Submitted to the Annals of Applied Statistics
arXiv: math.PR/0000000

ESTIMATING HETEROGENEOUS GRAPHICAL MODELS FOR DISCRETE DATA WITH AN APPLICATION TO ROLL CALL VOTING

By Jian Guo*, Jie Cheng†, Elizaveta Levina†, George Michailidis† and Ji Zhu†

**Harvard University and †University of Michigan*

We consider the problem of jointly estimating a collection of graphical models for discrete data, corresponding to several categories that share some common structure. An example for such a setting is voting records of legislators on different issues, such as defense, energy, and healthcare. We develop a Markov graphical model to characterize the heterogeneous dependence structures arising from such data. The model is fitted via a joint estimation method that preserves the underlying common graph structure, but also allows for differences between the networks. The method employs a group penalty that targets the common zero interaction effects across all the networks. We apply the method to describe the internal networks of the US Senate on several important issues. Our analysis reveals individual structure for each issue, distinct from the underlying well-known bipartisan structure common to all categories which we are able to extract separately. We also establish consistency of the proposed method both for parameter estimation and model selection, and evaluate its numerical performance on a number of simulated examples.

1. Introduction. The analysis of roll call data of legislative bodies has attracted a lot of attention both in the political science and statistical literature. For political scientists, such data allow to study broad issues such as party cohesion as well as more specific ones such as coalition formation; see for example the books [10, 24, 27, 29]. A popular tool in political science is the ideal point model [6] that posits a one-dimensional latent political space along which legislators and bills they vote for are aligned. A legislator's position corresponds to an ideal point, where bills coinciding with that position maximize his/her utility. These ideal points reveal legislators' preferences and it is of interest to infer them from roll call data. An extension of this model that incorporates information about the text of the bills being voted upon is discussed in [11], while the impact of absenteeism is examined in

AMS 2000 subject classifications: Primary 62H99

Keywords and phrases: graphical models, group penalty, high-dimensional data, ℓ_1 penalty, Markov network, binary data

[15].

A statistical challenge is how to best model and present the roll call data in a way that makes interesting patterns apparent and facilitates subsequent analyses. A number of techniques have been employed including principal components analysis (PCA) [8], multidimensional scaling (MDS) [9], Bayesian spatial voting models [6], and graphical models for binary data [3].

Dimension reduction techniques such as PCA and MDS aim at constructing a "map", with the members of the legislative body positioned relative to their peers according to their voting pattern. A typical example of such a map of the US Senate members in the 109th Congress (2005-06) using multidimensional scaling for selected votes is shown in Figure 1; for a detailed description of the data see Section 4. A clear separation between members of the two parties is seen (Republicans to the left of the map and Democrats to the right), together with some members exhibiting a voting pattern deviating from their party, e.g., Nelson (Democrat of Nebraska), and Collins and Snow (Republicans of Maine), while the independent Jeffords (shown in purple) votes like a Democrat. More interestingly, the voting patterns within both parties form distinct subclusters. While the nature of this division is impossible to infer from an MDS or a PCA representation such as the one shown in Figure 1, our subsequent analysis will show that this difference is driven by votes on defense/security and healthcare issues.

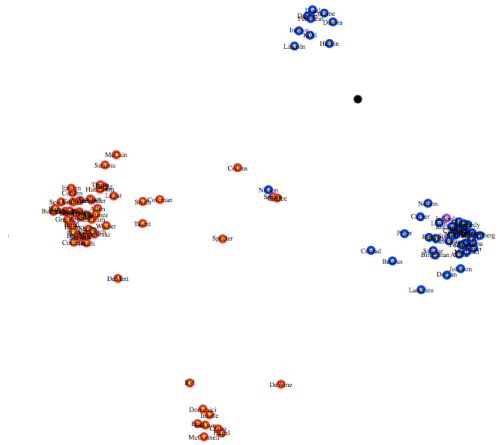

FIG 1. *Multidimensional scaling projection of roll call data of the US Senate for the period 2005-06 (Republicans shown in red and Democrats in blue).*

This finding suggests that treating all votes as homogeneous, i.e., assuming that they represent the same underlying relationship between senators, may mask more subtle patterns which depend on the issues being voted upon. Therefore, treating votes as heterogeneous is more accurate, and can provide further insight into the voting behavior of different groups of senators on different issues. In this paper, we focus on voting records on three types of bills: defense and national security, environment and energy, and healthcare issues. Voting on the latter category is typically more partisan than voting on defense and national security, and thus we expect to see different connections in different categories.

The voting records of the U.S. Senate from the 109th Congress covering the period 2005-2006 were obtained directly from the Senate's website (www.senate.gov). We chose the 109th Congress because its voting patterns have been previously analyzed in the literature (see e.g., [3]), but as we have discovered, the version of the data previously analyzed was contaminated with voting records from the 1990s (when the set of senators would have been different). Thus we collected the data ourselves, on all the 645 votes that the Senate deliberated and voted on during that period, which include bills, resolutions, motions, debates and roll call votes. To study the potential heterogeneity in the voting patterns, we focused on three largest meaningful (i.e. excluding purely procedural votes) categories of votes extracted from bills, resolutions and motions: 1) defense and security issues; 2) environment and energy issues; 3) health and medical care issues. The categories were extracted by a combination of text analysis of bill names and manual labeling. A complete analysis of this dataset will be presented in Section 4.

Our goal in this paper is to develop a statistical model for studying dependence patterns in such situations: there is some overall structure present (party affiliation, which affects everything) and there are also distinct categories with their own individual structures. Since we are dealing with voting data, we use Markov network models to capture the dependence structure of binary or categorical random variables. Similarly to Gaussian graphical models, nodes in a Markov network correspond to (categorical) variables, while edges represent dependence between nodes conditional on all other variables. Graphical models are an exploratory data analysis tool used in a number of application areas to explore the dependence structure between variables, including bioinformatics [1], natural language processing [19] and image analysis [22]. In the case of Gaussian graphical models, which assumes the variables are jointly normally distributed, the structure of the underlying graph can be fully determined from the corresponding inverse covariance (precision) matrix, the off-diagonal elements of which are proportional to

partial correlations between the variables. A number of methods have been recently proposed in the literature to fit *sparse* Gaussian graphical models (see for example, Banerjee, El Ghaoui and d'Aspremont [3], Meinshausen and Buhlmann [25], Peng *et al.* [28], Ravikumar *et al.* [32], Rothman *et al.* [33], Yuan and Lin [37] and references therein). Sparse Markov networks for binary data (Ising models) have been studied by Anandkumar *et al.* [2], Guo *et al.* [12], Höefling and Tibshirani [17], Ravikumar, Wainwright and Lafferty [31], Xue, Zou and Cai [35]. These methods do not allow for different categories within the data.

To allow for heterogeneity, we develop a framework for fitting different Markov models for each category that are nevertheless *linked*, sharing nodes and some common edges across all categories, while other edges are uniquely associated with a particular category. This will allow us to borrow strength across categories instead of fitting them completely separately. For the Gaussian case, this type of joint graphical model was first studied by Guo *et al.* [14], who proposed a joint likelihood based estimation method that borrowed strength across categories. Several other papers have proposed alternative algorithms for the Gaussian case [7, 16, 36]. We note that a context-specific graphical model was proposed for count data in the form of contingency tables by [18], but contingency tables are not suitable for high-dimensional data, and the context-specific model is not sparse.

The advantage of using a Markov graphical model in this context is that it quantifies the degree of conditional dependence between the senators based on their voting record and hence the obtained network and is directly interpretable. Techniques like multidimensional scaling and principal components analysis represent relative similarities between senators' voting records on the map, and hence the distance between any two senators can be interpreted as a quantitative measure of similarity between their voting records. However, unlike in a Markov network, these distances are not interpretable in the context of a generative probability model.

The remainder of the paper is organized as follows. Section 2 introduces the Markov network and addresses algorithmic issues, and Section 3 briefly illustrates the performance of the joint estimation method on simulated data. A detailed analysis of the US Senate's voting record from 109th Congress is presented in Section 4. Some concluding remarks are drawn in Section 5, and the Appendix presents results on the asymptotic properties of the method. The Electronic supplementary material contains a detailed investigation of missing data imputation methods for the Senate vote data.

2. Model and Estimation Algorithm. In this section, we present the Markov model for heterogeneous data, focusing on the special case of binary variables (also known as the Ising model). The extension to general categorical variables is briefly discussed in Section 5. We start by discussing estimation of *separate* models for each category and then develop a method for joint estimation.

The main technical challenge when estimating the likelihood of Markov graphical models is its computational intractability due to the normalizing constant. To overcome this difficulty, different methods employing computationally tractable approximations to the likelihood have been proposed in the literature; these include methods based on surrogate likelihood [3, 20] and pseudo-likelihood [13, 17, 31]. Höefling and Tibshirani [17] also proposed an iterative algorithm that successively approximates the original likelihood through a series of pseudo-likelihoods, while Ravikumar, Wainwright and Lafferty [31] and Guo *et al.* [13] established asymptotic consistency of their respective methods.

2.1. *Problem Setup and Separate Estimation.* We start from setting up notation and reviewing previous work on estimating a single Ising model, which can be used to estimate the graph for each category separately. Suppose that data have been collected on p *binary* variables in K categories, with n_k observations in the k-th category, $k = 1, \ldots, K$. Let $\boldsymbol{x}_i^{(k)} = (x_{i,1}^{(k)}, \ldots, x_{i,p}^{(k)})$ denote a p-dimensional row vector containing the data for the i-th observation in the k-th category and assume that it is drawn independently from an exponential family with the probability mass function

$$\mathrm{f}_k(X_1, \ldots, X_p) = \frac{1}{\mathrm{Z}(\boldsymbol{\Theta}^{(k)})} \exp\Big(\sum_{j=1}^{p} \theta_{j,j}^{(k)} X_j + \sum_{1 \le j < j' \le p} \theta_{j,j'}^{(k)} X_j X_{j'} \Big). \tag{2.1}$$

The partition function $\mathrm{Z}(\boldsymbol{\Theta}^{(k)}) = \sum_{X_j \in \{0,1\}, j} \exp(\theta_{j,j}^{(k)} X_j + \sum_{j<j'} \theta_{j,j'}^{(k)} X_j X_{j'})$ ensures that the probabilities in (2.1) add up to one. The parameters $\theta_{j,j}^{(k)}$, $1 \le j \le p$ correspond to the main effect for variable X_j in the k-th category, and $\theta_{j,j'}^{(k)}$ is the interaction effect between variables X_j and $X_{j'}$, $1 \le j < j' \le p$. The underlying network associated with the k-th category is determined by the symmetric matrix $\boldsymbol{\Theta}^{(k)} = (\theta_{j,j'}^{(k)})_{p \times p}$. Specifically, if $\theta_{j,j'}^{(k)} = 0$, then X_j and $X_{j'}$ are conditionally independent in the k-th category given all the remaining variables, and hence their corresponding nodes are *not* connected. For each category, (2.1) is referred to as the Markov network in the machine learning literature, and as the log-linear model in the statistics literature, where $\theta_{j,j'}^{(k)}$ is also interpreted as the conditional log-odds-ratio between X_j

and $X_{j'}$ given the other variables. Although general Markov networks allow higher order interactions (3-way, 4-way, etc), Ravikumar, Wainwright and Lafferty [31] pointed out that in principle one can consider only the pairwise interaction effects without loss of generality, since higher order interactions can be converted to pairwise ones by introducing additional variables [34]. For the rest of this paper, we only consider models with pairwise interactions of the original binary variables.

