Prediction Approach for Ising Model Estimation

Jinyu Li*, Pan Yu†, Hongfeng Yu†, Qi Zhang*
*Department of Statistics
University of Nebraska-Lincoln, Lincoln, US
†Department of Computer Science and Engineering
University of Nebraska-Lincoln, Lincoln, US
Email: qi.zhang@unl.edu

***Abstract*—We consider the graph estimation for Ising model from observed binary data. Popular approaches in the literature are largely penalized sparse selection procedures that depend on tuning parameters to be selected. The output of such procedures is usually one single sparse graph without any ranking information of the individual edges. In scientific practice, however, it is more desirable to be able to rank all potential edges based on their statistical significance, and select the sparse graph by thresholding. In this paper, we propose a novel PRediction Approach for Ising Model Estimation (PRAIME). The proposed framework reformulates Ising model estimation as the prediction of the observed data, and provides an estimate and a statistical significance measure of the Ising model parameter for each node pair using only the predicted values. Thus it enables the ranking all potential edges and the flexible sparse graph selection by thresholding, and allows the researchers to use the predictive algorithm of their choice. We implemented PRAIME using random forest, illustrated the advantage of PRAIME over the penalized sparse selection approaches in accuracy and flexibility using synthetic data, and applied it to a congress co-sponsorship dataset.**

***Index Terms*—graphical model, Ising model, random forest**

I. Introduction

Probabilistic graphical models have enabled the scientists from various domains to infer the dependency structure among variables. Very often, such dependencies can be conveniently modeled as Markov Random Fields (MRF). We focus on the binary MRF, i.e., the Ising model. Ising model with chain structure was first proposed in statistical mechanics for modeling ferromagnetism [1]. Later it was extended to lattice and arbitrary graphs, and became widely used in spatial statistics [2], imaging processing [3] and neural science [4], for which the graph structures are usually known.

Graph learning for Ising model has also drawn more attention in the last decade. For identifiability and interpretability, the unknown graph is preasumbly sparse. A popular approach of sparse graph estimation for Ising model is to perform separate node-wise penalized logistic regression, which is referred to as neighborhood selection in the literature [5], [6]. This can be considered as the binary extension of [7] which is for Gaussian graphical model (GGM). Similar to graphical lasso [8] that provides a penalized maximum likelihood estimator (MLE) for the sparse GGM, penalized maximum pseudo-likelihood estimators have also been proposed for sparse Ising model estimation [9], [10]. It can be seen as solving all node-wise logistic regressions jointly without giving up the symmetric constraints. A related line of research approximate penalized MLE using MCMC samples from Gibbs sampler where the sampling distributions are essentially the node-wise components of the pseudo-likelihood [11], [12]. All the above methods formulate the Ising model estimation as a sparse model selection problem, and propose penalized optimization procedures that require predefined regularization parameters. Despite the recent progress on penalty parameter selection, penalized model selection remains a difficult problem as these parameters are often un-intuitive, and the optimal selection of them may require additional "hyper" tuning parameters. Furthermore, penalized model selection approaches do not provide ranking of statistical significance of the edges, and there is no precise control over the sparsity of the resultant graph. An output graph with reasonable number of edges are often from numerous trial-and-error experiments with various penalty parameters. The cause of these problems lays in the sparsity constraints of optimization objectives, which cannot be solved by further development along this direction, even though it may be alleviated.

We propose to address these issues by a complete decoupling of the Ising model parameters, and the edge-wise estimation of them with associated uncertainty measures. Since all parameters are estimated separately, the algorithm is embarrassingly parallel. All the edges can be ranked based on their statistical significance, and further thresholding yields a sparse graph with appropriate number of edges. Our estimators only depend on the predicted conditional probabilities of the observed data. Thus there is no structural assumptions such as sparsity of the graph to be estimated, or parametric assumptions on the predictive models utilized. In this paper, we use random forest for predictions. But it can be replaced by other probabilistic predictive models such as neural network.

II. Background

A. Notations

We first introduce the default notations used in this paper. For a length q vector x, and set $A \subset \{1, \ldots, q\}$ with $|A| = s$, we use x_A to denote the length s subvectors of x with coordinates in A, and $x_{-A} = x_{A^c}$ where A^c is the complement set of A. Similarly, for a $q \times p$ matrix X, and sets $A \subset \{1, \ldots, q\}$ with $|A| = s$ and $B \subset \{1, \ldots, p\}$ with $|B| = r$, we use $X_{A,B}$ to denote the $s \times r$ submatrix of X with rows in A and columns in B, and $X_{-A,B} = X_{A^c,B}$,

$X_{A,-B} = X_{A,B^c}$ and $X_{-A,-B} = X_{A^c,B^c}$. In particular, we use X_{jk} to denote the element of X in its jth row and kth column, $X_{A,\star}$ as the $s \times p$ submatrix of X composed of the rows in A and $X_{\star,B}$ the $q \times r$ submatrix consisting of the columns in B.

B. Ising Models

Let $Y = (Y_1, Y_2, \ldots, Y_p)^T \in \{0,1\}^p$ be a length p binary random vector, and assume that it follows the Ising model with probability mass function

$$f_Y(y) = \frac{1}{Z(\Omega)} \exp\left(-y^T \Omega y\right) \tag{1}$$

where the symmetric parameter matrix $\Omega = (\Omega_{jk})_{p\times p}$, and $Z(\Omega) = \sum_{y\in\{0,1\}^p} \exp\left(-y^T\Omega y\right)$ is the normalization constant.

