_{i,j}])^{-1}$ on $\widehat{\beta}_i$ for a *fixed* index i with *repeated* network observations generated by the same true β^*. Our Theorem 3 is therefore a much stronger result than Wahlström et al. (2017). Lee and Courtade (2020) can be compared to the following conventional (i.e. non-local) lower bound result.

THEOREM 4 (NON-LOCAL LOWER BOUNDS). *Recall the definition of the parameter space* $\mathcal{S}_1$ *from the beginning of this section. Define* $b_{\mathcal{S}_1} = \sup_{\beta\in\mathcal{S}_1} \max_{1\le i<j\le n}(1 + e^{\beta_i+\beta_j})^2/e^{\beta_i+\beta_j}$*. We have*

$$\inf_{\widehat{\beta}} \sup_{\beta^*\in\mathcal{S}_1} n^{-1/2} \cdot \mathbb{E}[\|\widehat{\beta} - \beta^*\|_2] \gtrsim (b_{\mathcal{S}_1}^{-1} n)^{-1/2}. \tag{25}$$

Moreover, (25) remains valid if $n^{-1/2}\|\cdot\|_2$ *is replaced by* $n^{-1}\|\cdot\|_1$ *or* $\|\cdot\|_\infty$*.*

Define $c_{\mathcal{S}_1} = \inf_{\beta\in\mathcal{S}_1} \min_{1\le i<j\le n}(1 + e^{\beta_i+\beta_j})^2/e^{\beta_i+\beta_j}$. Using Stein and Leng (2020), we see that Theorem 7 in Lee and Courtade (2020) established a $C_{\mathcal{S}_1} n^{-1}$ lower bound on $n^{-1} \cdot \mathbb{E}[\|\widehat{\beta} - \beta^*\|_2^2]$, much looser than our (25). Also, their proof seems to specifically cater to ℓ_2 norm, and possible adaptation of their analysis for ℓ_1 and ℓ_∞ norms is unclear. Lee and Courtade (2020) also remarks that currently no matching upper bound is known for GLM's. In view of Theorem 4, our Theorem 3 may be an informative starting point for future study to investigate the open challenge of rate-optimal estimation for the larger (non-local) parameter space $\mathcal{S}_1$.

3.3. *High-dimensional asymptotic normality*

We present our novel asymptotic normality results. We discuss two representative cases: $\lambda = O(1)$ and $\lambda \to \infty$ as in Theorem 2. We discover that small and large λ's yield very different asymptotic dependency structures in $\widehat{\beta}_\lambda - \beta^*$. Therefore, it might be very difficult to build a unified asymptotic distribution theorem for any choice of λ.

THEOREM 5 (ASYMPTOTIC NORMALITY). *Suppose index set* $\mathcal{J} \subset \{1, \ldots, n\}$ *satisfies*

$$|\mathcal{J}| \ll c_n (b_n^{-1} n)^{1/2}. \tag{26}$$

Under the conditions of Theorem 1, we have

(i). Setting $\lambda = O(1)$*. Suppose*

$$\frac{b_n^3}{q_n^{5/2} n^{1/2}} \cdot \log n \to 0, \tag{27}$$

$$\frac{c_n^{-1}}{c_n^{-1} + b_n^{-1}(n-2)} \cdot \log n \ll \frac{q_n^2}{b_n^2}, \tag{28}$$

and define $D(\beta^*)$ *to be the diagonal matrix of* $V(\beta^*)$*. Then we have*

$$(\widehat{\beta}_\lambda - \beta^*)_{\mathcal{J}} \xrightarrow{d} N\Big[0, \{D(\beta^*)\}_{\mathcal{J},\mathcal{J}}^{-1}\Big]. \tag{29}$$

(ii). Setting $\lambda \gg n$. We have

$$\widehat{\beta}_\lambda - \frac{\mathbb{1}^T V(\breve{\beta})\beta^*}{\mathbb{1}^T V(\breve{\beta})\mathbb{1}} \cdot \mathbb{1} \xrightarrow{d} N\left(0, \text{Variance} = \frac{n(n-1)\mathbb{1}^T V(\beta^*)\mathbb{1}}{2\{\mathbb{1}^T V(\breve{\beta})\mathbb{1}\}^2}\right) \cdot \mathbb{1}, \tag{30}$$

where we define $\{V(\breve{\beta})\}_{i,j}$ as

$$\{V(\breve{\beta})\}_{i,j} = \frac{\bar{A} - \frac{e^{\beta_i^*+\beta_j^*}}{1+e^{\beta_i^*+\beta_j^*}}}{\log\frac{\bar{A}}{1-\bar{A}} - \beta_i^* - \beta_j^*} \quad \textit{for all } 1 \le \{i \ne j\} \le n, \tag{31}$$

and $\{V(\breve{\beta})\}_{i,i} = \sum_{j:1\le j\le n, j\ne i}\{V(\breve{\beta})\}_{i,j}$.

Parts (i) and (ii) of Theorem 5 illustrate the distinct asymptotic covariance structures in $\widehat{\beta}_\lambda$ with different λ's. Part (i) finds $\widehat{\beta}_\lambda$'s elements to be asymptotically independent for $\lambda = O(1)$; whereas Part (ii) discovers that $\widehat{\beta}_\lambda$'s elements will have correlation approaching 1 with a fast growing λ. This result is not surprising – to understand it, we can consider the special example of $\beta^* /\!/ \mathbb{1}$. In this case, $V(\beta^*)$ is proportional to the $(n-2)I + \mathbb{1}\mathbb{1}^T$. Thus the off-diagonal elements are at the order of $O(n^{-1})$ of the diagonal elements in $\{V(\beta^*)\}^{-1}$, matching the conclusion of (29).

Theorem 5 provides the first quantitative characterization for the large λ scenario. Notice that the conclusion of Part (ii) of Theorem 5 is nontrivial – although it is qualitatively clear that a large λ makes the intercept close to being parallel to $\mathbb{1}$, this intuition implies nothing about the dependency structure of $\widehat{\beta}_\lambda$ around its limiting center, since each element $\widehat{\beta}_{\lambda;i}$ might have positive or negative stochastic variations. Our Theorem 5 shows that asymptotically their randomness are also synchronized. This is verified by our numerical experiments. Theorem 5 is also the first *high-dimensional* (we allow $|\mathcal{J}|$ to grow with n) asymptotic normality result in the β-model literature; all other existing papers only study fixed-dimensional asymptotic normality (Yan and Xu, 2013; Yan et al., 2016; Fan et al., 2020a; Chen et al., 2021; Stein and Leng, 2020).

4. An AIC-type criterion for data-driven λ selection

Selecting the tuning parameter λ in our method is nontrivial. Notice that the type of regularization in our paper is different from ℓ_0 sparsity, so the convenient BIC variable selection in Chen et al. (2021) and Stein and Leng (2021) is not available in our setting.

In this section, we propose a novel AIC-type tuning criterion for selecting λ. Recall that the β-model has a logistic regression form with a design matrix $X \in \{0,1\}^{\binom{n}{2}\times n}$. Our approach is inspired by the AIC criterion for logistic regression. By Stein and Leng (2020), the design matrix, denoted by X, is $X = (X_{1,2}, \cdots, X_{i,j}, \cdots, X_{n-1,n})^T$, where $X_{i,j} \in \{0,1\}^{n\times 1}$ with ones on the ith and jth places and zeros everywhere else. Following Section 1.8.1 of van Wieringen (2015), we regard the trace of the hat matrix $H(\lambda)$ for GLM (see Equation (12.3) in McCullagh and Nelder (2019) and Equations (2.4.4), (2.4.7) and (2.4.13) in Lu et al. (1997)) as the effective dimensionality of the model and the AIC criterion. We have

$$H(\lambda) = \text{Tr}[\mathcal{W}^{1/2} X \Big(X^T \mathcal{W} X + \lambda I_{\binom{n}{2}\times\binom{n}{2}}\Big)^{-1} X^T \mathcal{W}^{1/2}], \tag{32}$$

where $\mathcal{W}$ is a diagonal matrix defined by $\mathcal{W}_{(i,j),(i,j)} = \{V(\beta^*)\}_{i,j}$. Next, we simplify (32). By definition, one can verify that $X^T\mathcal{W}X = V(\beta^*)$. We have

$$\begin{aligned} H(\lambda) &= \text{Tr}\{(X^T\mathcal{W}X + \lambda I)^{-1} X^T\mathcal{W}X\} = \text{Tr}\{\big(I + \lambda\{V(\beta^*)\}^{-1}\big)^{-1}\} \\ &\le n\lambda_{\max}\{\big(I + \lambda\{V(\beta^*)\}^{-1}\big)^{-1}\} = \frac{n}{\lambda_{\min}\big(I + \lambda\{V(\beta^*)\}^{-1}\big)} = \frac{n}{1 + \lambda\cdot\{\lambda_{\max}(V(\beta^*))\}^{-1}}. \end{aligned} \tag{33}$$

To ease computation, we make some unrigorous approximation to further simplify (33). First, in light of the observation that the upper bounds for $\|V(\beta^*)\|_\infty$ and $\|V(\beta^*)\|_{\text{op}}$ in Hillar et al. (2012) are similar, we roughly replace $\lambda_{\max}(V(\beta^*))$ by $\|V(\beta^*)\|_\infty = q_n^{-1}(n-1)$. Then we further replace $q_n^{-1}(n-1)$ by its estimator $d_{\max} = \max_{1\le i\le n} d_i$. This leads to our proposed AIC criterion:

$$\text{AIC}(\lambda) = \frac{nd_{\max}}{d_{\max}+\lambda} + \mathcal{L}_\lambda(\widehat{\beta}_\lambda). \tag{34}$$

Our proposed AIC-type criterion shows promising performance and usefulness under various settings. For empirical evidences, please refer to the simulation results in Section 5.3 and data applications in 6.1 and 6.2.