The simplest way to deal with heterogenous data is to estimate K separate Markov models, one for each category. If one further assumes sparsity for the k-th category, the structure of the underlying graph can be estimated by regularizing the log-likelihood using an ℓ_1 penalty:

$$\max_{\mathbf{\Theta}^{(k)}} \frac{1}{n_k} \sum_{i=1}^{n_k} \Big\{ \sum_{j=1}^{p} \theta_{j,j}^{(k)} x_{i,j}^{(k)} + \sum_{j<j'} \theta_{j,j'}^{(k)} x_{i,j}^{(k)} x_{i,j'}^{(k)} \Big\} - \log \mathrm{Z}(\mathbf{\Theta}^{(k)}) - \lambda \sum_{j<j'} |\theta_{j,j'}^{(k)}|. \tag{2.2}$$

The ℓ_1 penalty shrinks some of the interaction effects $\theta_{j,j'}^{(k)}$ to zero and λ controls the degree of sparsity. However, estimating (2.2) directly is computationally infeasible due to the nature of the partition function. A standard approach in such situation is to replace the likelihood with pseudo-likelihood [5], which has been shown to work well in a range of situations. Here, we use a pseudo-likelihood estimation method for Ising models [13, 17], based on

$$\begin{aligned} \max_{\mathbf{\Theta}^{(k)}} \quad & \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{p} \Big[x_{i,j}^{(k)} \Big(\theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \Big) \\ & - \log \Big\{ 1 + \exp \Big(\theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \Big) \Big\} \Big] \\ & - \lambda \sum_{j<j'} |\theta_{j,j'}^{(k)}| \ , \end{aligned} \tag{2.3}$$

where $\mathbf{\Theta}^{(k)}$ is restricted to be symmetric. Criterion (2.3) can be efficiently maximized using the modified coordinate descent algorithm of Höefling and Tibshirani [17].

2.2. *Joint Estimation of Heterogeneous Networks.* The separate estimation methods reviewed in the previous section do not take advantage of the shared nodes among the categories and potential common structure. Our goal here is to explicitly include this into the estimation procedure. We start by reparameterizing each $\theta_{j,j'}^{(k)}$ as

$$\theta_{j,j'}^{(k)} = \phi_{j,j'} \gamma_{j,j'}^{(k)}, \quad 1 \leq j \neq j' \leq p; 1 \leq k \leq K. \tag{2.4}$$

To avoid sign ambiguities between $\phi_{j,j'}$ and $\gamma_{j,j'}^{(k)}$, we restrict $\phi_{j,j'} \geq 0$, $1 \leq j < j' \leq p$. To preserve the symmetry of $\mathbf{\Theta}^{(k)}$, we also require $\phi_{j,j'} = \phi_{j',j}$ and $\gamma_{j,j'}^{(k)} = \gamma_{j',j}^{(k)}$, for all $1 \leq j < j' \leq p$ and $1 \leq k \leq K$. Moreover, for identifiability reasons, we restrict the diagonal elements $\phi_{j,j} = 1$ and $\gamma_{j,j}^{(k)} = \theta_{j,j}^{(k)}$. Note that $\phi_{j,j'}$ is a common factor across all K categories that controls the occurrence of common links shared across categories, while $\gamma_{j,j'}^{(k)}$ is an individual factor specific to the k-th category. The proposed joint estimation method maximizes the following penalized criterion:

$$\begin{aligned} \max_{\{\mathbf{\Phi}^{(k)},\mathbf{\Gamma}^{(k)}\}_{k=1}^{K}} \quad & \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{p} \Big[x_{i,j}^{(k)} \Big(\theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \Big) \\ & - \log \Big\{ 1 + \exp \Big(\theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \Big) \Big\} \Big] \\ & - \eta_1 \sum_{j<j'} \phi_{j,j'} - \eta_2 \sum_{j<j'} \sum_{k=1}^{K} |\gamma_{j,j'}^{(k)}| \ , \end{aligned} \tag{2.5}$$

where $\mathbf{\Phi}^{(k)} = (\phi_{j,j'})_{p \times p}$ and $\mathbf{\Gamma}^{(k)} = (\gamma_{j,j'}^{(k)})_{p \times p}$. The tuning parameter η_1 controls sparsity of the common structure across the K networks. Specifically, if $\phi_{j,j'}$ is shrunk to zero, all $\theta_{j,j'}^{(1)}, \ldots, \theta_{j,j'}^{(K)}$ are also zero, and hence there is no link between nodes j and j' in any of the K graphs. Similarly, η_2 is a tuning parameter controlling sparsity of links in individual categories. Due to the nature of the ℓ_1 penalty, some of $\gamma_{j,j'}^{(k)}$'s will be shrunk to zero, resulting in a collection of graphs with individual differences. Note that this two-level penalty was originally proposed by Zhou and Zhu [38] for group variable selection in linear regression.

The criterion (2.5) achieves the stated goal of estimating common structure and hence borrow strength across the K data categories, but requires the selection of two tuning parameters. However, there is an equivalent criterion presented next, that only involves a single tuning parameter, thus

simplifying the estimation task.

$$\max_{\{\boldsymbol{\Theta}^{(k)}\}_{k=1}^K} \quad \sum_{k=1}^K \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^p \Big[x_{i,j}^{(k)} \Big(\theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \Big) - \log \Big\{ 1 + \exp \Big(\theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \Big) \Big\} \Big] - \lambda \sum_{1 \leq j < j' \leq p} \sqrt{\sum_{k=1}^K |\theta_{j,j'}^{(k)}|}\,, \tag{2.6}$$

where $\lambda = 2\sqrt{\eta_1 \eta_2}$. The optimization problems given by (**??**) and (2.6) are equivalent in the sense that for each pair of (η_1, η_2) there is a λ that gives the same solution, and vice versa. Their equivalence can be formalized as follows (here $\boldsymbol{A} \cdot \boldsymbol{B}$ denotes the Schur-Hadamard element-wise product of two matrices):

PROPOSITION 1. *Let $\{\widehat{\boldsymbol{\Theta}}^{(k)}\}_{k=1}^K$ be a local maximizer of (2.6). Then there exists a local maximizer of (2.5), $(\widehat{\boldsymbol{\Phi}}, \{\widehat{\boldsymbol{\Gamma}}^{(k)}\}_{k=1}^K)$, such that $\widehat{\boldsymbol{\Theta}}^{(k)} = \widehat{\boldsymbol{\Phi}} \cdot \widehat{\boldsymbol{\Gamma}}^{(k)}$, for all $1 \leq k \leq K$. On the other hand, if $(\widehat{\boldsymbol{\Phi}}, \{\widehat{\boldsymbol{\Gamma}}^{(k)}\}_{k=1}^K)$ is a local maximizer of (2.5), then there also exists a local maximizer of (2.6), $\{\widehat{\boldsymbol{\Theta}}^{(k)}\}_{k=1}^K$, such that $\widehat{\boldsymbol{\Theta}}^{(k)} = \widehat{\boldsymbol{\Phi}} \cdot \widehat{\boldsymbol{\Gamma}}^{(k)}$, for all $1 \leq k \leq K$.*

The proof of this proposition is similar to the proofs of Lemma 1 and Theorem 1 in [38] and is omitted here. Note that even though choosing a single tuning parameter λ corresponds to a particular path in the (η_1, η_2) space, this restriction affects only the individual estimates $\phi_{j,j'}$ and $\gamma_{j,j'}$, but not their product $\theta_{j,j'}$.

2.3. *Algorithm and Model Selection.* Criterion (2.6) leads to an efficient estimation algorithm based on the local linear approximation. Specifically, letting $(\theta_{j,j'}^{(k)})^{[t]}$ denote the estimates from the t-th iteration, we approximate $\sqrt{\sum_{k=1}^K |\theta_{j,j'}^{(k)}|} \approx \sum_{k=1}^K |\theta_{j,j'}^{(k)}| / \sqrt{\sum_{k=1}^K |(\theta_{j,j'}^{(k)})^{[t]}|}$, when $\theta_{j,j'}^{(k)} \approx (\theta_{j,j'}^{(k)})^{[t]}$. Thus, at the $(t+1)$-th iteration, problem (2.6) is decomposed into K individual

optimization problems:

$$\max_{\mathbf{\Theta}^{(k)}} \quad \frac{1}{n_k}\sum_{i=1}^{n_k}\sum_{j=1}^{p}\Big[x_{i,j}^{(k)}\Big(\theta_{j,j}^{(k)}+\sum_{j'\neq j}\theta_{j,j'}^{(k)}x_{i,j'}^{(k)}\Big) - \log\Big\{1+\exp\Big(\theta_{j,j}^{(k)}+\sum_{j'\neq j}\theta_{j,j'}^{(k)}x_{i,j'}^{(k)}\Big)\Big\}\Big] - \lambda\sum_{1\le j<j'\le p}\Big(\sum_{k=1}^{K}|(\theta_{j,j'}^{(k)})^{[t]}|\Big)^{-1/2}|\theta_{j,j'}^{(k)}|\,. \tag{2.7}$$

Note that criterion (2.7) is a variant of criterion (2.3) with a weighted ℓ_1 penalty and hence can be solved by the algorithm of [17]. For numerical stability, we threshold $\sqrt{\sum_{k=1}^{K}|(\theta_{j,j'}^{(k)})^{[t]}|}$ at 10^{-10}. The algorithm is summarized as follows:

Step 1. Initialize $\widehat{\theta}_{j,j'}^{(k)}$'s $(1\le j,j'\le p; 1\le k\le K)$ using the estimates from the separate estimation method;

Step 2. For each $1\le k\le K$, update $\widehat{\theta}_{j,j'}^{(k)}$'s by solving (2.7) using the pseudo-likelihood algorithm [13, 17].

Step 3. Repeat Step 2 until convergence.

The tuning parameter λ in (2.6) controls the sparsity of the resulting estimator and can be selected using cross-validation. Specifically, for each $1\le k\le K$, we randomly split the data in the k-th category into D subsets of similar sizes and denote the index set of the observations in the d-th subset as $\mathcal{T}_d^{(k)}$, $1\le d\le D$. Then λ is selected by maximizing

$$\frac{1}{D}\sum_{d=1}^{D}\sum_{k=1}^{K}\frac{1}{|\mathcal{T}_d^{(k)}|}\sum_{i\in\mathcal{T}_d^{(k)}}\sum_{j=1}^{p} x_{i,j}^{(k)}\Big\{(\widehat{\theta}_{j,j}^{(k)})^{[-d]}(\lambda)+\sum_{j'\neq j}(\widehat{\theta}_{j,j'}^{(k)})^{[-d]}(\lambda)x_{i,j'}^{(k)}\Big\} - \log\Big[1+\exp\Big\{(\widehat{\theta}_{j,j}^{(k)})^{[-d]}(\lambda)+\sum_{j'\neq j}(\widehat{\theta}_{j,j'}^{(k)})^{[-d]}(\lambda)x_{i,j'}^{(k)}\Big\}\Big], \tag{2.8}$$

where $|\mathcal{T}_d^{(k)}|$ is the cardinality of $\mathcal{T}_d^{(k)}$ and $(\widehat{\theta}_{j,j'}^{(k)})^{[-d]}(\lambda)$ is the joint estimate of $\theta_{j,j'}^{(k)}$ based on all observations except those in $\mathcal{T}_d^{(1)}\bigcup\ldots\bigcup\mathcal{T}_d^{(K)}$, as well as the tuning parameter λ.

3. Simulation Study. Before turning our attention to examining the US Senate voting patterns, we evaluate the performance of the joint estimation method on three synthetic examples, each with $p=100$ variables

and $K = 3$ categories. The network structure in each example is composed of two parts: the common structure across all categories and the individual structure specific to a category. The common structures in these examples are a chain graph, a nearest neighbor graph and a scale-free graph. These graphs are generated as follows:

Example 1: Chain Graph. A chain graph is generated by connecting nodes 1 to p in increasing order, as shown in Figure 2 (A1).