Let $G(\Omega) = (V, E)$ be the sparse graph among the p coordinates of Y induced by Ω, where $V = \{1, \ldots, p\}$ and E is the set of the pairs of the row and column ID's of the nonzero off-diagonal elements of Ω. Since the graph is undirected, these pairs are un-ordered. The conditional independence structure among the p variables is encoded in this graph, and there is

$$\Omega_{jk} = 0 \Leftrightarrow Y_j \perp Y_k | Y_{-\{j,k\}}.$$

Suppose we observe $y^{(1)}, y^{(2)}, \ldots, y^{(n)}$, n iid samples from the above Ising model. The goal of this paper is to use these observed data to estimate the Ising model parameter matrix Ω, especially recovering its sparse graph structure.

C. Node-wise Logistic Regression

This graph estimation problem is equivalent to identifying the neighboring vertex set for each node. This observation leads to the neighborhood selection methods for the Ising model graph estimation [5], [6].

For an arbitrary node $u \in \{1, \ldots, p\}$, the conditional distribution of $Y_u|Y_{-u}$ satisfies

$$logit\left(P(Y_u = 1|Y_{-u} = y_{-u})\right) = -\Omega_{uu} - 2\sum_{w\in ne(u)} \Omega_{wu} y_w$$

where $ne(u) = \{v \in V : \{v, u\} \in E, v \neq u\}$ is the neighborhood of the node u.

Neighborhood selection approaches reformulate the graph estimation as p separate variable selection problems, one for each node. They can be solved using penalized regression techniques such as the lasso [13]. The variable selection consistency of the lasso for logistic regression assures the graph selection consistency of the neighborhood selection procedures. The idea of neighborhood selection was first arisen in the context of GGM estimation [7]. [5] can be regarded as its extension to the binary MRF, and it was later generalized to the a wider range of MRFs with node-wise conditional distributions from certain exponential families [14], [15]. Node-wise regression ignores the fact that Ω, the parameter matrix of the Ising model, is symmetric. Thus the output parameter matrix needs to be symmetrized, and the final estimate of Ω_{jk} could be the minimum or the maximum of the corresponding estimates from the two logistic regression models for nodes j and k.

D. Classical Pseudo-likelihood

The Ising model likelihood is notoriously intractable. Even when the graph structure is known, parameter estimation via maximizing the likelihood is difficult. A computationally tractable alternative without much loss in accuracy is maximizing the following pseudo-likelihood based on the node-wise conditional distributions

$$\breve{\ell}(\Omega; y^{(1)}, \ldots, y^{(n)}) = \sum_{j=1}^{p}\left[\sum_{i=1}^{n} \log\left(P(Y_j = y_j^{(i)}|Y_{-j} = y_{-j}^{(i)})\right)\right] \tag{2}$$

where

$$\begin{aligned}&\log\left(P(Y_j = y_j^{(i)}|Y_{-j} = y_{-j}^{(i)})\right)\\ =& -y_j^{(i)}\left(\Omega_{jj} + 2\sum_{k\neq k} y_k^{(i)}\Omega_{jk}\right) - \Phi(y_{-j}^{(i)}, \Omega)\end{aligned}$$

with $\Phi(y_{-j}^{(i)}, \Omega) = \log\left(1 + \exp(-\Omega_{jj} - 2\sum_{k\neq j} y_k^{(i)}\Omega_{jk})\right)$ as the log-normalization constant.

Pseudo-likelihood based approach was first proposed by [16] for Ising model parameter estimation on lattice graph, and was later extended to sparse graph estimation with a graph sparsity penalty imposed [9], [10]. These pseudo-likelihood based methods for sparse graph estimation takes advantage of the simple structure of the conditional distributions so that more computationally efficient estimation algorithms become possible.

The conditional likelihood of each node only depends on one row (or column) of the parameter matrix Ω. The node-wise logistic regression procedures are in fact approximates of the pseudo-likelihood estimator by maximizing the p pieces of the pseudo-likelihood for each node separately to estimate the rows (or columns) of Ω without the symmetry constraint. The pseudo-likelihood methods can be seen as solving the p logistic regressions together. This connection between them is similar to that between graphical lasso and [7], with the only difference being that the exact penalized MLE is feasible for GGM but not for Ising model due to the computational burden in evaluating the partition function as discussed above.

III. Proposed Method: PRAIME

A. Pairwise conditional likelihood

The pseudo-likelihood (2) belongs to a wider class of composite likelihood methods aiming at performing statistical inference based on the product of a collection of simpler component likelihoods instead of the full likelihood [17]. The individual components of the composite likelihood could be the conditional distributions of the individual variables given the values of all other variables as in (2), the marginal distribution of the individual variables as commonly used in variational Bayes methods, the pairwise marginal distributions

of the variables, or the marginal distributions of some other simple functions of pairs of variables.