5. Simulations

5.1. *Simulation 1: Convergence and consistency of our method and validation of our theory on $\widehat{\beta}_\lambda$'s asymptotic normality*

Our first simulation checks the correctness of our method and the verifies our theory's prediction. Set

$$\beta_1^* = \cdots = \beta_{[n/5]}^* = \widetilde{\gamma}\log n, \quad \beta_{[n/5]+1}^* = \cdots = \beta_n^* = \widetilde{\alpha}\log n.$$

We consider four different configurations (i) – (iv) for assessing the performance of our method on networks with different sparsity levels. The first two simulation settings (i) and (ii) consider

Setting	(i)	(ii)	(iii)	(iv)	(v)	(vi)
True $\widetilde{\gamma}$	$-1/3$	$-1/2$	$-2/3$	$-2/3$	$-2/3$	$-1/3$
True $\widetilde{\alpha}$	$1/5$	$1/5$	$1/3$	$1/3$	$1/3$	$-1/3+0.05$
Working λ	0.1	0.1	0.1	10	200	$2n$
Network sparsity	Dense $\longleftrightarrow$ Sparse					
Degree heterogeneity	Low $\longleftrightarrow$ High					Small $\|\mathcal{P}_\perp\beta^*\|_\infty$
Regularization λ	Small $\longleftrightarrow$ Large					

Table 3: Set up for Simulation 1

relatively dense networks, where we impose a small λ value; whereas the other two settings (iii) and (iv) consider a sparser network, with small and large λ's, respectively. We record the following aspects of measurements: (1) convergence speed: we record $n^{-1/2}\|\widehat{\beta}^{(t)} - \widehat{\beta}_\lambda\|_2$ versus iteration, where $\widehat{\beta}^{(t)}$ is the estimated $\widehat{\beta}$ at the tth iteration; (2) relative error $\|\widehat{\beta}_\lambda - \beta^*\|_2/\|\beta^*\|_2$: we record and present the change of log-relative error as a $\log n$ increases linearly; and (3) computation time. we implement our accelerated algorithm in Section 2.2, equipped with gradient and Newton's methods, respectively.

Figure 1 reports the simulation results. Row 1 shows that as expected, Newton's method typically converges much faster than gradient descent. Row 2 shows the decaying speed of the relative error $\|\widehat{\beta}_\lambda - \beta^*\|_2/\|\beta^*\|_2$. In most cases, the two methods output the same result. The problem becomes more difficult as one travels from setting (i) to setting (iii), and the error readings from Y-axis in plots in Row 2 confirm this understanding. Row 3 confirmed the theoretical $O(\rho_n n^2)$ computational complexity of our method. Another observation is that although Newton's method needs much less iterations than the gradient method, each of its iteration requires a matrix inversion. Overall it does not appear much faster.

Next, we show simulation plots that validate our Theorem 5. We first check the intercept term. For this purpose, we plot the one-dimensional marginal empirical distributions of $\widehat{\beta}_{\lambda;1}$ under our simulation settings (i) and (vi) in Table 3, respectively. We range $n \in \{100, 400, \cdots, 6400\}$. The results are reported in Figure 2. Row 1 of Figure 2 corresponds to dense networks, where we set a small $\lambda = 0.1$; and in row 2, we set a large $\lambda = 2n$. The results verify that our prediction of the intercept in Theorem 5 is accurate.

Next, we validate our Theorem 5's prediction of the variance term. Then, we illustrate the dependency structure between $\widehat{\beta}_{\lambda;n-1}$ and $\widehat{\beta}_{\lambda;n}$ under our settings (iii) – (v). The simulation results

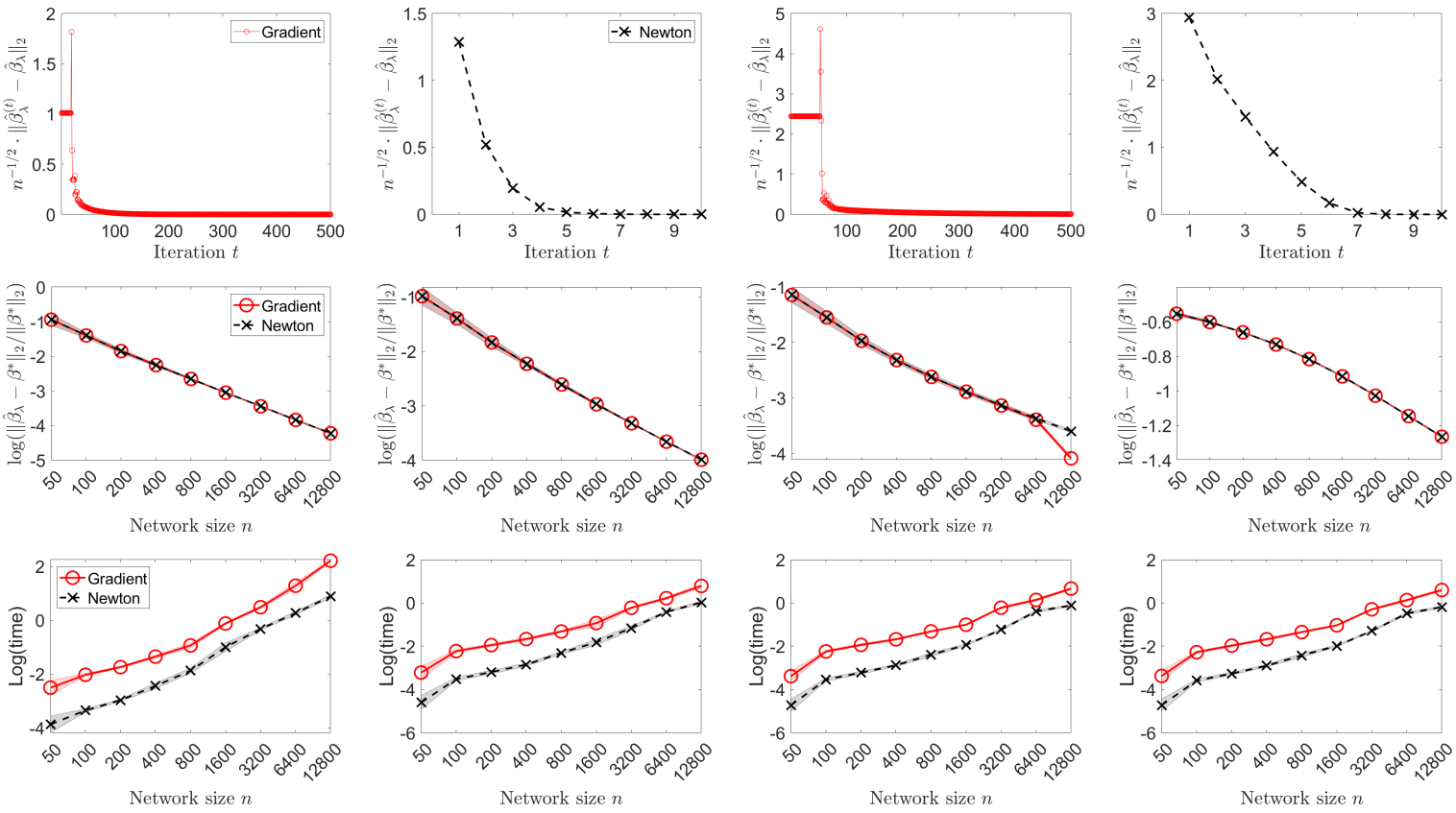

Fig. 1: Row 1: convergence speed, setting (i), plots 1 & 2: $n = 50$, 3 & 4: $n = 12800$; plots 1 & 3: gradient, 2 & 4: Newton; Row 2: log average ℓ_2 errors, from left: settings (i)–(iv); Row 3: computation time, from left: settings (i)–(iv).

presented in Figure 3 provide clear empirical evidences that well-match our novel normality result Theorem 5. Traveling from the left to the right in Row 1 of Figure 3, as λ increases with other settings fixed, we see an increasing correlation between $\widehat{\beta}_{\lambda;n-1}$ and $\widehat{\beta}_{\lambda;n}$, exactly as Part (ii) of our Theorem 5 predicts. Traveling from the right-most plot in Row 1 and then continue from the left in Row 2, we are increasing n with a fixed λ. This is roughly equivalent to decreasing λ and traveling from large λ back to small λ. In this comparison sequence, we clearly see an increasing independence, which again confirms the prediction of our Theorem 5.