Example 2: Nearest Neighbor Graph. The data generating mechanism of the nearest neighbor graph is adapted from Li and Gui [23]. Specifically, we generate p points randomly on a unit square, calculate all $p(p-1)/2$ pairwise distances, and find three nearest neighbors of each point in terms of these distances. The nearest neighbor network is obtained by linking any two points that are nearest neighbors of each other. Figure 2 (B1) illustrates a nearest-neighbor graph.

Example 3: Scale-free Graph. A scale-free graph has a power-law degree distribution and can be simulated by the Barabasi-Albert algorithm [4]. A realization of a scale-free network is depicted in Figure 2 (C1).

In each example, the network for the k-th category ($k = 1, \ldots, K$) is created by randomly adding links to the common structure. The individual links in different categories are disjoint and have the same degree of sparsity, measured by ρ, the ratio of the number of individual links to the number of common links. In particular, $\rho = 0$ corresponds to identical networks for all three categories. In the simulation study, we consider ρ=0, 1/4 and 1, gradually increasing the proportion of individual links (Figure 2). Given the graphs, the symmetric parameter matrix $\mathbf{\Theta}^{(k)}$ is generated as follows. Each $\theta^{(k)}_{j,j'} = \theta^{(k)}_{j',j}$ corresponding to an edge between nodes j and j' is uniformly drawn from $[-1, -0.5] \cup [0.5, 1]$, whereas all other elements are set to zero. Then we generate the data using Gibbs sampling. Specifically, suppose the i-th iteration sample has been drawn and is denoted as $(x^{(k)}_1)^{[t]}, \ldots, (x^{(k)}_p)^{[t]}$; then, in the $(t+1)$-th iteration, we draw $(x^{(k)}_j)^{[t+1]}$, $1 \leq j \leq p$, from the Bernoulli distribution:

$$(x^{(k)}_j)^{[t+1]} \sim \text{Bernoulli}\Big(\frac{\exp(\theta^{(k)}_{j,j} + \sum_{j' \neq j} \theta^{(k)}_{j,j'} (x^{(k)}_{j'})^{[t]})}{1 + \exp(\theta^{(k)}_{j,j} + \sum_{j' \neq j} \theta^{(k)}_{j,j'} (x^{(k)}_{j'})^{[t]})}\Big). \tag{3.1}$$

To ensure that the simulated observations are close to i.i.d. samples from the target distribution, the first 1,000,000 rounds are discarded (burn-in) and the data are collected every 100 iterations from the sampler. In the simulation

study, we consider a balanced scenario and an unbalanced scenario. The former consists of $n_k = 300$ observations in each category, whereas the latter has three unbalanced categories with sample sizes $n_1 = 200$, $n_2 = 300$ and $n_3 = 400$.

We compared the structure estimation results of the joint estimation method and the separate estimation method using ROC curves, which dynamically characterize the sensitivity (proportion of correctly identified links) and the specificity (proportion of correctly excluded links) by varying the tuning parameter λ. Figure 3 shows the ROC curves averaged over 10 replications from the three examples in the balanced scenario, where the joint estimation method dominates separate estimation when the proportion of individual links is low. As ρ increases, the structures become more different, and the joint and separate methods move closer together. This is expected, since the joint estimation method is designed to take advantage of common structure. The results in the unbalanced scenario exhibit a similar pattern (Figure 4).

4. Analysis of the U.S. Senate voting records. We applied the proposed joint estimation method to the voting records of the U.S. Senate from the 109th Congress covering the period 2005-2006. The $p = 100$ variables correspond to the senators. The Senate held 645 votes in that period, from which we extracted $n = 222$ votes in the three largest categories, namely defense and security (141), environment and energy (34), and healthcare (47). The votes are recorded as "yes" (encoded as "1") and "no" (encoded as "0"). The assumption of our model is that bills within a category are an i.i.d. sample from the same underlying Ising model. In reality, the voting process may be more complex, with possible temporal factors and further dependencies amongst bill, possibly reflecting backroom deals. Neverthless, this is an improvement on previous analyses of such data, which treated all bills in all categories as i.i.d. [3], and is a reasonable trade-off for an exploratory data analysis tool.

There were missing observations, as not all senators vote on all bills. The number of bills containing at least one missing vote was 98 out of 141 for defense and security, missing the total of 2.26% of all votes; 24 out of 34 for environment and energy, missing the total of 3.23% of votes; and 20 out of 47 for healthcare, missing 2.38% of all votes. While the number of bills that are missing at least one Senator's vote is relatively high, the overall proportion of missing observations is quite low, and thus we do not expect to create a major problem in the analysis. Nevertheless, we have investigated multiple strategies for imputing the missing data in the electronic supple-

ment; specifically, we considered replacing the missing vote by the party's majority, by the majority vote of the five most similar Senators, and, to test robustness to the imputation method, also by the opposite party's majority and at random. We found that the main conclusions of the analysis are not very sensitive to missing data imputation methods. In the subsequent analysis, we replace a missing vote for a Senator by his/her party's majority vote on the bill; for the Independent Senator Jeffords, we take the Democratic majority vote. After the imputation, the bills with a "yes/no" proportion greater than 90% or less than 10% were excluded from the analysis, as these typically correspond to procedural votes. This left 97, 29, and 40 bills in the three categories, respectively. Given that two of the sample sizes are fairly small (29 and 40), we added an ℓ_2 penalty with a small tuning parameter $\lambda_2 = 0.01$ This approach, known as the elastic net, has been shown to help avoid extremely sparse networks in such situations [39].

The main tuning parameter for our method was selected through cross-validation. Following Li and Gui [23], we used a bootstrap procedure for final edge selection, estimating the network for 100 bootstrap samples of the same size, and only retained edges that appeared more that α percent of the time. This procedure is similar to stability selection [26].

The network representation, depicting both the common and the individual structures with a cut-off value for inclusion $\alpha = 0.4$ and a value of $\lambda = 0.05$ is depicted in Figure 5. Note that unlike techniques such as principal components analysis and multidimensional scaling that directly embed the senators in a two-dimensional map, the proposed method estimates the edges and constructs the adjacency matrix of the graph of Senators; subsequently, we employed a graph drawing program to visualize this graph. The common network structure estimated by the joint estimation method is shown in the top left panel of Figure 5. For the individual categories, we only plot the edges associated with the category that are not part of the common network, to enhance the readability of the graphs. As expected, members of the two political parties are clearly separated. For both tuning parameter values, there are strong positive associations between senators of the same party and selected strong negative associations between senators of opposite parties. Obviously, at the higher tuning parameter value the common dependence structure becomes sparser. Of particular interest is the finding that at both tuning values, there are many more associations between Democratic senators than Republican ones and this pattern holds for both the common and individual structures. One possible explanation may be that during that period the Democrats were in the minority and thus voting more frequently as a block. Further, the Independent Senator Jeffords

is associated with the Democrats, while the moderate Republicans Collins, Snowe, Chafee and Specter (who switched to the Democratic party in early 2009) are not strongly associated with their Republican colleagues, thus confirming results of previous analyses by Clinton, Jackman and Rivers [6] and de Leeuw [8] (albeit based on data from the 105th Congress). The conservative Democrat Nelson (Nebraska) is also not closely associated with his party, as well as the very conservative Republican de Mint (South Carolina). Also, the analysis suggests that Senator Lieberman had a solid Democratic voting record before becoming an Independent in 2008.

Other interesting patterns emerging from the analysis are that the more moderate members of two parties are located closer to the center of their respective "clouds" (e.g. Warner, Frist, Voinovich, Smith on the Republican side and Levin, Reid, Mikulski, Rockefeller on the Democratic side), the cluster of economic conservatives on the Republican side (McConnell, Domenici, Crapo, Inhofe), the close ties of the liberal Democrats Kennedy, Boxer and Nelson (Florida), the close voting records of senators from the same state (Schumer and Clinton from New York, Murkowski and Stevens from Alaska, Snowe and Collins from Maine, Cantwell and Murray from Washington). There is also a strong dependence between Durbin, Corzine, Lincoln, Harkin and Dodd on the Democratic side.

Examining the individual networks for the three categories shown in Figure 5, we note that additional positive associations among Democrats emerge, primarily for defense and healthcare categories, thus indicating a stronger ideological cohesion on these issues. Further, a number of stable negative associations emerge in the environment and healthcare categories, indicating a stronger ideological divide between senators.

On defense, some additional strong ties emerge between more liberal leaning Democrats (Stabenow, Biden, Leahy, Kerry, Boxer), while a strong cluster on environmental issues arises between Republican senators from energy producing states (Murkowski and Stevens from Alaska, Thune from South Dakota, Hutchison from Texas, but also Bond from Missouri, Chambliss from Georgia, Craig from Idaho and Roberts from Kansas with their unwavering support for offshore drilling). On health and medical issues, a number of additional strong positive associations emerge amongst Democratic senators, possibly reflecting the fact that the 109th Congress dealt with issues ranging from veterans affairs, to medical malpractice to food safety and especially on health savings accounts legislation to reduce medical insurance costs.

Different imputation strategies for missing data were also examined and the analysis results are given in Figures 1-3 in the Supplement, for the same

values of the cutoff α and tuning parameter λ. It can be seen that similar patterns emerge, although alternative methods of imputation may lead to the emergence of a few more associations. Nevertheless, the main findings seem to be robust to the examined choices of the imputation mechanism, although at very high levels of absenteeism this may not hold [15].

For comparison purposes, separate multidimensional scaling analyses are shown in Figure 6 for all the votes together and for the three categories separately. MDS (or PCA or factor analysis) is one of the commonly taken approaches in social sciences when graphical modeling is not considered. Figure 6 suggests that the overall vote clustering in the two parties is driven to a large extent by the corresponding clustering in the defense and health categories. On the other hand, voting on environmental issues creates a clear separation between the two parties, although the moderate Republicans Chafee, Collins and Snowe are shown to have a voting record similar to the Democrats, while the Democrats Nelson (Nebraska) and Landrieu are closer to the Republicans. At a high level, MDS-based findings are similar to ours, which is a satisfactory result, but they do not provide explicit clusters or edges, nor do they provide a way to quantify the amount of dependence between individual pairs (visualized via edge thickness in Figure 5).

Another relevant comparison is to fitting a separate graphical model to each of the three categories, as could have been done with any of the previously developed methods for fitting the Ising model. The results are shown in Figure 7, in the same format as in Figure 5, with edges common to all three categories shown under "common structure", and all other edges under their own category. We followed the same tuning procedure as we did for joint estimation, bootstrapping the data 100 times for stability selection and selecting the value of the tuning parameter on a validation dataset. Even with the cut-off set at 1 (we included only the edges appearing in all the bootstrap replications), the graphs are dense and difficult to interpret. Similarly to MDS, they capture party cohesion through strong positive associations between members of the same party for all three categories and some negative associations between members of opposite parties. However, different voting patterns between categories are not clear, although the results suggest a more cohesive voting record for both parties for the defense category. Note that since this is exploratory data analysis, it is hard to verify which set of results is "better". Nevertheless, those obtained from the joint estimation method are more nuanced and interpretable and therefore provide better insights into voting strategies of members of Congress..

5. Concluding Remarks. We have proposed a joint estimation method for the analysis of heterogenous Markov networks motivated by the need to jointly estimate heterogeneous networks, such as those of the Senate voting patterns. The method improves estimation of the networks' common structure by borrowing strength across categories, and allows for individual differences. Asymptotic properties of the method have been established. In particular, we show that the convergence rate is similar to the rate for Gaussian graphical models in a similar context [13]. The proposed method can be extended to deal with general categorical data with more than two levels using the strategy described in Ravikumar, Wainwright and Lafferty [31] and Guo *et al.* [13]. The most interesting feature emerging from the analysis of the Senate voting records is the existence of more stable associations for the Democrats, both in terms of the common structure and in the healthcare and defense categories.