In this paper, we propose to consider the following composite likelihood for Ising model based on pairwise conditional distributions.

$$\begin{aligned}&\tilde{\ell}(\Omega; y^{(1)}, \ldots, y^{(n)}) \\ =&\sum_{j=1}^{p-1}\sum_{k=j+1}^{p}\left[\sum_{i=1}^{n}\log\left(P(Y_{(j,k)} = y_{(j,k)}^{(i)} | Y_{-(j,k)} = y_{-(j,k)}^{(i)})\right)\right]\end{aligned} \tag{3}$$

Define $r_{10} = \Omega_{j,j} + 2\Omega_{j,-(j,k)}y_{-(j,k)}$, $r_{01} = \Omega_{k,k} + 2\Omega_{k,-(j,k)}y_{-(j,k)}$, and there is

$$\begin{aligned}&P(Y_{(j,k)} = \delta | Y_{-(j,k)} = y_{-(j,k)}) \\ \propto&\begin{cases} 1 & \delta = (0,0) \\ \exp[-r_{10}] & \delta = (1,0) \\ \exp[-r_{01}] & \delta = (0,1) \\ \exp[-r_{10} - r_{01} - 2\Omega_{jk}] & \delta = (1,1) \end{cases}\end{aligned} \tag{4}$$

Let

$$\pi_{jk}^{\delta}(y_{-(j,k)}) \equiv P(Y_{(j,k)} = \delta | Y_{-(j,k)} = y_{-(j,k)}) \tag{5}$$

for $\delta \in S = \{(0,0),(0,1),(1,0),(1,1)\}$, and it immediately follows that

$$\Omega_{jk} = -\frac{1}{2}\log\left(\frac{\pi_{jk}^{(1,1)}(y_{-(j,k)})\pi_{jk}^{(0,0)}(y_{-(j,k)})}{\pi_{jk}^{(0,1)}(y_{-(j,k)})\pi_{jk}^{(1,0)}(y_{-(j,k)})}\right),$$

which only depends on the conditional distribution of $Y_{(j,k)}$, but not the other components of the composite likelihood (3).

B. Pairwise Classification

We propose to take advantage the above observation, and estimate each element of the parameter matrix completely separately using an individual component pairwise conditional likelihood.

For each distinct node pair (j,k), we treat the bivariate random vector $Y_{(j,k)}$ as a categorical variable with four classes $S = \{(0,0),(0,1),(1,0),(1,1)\}$, and estimate the conditional distribution $P(Y_{(j,k)}|Y_{-(j,k)} = y_{-(j,k)})$ using a probabilistic multi-class classification algorithm such as logistic regression or random forest. We use $\hat{\pi}_{jk}^{\delta}(y_{-(j,k)}^{(i)})$ for $\delta \in S$ to denote the predicted values of the conditional probabilities (5) for sample i.

We propose a PRediction Approach for Ising Model Estimation (PRAIME) which estimates Ω_{jk} with

$$\widehat{\Omega}_{jk} = n^{-1}\sum_{i=1}^{n} h_{jk}^{(i)} \tag{6}$$

where

$$h_{jk}^{(i)} = -\frac{1}{2}\log\left(\frac{\hat{\pi}_{jk}^{(1,1)}(y_{-(j,k)}^{(i)})\hat{\pi}_{jk}^{(0,0)}(y_{-(j,k)}^{(i)})}{\hat{\pi}_{jk}^{(0,1)}(y_{-(j,k)}^{(i)})\hat{\pi}_{jk}^{(1,0)}(y_{-(j,k)}^{(i)})}\right)$$

The PRAIME outputs are estimates of Ω_{jk} for all node pairs. A statistical significance measure for each can be calculated according to the properties of the probabilistic classification algorithm used. For example, if implemented with multi-class logistic regressions, $\widehat{\Omega}_{jk}$ is simply a linear combination of the estimated class-specific intercepts, and the corresponding Wald statistic can be used for ranking all potential edges.

The theoretical properties of PRAIME also depend on the properties of the predictive algorithm used, and a comprehensive case-by-case investigation is beyond the scope of this paper. The following proposition on the consistency of PRAIME holds in general.

Proposition 1 (Consistency)**.** *The estimator* (6) *is consistent if the mean square prediction error of the probabilistic multi-class classifier goes to zero and if the true and the predicted conditional probabilities are bounded above zero.*

Proof. If the true and the predicted conditional probabilities for all samples are all abounded above $d > 0$, then for some constant $C_d > 0$ that depends on d, there is

$$(h_{jk}^{(i)} - \Omega_{jk})^2 \le C_d \cdot \sum_{\delta \in S}[\hat{\pi}_{jk}^{\delta}(y_{-(j,k)}^{(i)}) - \pi_{jk}^{\delta}(y_{-(j,k)}^{(i)})]^2.$$

Aggregating both sides of the above over $i = 1, \ldots, n$ gives the following upper bound for the estimation error of (6)

$$E(\widehat{\Omega}_{jk} - \Omega_{jk})^2 \le E\left[\frac{1}{n}\sum_{i=1}^{n}(h_{jk}^{(i)} - \Omega_{jk})^2\right] \le C_d \cdot MSPE_{jk},$$

where the mean square prediction error of the multi-class probabilistic classifier is defined as

$$MSPE_{jk} = E\left\{n^{-1}\sum_{i=1}^{n}\sum_{\delta \in S}[\hat{\pi}_{jk}^{\delta}(y_{-(j,k)}^{(i)}) - \pi_{jk}^{\delta}(y_{-(j,k)}^{(i)})]^2\right\}$$

So the proposed estimator is consistent if $MSPE_{jk} \to 0$ for the probabilistic classifier used.