5.2. Simulation 2: performance comparison to ℓ_0 and ℓ_1 regularization methods

In this simulation, we compare our proposed ℓ_2-regularized MLE to the ℓ_0 (Chen et al., 2021) and ℓ_1 Stein and Leng (2020) regularization methods. We consider the following settings with different network sparsity levels:

$$\beta_i^* = (-0.2\log n)/2 + \mathbb{1}_{[i\in\{1,\cdots,[n/10]\}]}(0.4\log n) + \mathbb{1}_{[i\in\{[n/10]+1,\cdots,[n/5]\}]}(-0.4\log n), \tag{35}$$

$$\beta_i^* = (-0.6\log n)/2 + \mathbb{1}_{[i\in\{1,\cdots,[n/20]\}]}(0.2\log n) + \mathbb{1}_{[i\in\{[n/20]+1,\cdots,[n/10]\}]}(-0.2\log n). \tag{36}$$

We varied network size $n \in \{100, 200, \cdots, 3200\}$ and evaluate the performance of the $\widehat{\beta}$'s produced by all three methods in the following five aspects: (1) $n^{-1}\|\widehat{\beta} - \beta^*\|_1$; (2) $n^{-1/2}\|\widehat{\beta} - \beta^*\|_2$; (3) $\|\widehat{\beta} - \beta^*\|_\infty$; (4) the relative error in estimating the "active set" $\mathcal{S}$ ($\mathcal{S} = [1 : [n/5]]$ in the dense network setting, and $\mathcal{S} = [1 : [n/10]]$ in the sparse network setting), denoted by the Hamming distance between $\widehat{\mathcal{S}}$ and $\mathcal{S}$, divided by n; and (5) computation time; where in (4), we employ the post-estimation thresholding method described in our Corollary 1 to obtain $\widehat{\mathcal{S}}$ for our method. In each setting, we repeated the

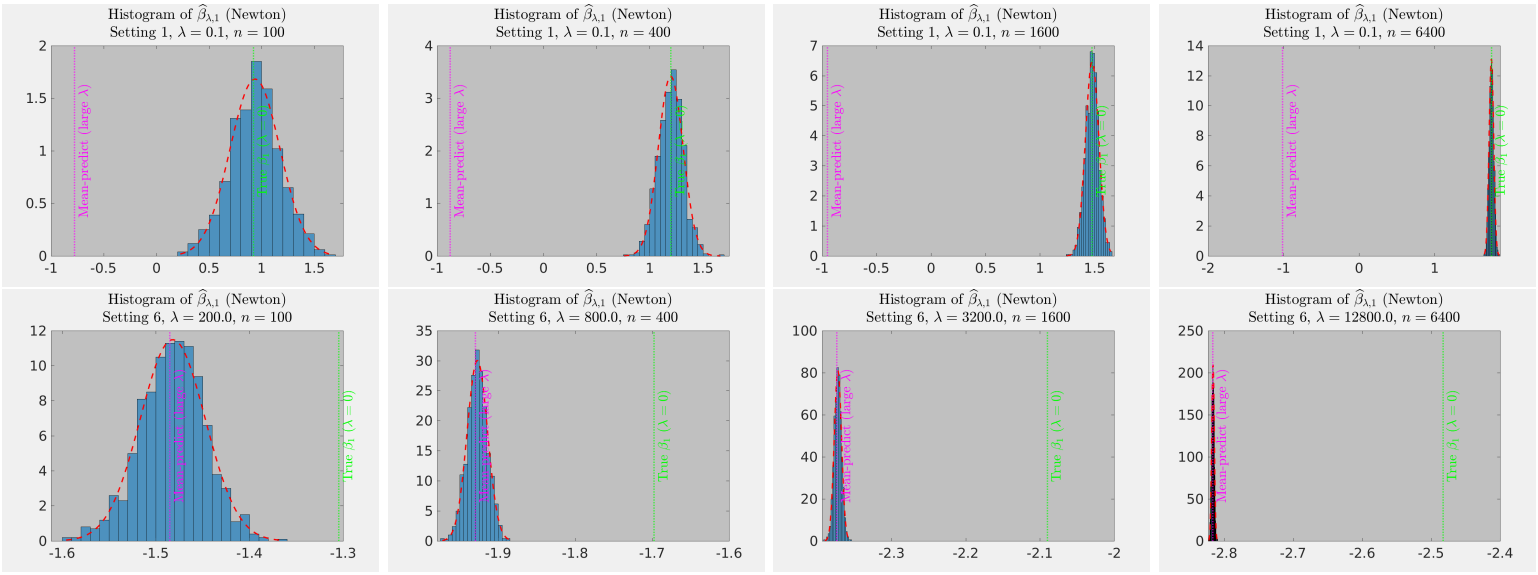

Fig. 2: Asymptotic distributions of $\widehat{\beta}_{\lambda;1}$ under various n for tuning parameter λ. Row 1: setting 1 (moderately dense network, small λ); row 2: setting 6 (sparse network, large λ). Green line: mean prediction for small λ case; magenta line: mean prediction for large λ case.

experiment 1000 times for our method and 100 times for the other two methods due to their much higher computation costs.

Figure 4 shows the result. Row 1 corresponds to the denser network setting (35), where we set $\lambda = 0$. Our method shows steadily achieves the best or a competitive accuracy across all settings. Specifically, unlike Chen et al. (2021) and Stein and Leng (2020), our thresholding method for estimating $\mathcal{S}$ does not require that $\beta_i^* - \beta_j^* > 0$ for all $i \in \mathcal{S}$, $j \notin \mathcal{S}$. This explains the advantage of our method in estimating the sparsity structure of β^* compared to these methods. Notice that the parameter estimation accuracy of both Chen et al. (2021) and Stein and Leng (2020) depends on their accurate specification of the sparsity structure in β^*, whereas our method does not. The fifth plot shows a clear speed advantage of our method. Chen et al. (2021) and Stein and Leng (2020) could not handle networks over $n = 10^3$ effectively and require an infeasible amount of memory when $n \approx 10^4$ and could not finish even one run. In our repeated experiments, they start to time out earlier as $n \asymp 10^3$.

The results shown in row 2 can be interpreted similarly to row 1. In row 2, we tested different choices of λ for our method. Despite the higher network sparsity level compared to row 1, choosing a small positive λ still yields the best performance across all measurements in row 2. This echoes the interpretation of our Theorems 1 and 2. We generally do not recommend choosing a large λ, unless the network is extremely sparse and we believe the true β^* is approximately parallel to $\mathbb{1}$. Indeed, our method with a very large $\lambda = n \log n$ effectively fits a nearly Erdos-Renyi model to the data. The results here also matches the prediction of our main theorems. For example, all ℓ_p errors diverge, and our theorem predicts the ℓ_p, $p = 2, \infty$ upper bounds to be $n^{-1/p} \cdot \|\mathcal{P}_\perp \beta^*\|_p$, $p = 2, \infty$, both of which also diverge at the rate of $\log n$ in this simulation.

5.3. *Simulation 3: performance of our AIC-type criterion for tuning λ*

Here, we assess the correctness of our AIC-type criterion, proposed in Section 4, in automatically tuning $\widehat{\beta}_\lambda$. We generate data from dense and sparse networks with the following specifications:

$$\beta_1^* = \cdots = \beta_{[n/10]}^* = -1/5 \log n, \quad \beta_{[n/10]+1}^* = \cdots = \beta_n^* = 1/2 \log n,$$
$$\beta_1^* = \cdots = \beta_{[n/3]}^* = -1/3 \log n, \quad \beta_{[n/3]+1}^* = \cdots = \beta_n^* = (-1/3 + 0.05) \log n.$$