There are other techniques suitable for analyzing roll call data. Dimension reduction techniques create maps, where the relative positioning of the senators allows one to infer similarity in their voting patterns. They provide a useful visual tool to capture broad patterns and relationships. On the other hand, a Markov network model aims directly at estimating the associations between the senators and thus provides an alternative view of the voting patterns, which together with the thresholding technique employed gives a measure of the stability of such associations. Further, the joint estimation method allows one to separately study the overall voting patterns and those driven by specific issues. In our view, both set of techniques are useful, with dimension reduction providing a global perspective, and the Markov model revealing more nuanced patterns.

Acknowledgements. E. Levina's research is partially supported by NSF grants DMS-01106772 and DMS-1159005. G. Michailidis's research is partially supported by NIH grant 1RC1CA145444-0110 and NSF grant DMS-1228164. J. Zhu's research is partially supported by NSF grants DMS-0705532 and DMS-0748389.

References.

[1] Airoldi, E. M. (2007). Getting Started in Probabilistic Graphical Models. *PLoS Computational Biology* **3** e252.

[2] Anandkumar, A., Tan, V. Y. F., Huang, F. and Willsky, A. S. (2012). High-dimensional structure estimation in Ising models: Local separation criterion. **40**.

[3] Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research* **9** 485–516.

[4] Barabasi, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512.

[5] Besag, J. E. (1986). On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc., Ser. B* **48** 259–302.

[6] Clinton, J., Jackman, S. and Rivers, D. (2004). The statistical analysis of roll call data. *American Political Science Review* **98** 355–370.

[7] Danaher, P., Wang, P. and Witten, D. M. (2011). The joint graphical lasso for inverse covariance estimation across multiple classes. *ArXiv e-prints*.

[8] de Leeuw, J. (2006). Principal component analysis of senate voting patterns. In *Real Data Analysis* (S. Sawilowski, ed.) 405–411. Information Age Publishing, North Carolina.

[9] Diaconis, P., Goel, S. and Holmes, S. (2008). Horseshoes in multidimensional scaling and local kernel methods. *Annals of Applied Statistics* 777–807.

[10] Enelow, J. M. and Hinich, M. J. (1984). *The spatial theory of voting: an introduction*. Cambridge University Press, Cambridge.

[11] Gerrish, S. M. (2011). Predicting Legislative Roll Calls from Text. *Young* 489–496.

[12] Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2009). Joint structure estimation of Markov network Technical Report, Department of Statistics, University of Michigan, Ann Arbor.

[13] Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2010). Joint structure estimation for categorical Markov networks Technical Report, Department of Statistics, University of Michigan, Ann Arbor.

[14] Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98** 1-15.

[15] Han, J. H. (2007). Analysing Roll Calls of the European Parliament: A Bayesian Application. *European Union Politics* **8** 479–507.

[16] Hara, S. and Washio, T. (2013). Learning a common substructure of multiple graphical Gaussian models. *Neural Networks* **38** 23 - 38.

[17] Höefling, H. and Tibshirani, R. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *Journal of Machine Learning Research* **10** 883–906.

[18] Højsgaard, S. (2004). Statistical Inference in Context Specific Interaction Models for Contingency Tables. *Scandinavian Journal of Statistics* **31** 143–158.

[19] Jung, S. Y., Park, Y. C., Choi, K. S. and Kim, Y. (1996). Markov random field based English part-of-speech tagging system. In *Proceedings of the 16th Conference on Computational Linguistics* 236–242.

[20] Kolar, M. and Xing, E. P. (2008). Improved estimation of high-dimensional Ising models. In *Eprint arXiv:0811.1239*.

[21] Leeb, H. and Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics* **142** 201-211.

[22] Li, S. Z. (2001). *Markov Random Field Modeling in Image Analysis*. Springer, New York.

[23] Li, H. and Gui, J. (2006). Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics* **7** 302–317.

[24] Matthews, D. R. and A., S. J. (1975). *Yeas and nays: normal decision-making in the U.S. House of Representatives*. Wiley Press, New York.

[25] Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs with the lasso. *Annals of Statistics* **34** 1436–1462.

[26] Meinshausen, N. and Bühlmann, P. (2010). Stability Selection (with discussion). *Journal of the Royal Statistical Society, Series B* **72** 417–473.

[27] Morton, R. B. (1999). *Methods and models: a guide to the empirical analysis of*

formal models in political science. Cambridge University Press, Cambridge.

[28] Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial correlation estimation by joint sparse regression model. *Journal of the American Statistical Asscociation* **104** 735–746.

[29] Poole, K. T. and Rosenthal, H. (1997). *Congress: a political-economic history of roll-call voting*. Oxford University Press, Oxford.

[30] Pötscher, B. M. and Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis* **100** 2065-2082.

[31] Ravikumar, P., Wainwright, M. J. and Lafferty, J. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics* **38** 1287-1319.

[32] Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronical Journal of Statistics* **5** 935-980.

[33] Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515.

[34] Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **1** 1–305.

[35] Xue, L., Zou, H. and Cai, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse Ising models. *Annals of Statistics* **40**.

[36] Yang, S., Pan, Z., Shen, X., Wonka, P. and Ye, J. (2012). Fused Multiple Graphical Lasso. *CoRR* **abs/1209.2139**.

[37] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35.

[38] Zhou, N. and Zhu, J. (2007). Group variable selection via a hierarchical lasso and its oracle property Technical Report, Department of Statistics, University of Michigan.

[39] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* **67** 301–320.

Appendix: Asymptotic Properties. In this section, we study the asymptotic properties of the proposed joint estimation method. Since the structure of the underlying network only depends on the interaction effects, we focus on a variant of the model without main effects. Specifically, we solve

$$\max_{\{\mathbf{\Theta}^{(k)}\}_{k=1}^K} \quad \sum_{k=1}^K \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^p \Big[x_{i,j}^{(k)} \Big(\sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \Big) - \log \Big\{ 1 + \exp \Big(\sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \Big) \Big\} \Big]$$

$$-\lambda \sum_{j<j'} \sqrt{\sum_{k=1}^K |\theta_{j,j'}^{(k)}|} \ . \tag{5.1}$$

We will show that the estimator in criterion (5.1) is consistent in terms of both parameter estimation and model selection, when p and n go to infinity and the tuning parameter λ goes to zero at some appropriate rate. We note that our results are pointwise rather than uniform in Θ, as is standard

in the literature. Some interesting implications of non-uniform bounds for sparse estimators in linear regression have recently been discussed by [21, 30], although their conclusions do not apply to graphical models.

Before stating the main results, we introduce necessary notation and regularity conditions. For each $k = 1, \ldots, K$, denote $\boldsymbol{\theta}^{(k)} = (\theta^{(k)}_{1,2}, \ldots, \theta^{(k)}_{j,j'}, \ldots, \theta^{(k)}_{p-1,p})$ as a $p(p-1)/2$-dimensional vector, recording all upper triangular elements in $\boldsymbol{\Theta}^{(k)}$. Let $\overline{\boldsymbol{\theta}}^{(k)}$ be the true value of $\boldsymbol{\theta}^{(k)}$. Let $\overline{\boldsymbol{Q}}^{(k)}$ be the population Fisher information matrix of the model in criterion (5.1) (see the Appendix for a precise definition) and let $\mathcal{X}^{(k)}_{(i)}$ be a matrix with p rows and $p(p-1)/2$ columns, whose (j, j')-th column is composed of zeros except the j-th (j'-th) component being $x_{i,j'}$ ($x_{i,j}$). In addition, we define $\overline{\boldsymbol{U}}^{(k)} = E[\boldsymbol{\mathcal{X}}^{(k)\mathsf{T}}_{(i)} \boldsymbol{\mathcal{X}}^{(k)}_{(i)}]$. To index the zero and nonzero elements, let $S_k = \{(j, j') : \theta^{(k)}_{j,j'} \neq 0, 1 \leq j < j' \leq p\}$ and $S^c_k = \{(j, j') : \theta^{(k)}_{j,j'} = 0, 1 \leq j < j' \leq p\}$, and let $S_\cap = \bigcap_{k=1}^K S_k$, $S_\cup = \bigcup_{k=1}^K S_k$. The cardinalities of S_k and $S_\cup$ are denoted by q_k and q, respectively. For any matrix $\boldsymbol{W}$ and subsets of row and column indices $\mathcal{U}$ and $\mathcal{V}$, let $\boldsymbol{W}_{\mathcal{U},\mathcal{V}}$ be the matrix consisting of rows $\mathcal{U}$ and columns $\mathcal{V}$ in $\boldsymbol{W}$. Finally, let $\Lambda_{\min}(\cdot)$ and $\Lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalue of a matrix, respectively.

The asymptotic properties of the joint estimation method rely on the following regularity conditions:

(A) Nonzero elements bounds: There exist positive constants $\gamma_{\min}$ and $\gamma_{\max}$ such that

(i) $\min_{1\leq k\leq K} \min_{(j,j')\in S_k} |\overline{\theta}^{(k)}_{j,j'}| \geq \gamma_{\min}$;

(ii) $\max_{1\leq k\leq K} \max_{(j,j')\in S_k\setminus S_\cap} |\overline{\theta}^{(k)}_{j,j'}| \leq \gamma_{\max}$.

(B) Dependency: There exist positive constants $\tau_{\min}$ and $\tau_{\max}$ such that for any $k = 1, \ldots, K$,

$$\Lambda_{\min}(\overline{\boldsymbol{Q}}^{(k)}_{S_k,S_k}) \geq \tau_{\min} \quad \text{and} \quad \Lambda_{\max}(\overline{\boldsymbol{U}}^{(k)}_{S_k,S_k}) \leq \tau_{\max} \ . \tag{5.2}$$

(C) Incoherence: There exists a constant $\tau \in (1 - \sqrt{\gamma_{\min}/4\gamma_{\max}}, 1)$ such that for any $k = 1, \ldots, K$,

$$\|\overline{\boldsymbol{Q}}^{(k)}_{S^c_k,S_k}(\overline{\boldsymbol{Q}}^{(k)}_{S_k,S_k})^{-1}\|_\infty \leq 1 - \tau \ . \tag{5.3}$$

Condition (A) enforces a lower bound on the magnitudes of all nonzero elements, as well as an upper bound on the magnitudes of those nonzero elements associated with individual links. Conditions (B) and (C) bound

the amount of dependence and the influence that the non-neighbors can have on a given node, respectively. Conditions similar to (B) and (C) were also assumed by Meinshausen and Buhlmann [25], Ravikumar, Wainwright and Lafferty [31], Peng *et al.* [28] and Guo *et al.* [13]. Our conditions are most closely related to those of Guo *et al.* [13], but here they are extended to the heterogenous data setting.

THEOREM 1. *(Parameter estimation). Suppose all regularity conditions hold. If the tuning parameter* $\lambda = C_\lambda \sqrt{(\log p)/n}$ *for some constant* $C_\lambda > (8-4\tau)\sqrt{\gamma_{\min}}/(1-\tau)$ *and if* $\min\{n/q^3, n_1/q_1^3, \ldots, n_K/q_K^3\} > (4/C)\log p$ *for some constant* $C = \min\{\tau_{\min}^2\tau^2/288(1-\tau)^2, \tau_{\min}^2\tau^2/72, \tau_{\min}\tau/48\}$*, then there exists a local maximizer of the criterion* (5.1)*,* $\{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K$*, such that, with probability tending to 1,*

$$\sum_{k=1}^K \|\widehat{\boldsymbol{\theta}}^{(k)} - \overline{\boldsymbol{\theta}}^{(k)}\|_2 \le M\sqrt{\frac{q\log p}{n}}\ , \tag{5.4}$$

for some constant $M > (2KC_\lambda/\tau_{\min}\sqrt{\gamma_{\min}})(3-2\tau)/(2-\tau)$.