□

C. Random Forest Graph Estimator

We propose to implement PRAIME using random forest [18] as the classifier due to its overall superior empirical performance and simplicity in training and tuning. We use the out-of-bag samples for prediction. We refer to this version of PRAIME as PRAIME-RF. In the literature of conditional independence graph estimation, random forest has been used within the node-wise regression framework where the neighborhood for each node is selected based on random forest's variable importance measure [19]. In contrast, PRAIME-RF only relies on the prediction performance of random forest.

For PRAIME-RF, we propose to rank the potential edges using an empirical Bayes framework. In detail, for node pair (j,k), we define its t statistic as

$$Z_{jk} = \frac{\widehat{\Omega}_{jk}}{sd_{jk}/\sqrt{n}} \tag{7}$$

where sd_{jk} is the standard deviation of sequence $\{h_{jk}^{(i)}\}_{i=1}^{n}$.

In principle, all potential edges can be ranked based on Z_{jk}'s, and there is no edge between (j,k) if it is small. In practice, we apply further empirical Bayes adjustment, as it is unclear whether these t statistics have an appropriate theoretical null distribution (the distribution when there is no true edge). Using all $\{Z_{jk} : 1 \leq j < k \leq p\}$ as the input z scores, we calculated their local false discovery rates following Efron's framework [20, Chapter 5]. For node pair (j,k), we refer to its local false discovery rate as $locfdr_{jk}$. Roughly speaking, $locfdr_{jk}$ has the interpretation of $P(\Omega_{jk} = 0|y^{(1)}, \ldots, y^{(n)})$. So the statistical significance of the edge decreases as locfdr increases. Ranking all edges (j,k) with $j < k$ in the ascending order of $locfdr_{jk}$, and let $locfdr^{\star}_m$ be the mth value in this sequence, the False Discovery Rate (FDR) for the graph with exactly M edges is

$$FDR^{\star}_M = M^{-1} \sum_{m=1}^{M} locfdr^{\star}_m$$

Thus the user can threshold the graph based on prespecified FDR control or domain knowledge.

We summarize PRAIME-RF in Algorithm 1.

Algorithm 1 PRAIME-RF

Input: $y^{(1)}, y^{(2)}, \ldots, y^{(n)}$, n iid observations from an Ising model with unknown parameter matrix Ω

Output: Estimated graph structure of Ω.

1: **for** `each node pair` $1 \leq j < k \leq p$ **do**
2: Use random forest to predict $Y_{(j,k)}$ as a multi-class response, and obtain $\hat{\pi}^{\delta}_{jk}(y^{(i)}_{-(j,k)})$, the out-of-bag probabilistic predictions of $P(Y_{(j,k)} = \delta | Y_{-(j,k)} = y^{(i)}_{-(j,k)})$ for $\delta \in S$ and $i = 1, \ldots, n$.
3: Estimate Ω_{jk} by (6)
4: Calculate the t statistic by (7)
5: **end for**
6: Calculate the local false discovery rates $locfdr_{jk}$ for $1 \leq j < k \leq p$ according to [20, Chapter 5].
7: Rank the edges in the ascending order of locfdr and select the sparse graph by hard thresholding.

IV. Experiments

In this section, we investigate PRAIME-RF's ranking performance and the properties of the resultant graphs in experiments using both synthetic and real world datasets. We compared PRAIME-RF with many other competitors in synthetic data evaluation, including: (1) an efficient pseudo-likelihood method with L1 sparsity penalty (EPL) [10]; and (2) a neighborhood selection method [6] where extended BIC [21] is used to select the penalty parameters of the lasso logistic regressions for neighborhood selection. We refer to this neighborhood selection method as NS-BIC(γ) where γ is the tuning parameter of the extended BIC. We consider $\gamma = 0, 0.5, 1$, same as in [6]. For the real data evaluation, however, we have to skip EPL due to the limitation in computational speed and memory, even though it is already much more computationally efficient than the previous pseudo-likelihood algorithms [9].

A. Synthetic Data Evaluation

We investigate the ranking performance using synthetic data. We simulate data using R package *IsingSampler* [22] with the Ising model parameter matrix $\Omega = \tau I_p - \theta[G - diag(G\mathbf{1_p})]$ where G is a binary symmetric sparse matrix whose induced graph is a $\sqrt{p} \times \sqrt{p}$ 2D lattice. In this simulation model, the strength of the edges increases with θ and the proportion of 1's in the simulated outcomes decreases as τ increases. The existing literature of Ising model estimation (e.g., [6]) predominantly focuses on the cases where the two classes of the outcomes are balanced (roughly 50% of 0 and 1, respectively). This is equivalent to $\tau = 0$ in our simulation setup, and we also include more realistic settings where the data is imbalanced.

In our simulations, we fixed $n = 1000$, $p = 64$, and the $\sqrt{p} \times \sqrt{p}$ lattice as the induced graph. We remark that there are $p(p-1)/2 = 2016$ node pairs in the graph. Hence the number of parameters is actually greater than the sample size n. There are 112 true edges in the graph, roughly 5.6% of all node pairs. We consider all nine combinations of $\theta \in \{0.25, 0.5, 1\}$ and $\tau \in \{0, 0.025, 0.05\}$, and repeated each setting for $N = 40$ times.