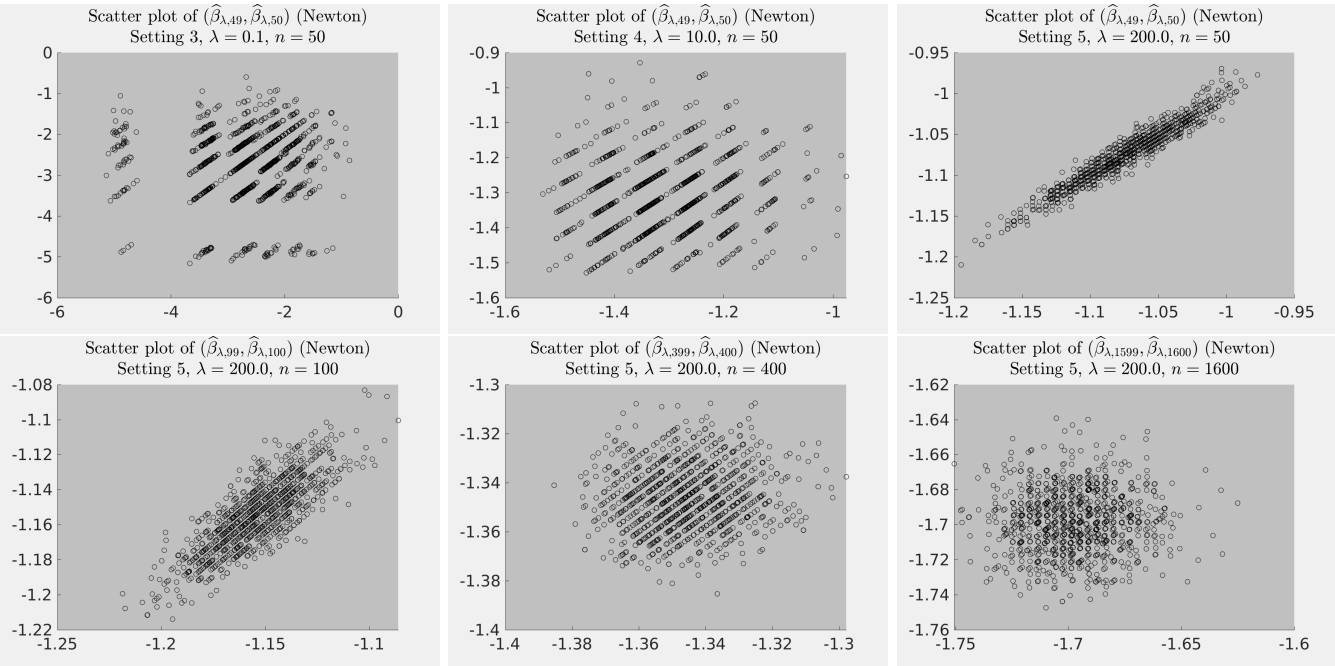

Fig. 3: Joint distribution of $(\widehat{\beta}_{\lambda;n-1}, \widehat{\beta}_{\lambda;n})$. Row 1: fix $n = 50$ and increase $\lambda \in \{0.1, 10, 200\}$; row 2: fix $\lambda = 200$ and increase $n \in \{100, 400, 1600\}$; other simulation set-ups are identical across all 6 plots.

Figure 5 shows that the tracks of the rescaled versions of our AIC-type criterion's values is consistent with the relative ℓ_2 errors of $\widehat{\beta}_\lambda$ under different λ choices. For the first setting where the network is dense, it suggests $\lambda = 0$; whereas in the latter setting with a sparse network, it selects a large λ.

6. Data examples

6.1. Data example 1: impact of COVID-19 on Swiss student mental health

The data set Elmer et al. (2020) contains social and psychological measurements on a local group of Swiss students to assess the impact of COVID-19 pandemic lockdown on their mental health. The original data set contains two cohorts of students dated 2019-04 (cohort 1), 2019-09 (cohort 2) and 2020-04 (cohort 2). For better comparability, we select the 2019-09 and 2020-04 subgroups since they share many individuals in common; while the 2019-04 data were surveyed on a distinct group of students longer before the pandemic. The variables fall into two main categories: 1. mental health, including *depression*, *anxiety*, *stress* and *loneliness*; 2. sociality, encoded by students' self-reported number of other students, with whom they have the following types of relations: *friendship*, *pleasant interactions*, *emotional support*, *informational support* and *co-study*.

First, as aforementioned, the released data only disclose node degrees, rather than adjacency matrices. This renders GLM-based methods such as Yan et al. (2019); Stein and Leng (2020) inapplicable. Second, not all students show up in both subgroups. The two subgroups contain 207 (2019-09) and 271 (2020-04) students, respectively, sharing 202 students in common. To make the most use of available data, we perform two marginal analysis on the two subgroups, respectively, then perform a differential analysis to compare the estimated parameters on the common students.

The third consideration is that both networks, despite their small sizes, are very sparse, as shown in Table 4, especially the *emotional support*, *informational support* and *co-study* networks. To alleviate this sparsity, we combined the three networks by summing up the degrees of each node in each subgroup as *support* network, resulting average degrees of 5.51(4.33) (2019-09) and 5.46(4.63) (2020-04), respectively. Degree distributions suggest that β_i's might not be ℓ_0-sparsity, and our proposed ℓ_2-regularized MLE with a small λ seems appropriate. We removed isolated nodes in each network and fit β-models to the social networks, with candidate choices of $\lambda \in \{0, e^{0.5} - 1, \cdots, e^6 - 1\}$.

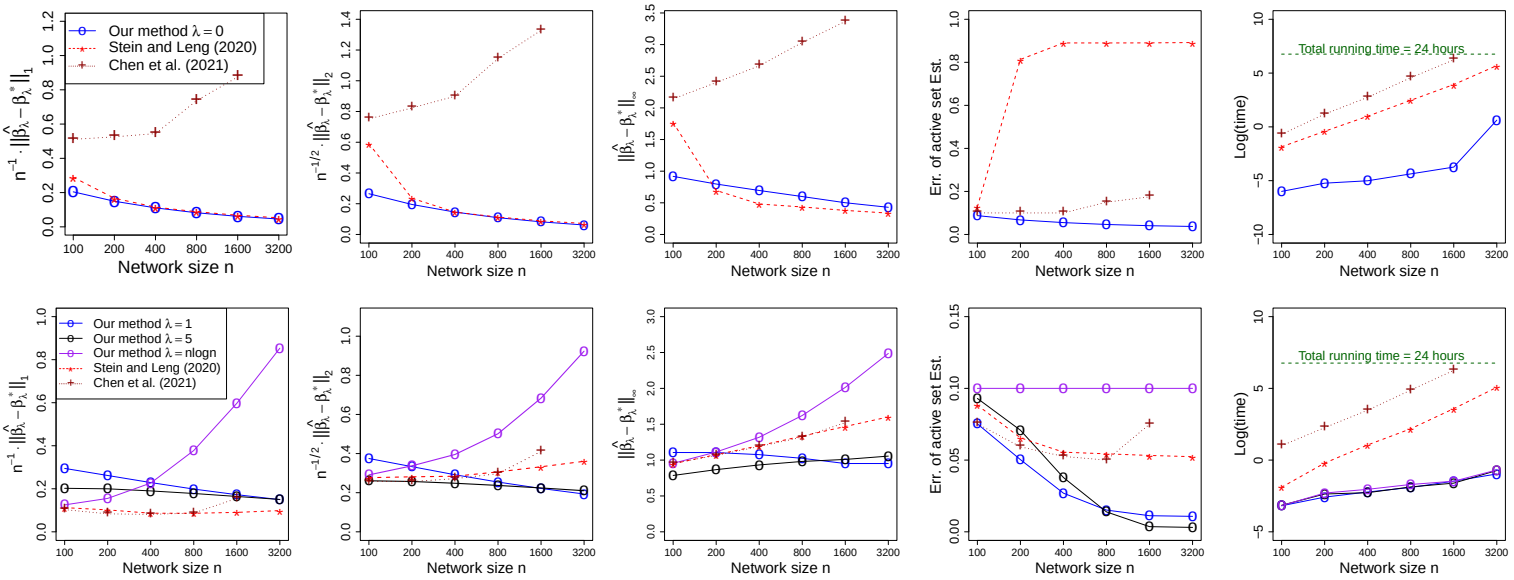

Fig. 4: Performance comparison with benchmark methods. Row 1: setting (35); row 2: setting (36).

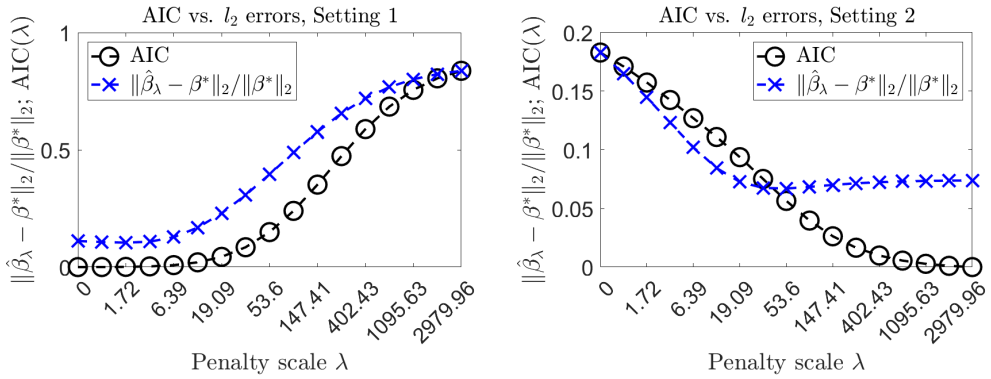

Fig. 5: Tuning λ under large (left) and small (right) $\|\mathcal{P}_\perp \beta^*\|_2$ values. Blue curve: true estimation error $\|\widehat{\beta}_\lambda - \beta^*\|_2/\|\beta^*\|_2$; black dashed curve: AIC-type criterion AIC(λ). The black curve has been rescaled, such that it has the same highest point as the blue curve. The original heights of black and blue curves are very different.

The first three columns in Figure 6 show the AIC curves for each type of social network in the two subgroups respectively. The *friendship* network turns out to be too sparse that our AIC criterion suggests choosing a large λ, and this will lead to an estimation with similar $\widehat{\beta}_{\lambda;i}$'s for all nodes. It is also conceptually difficult to combine with other networks, so we do not consider the friend network in our following analysis. As for the other two networks, the AIC plots suggest choosing $\lambda = 0.65$ for the *interaction* network and $\lambda = 0$ for the *support* network, respectively.