THEOREM 2. *(Structure selection). Under conditions of Theorem 1, with probability tending to 1, the maximizer* $\{\widehat{\boldsymbol{\theta}}^{(k)}\}_{k=1}^K$ *from Theorem 1 satisfies*

$$\begin{aligned} \widehat{\theta}_{j,j'}^{(k)} &\neq 0, \quad \textit{for all} \quad (j,j') \in S_k, k=1,\ldots,K; \\ \widehat{\theta}_{j,j'}^{(k)} &= 0, \quad \textit{for all} \quad (j,j') \in S_k^c, k=1,\ldots,K\ . \end{aligned}$$

Theorems 1 and 2 establish the consistency in terms of parameter estimation and structure selection, respectively.

The main idea of the proofs is closely related to [13], and some strategies for dealing with the joint estimation are borrowed from [14]. We introduce notation first. For the k-th category, we define the log-likelihood as

$$l(\boldsymbol{\theta}^{(k)}) = \frac{1}{n_k}\sum_{i=1}^{n_k}\sum_{j=1}^{p}[x_{i,j}^{(k)}(\sum_{j'\neq j}\theta_{j,j'}^{(k)}x_{i,j'}^{(k)}) - \log\{1+\exp(\sum_{j'\neq j}\theta_{j,j'}^{(k)}x_{i,j'}^{(k)})\}]\ ,$$

whose first derivative and second derivative are denoted by $\nabla l(\boldsymbol{\theta}^{(k)})$ and $\nabla^2 l(\boldsymbol{\theta}^{(k)})$, respectively. Note that $\nabla l(\boldsymbol{\theta}^{(k)})$ is a $p(p-1)/2$-dimensional vector and $\nabla^2 l(\boldsymbol{\theta}^{(k)})$ is a $p(p-1)/2 \times p(p-1)/2$ matrix. Then, the population Fisher information matrix of the model in (5.1) at $\overline{\boldsymbol{\theta}}$ can be defined as $\overline{\boldsymbol{Q}}^{(k)} = -\mathrm{E}[\nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)})]$, and its sample counterpart is $\widehat{\boldsymbol{Q}}^{(k)} = -\nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)})$.

We also write $\widehat{\boldsymbol{U}}^{(k)} = 1/n \sum_{i=1}^{n} \boldsymbol{\mathcal{X}}_{(i)}^{(k)\mathsf{T}} \boldsymbol{\mathcal{X}}_{(i)}^{(k)}$ for the sample counterpart of $\overline{\boldsymbol{U}}^{(k)}$. Let $\underline{\boldsymbol{\theta}}^{(k)} = (\underline{\theta}_{1,2}^{(k)}, \ldots, \underline{\theta}_{j,j'}^{(k)}, \ldots, \underline{\theta}_{p-1,p}^{(k)})$ be the same as $\boldsymbol{\theta}^{(k)}$ except that all elements in S_k^c are set to zero and write $\boldsymbol{\delta}^{(k)} = \boldsymbol{\theta}^{(k)} - \overline{\boldsymbol{\theta}}^{(k)}$ and $\underline{\boldsymbol{\delta}}^{(k)} = \underline{\boldsymbol{\theta}}^{(k)} - \overline{\boldsymbol{\theta}}^{(k)}$. Finally, let $\mathcal{W}$ be a subset of the index set $\{1, 2, \ldots, p(p-1)/2\}$. For a $p(p-1)/2$-dimensional vector $\boldsymbol{\beta}$, we define $\boldsymbol{\beta}_{\mathcal{W}}$ as the vector consisting of the elements of $\boldsymbol{\beta}$ associated with $\mathcal{W}$.

Next, we introduce a variant of criterion (5.1) by restricting all true zeros in $\{\boldsymbol{\theta}^{(k)}\}_{k=1}^{K}$ to be estimated as zero. Specifically, the restricted criterion is formulated as follows:

$$\max_{\{\underline{\boldsymbol{\theta}}^{(k)}\}_{k=1}^{K}} \sum_{k=1}^{K} l(\underline{\boldsymbol{\theta}}^{(k)}) - \lambda \sum_{1 \le j < j' \le p} \sqrt{\sum_{k=1}^{K} |\underline{\theta}_{j,j'}^{(k)}|} \ , \tag{5.5}$$

and its maximizer is denoted by $\{\widehat{\underline{\boldsymbol{\theta}}}^{(k)}\}_{k=1}^{K}$. In addition, we consider the sample versions of regularity conditions (B) and (C).

(B′) Sample dependency: There exist positive constants $\tau_{\min}$ and $\tau_{\max}$ such that for any $k = 1, \ldots, K$,

$$\Lambda_{\min}(\widehat{\boldsymbol{Q}}_{S_k,S_k}^{(k)}) \ge \tau_{\min} \quad \text{and} \quad \Lambda_{\max}(\widehat{\boldsymbol{U}}_{S_k,S_k}^{(k)}) \le \tau_{\max} \ . \tag{5.6}$$

(C′) Sample incoherence: There exists a constant $\tau \in (1 - \sqrt{\gamma_{\min}/4\gamma_{\max}}, 1)$ such that for any $k = 1, \ldots, K$,

$$\|\widehat{\boldsymbol{Q}}_{S_k^c,S_k}^{(k)} (\widehat{\boldsymbol{Q}}_{S_k,S_k}^{(k)})^{-1}\|_{\infty} \le 1 - \tau \ . \tag{5.7}$$

For convenience of the readers, the proof of our main result is divided into two parts: Part I presents the main idea of the proof by listing the important propositions and the proofs of Theorems 1 and 2, whereas Part II contains additional technical details and proofs of propositions in Part I.

Part I: Propositions and Proof of Theorems 1 and 2. The proof consists of the following steps. Proposition 2 shows that, under sample regularity conditions (B′) and (C′), the conclusions of Theorems 1 and 2 hold for the local maximizer of the restricted problem (5.5). Next, Proposition 3 proves that the population regularity conditions (B) and (C) give rise to their sample counterparts (B′) and (C′) with probability tending to one; hence the conclusions of Proposition 2 also hold with the population regularity conditions. Lastly, we show that the local maximizer of (5.5) is also a local maximizer of the original model (5.1). This is established via Proposition

4, which sets out the Karush-Kuhn-Tucker (KKT) conditions for the local maximizer of criterion (5.1), and Proposition 5, which shows that, with probability tending to one, the local maximizer of (5.5) satisfies these KKT conditions.

PROPOSITION 2. *Suppose condition (A) and the sample conditions (B′) and (C′) hold. If the tuning parameter* $\lambda = C_\lambda\sqrt{(\log p)/n}$ *for some constant* $C_\lambda > (8-4\tau)\sqrt{\gamma_{\min}}/(1-\tau)$ *and* $q\sqrt{(\log p)/n} = o(1)$*, then with probability tending to one, there exists a local maximizer of the restricted criterion,* $\{\underline{\widehat{\boldsymbol{\theta}}}^{(k)}\}_{k=1}^K$*, satisfying*

(i) $\sum_{k=1}^K \|\underline{\widehat{\boldsymbol{\theta}}}^{(k)} - \overline{\boldsymbol{\theta}}^{(k)}\|_2 \le M\sqrt{q(\log p)/n}$ *for some constant* $M > (2KC_\lambda/\tau_{\min}\sqrt{\gamma_{\min}})[(3-2\tau)/(2-\tau)]$*;*

(ii) *For each* $k = 1,\ldots,K$*,* $\underline{\widehat{\theta}}^{(k)}_{j,j'} \neq 0$ *for all* $(j,j') \in S_k$ *and* $\underline{\widehat{\theta}}^{(k)}_{j,j'} = 0$ *for all* $(j,j') \in S_k^c$*.*

PROPOSITION 3. *Suppose the regularity conditions (B) and (C) hold, then for any* $\epsilon > 0$*, the following inequalities hold with probability tending to one for all* $k = 1,\ldots,K$*:*

(i) $\mathrm{P}\{\Lambda_{min}(\widehat{\boldsymbol{Q}}^{(k)}_{S_k,S_k}) \le \tau_{min} - \epsilon\} \le 2\exp\{-(\epsilon^2/2)(n_k/q_k^2) + 2\log q_k\}$*;*

(ii) $\mathrm{P}\{\Lambda_{\max}(\widehat{\boldsymbol{U}}^{(k)}_{S_k,S_k}) \ge \tau_{\max} + \epsilon\} \le 2\exp\{-(\epsilon^2/2)(n_k/q_k^2) + 2\log q_k\}$*;*

(iii) $\mathrm{P}[\|\widehat{\boldsymbol{Q}}^{(k)}_{S_k^c,S_k}(\widehat{\boldsymbol{Q}}^{(k)}_{S_k,S_k})^{-1}\|_\infty \ge 1-\tau/2] \le 12\exp(-Cn_k/q_k^3 + 4\log p)$*, for some constant* $C = \min\{\tau_{min}^2\tau^2/288(1-\tau)^2, \tau_{min}^2\tau^2/72, \tau_{min}\tau/48\}$*.*

PROPOSITION 4. $\{\widehat{\boldsymbol{\theta}}\}_{k=1}^K$ *is a local maximizer of problem (5.1) if and only if the following conditions hold for all* $k = 1,\ldots,K$*:*

$$\begin{array}{rcll} \nabla_{j,j'} l(\widehat{\boldsymbol{\theta}}^{(k)}) & = & \lambda \mathrm{sgn}(\widehat{\theta}^{(k)}_{j,j'})/(\sum_{k=1}^K |\widehat{\theta}^{(k)}_{j,j'}|)^{1/2}, & \text{if } \widehat{\theta}^{(k)}_{j,j'} \neq 0; \\ |\nabla_{j,j'} l(\widehat{\boldsymbol{\theta}}^{(k)})| & < & \lambda/(\sum_{k=1}^K |\widehat{\theta}^{(k)}_{j,j'}|)^{1/2}, & \text{if } \widehat{\theta}^{(k)}_{j,j'} = 0. \end{array} \tag{5.8}$$

PROPOSITION 5. *Under all conditions of Proposition 2, with probability tending to one, we have, for each* $k = 1,\ldots,K$*,*

$$\begin{array}{rcll} \nabla_{j,j'} l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)}) & = & \lambda \mathrm{sgn}(\underline{\widehat{\theta}}^{(k)}_{j,j'})/(\sum_{k=1}^K |\underline{\widehat{\theta}}^{(k)}_{j,j'}|)^{1/2}, & \text{for all } (j,j') \in S_k; \\ |\nabla_{j,j'} l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})| & < & \lambda/(\sum_{k=1}^K |\underline{\widehat{\theta}}^{(k)}_{j,j'}|)^{1/2}, & \text{for all } (j,j') \in S_k^c. \end{array} \tag{5.9}$$

Proof of Theorems 1 and 2. The condition $\min\{n/q^3, n_1/q_1^3, \ldots, n_K/q_K^3\} > (4/C)\log p$ implies that, for each $k = 1,\ldots,K$, we have $-Cn_k/q_k^3 + 4\log p <$

0 and $-(\epsilon^2/2)(n_k/q_k^2) + 2\log q_k < 0$ when q_k is large enough. This condition also implies $q\sqrt{(\log p)/n} = o(1)$. In addition, by Proposition 3, the sample conditions (B′) and (C′) hold with probability tending to one when regularity conditions (B) and (C) hold. Therefore, by Proposition 2, with probability tending to one, the solution of the restricted problem $\{\widehat{\underline{\boldsymbol{\theta}}}^{(k)}\}_{k=1}^K$ satisfies both parameter estimation consistency and structure selection consistency. On the other hand, by Proposition 5, with probability tending to one, $\{\widehat{\underline{\boldsymbol{\theta}}}^{(k)}\}_{k=1}^K$ also satisfies the KKT conditions in Proposition 4, thus it is a local maximizer of criterion (5.1). This proves Theorems 1 and 2. □

Part II: Proofs of Propositions. Before proving the propositions, we state a few lemmas which will be used in the proofs. These lemmas are variants of Lemmas 1, 2 and 5 in Guo *et al.* [13], adapted to the settings of the heterogenous model and thus the proofs are omitted here. Likewise, the proof of Proposition 3 is very similar to the proof of Propositions 3 and 4 in Guo *et al.* [13] and is omitted.