We first investigate how the accuracy (the proportion of true edges among the selected) of PRAIME-RF change with the proportion of node pairs selected (Figure 1). Since only 5.6% of the node pairs are true edges in this graph, the accuracy of the top candidates is a more relevant measure of ranking performance than measures of the whole ranking list such as area under ROC. We find that the top candidates selected by PRAIME-RF are very accurate, and selecting the top 5.6% leads to almost perfect graph selection in many settings. While EPL does not provide a ranking of all edges directly, its regularization trace does output a sequence of graphs with increasing density as a discrete approximate ranking. We plot the accuracy of these graphs as curve in these figures, and found that it does not perform as well as PRAIME-RF, since the curve accuracy of EPL is always lower than the corresponding one for PRAIME-RF.

Neighborhood selection methods cannot return exact or approximate ranking of graphs, as p regularization paths are involved, and their optimal penalty parameters may be different. Instead, neighborhood selection methods typically only output one single graph. For each simulation setting, and each $\gamma = 0, 0.5, 1$ in NS-BIC(γ), we also plotted a marker in Figure 1 representing its average proportion of node pairs selected (X axis value) and the average accuracy (Y axis value) of its output graphs across simulation replicates. We first find that NS-BIC(γ) tends to select graphs much sparser than the true graphs. Even for the same proportion of edges selected, the accuracy of NS-BIC(γ) are lower than PRAIME-RF in the most of the cases, except when its selected graph is almost empty (e.g., when $\theta = 0.25$). In contrast, PRAIME-RF enables

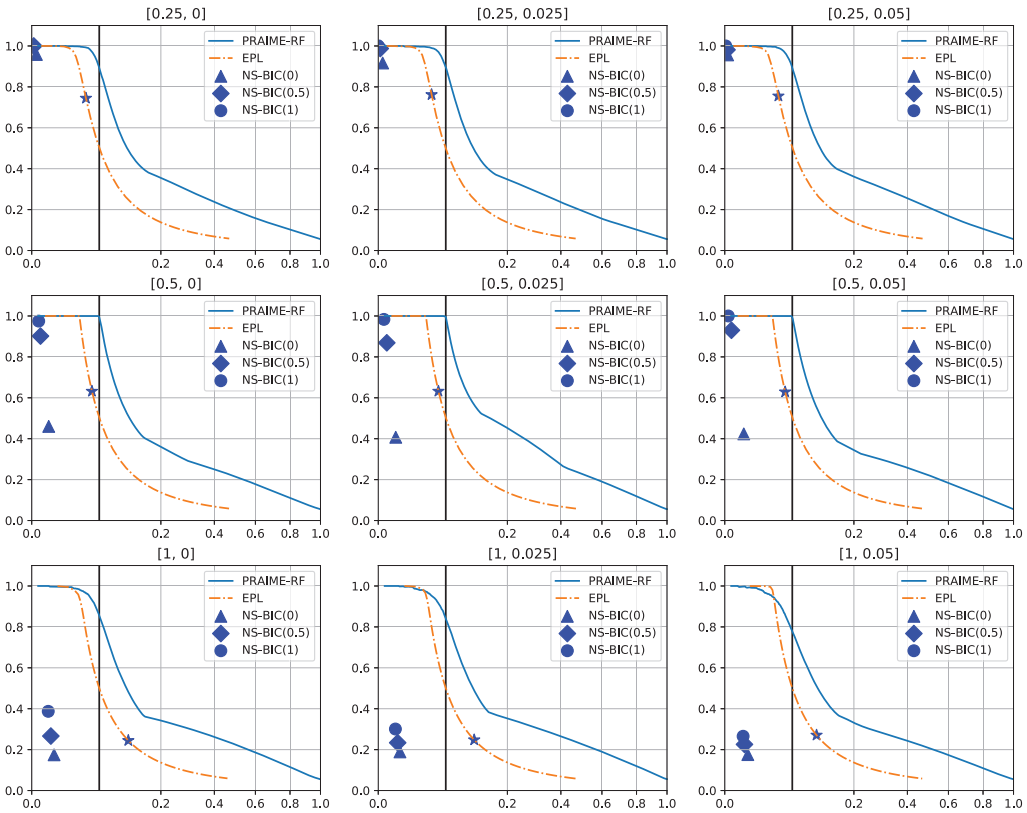

Fig. 1. Ranking accuracy in simulations. The titles of the panels are the values of the simulation parameters (θ, τ). The X-axis represents the proportion of node pairs selected (in square-root scale), and the Y-axis represents the accuracy, i.e., the proportion of true edges in the selected node pairs. We present the mean accuracy curve for PRAIME-RF and EPL methods. The vertical lines represent the sparsity level of the true graph (5.6% of node pairs are true edges). The markers in each panel show the average proportion of selected edges and the average accuracy for NS-BIC(γ), a neighborhood selection method [6]. The markers triangle, diamond and circle represent the performance of BIC with the parameter $\gamma = 0, 0.5, 1$ respectively. The star markers represent the the average proportion of selected edges and the average accuracy for EPL with optimal λ selected by cross-validation. The position of the markers are determined by the mean accuracy (y-axis) and the proportion of detected edges (x-axis).

the flexibility of selecting sparse graphs with high accuracy based on user-provided thresholds in all cases.

We also compared the computational costs of these methods (Table I), and find that NS-BIC is the fastest, but it does not compensate the loss in accuracy as shown in Figure 1. In the remaining two methods, PRAIME-RF is faster than EPL.