Recall our goal is to study the relationship between mental health and sociality. We treat the $\widehat{\beta}_\lambda$'s estimated from different networks as covariates and perform a sparse canonical correlation analysis (CCA) (Witten et al., 2009) between mental health variables and our estimated sociality β parameters on three data sets: 2019-04, and 2020-09 and the difference in their corresponding covariate values over their common nodes. For this part, we employ the `CCA` function in the R package `PMA` (Witten et al., 2009) and let it automatically tune the penalty term for its ℓ_1-penalized sparse CCA procedure. To assess the significance of each individual CCA coefficient, we employ a bootstrap procedure that repeatedly randomly shuffles the rows of mental health variables, while keeping the row order of sociality variables, and compute its empirical p-values.

Variable	friend	p.interaction	e.support	inf.support	co-study
Mean degree (2019-09)	3.928	6.454	1.444	2.019	2.048
Std. dev. (2019-09)	(2.554)	(4.225)	(1.503)	(1.692)	(2.004)
Mean degree (2020-04)	4.007	5.657	1.590	2.173	1.694
Std. dev. (2020-04)	(2.878)	(3.810)	(1.749)	(1.835)	(2.071)

Table 4: Network degree statistics, each column represents a different sociality network.

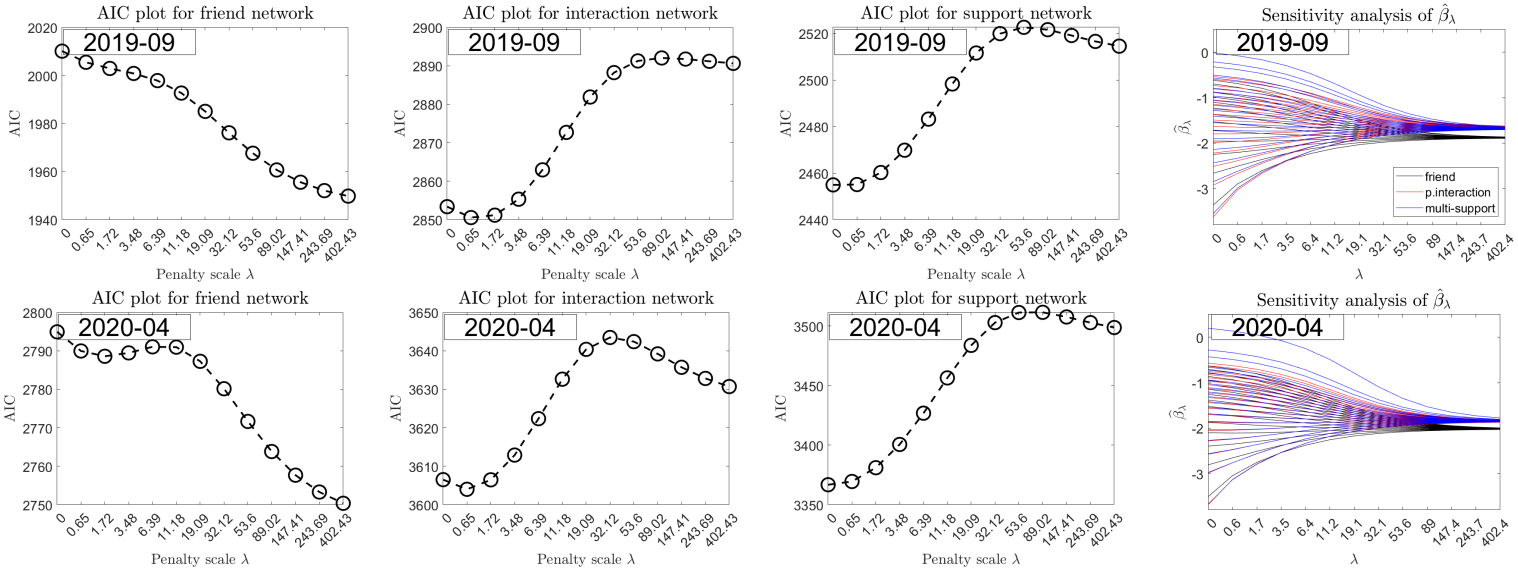

Fig. 6: Left three panels: AIC plots for tuning λ; right panel: tracks of $\widehat{\beta}_\lambda$ over different λ choices. The left-most panel suggests that the friendship networks are too sparse for informative β-model fits.

The results are reported in Table 5, where we print the first two canonical components. They mainly capture the connections between each of the two social networks (*support* and *pleasant interaction*) and mental health covariates, respectively. Before the lockdown, *loneliness* was the most prominent factor. Its almost opposite correlation directions with the two types of social networks might possibly be explained by that students tend to prefer pleasant interactions with friends rather than information support or co-study interactions, when they feel lonely. After the lockdown, *stress* became the most important factor that positively correlates with both *pleasant interaction* and *support* networks; while *loneliness* faded out. This is understandable, since many students might have to worry more about practical problems including study and job hunting, therefore might pay less attention to loneliness. The third main column in Table 5 shows the canonical correlations between the difference in mental health and the difference in sociality variables. Combined with Figure 2 in Elmer et al. (2020), we understand the result in this part as that increased *loneliness* and *anxiety* levels led to higher demands for *pleasant interaction* and *support*, respectively.

	Marginal								Common nodes			
	2019-09				2020-04				Difference			
Variable	CC1	p-val.	CC2	p-val.	CC1	p-val.	CC2	p-val.	CC1	p-val.	CC2	p-val.
depression	−0.005	(0.424)	0.183	(0.316)	0.000	(0.420)	0.049	(0.371)	0.399	(0.278)	0.000	(0.459)
anxiety	0.453	(0.288)	0.139	(0.377)	0.529	(0.289)	0.181	(0.361)	0.000	(0.460)	0.941	(0.179)
stress	−0.051	(0.404)	−0.065	(0.396)	0.848	(0.217)	0.960	(0.161)	−0.089	(0.425)	0.295	(0.353)
loneliness	−0.890	(0.244)	0.971	(0.149)	−0.023	(0.448)	0.210	(0.385)	0.913	(0.193)	0.164	(0.409)
p.interaction	0.000	(0.506)	1.000	(0.000)	0.000	(0.508)	1.000	(0.000)	1.000	(0.000)	0.000	(0.505)
support	1.000	(0.000)	0.000	(0.494)	1.000	(0.000)	0.000	(0.492)	0.000	(0.495)	1.000	(0.000)

Table 5: Estimated sparse CCA coefficients with empirical p-values.

6.2. Data example 2: analysis of two very large COVID-19 knowledge graphs

In this subsection, we apply our fast algorithm in Section 2.2 to two very large COVID-19 knowledge graphs. These applications demonstrate our method's significant superiority in speed and memory.

The first data set Steenwinckel et al. (2020) was transcribed from the well-known Kaggle CORD-19 data challenge in 2020. We downloaded the data from https://www.kaggle.com/group16/covid19-literature-knowledge-graph, which contains $n = 1304155 \approx 1.3$ million nodes. The second data set Wise et al. (2020) is part of the open-access Amazon data lake that is still updating real-time at the frequency of several times per hour. We analyze the version downloaded at 20:06pm UTC on 15th September, 2021. After cleaning up, the citation network contains $n = 57312$ papers (nodes). The reason we choose to study these two particular COVID-19 knowledge graph databases is that they are well-documented and maintained. The sizes of these data are prohibitive for the conventional GLM-based algorithms for fitting β-models (Yan et al., 2015; Chen et al., 2021; Stein and Leng, 2020, 2021), which typically need hours to compute networks of a few thousand nodes. In sharp contrast, our code costs less than 10 minutes on a personal computer to estimate for Steenwinckel et al. (2020).

Both data sets are structured following the typical knowledge graph fashion. The raw data are formatted in such 3-tuples: (entity 1, relation, entity 2), for example, (paper 1, cites, paper 2), (paper 1, authored by, author 3), (author 4, membership, institution 7) and so on. In this study, we focus on the paper citation network (paper 1, cites, paper 2) and ignore edge directions like that has been done in Li and Yang (2021) and Liu et al. (2019). Indeed, there are many interesting scientific problems that can potentially be addressed with these data sets. Due to page and scope limits, in this paper, we focus on findings based on β-model fits.

6.2.1. Data set 2a: the ISWC 2020 transcription of Kaggle CORD-19 data challenge

We first analyze the transcribed Kaggle data set (Steenwinckel et al., 2020). Compared to the Amazon data set Wise et al. (2020), Steenwinckel et al. (2020) has a larger citation network but fewer nodal covariates and no map between papers and topics (keywords). Therefore, its analysis is comparatively simpler. However, in this ISWC transcription of Kaggle data, most of the 1.3 million nodes in the complete network here are not on COVID-19 but general medical literature. We would use the complete network to estimate $\widehat{\beta}_\lambda$; afterwards, we use `metadata.csv` in Kaggle CORD-19 open challenge downloaded from https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge to filter and only keep COVID-19 papers, similar to the treatment in Rausch (2020).