LEMMA 1. *For each $k = 1, \ldots, K$, with probability tending to 1, we have $\|\nabla l(\overline{\boldsymbol{\theta}}^{(k)})\|_\infty \le C_\nabla \sqrt{(\log p)/n}$ for some constant $C_\nabla > 4$.*

LEMMA 2. *If the sample dependency condition (B′) holds and $q\sqrt{(\log p)/n} = o(1)$, then for any $\alpha_k \in [0,1]$, $k = 1, \ldots, K$, the following inequality holds with probability tending to 1:*

$$-\sum_{k=1}^K \boldsymbol{\delta}_{S_k}^{(k)\mathsf{T}} [\nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)} + \alpha_k \underline{\boldsymbol{\delta}}^{(k)})]_{S_k,S_k} \boldsymbol{\delta}_{S_k}^{(k)} \ge \frac{1}{2}\tau_{\min} \sum_{k=1}^K \|\underline{\boldsymbol{\delta}}^{(k)}\|_2^2 \ . \tag{5.10}$$

LEMMA 3. *Suppose the sample dependency condition (B) holds. For any $\alpha_k \in [0,1]$, $k = 1, \ldots, K$, the following inequality holds with probability tending to one:*

$$\|[\nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)} + \alpha_k \underline{\boldsymbol{\delta}}^{(k)}) - \nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)})]\underline{\boldsymbol{\delta}}^{(k)}\|_\infty \le \tau_{\max} \|\underline{\boldsymbol{\delta}}^{(k)}\|_2^2 \ . \tag{5.11}$$

Proof of Proposition 2. The main idea of the proof was first introduced in this context in Rothman *et al.* [33] and has since been used by many authors. Define

$$\mathrm{G}(\{\underline{\boldsymbol{\delta}}^{(k)}\}_{k=1}^K) = -\sum_{k=1}^K [l(\overline{\boldsymbol{\theta}}^{(k)} + \underline{\boldsymbol{\delta}}^{(k)}) - l(\overline{\boldsymbol{\theta}}^{(k)})] + \lambda \sum_{1\le j<j'\le p} \{(\sum_{k=1}^K |\overline{\theta}_{j,j'}^{(k)} + \underline{\delta}_{j,j'}^{(k)}|)^{1/2} - (\sum_{k=1}^K |\overline{\theta}_{j,j'}^{(k)}|)^{1/2}\}. \tag{5.12}$$

It can be seen from (5.5) that, $\{\widehat{\underline{\boldsymbol{\delta}}}^{(k)}\}_{k=1}^K$ minimizes $\mathrm{G}(\{\underline{\boldsymbol{\delta}}^{(k)}\}_{k=1}^K)$ and $\mathrm{G}(\{\mathbf{0}\}_{k=1}^K) = 0$. Thus we must have $\mathrm{G}(\{\widehat{\underline{\boldsymbol{\delta}}}^{(k)}\}_{k=1}^K) \leq 0$. If we take a closed set $\mathcal{A}$ which contains $\{\mathbf{0}\}_{k=1}^K$, and show that G is strictly positive everywhere on the boundary $\partial\mathcal{A}$, then it implies that G has a local minimum inside $\mathcal{A}$, since G is continuous and $\mathrm{G}(\{\mathbf{0}\}_{k=1}^K) = 0$. Specifically, we define $\mathcal{A} = \{\{\underline{\boldsymbol{\delta}}^{(k)}\}_{k=1}^K : \sum_{k=1}^K \|\underline{\boldsymbol{\delta}}^{(k)}\|_2 \leq Ma_n\}$, with boundary $\partial\mathcal{A} = \{\{\underline{\boldsymbol{\delta}}^{(k)}\}_{k=1}^K : \sum_{k=1}^K \|\underline{\boldsymbol{\delta}}^{(k)}\|_2 = Ma_n\}$, for some constant $M > (2KC_\lambda/\tau_{\min}\sqrt{\gamma_{\min}})[(3-2\tau)/(2-\tau)]$ and $a_n = \sqrt{q(\log p)/n}$. For any $\{\underline{\boldsymbol{\delta}}^{(k)}\}_{k=1}^K \in \partial\mathcal{A}$, the Taylor series expansion gives $\mathrm{G}(\{\underline{\boldsymbol{\delta}}^{(k)}\}_{k=1}^K) = I_1 + I_2 + I_3$, where

$$
\begin{aligned}
I_1 &= -\sum_{k=1}^K [\nabla l(\overline{\boldsymbol{\theta}}^{(k)})]_{S_k}{}^{\mathsf{T}} \boldsymbol{\delta}_{S_k}^{(k)} \ , \\
I_2 &= -\sum_{k=1}^K \boldsymbol{\delta}_{S_k}^{(k)\mathsf{T}} [\nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)} + \alpha_k \underline{\boldsymbol{\delta}}^{(k)})]_{S_k,S_k} \boldsymbol{\delta}_{S_k}^{(k)}, \text{ for some } \alpha_k \in [0,1] \ , \\
I_3 &= \lambda \sum_{(j,j')\in S_\cup} \{(\sum_{k=1}^K |\overline{\theta}_{j,j'}^{(k)} + \underline{\delta}_{j,j'}^{(k)}|)^{1/2} - (\sum_{k=1}^K |\overline{\theta}_{j,j'}^{(k)}|)^{1/2}\} \ . \qquad (5.13)
\end{aligned}
$$

Since $C_\lambda > (8-4\tau)\sqrt{\gamma_{\min}}/(1-\tau)$, we have $[(1-\tau)/(2-\tau)]C_\lambda/\sqrt{\gamma_{\min}} > 4$. By Lemma 1,

$$
|I_1| \leq \sum_{k=1}^K \|[\nabla l(\overline{\boldsymbol{\theta}}^{(k)})]_{S_k}\|_\infty \|\boldsymbol{\delta}_{S_k}^{(k)}\|_1 \leq [(1-\tau)C_\lambda M \gamma_{\min}^{-1/2}/(2-\tau)](q\log p)/n \ . \qquad (5.14)
$$

In addition, by condition $q\sqrt{(\log p)/n} = o(1)$, Lemma 2 holds and thus

$$
I_2 \geq (\tau_{\min}/2)\sum_{k=1}^K \|\underline{\boldsymbol{\delta}}^{(k)}\|_2^2 \geq [\tau_{\min}/(2K)]M^2 q(\log p)/n \ . \qquad (5.15)
$$

Finally, by the triangular inequality and regularity condition (A),

$$
\begin{aligned}
|I_3| &\leq \lambda \sum_{(j,j')\in S_\cup} \sum_{k=1}^K \frac{||\overline{\theta}_{j,j'}^{(k)} + \underline{\delta}_{j,j'}^{(k)}| - |\overline{\theta}_{j,j'}^{(k)}||}{(\sum_{k=1}^K |\overline{\theta}_{j,j'}^{(k)} + \underline{\delta}_{j,j'}^{(k)}|)^{1/2} + (\sum_{k=1}^K |\overline{\theta}_{j,j'}^{(k)}|)^{1/2}} \\
&\leq (\lambda\gamma_{\min}^{-1/2}) \sum_{k=1}^K \sum_{(j,j')\in S_\cup} |\underline{\delta}_{j,j'}^{(k)}| \leq (\lambda q^{1/2}\gamma_{\min}^{-1/2}) \sum_{k=1}^K \|\underline{\boldsymbol{\delta}}^{(k)}\|_2 \\
&\leq (MC_\lambda\gamma_{\min}^{-1/2})\{q(\log p)/n\} \qquad (5.16)
\end{aligned}
$$

Then we have

$$\mathrm{G}(\{\underline{\boldsymbol{\delta}}^{(k)}\}_{k=1}^K) \geq M^2 \frac{q\log p}{n}\Big(\frac{\tau_{\min}}{2K} - \frac{(1-\tau)C_\lambda}{(2-\tau)M\gamma_{\min}^{1/2}} - \frac{C_\lambda}{M\gamma_{\min}^{1/2}}\Big) > 0. \tag{5.17}$$

The last inequality uses the condition $M > (2KC_\lambda/\tau_{\min}\sqrt{\gamma_{\min}})[(3-2\tau)/(2-\tau)]$. Therefore, with probability tending to 1, we have $\sum_{k=1}^K \|\underline{\widehat{\boldsymbol{\theta}}}^{(k)} - \overline{\boldsymbol{\theta}}^{(k)}\|_2 \leq M\sqrt{q(\log p)/n}$, and consequently claim (i) in Proposition 2 holds.

On the other hand, by the definition of $\underline{\widehat{\boldsymbol{\theta}}}^{(k)}$, we have $\underline{\widehat{\theta}}_{j,j'}^{(k)} = 0$ for all $(j,j') \in S_k^c$. By regularity condition (A) and Proposition 2 (i), for any $(j,j') \in S_k$, $k = 1,\ldots,K$, we have $|\underline{\widehat{\theta}}_{j,j'}^{(k)}| \geq |\overline{\theta}_{j,j'}^{(k)}| - |\underline{\widehat{\theta}}_{j,j'}^{(k)} - \overline{\theta}_{j,j'}^{(k)}| \geq \gamma_{\min}/2 > 0$, when n is large enough. □

Proof of Proposition 5. By Proposition 2, with probability tending to one, we have $\underline{\widehat{\theta}}_{j,j'} \neq 0$ for all $(j,j') \in S_k$. Since $\{\underline{\widehat{\boldsymbol{\theta}}}^{(k)}\}_{k=1}^K$ is a local maximizer of the restricted problem (5.5), with probability tending to one, $\nabla_{j,j'} l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)}) = \lambda \mathrm{sgn}(\underline{\widehat{\theta}}_{j,j'}^{(k)})/(\sum_{k=1}^K |\underline{\widehat{\theta}}_{j,j'}^{(k)}|)^{1/2}$, for all $(j,j') \in S_k$.