We further studied the False Discovery Rate (FDR) control of PRAIME-RF (Table II), and find that the empirical FDR are reasonably close to the nominal level in the majority of the case. We remark that the pseudo-likelihood methods and neighborhood selection methods cannot provide any natural FDR control at all.

TABLE I
THE MEAN AND STANDARD DEVIATION (IN PARENTHESES) OF COMPUTATION TIMES (IN SECONDS) ACROSS 40 SIMULATIONS.

(θ, τ)	PRAIME-RF	NS-BIC(0)	NS-BIC(0.5)	NS-BIC(1)	EPL
(0.25, 0)	46.24(0.23)	0.07(0.01)	0.07 (0.01)	0.07 (0.03)	101.83(7.12)
(0.25, 0.025)	46.36(0.24)	0.06 (0.01)	0.06 (0.01)	0.06 (0.01)	102.43(6.85)
(0.25, 0.05)	46.34(0.25)	0.06 (0.01)	0.06 (0.01)	0.06 (0.01)	101.75(7.12)
(0.5, 0)	46.20(0.25)	0.08 (0.01)	0.08 (0.01)	0.08 (0.01)	101.10(5.12)
(0.5, 0.025)	46.45(0.26)	0.08 (0.01)	0.08 (0.01)	0.08 (0.01)	101.20(4.71)
(0.5, 0.05)	46.18(0.28)	0.08 (0.01)	0.08 (0.01)	0.08 (0.01)	97.55(4.54)
(1,0)	45.62(0.24)	0.14 (0.02)	0.14 (0.02)	0.14 (0.02)	157.25(6.62)
(1, 0.025)	45.54(0.25)	0.12 (0.02)	0.12 (0.02)	0.12 (0.02)	147.62(5.62)
(1, 0.05)	45.82(0.26)	0.10 (0.01)	0.10 (0.01)	0.10 (0.01)	134.94(6.01)

B. *Analysis of House Co-sponsorship data*

We analyze the US House of Representatives cosponsorship dataset [23] for the 109th (January 3, 2005 - January 3, 2007)

TABLE II

THE MEAN AND STANDARD DEVIATION (IN PARENTHESES) OF PROPORTION OF CANDIDATE EDGES SELECTED (% EDGES) BY PRAIME-RF AND THE EMPIRICAL FDR (EFDR) ACROSS 40 SIMULATIONS AT THE NOMINAL FDR LEVEL 0.05, 0.1 AND 0.2.

	FDR=0.05		FDR=0.10		FDR=0.20	
(θ, τ)	% edges	EFDR	% edges	EFDR	% edges	EFDR
(0.25, 0)	0.0491(0.0038)	0.0651(0.0282)	0.0559(0.0037)	0.1313(0.0396)	0.0697(0.0047)	0.2500(0.0446)
(0.25, 0.025)	0.0484(0.0033)	0.0527(0.0287)	0.0562(0.0037)	0.1224(0.0389)	0.0689(0.0054)	0.2414(0.0526)
(0.25, 0.05)	0.0495(0.0031)	0.0703(0.0250)	0.0578(0.0038)	0.1426(0.0410)	0.0712(0.0054)	0.2647(0.0479)
(0.5, 0)	0.0615(0.0017)	0.1068(0.0253)	0.0662(0.0027)	0.1683(0.0342)	0.0754(0.0054)	0.2676(0.0568)
(0.5, 0.025)	0.0616(0.0022)	0.1056(0.0332)	0.0660(0.0037)	0.1624(0.0511)	0.0746(0.0075)	0.2534(0.0862)
(0.5, 0.05)	0.0611(0.0014)	0.1019(0.0201)	0.0661(0.0022)	0.1645(0.0289)	0.0750(0.0045)	0.2708(0.0491)
(1,0)	0.0451(0.0062)	0.0611(0.0312)	0.0547(0.0060)	0.1244(0.0404)	0.0349(0.0034)	0.2539(0.0559)
(1, 0.025)	0.0369(0.0097)	0.0434(0.0293)	0.0479(0.0097)	0.1102(0.0388)	0.0641(0.0102)	0.2326(0.0627)
(1, 0.05)	0.0164(0.0089)	0.0375(0.0425)	0.0284(0.0099)	0.0605(0.0529)	0.0454(0.0117)	0.1507(0.0755)

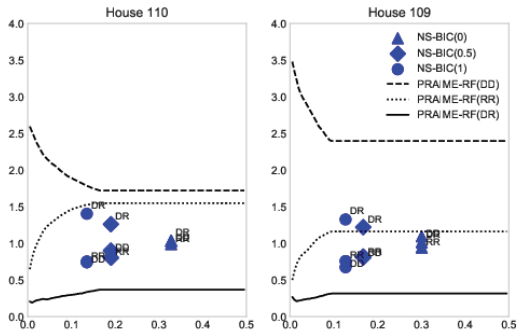

Fig. 2. Edge densities within each party and between party, normalized by the overall density of the corresponding network. The curves are for PRAIME-RF, and the markers are for the neighborhood selection method NS-BIC(γ). For each of $\gamma = 0, 0.5, 1$, the points for the edge densities within Democrats, within Republications and between parties are annotated with "DD", "RR" and "DR", respectively.

and 110th (January 3, 2007 - January 3, 2009) Congresses. For each bill introduced to the house of representatives, there must be one sponsor congressperson. Then the other members of the house can express their support to the bill by signing as cosponsors. A house member may sponsor/cosponsor a bill due to its ideological appealingness, or his/her social relationship with the other congresspersons supporting it.