Here, we focus on the nodal covariate *country* and compare the empirical distributions of $\widehat{\beta}_\lambda$ entries corresponding to different countries. We selected 6 representative countries/regions in the study of pandemic: UK, China, USA, EU (we counted France, Germany, Italy, Spain, Switzerland and Netherlands, which constitute the overwhelming majority of papers from EU), Japan plus South Korea, and India. All other papers are collected by the "Other" category. To choose a proper tuning parameter, we vary $\lambda \in \{0, 1, 2.5, \cdots, 1280\}$ and plot the track of AIC(λ) and $\widehat{\beta}_\lambda$ in Figure 7. Our AIC-type criterion suggests choosing $\lambda = 0$. Then we run our accelerated Newton's method on the complete network, and filter $\widehat{\beta}_\lambda$ entries using CORD-19 metadata as aforementioned.

Region	Other	China	US	UK	EU	JpKr	India
Entry count	53929	6096	15169	4747	11180	2394	1450
mean($\widehat{\beta}_\lambda$)	−5.426	−4.426	−4.477	−4.590	−4.553	−4.507	−4.737
std($\widehat{\beta}_\lambda$)	(1.417)	(1.018)	(1.104)	(1.099)	(1.084)	(1.048)	(1.172)

Table 6: Estimated $\widehat{\beta}_\lambda$ by region

Table 6 reports the numerical summary of estimation results and Figure 8 shows the region-wise histograms. Second, despite different total paper counts, the $\widehat{\beta}_\lambda$ distributions across the 6 regions we studied show similar marginal distributions. Inspecting the raw data, we understand that this can be partially attributed to the active international collaboration and mutual citation. Overall, we

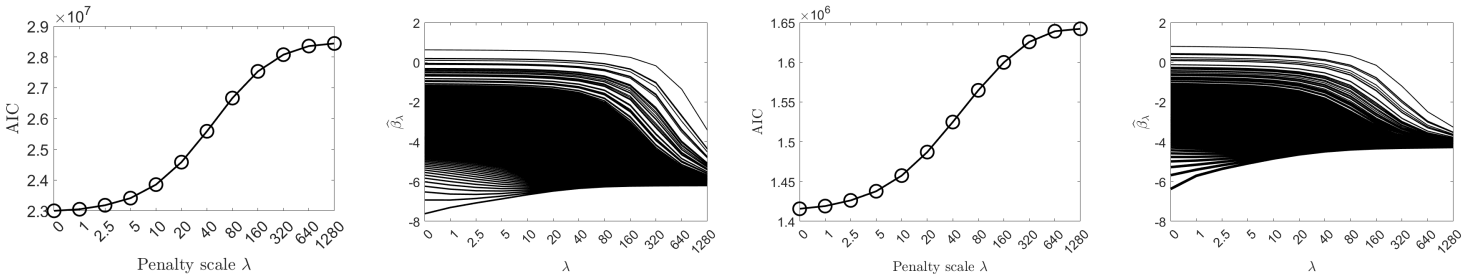

Fig. 7: Tracks of AIC(λ) and $\widehat{\beta}_\lambda$, Left two panels: Steenwinckel et al. (2020); right two panels: Wise et al. (2020)

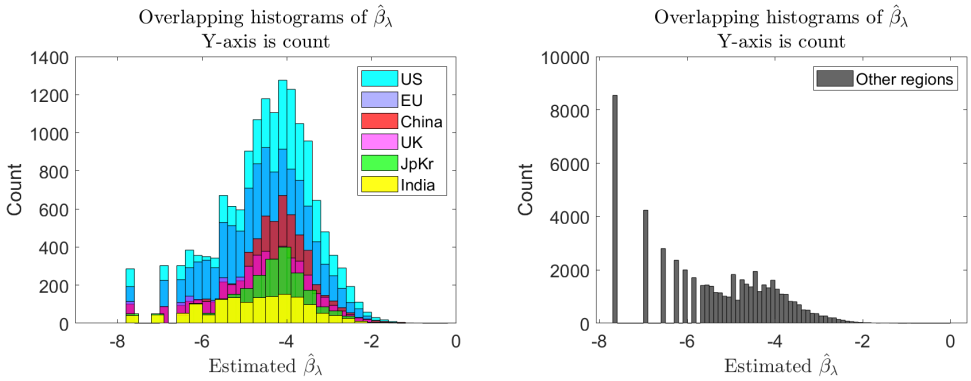

Fig. 8: Overlapping histograms of estimated $\widehat{\beta}_\lambda$ by region

see from Figure 6 the clear evidence of solidarity and impartiality among scientists and researchers studying COVID-19.

6.2.2. Data set 2b: the Amazon public COVID-19 data lake, knowledge graph section

The Amazon public COVID-19 data lake Wise et al. (2020) documents less papers ($n = 57312$) than Steenwinckel et al. (2020), but provides richer details on each entry, including a list of *"topics"*, which can be understood as key words that are not specified by the authors but automatically learned by a latent Dirichlet allocation (Blei et al., 2003) text analysis (see the "Graph Structure" section of Kulkarni et al. (2020)). The outcome can be represented as a list for paper i: $\{\text{topic}_j^{(i)} : s_j^{(i)}\}_{j=1,\ldots,k_i}$, where $s_j^{(i)} \in [0,1]$ is a score indicating the relevance of the topic. Here, we are interested in finding the "core" and "peripheral" topics in current literature. Our approach is to propose a *"Weighted Accumulative Beta Score (WABS)"*. For each concept j, let $\mathcal{A}_j \subseteq \{1,\ldots,n\}$ be the index set of all papers that specify j in its relevant topic list, and $s_j^{(i)}$ be the corresponding relevance score reported in the data set. We define

$$\text{WABS}_j = \sum_{i \in \mathcal{A}_j} e^{\widehat{\beta}_{\lambda;i}} \times s_j^{(i)}. \tag{37}$$

Now we briefly explain the intuition behind the definition of (37). On one hand, we design it as a sum, rather than an average, because the total count carries useful information and should be reflected in the measure. On the other hand, however, we also want to prevent "quantity over quality" by re-weighting each relevance score $s_j^{(i)}$ by the paper's transformed global popularity $e^{\widehat{\beta}_{\lambda;i}}$. In plain words, for a topic to sit at the center of the knowledge base, not only it should be a closely relevant topic of many papers, but further, it needs to be associated with many influential papers. One influential paper's contribution to the right hand side of (37) could easily outweigh connections to many peripheral works. Similar to the aforementioned point in our analysis of data example 1

(Elmer et al., 2020), our proposed WABS measure inherits the scale-free advantage from $\widehat{\beta}_\lambda$, thus is comparable across networks of potentially very different sizes.