To show the second claim, we apply the mean value theorem and write $\nabla l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)}) = \nabla l(\overline{\boldsymbol{\theta}}^{(k)}) + \boldsymbol{r}^{(k)} - \widehat{\boldsymbol{Q}}^{(k)}\underline{\widehat{\boldsymbol{\delta}}}^{(k)}$, where $\boldsymbol{r}^{(k)} = \{\nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)} + \alpha_k \underline{\widehat{\boldsymbol{\delta}}}^{(k)}) - \nabla^2 l(\overline{\boldsymbol{\theta}}^{(k)})\}\underline{\widehat{\boldsymbol{\delta}}}^{(k)}$. After some simplifications, we have

$$[\nabla l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})]_{S_k^c} = [\nabla l(\overline{\boldsymbol{\theta}}^{(k)})]_{S_k^c} + \boldsymbol{r}_{S_k^c}^{(k)} - [\widehat{\boldsymbol{Q}}_{S_k^c,S_k}^{(k)}(\widehat{\boldsymbol{Q}}_{S_k,S_k}^{(k)})^{-1}]\{[\nabla l(\overline{\boldsymbol{\theta}}^{(k)})]_{S_k} + \boldsymbol{r}_{S_k}^{(k)} - [\nabla l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})]_{S_k}\} \tag{5.18}$$

and thus,

$$\begin{aligned}
\|[\nabla l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})]_{S_k^c}\|_\infty &\leq \|[\nabla l(\overline{\boldsymbol{\theta}}^{(k)})]_{S_k^c}\|_\infty + \|\boldsymbol{r}_{S_k^c}^{(k)}\|_\infty \\
&\quad + \|\widehat{\boldsymbol{Q}}_{S_k^c,S_k}^{(k)}(\widehat{\boldsymbol{Q}}_{S_k,S_k}^{(k)})^{-1}\|_\infty\{\|[\nabla l(\overline{\boldsymbol{\theta}}^{(k)})]_{S_k}\|_\infty + \|\boldsymbol{r}_{S_k}^{(k)}\|_\infty + \|[\nabla l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})]_{S_k}\|_\infty\} \\
&\leq (2-\tau)\|\nabla l(\overline{\boldsymbol{\theta}}^{(k)})\|_\infty + (2-\tau)\|\boldsymbol{r}^{(k)}\|_\infty + (1-\tau)\|[\nabla l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})]_{S_k}\|_\infty \\
&\leq [(1-\tau)C_\lambda/\sqrt{\gamma_{\min}}]\sqrt{(\log p)/n} + (2-\tau)\tau_{\max}M^2 q(\log p)/n \\
&\quad + (1-\tau)\lambda/\min_{(j,j')\in S_k}[\sum_{k=1}^K |\underline{\widehat{\theta}}_{j,j'}|]^{1/2} \\
&\leq [2(1-\tau)/\sqrt{\gamma_{\min}}]\lambda + o_p(\lambda).
\end{aligned} \tag{5.19}$$

On the other hand, $\lambda/[\sum_{k=1}^K |\underline{\widehat{\theta}}_{j,j'}^{(k)}|]^{1/2} = +\infty$ when $(j,j') \in S_\cup^c$. Otherwise, if $(j,j') \in S_\cup \backslash S_k$, then

$$\lambda/(\sum_{k=1}^K |\underline{\widehat{\theta}}_{j,j'}|)^{1/2} \geq \lambda/\{\sum_{k=1}^K (|\underline{\widehat{\theta}}_{j,j'} - \overline{\theta}_{j,j'}| + |\overline{\theta}_{j,j'}|\}^{1/2} \geq \lambda/\sqrt{\gamma_{\max}} \geq (2-2\tau)\lambda/\sqrt{\gamma_{\min}}\,.$$

Thus, for any $(j, j') \in S_k^c$ $(k = 1, \ldots, K)$, we have

$$
\begin{aligned}
|\nabla_{j,j'} l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})| &\leq \max_{1 \leq k \leq K} \max_{(j,j') \in S_k^c} |\nabla_{j,j'} l(\underline{\widehat{\boldsymbol{\theta}}}^{(k)})| \\
&< \min_{1 \leq k \leq K} \min_{(j,j') \in S_k^c} \lambda \Big/ \sqrt{\sum_{k=1}^{K} |\underline{\widehat{\theta}}_{j,j'}^{(k)}|} \;\leq\; \lambda \Big/ \sqrt{\sum_{k=1}^{K} |\underline{\widehat{\theta}}_{j,j'}^{(k)}|} \,. \qquad (5.20)
\end{aligned}
$$

□

Department of Biostatistics
655 Huntington Avenue
Boston, MA 02115
E-mail: jguo@hsph.harvard.edu

Department of Statistics,
1085 South University,
Ann Arbor, MI 48109-1107
E-mail: jieche@umich.edu, E-mail: elevina@umich.edu
E-mail: gmichail@umich.edu, E-mail: jizhu@umich.edu

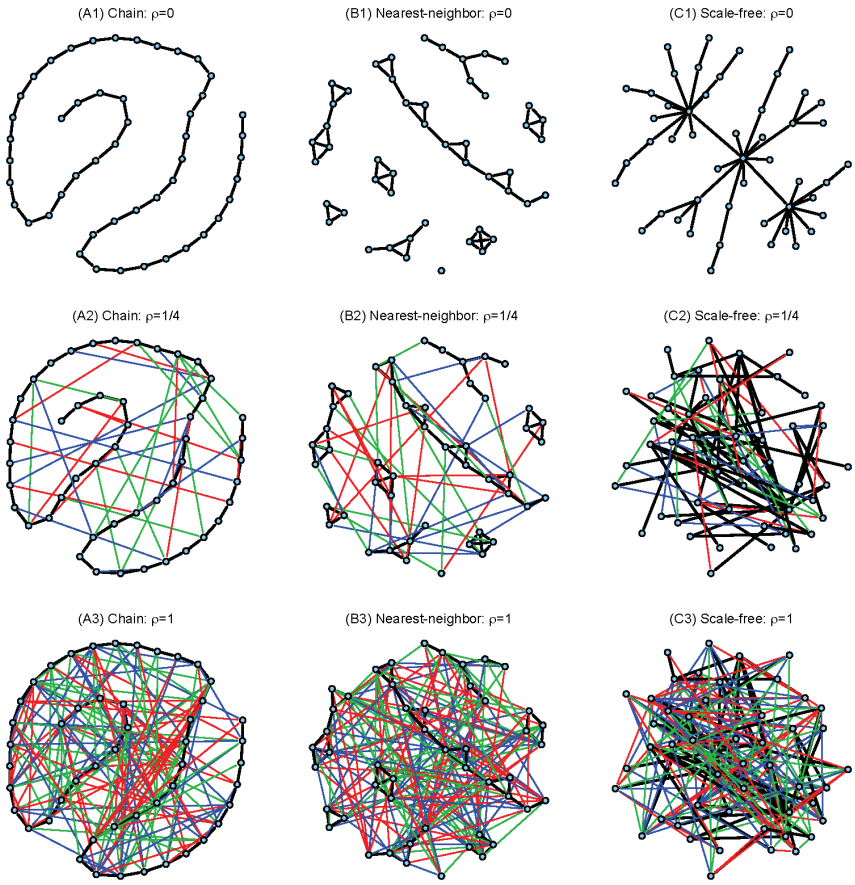

FIG 2. *The networks used in three simulated examples. The black lines represent the common structure, whereas the red, blue and green lines represent the individual links in the three categories.* ρ *is the ratio of the number of individual links to the number of common links.*

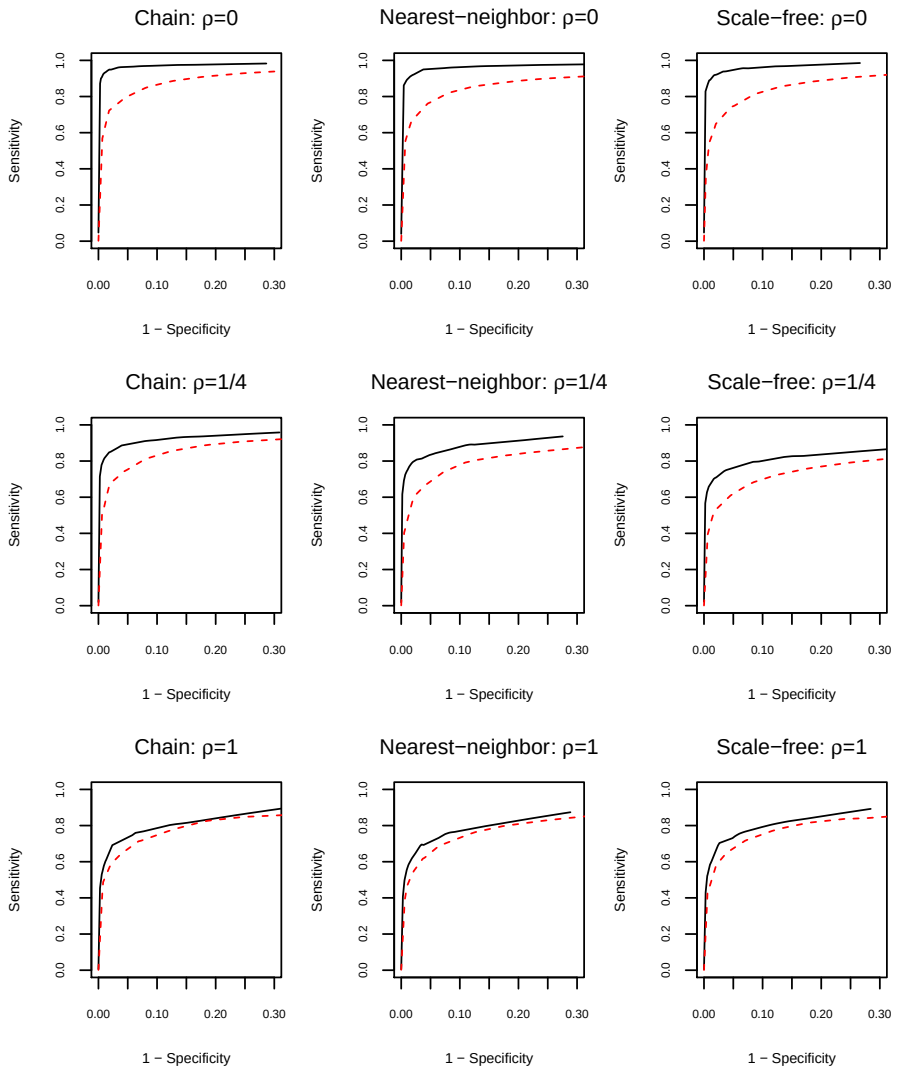

FIG 3. *Results for the balanced scenario ($n_1 = n_2 = n_3 = 300$) and dimension $p = 100$. Black solid curve: joint estimation; red dashed curve: separate estimation. The ROC curves are averaged over 10 replications. ρ is the ratio between the number of individual links and the number of common links.*