Let p be the number of the members of the house, and n be the number of bills introduced. We observe $y^{(i)} = (y_1^{(i)}, \ldots, y_p^{(i)})^T \in \{0,1\}^p$ for $i = 1, \ldots, n$ where $y_j^{(i)} = 1$ if the congressperson j sponsor/cosponsor bill j. We model $y^{(i)}$ for $i = 1, \ldots, n$ as independent samples from an Ising model. The sparse graph structure induced by the parameter matrix of this Ising model contains the information on the interdependence among the members of the house. Each bill usually only receives cosponsorship from less than 5% of the congress members, and typically each member of the house only sponsors or cosponsors no more than 4% of the bills introduced. So this cosponsorship dataset is very imbalanced as there are much less 1's than 0's in the outcomes.

Since congresspersons in the same party collaborate more often, they are more likely to cosponsor the same bills. Previous studies on cosponsorhip network have also confirmed that the initial two large community detected in this network are roughly along the party line [24]. In the absence of the ground truth in real data analysis, we use the party affiliation as an approximate, i.e., we expect higher edge densities within each party than that between the two parties in the sparse graph induced by the estimated Ising model parameter matrix.

For PRAIME-RF, we rank all node pairs by locfdr, and investigated how the within party and between party edge densities change as the threshold change (Figure 2). As expected, we find that the edge densities within each party are always higher than that between the two parties. But this difference starts vanishing as higher proportions of edges are introduced in the network, suggesting that the network may become less informative if the threshold is too loose. We also find that the density within Democrats is higher than that within Republicans, especially that the strongest edges are mostly between Democrats. This is consistent with the findings in the literature of political science that liberals may have more intensive cosponsorship activities as they believe the government should take more extensive responsibilities [25]. We also analyze the cosponsorship data using the neighborhood selection method NS-BIC(γ). For each of $\gamma = 0, 0.5, 1$, we calculate and plot their edge intensities in Figure 2. Surprisingly, we find in these graphs that the edges are denser among the congresspersons in different parties, contradicting to the conventional wisdom.

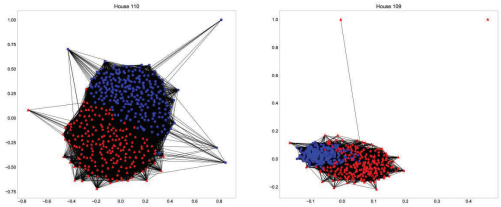

Fig. 3. PRAIME-RF output networks using $1-$locfdr as the edge weights. Democrats are plotted as blue circles, and Republicans red triangles.

We used $1-$locfdr as the edge weights of the congress member networks (Figure 3). The large scale community

TABLE III
COMMUNITY DETECTION ERROR RATES USING PARTY AFFILIATION AS THE TRUE LABELS.

Graph	House 110	House 109
PRAIME-RF	0.080	0.097
NS-BIC(0)	0.235	0.200
NS-BIC(0.5)	0.228	0.180
NS-BIC(1)	0.210	0.160

structure appears to be along the party line. To further quantitatively validate this, we apply spectral clustering based community detection [27] to these PRAIME-RF output graphs, and evaluate the cluster assignments using the party affiliation as the true labels. We find that PRAIME-RF output networks lead to much lower mis-classification rates than the results from the neighborhood selection outputs (Table III).

These conditional independence graphs among congresspersons could help political scientists to gain insights in various aspects of congressional politics. We present one such example to illustrate the interpretability of PRAIME-RF outputs. In the literature of network analysis, it is commonly believed that the "hubs", i.e., highly connected nodes, within a community is more likely to be influential or scientifically interesting than the hubs of the whole network (e.g., [28]). Intuitively, a congressperson with no direct association with the members in the opposite party has higher chance to be ideological extreme. Combining the above two thoughts, we examine the political positions of the congress members who have high within-party degree (above 80% quantile of the party) and low across-party degree (no more than 0.2) in both of the 109th and 110th congresses (Table IV). We find that all Democrats satisfying these criteria are among the most prominent progressive figures in the house, and their Republican counterparts are regarded as the most conservative members or the leaders in the conservative wing, including the then-Congressman and today's Vice President, Mike Pence.

V. DISCUSSION AND CONCLUSION

The literature has predominantly treated the sparse graph estimation for Ising models as a sparse variable selection problem. The existing popular methods for Ising model estimation are largely based on penalized regression with un-intuitive tuning parameters, and return one single graph with no full ranking of all edges or guarantees of the desired sparsity level.

We propose a PRediction Approach for Ising Model Estimation (PRAIME). PRAIME is based on the pairwise complete decouping of the Ising model parameters. It estimates the individual parameters for each potential edge using the probabilistic predictions of the observed data from an arbitrary probabilistic predictive model, provides the ranking of all node pairs by statistical significance, and enables flexible selection of the sparse graph by thresholding with the threshold chosen by the analysts. We implemented PRAIME using random forest, and illustrated its advantage in accuracy and flexibility over the neighborhood selection method using synthetic and real data.

REFERENCES

[1] E. Ising, "Beitrag zur theorie des ferromagnetismus," *Zeitschrift für Physik*, vol. 31, no. 1, pp. 253–258, 1925.