Top 50					Bottom 50				
Concept	Category	Paper #	Avg(d_i)	WABS	Concept	Category	Paper #	Avg(d_i)	WABS
infection	dx name	18558	10.66	319.85	méthicilline	dx name	1	3.00	0.01
respiratory syndrome	dx name	6951	13.49	158.09	icer	dx name	2	2.00	0.01
death	dx name	7689	10.92	126.06	colostrum	dx name	3	1.67	0.01
lung	system organ site	6051	12.50	121.56	economic injury	system organ site	2	2.50	0.01
respiratory tract	system organ site	5094	14.53	117.10	cellmediated immunity	system organ site	2	2.50	0.01
pneumonia	dx name	4696	14.40	115.15	mesenteric lymphatic	dx name	2	2.00	0.01
fever	dx name	5531	12.11	111.01	perianal infection	dx name	2	1.50	0.01
viral infection	dx name	6669	11.02	109.37	psychiatric treatment	dx name	3	1.67	0.01
culture	test name	5739	11.02	92.63	potassium ion	test name	1	5.00	0.01
cough	dx name	3794	13.37	85.58	lysine decarboxylase	dx name	2	2.00	0.01
die	dx name	4194	12.46	84.92	fibrosis progression	dx name	3	1.67	0.01
vaccine	treatment name	5328	11.22	83.53	gldh	treatment name	2	1.50	0.01
infect	dx name	5852	11.90	82.11	neurobehavioral disorder	dx name	1	3.00	0.01
liver	system organ site	3910	11.09	69.14	nr2b	system organ site	2	2.00	0.01
diarrhea	dx name	3244	12.59	67.96	TRA	dx name	2	2.50	0.01
hand	system organ site	5347	9.37	67.51	hcv replicon assay	system organ site	2	2.50	0.01
kidney	system organ site	3213	13.31	66.10	Methylprednisolon	system organ site	3	1.00	0.01
HIV	dx name	4678	9.16	65.55	platelet index	dx name	2	1.50	0.01
respiratory infection	dx name	3081	13.46	65.07	immune tissue	dx name	2	2.00	0.01
respiratory disease	dx name	2686	15.10	61.62	herpetic uveitis	dx name	2	1.50	0.01
respiratory syncytial virus	dx name	2900	12.91	54.82	cadpr	dx name	2	2.50	0.01
chest	system organ site	2159	14.99	53.93	rna expression profiling	system organ site	3	1.67	0.01
rt-pcr	test name	2613	15.71	53.55	demostraron	test name	2	2.00	0.01
infectious disease	dx name	4766	8.55	51.79	tnfsf4	dx name	1	3.00	0.01
throat	system organ site	2201	14.78	50.78	boutonneuse fever	system organ site	2	2.50	0.01
heart	system organ site	2955	10.64	50.61	mucopolysaccharide	system organ site	3	1.33	0.01
pcr	test name	3381	11.65	50.27	urethral mucosa	test name	2	1.50	0.01
lesion	dx name	2586	11.38	47.47	covariance analysis	dx name	3	1.33	0.01
titer	test name	2681	12.25	47.05	transfusion-associated circulatory overload	test name	2	2.50	0.01
influenza virus	dx name	3401	10.20	45.76	Chondrex	dx name	3	1.33	0.01
inflammation	dx name	3500	8.51	45.29	tea	dx name	3	1.67	0.01
phylogenetic analysis	test name	1825	16.37	43.70	oxidovanadium	test name	2	1.50	0.01
vomiting	dx name	1869	13.33	42.37	anti-tnfα drug	dx name	1	4.00	0.01
respiratory tract infection	dx name	2076	13.84	41.19	control assay	dx name	2	2.50	0.01
penicillin	generic name	2534	9.63	40.83	avanzadas	generic name	3	1.33	0.01
adenovirus	dx name	2275	12.65	40.71	dystrophic neurite	dx name	3	1.67	0.01
respiratory distress syndrome	dx name	1248	19.52	39.50	gastric erosion	dx name	2	1.50	0.01
ribavirin	generic name	1254	18.20	39.21	foetal loss	generic name	3	1.33	0.01
antibiotic	generic name	3094	9.42	38.68	alloreactive t cell	generic name	2	2.50	0.01
rsv	dx name	2345	11.79	38.49	cefuroxima-axetilo	dx name	2	2.00	0.01
serum sample	test name	1822	13.51	37.97	cerebrovascular complication	test name	1	3.00	0.01
respiratory illness	dx name	1668	15.17	37.39	inadequate tissue oxygenation	dx name	2	2.50	0.01
respiratory virus	dx name	2101	13.47	36.86	il-1β concentration	dx name	1	3.00	0.01
streptomycin	generic name	2139	9.93	35.78	hypertransfusion	generic name	1	3.00	0.01
brain	system organ site	2556	8.88	35.67	vascular constriction	system organ site	2	2.50	0.01
respiratory symptom	dx name	1561	15.74	34.37	flu peptide	dx name	3	1.33	0.01
outbreak	dx name	2472	12.09	33.75	collagenous colitis	dx name	2	1.50	0.01
membrane	system organ site	2187	12.52	33.67	facial nucleus	system organ site	2	1.50	0.01
respiratory failure	dx name	1143	17.85	33.39	tricyclic compound	dx name	2	1.50	0.01
bacterial infection	dx name	2123	9.81	33.37	ischemic heart failure	dx name	2	2.00	0.01

Table 7: Summary statistics for top- and bottom-50 concepts, ranked by *Weighted Accumulative Beta Score* (WABS)

Following the suggestion of the AIC track in Figure 7, we select $\lambda = 0$ for method to obtain $\widehat{\beta}_\lambda$ and then compute WABS scores. Table 7 reports the top- and bottom-50 topics ranked by their WABS ratings. The outcome well-matches our intuitive understandings. For instance, the top-ranked list contains relevant organs such as *respiratory tract*, *lung* and *throat* that are directly related to COVID-19 as a respiratory disease, but also includes organs like *liver* and *kidney* that are now widely-believed also main attack objectives of the virus (Zhang et al., 2020; Fan et al., 2020b; Hirsch et al., 2020; Pei et al., 2020). Top-ranked concepts related to testing methods and treatments, including *culture* (meaning "viral culturing"), *PCR*, *phylogenetic analysis* (related to backtracking ancestors of the virus and monitoring latest variants) and *vaccine*, also reflect the current mainstream approaches. The other top entries cover important symptoms and related viruses. In contrast, most concepts ranked in the bottom seem to lack either specificity. Comparing the top- and bottom-lists, we see that our proposed WABS score yields evidently more meaningful result than several potential alternative

approaches, such as simply ranking concepts by counting the number of papers they present.

7. Discussion

One major question that our paper does not address is assessing goodness-of-fit. For stochastic block models, this problem has been well-solved (Lei, 2016). The common approach is based on the fact that the largest singular value of a matrix $\widetilde{A}$, denoted by $\sigma_1(\widetilde{A})$, where

$$\widetilde{A}_{i,j} = \frac{A_{i,j} - P_{i,j}}{\sqrt{(n-1)P_{i,j}(1-P_{i,j})}}, \tag{38}$$

and $\widetilde{A}_{i,i} = 0$, satisfies that $T = n^{2/3}(\sigma_1(\widetilde{A}) - 2) \xrightarrow{d} \mathrm{TW}_1$, where TW_1 is a Tracy-Widom distribution indexed 1 (Erdős et al., 2012; Lee and Yin, 2014). The key result that makes this work for stochastic block models is that when the number of communities K is moderately small, one can estimate the community structure very accurately, so that there are only $O(K^2)$ different $P_{i,j}$ values, each corresponding to $O(n^2/K^2)$ independent observations. Therefore the $P_{i,j}$'s in (38) can be replaced by $\widehat{P}_{i,j}$ without altering the limiting distribution of $\sigma_1(\widetilde{A})$. This is unfortunately not true for β-models. Now there are n parameters, each corresponding to $O(n)$ independent observations. Lubold et al. (2021) suggests replacing $P_{i,j}$ in (38) by $\widehat{P}_{i,j} = e^{\widehat{\beta}_{\lambda;i}+\widehat{\beta}_{\lambda;j}}/(1+e^{\widehat{\beta}_{\lambda;i}+\widehat{\beta}_{\lambda;j}})$ anyways. To assess the accuracy of this approach, we generated data with $\beta_i^* = b\log n$ for $i \le [0.4n]$ and $\beta_j^* = 0.1\log n$ for $j \ge [0.4n]+1$. We varied $b \in \{-0.1, -0.4\}$ and $n \in \{100, 200, \cdots, 1600\}$, and set $\lambda = 0$. For each (n, β^*), we compared the empirical distribution of the test statistic T (with P replaced by $\widehat{P}$) with TW_1 via $n_{MC} = 10^4$ Monte-Carlo repetitions. Figure 9 shows a non-vanishing discrepancy between these two distributions. Therefore, we can at least conclude that the approach of Lubold et al. (2021) is not applicable to β-models in large and sparse networks. The goodness-of-fit test remains an open challenge for β-models.

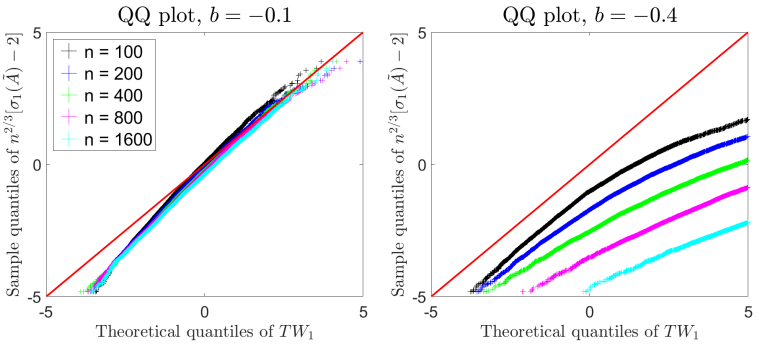

Fig. 9: QQ plot of empirical distribution of T using $\widehat{P}$ in (38) vs. TW_1 distribution.

In this paper, we focused exclusively on analyzing network data. We analyzed nodal covariates in data examples, but the covariates do not participate in the network generation model. The consideration is mainly three-fold. First, we pursue a good understanding of a simple model, and then gradually push forward towards more complex ones. Research on the β-model without covariates still faces several open problems that would require considerable future effort to resolve. Second, the joint modeling approach (Yan et al., 2019; Stein and Leng, 2020, 2021) may encounter substantive challenge in large and sparse networks, where the response is highly extreme imbalanced, with most 0's and few 1's. Some treatments may be necessary to properly address this issue, see analogous discussions in classification (Sun et al., 2009). The third consideration is computation. As pointed out by Stein and Leng (2021), the monotonicity lemma would not hold for a joint model involving covariates. Consequently, the joint model could not yet effectively scale up beyond $n \asymp 10^3$ nodes. The data examples we studied in this paper have up to 10^7 nodes and are typically very sparse.

Another interesting question is whether our work can be extended to bipartite and directed networks. We believe that the extension to bipartite networks would be quite natural, as Lemma 2 easily extends to the bipartite case, after slight adaptions. Extension to direct networks, however, is nontrivial and would require novel treatment, because there, the lack of symmetry $A_{i,j} \neq A_{j,i}$ breaks Lemma 2. Due to page and scope limits, we will not discuss this in greater detail.

Finally, our paper exclusively focuses on analyzing β-models with *independent* edge generation. There exist a line of fine works that address dependent edges (Frank and Strauss, 1986; Hunter et al., 2012; Schweinberger and Stewart, 2020). With the introduction of edge dependency, not only estimation and inference, but even data generation from a given model would become much more difficult and costly. We feel that this is an interesting but also challenging future direction.