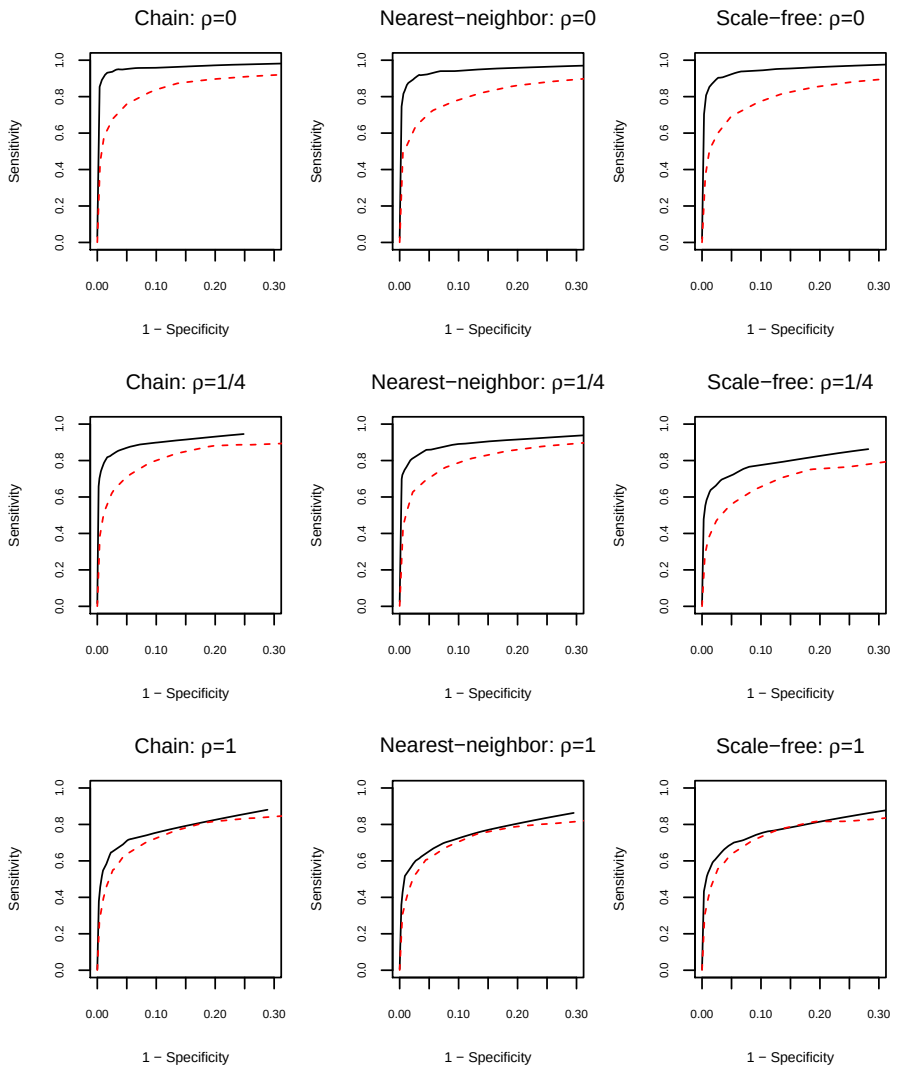

FIG 4. *Results for the unbalanced scenario ($n_1 = 200$, $n_2 = 300$, $n_3 = 400$) and dimension $p = 100$. Black solid curve: joint estimation; red dashed curve: separate estimation.The ROC curves are averaged over 10 replications. ρ is the ratio between the number of individual links and the number of common links.*

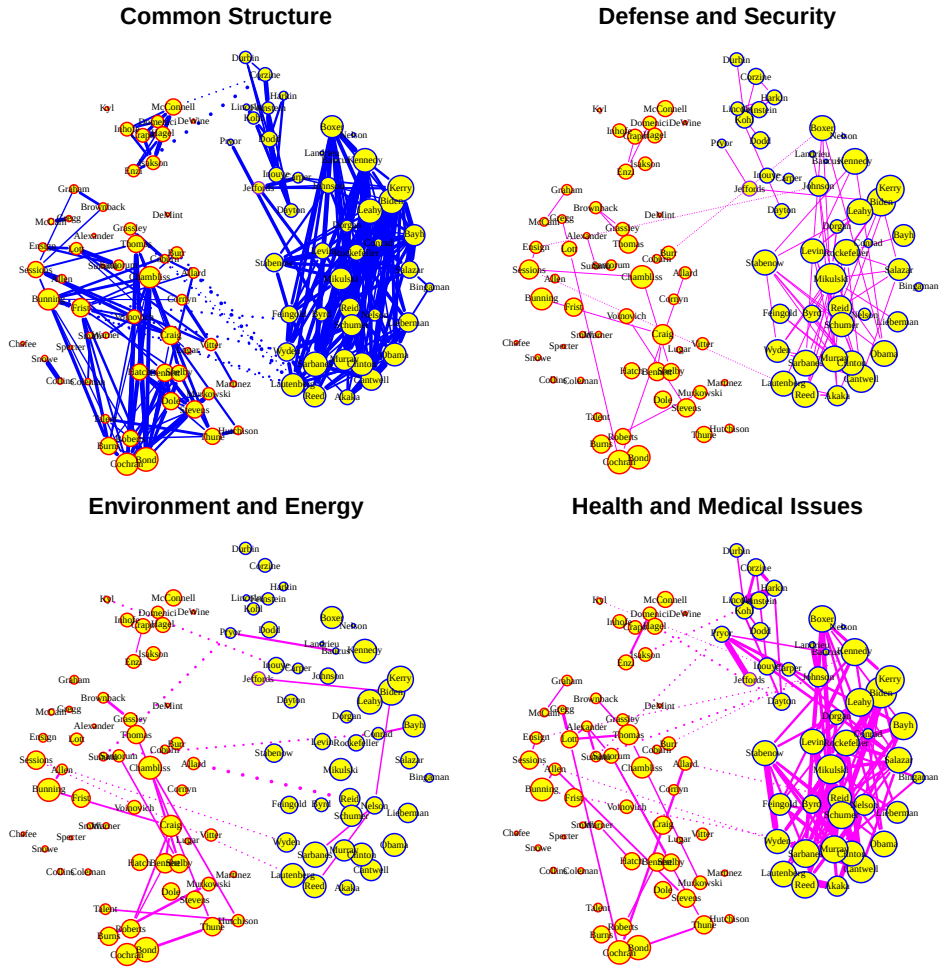

FIG 5. *The estimated graphical models for the three categories in the Senate voting data with an inclusion cutoff value of 0.4 and tuning parameter value of 0.5. Edges common to all three categories are shown under the heading "common structure"; all other edges are shown on category-specific graphs. The nodes represent the 100 senators, with red, blue and purple node colors corresponding to Republican, Democrat, or Independent (Senator Jeffords), respectively. A solid line corresponds to a positive interaction effect and a dashed line to a negative interaction effect. The width of a link is proportional to the magnitude of the corresponding overall interaction effect.*

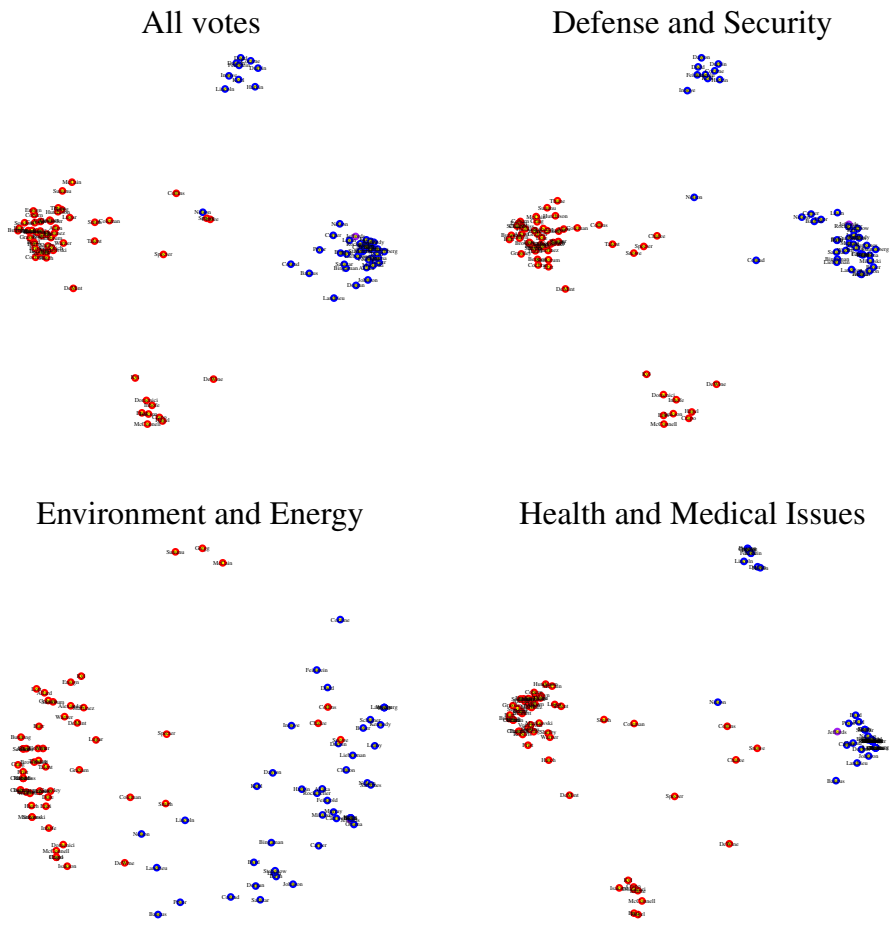

FIG 6. *Multidimensional scaling analysis for all the votes together, and the three individual categories. The nodes represent the 100 senators, with red, blue and purple node colors corresponding to Republican, Democrat, or Independent (Senator Jeffords), respectively.*

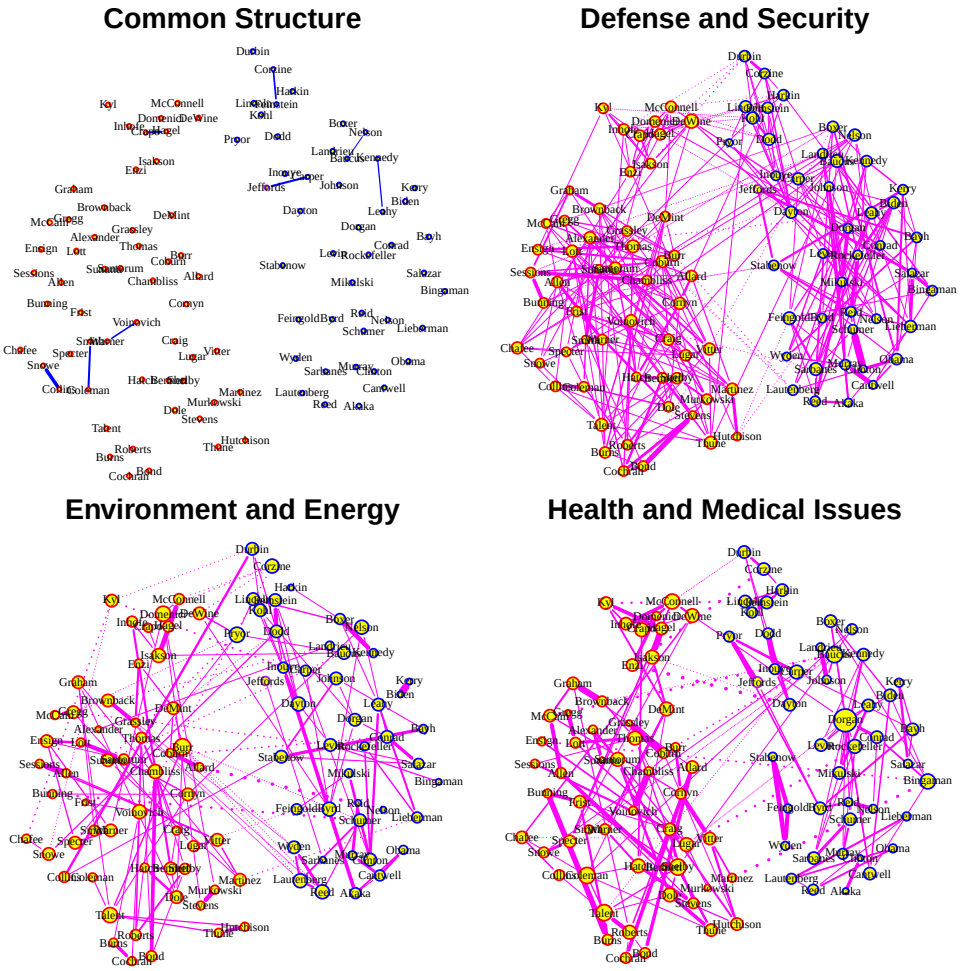

FIG 7. *The estimated graphical models for the three categories in the Senate voting data fitted via separate estimation. Edges common to all three categories are shown under the heading "common structure"; all other edges are shown on category-specific graphs. The cut-off value is 1 (only edges appearing in all bootstrap replications are included). The nodes represent the 100 senators, with red, blue and purple node colors corresponding to Republican, Democrat, or Independent (Senator Jeffords), respectively. A solid line corresponds to a positive interaction effect and a dashed line to a negative interaction effect. The width of a link is proportional to the magnitude of the corresponding overall interaction effect.*