Computer code

The computer code, composed by author Meijia Shao, including the full instructions for reproducing the simulation and data analysis results in this paper, is available at `https://github.com/MjiaShao/L2-beta-model`. It does not include the original code for Chen et al. (2021) and Stein and Leng (2020) that we obtained from Professor Chenlei Leng, and data sets. See its `README` for more details.

Acknowledgements

The authors wish to express sincere thanks to the Editor, the Associate Editor and two anonymous referees for their insightful comments that led to very significant improvements in both the scientific contents and the presentation of this paper. We thank Professor Chenlei Leng for sharing the code files for his ℓ_0- and ℓ_1-regularized β-model papers; and thank him and Mr. Stefan Stein for advising us on selecting the tuning parameter in Stein and Leng (2020). We thank Professors David S. Choi, Yoonkyung Lee, Elizaveta Levina and Subhabrata Sen for their constructive discussions that helped us enrich our paper's contents. Finally, we thank Professors Steven MacEachern and Ji Zhu for their kind advice and warm encouragements.

References

Amazon H2O (2021) Amazon H2O AI platform documentation: Generalized Linear Model (GLM). `http://h2o-release.s3.amazonaws.com/h2o/rel-jordan/3/docs-website/datascience/glm.html`. Online, Accessed 08-October-2021.

Babai, L., Erdos, P. and Selkow, S. M. (1980) Random graph isomorphism. *SIaM Journal on computing*, **9**, 628–635.

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

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) Latent dirichlet allocation. *the Journal of machine Learning research*, **3**, 993–1022.

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

Chen, M., Kato, K. and Leng, C. (2021) Analysis of networks via the sparse β-model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **83**, 887–910.

Chen, N. and Olvera-Cravioto, M. (2013) Directed random graphs with given degree distributions. *Stochastic Systems*, **3**, 147–186.

Elmer, T., Mepham, K. and Stadtfeld, C. (2020) Students under lockdown: Comparisons of students' social networks and mental health before and during the covid-19 crisis in switzerland. *Plos one*, **15**, e0236337.

Erdős, L., Yau, H.-T. and Yin, J. (2012) Rigidity of eigenvalues of generalized wigner matrices. *Advances in Mathematics*, **229**, 1435–1515.

Fan, Y., Zhang, H. and Yan, T. (2020a) Asymptotic theory for differentially private generalized β-models with parameters increasing. *arXiv preprint arXiv:2002.12733*.

Fan, Z., Chen, L., Li, J., Cheng, X., Yang, J., Tian, C., Zhang, Y., Huang, S., Liu, Z. and Cheng, J. (2020b) Clinical features of covid-19-related liver functional abnormality. *Clinical Gastroenterology and Hepatology*, **18**, 1561–1566.

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.

Frank, O. and Strauss, D. (1986) Markov graphs. *Journal of the american Statistical association*, **81**, 832–842.

Gao, W. Y. (2020) Nonparametric identification in index models of link formation. *Journal of Econometrics*, **215**, 399–413.

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

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning*. Springer New York. URL: https://doi.org/10.1007/978-0-387-84858-7.

Hillar, C. and Wibisono, A. (2013) Maximum entropy distributions on graphs. *arXiv preprint arXiv:1301.3321*.

Hillar, C. J., Lin, S. and Wibisono, A. (2012) Inverses of symmetric, diagonally dominant positive matrices and applications. *arXiv preprint arXiv:1203.6812*.

Hirsch, J. S., Ng, J. H., Ross, D. W., Sharma, P., Shah, H. H., Barnett, R. L., Hazzan, A. D., Fishbane, S., Jhaveri, K. D., Abate, M. et al. (2020) Acute kidney injury in patients hospitalized with covid-19. *Kidney international*, **98**, 209–218.

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

Hunter, D. R., Krivitsky, P. N. and Schweinberger, M. (2012) Computational statistical methods for social network models. *Journal of Computational and Graphical Statistics*, **21**, 856–882.

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

Kulkarni, N., Wise, C., Price, G. and Romero, M. (2020) Amazon website services (AWS) database blog: Building and querying the AWS COVID-19 knowledge graph. https://aws.amazon.com/cn/blogs/database/building-and-querying-the-aws-covid-19-knowledge-graph/. Published 01-Jul-2020, accessed 14-Sep-2021.

Lee, J. O. and Yin, J. (2014) A necessary and sufficient condition for edge universality of wigner matrices. *Duke Mathematical Journal*, **163**, 117–173.

Lee, K.-Y. and Courtade, T. (2020) Minimax bounds for generalized linear models. *Advances in Neural Information Processing Systems*, **33**, 9372–9382.

Lei, J. (2016) A goodness-of-fit test for stochastic block models. *The Annals of Statistics*, **44**, 401–424.

Li, B. and Yang, Y. (2021) Undirected and directed network analysis of the chinese stock market. *Computational Economics*, 1–19.

Li, Y.-H., Scarlett, J., Ravikumar, P. and Cevher, V. (2015) Sparsistency of\ell_1-regularized m-estimators. In *Artificial Intelligence and Statistics*, 644–652. PMLR.

Liu, H., Kou, H., Yan, C. and Qi, L. (2019) Link prediction in paper citation network to construct paper correlation graph. *EURASIP Journal on Wireless Communications and Networking*, **2019**, 1–12.

Lu, J., Ko, D. and Chang, T. (1997) The standardized influence matrix and its applications. *Journal of the American Statistical Association*, **92**, 1572–1580.

Lubold, S., Liu, B. and McCormick, T. H. (2021) Spectral goodness-of-fit tests for complete and partial network data. *arXiv preprint arXiv:2106.09702*.

McCullagh, P. and Nelder, J. A. (2019) *Generalized linear models*. Routledge.

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

Park, J. and Newman, M. E. (2004) Statistical mechanics of networks. *Physical Review E*, **70**, 066117.

Pei, G., Zhang, Z., Peng, J., Liu, L., Zhang, C., Yu, C., Ma, Z., Huang, Y., Liu, W., Yao, Y. et al. (2020) Renal involvement and early prognosis in patients with covid-19 pneumonia. *Journal of the American Society of Nephrology*, **31**, 1157–1165.

Rausch, I. (2020) Covid-19: Knowledge graph - a network analysis. https://www.kaggle.com/iljara/covid-19-knowledge-graph-a-network-analysis. Published 21-May-2020, accessed, 22-Sep-2021.

Rinaldo, A., Petrović, S. and Fienberg, S. E. (2013) Maximum lilkelihood estimation in the β-model. *The Annals of Statistics*, 1085–1110.

Schweinberger, M. and Stewart, J. (2020) Concentration and consistency results for canonical and curved exponential-family models of random graphs. *The Annals of Statistics*, **48**, 374–396.

Steenwinckel, B., Vandewiele, G., Rausch, I., Heyvaert, P., Colpaert, P., Simoens, P., Dimou, A., Turkc, F. D. and Ongenae, F. (2020) Facilitating covid-19 meta-analysis through a literature knowledge graph. In *Accepted in Proc. of 19th International Semantic Web Conference (ISWC)*.

Stein, S. and Leng, C. (2020) A sparse β-model with covariates for networks. *arXiv preprint arXiv:2010.13604*.

— (2021) A sparse random graph model for sparse directed networks. *arXiv preprint arXiv:2108.09504*.

Su, L., Qian, X. and Yan, T. (2018) A note on a network model with degree heterogeneity and homophily. *Statistics & Probability Letters*, **138**, 27–30.

Sun, Y., Wong, A. K. and Kamel, M. S. (2009) Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, **23**, 687–719.

Wahlström, J., Skog, I., La Rosa, P. S., Händel, P. and Nehorai, A. (2017) The β-model—maximum likelihood, cramér–rao bounds, and hypothesis testing. *IEEE Transactions on Signal Processing*, **65**, 3234–3246.

van Wieringen, W. N. (2015) Lecture notes on ridge regression. *arXiv preprint arXiv:1509.09169*.

Wise, C., Ioannidis, V. N., Calvo, M. R., Song, X., Price, G., Kulkarni, N., Brand, R., Bhatia, P. and Karypis, G. (2020) Covid-19 knowledge graph: accelerating information retrieval and discovery for scientific literature. *arXiv preprint arXiv:2007.12731*.

Witten, D. M., Tibshirani, R. and Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.

Yan, T., Jiang, B., Fienberg, S. E. and 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. and Zhu, J. (2016) Asymptotics in directed exponential random graph models with an increasing bi-degree sequence. *The Annals of Statistics*, **44**, 31–57.

Yan, T. and 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.

Yan, T., Zhao, Y. and Qin, H. (2015) Asymptotic normality in the maximum entropy models on graphs with an increasing number of parameters. *Journal of Multivariate Analysis*, **133**, 61–76.

Zhang, C., Shi, L. and Wang, F.-S. (2020) Liver injury in covid-19: management and challenges. *The lancet Gastroenterology & hepatology*, **5**, 428–430.

Zhang, Y. and Xia, D. (2022) Edgeworth expansions for network moments. *The Annals of Statistics*, **50**, 726–753.

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