[2] J. Besag and P. J. Green, "Spatial statistics and bayesian computation," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 25–37, 1993.

[3] N. N. Schraudolph and D. Kamenetsky, "Efficient exact inference in planar ising models," in *Advances in Neural Information Processing Systems*, 2009, pp. 1417–1424.

[4] E. Schneidman, M. J. Berry II, R. Segev, and W. Bialek, "Weak pairwise correlations imply strongly correlated network states in a neural population," *Nature*, vol. 440, no. 7087, p. 1007, 2006.

[5] P. Ravikumar, M. J. Wainwright, J. D. Lafferty *et al.*, "High-dimensional ising model selection using 1-regularized logistic regression," *The Annals of Statistics*, vol. 38, no. 3, pp. 1287–1319, 2010.

[6] R. F. Barber, M. Drton *et al.*, "High-dimensional ising model selection with bayesian information criteria," *Electronic Journal of Statistics*, vol. 9, no. 1, pp. 567–607, 2015.

[7] N. Meinshausen, P. Bühlmann *et al.*, "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[8] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[9] H. Höfling and R. Tibshirani, "Estimation of sparse binary pairwise markov networks using pseudo-likelihoods," *Journal of Machine Learning Research*, vol. 10, no. Apr, pp. 883–906, 2009.

[10] S. Geng, Z. Kuang, and D. Page, "An efficient pseudo-likelihood method for sparse binary pairwise markov network estimation," *arXiv preprint arXiv:1702.08320*, 2017.

[11] S. Geng, Z. Kuang, J. Liu, S. Wright, and D. Page, "Stochastic learning for sparse discrete markov random fields with controlled gradient approximation error," in *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, vol. 2018. NIH Public Access, 2018, p. 156.

[12] B. lazej Miasojedow and W. Rejchel, "Sparse estimation in ising model via penalized monte carlo methods," *Journal of Machine Learning Research*, vol. 19, pp. 1–26, 2018.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[14] E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu, "Graphical models via univariate exponential family distributions," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 3813–3847, 2015.

[15] Y.-W. Wan, G. I. Allen, Y. Baker, E. Yang, P. Ravikumar, M. Anderson, and Z. Liu, "Xmrf: an r package to fit markov networks to high-throughput genetics data," *BMC systems biology*, vol. 10, no. 3, p. 69, 2016.

[16] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 192–236, 1974.

[17] C. Varin, N. Reid, and D. Firth, "An overview of composite likelihood methods," *Statistica Sinica*, pp. 5–42, 2011.

[18] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] B. Fellinghauer, P. Bühlmann, M. Ryffel, M. Von Rhein, and J. D. Reinhardt, "Stable graphical model estimation with random forests for discrete, continuous, and mixed variables," *Computational Statistics & Data Analysis*, vol. 64, pp. 132–152, 2013.

[20] B. Efron, *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press, 2012, vol. 1.

[21] J. Chen and Z. Chen, "Extended bic for small-n-large-p sparse glm," *Statistica Sinica*, pp. 555–574, 2012.

[22] S. Epskamp, "Isingsampler: Sampling methods and distribution functions for the ising model," *R package version 0.1*, vol. 1, 2014.

[23] J. H. Fowler, "Legislative cosponsorship networks in the us house and senate," *Social Networks*, vol. 28, no. 4, pp. 454–465, 2006.

[24] Y. Zhang, A. J. Friend, A. L. Traud, M. A. Porter, J. H. Fowler, and P. J. Mucha, "Community structure in congressional cosponsorship networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 7, pp. 1705–1712, 2008.

[25] J. E. Campbell, "Cosponsoring legislation in the us congress," *Legislative Studies Quarterly*, pp. 415–422, 1982.

TABLE IV

CONGRESSPERSONS WITH LESS THAN 0.2 DEGREE CROSS THE PARTY LINE, AND HIGH WITHIN PARTY TIES IN BOTH HOUSE 109 AND 110. THE FIRST COLUMN SHOWS THEIR NAME (PARTY). THE SECOND COLUMN DISPLAYS THEIR (WITHIN-PARTY DEGREE, BETWEEN-PARTY DEGREE) IN HOUSE 109 AND HOUSE 110. THE LAST COLUMN LISTS INFORMATION RELATED TO THEIR IDEOLOGY, INCLUDING THEIR LEADERSHIP ROLES IN THE CONGRESSIONAL PROGRESSIVE CAUCUS (CPC), THE REPUBLICAN STUDY COMMITTEE (RSC) AND OTHER INFORMATION FROM WIKIPEDIA.

Congressperson	**Degrees**	**Ideological position**
Barbara Lee (D)	(43.95,0.00);(49.16,0.16)	chair of CPC (2005-2009)
Pete Stark (D)	(17.12,0.05);(27.70,0.02)	the only open atheist in Congress
Jeb Hensarling (R)	(8.17,0.08);(14.62,0.16)	chair of RSC (2007-2009)
Mike Pence (R)	(11.38,0.10);(14.40,0.02)	chair of RSC (2005-2007)
Trent Franks (R)	(7.43,0.02);(14.05,0.05)	among the most conservative members [26]

[26] Wikipedia. (2019) Wikipedia page of Trent Franks. [Online]. Available: https://en.wikipedia.org/wiki/Trent$_{Franks}$

[27] K. Rohe, S. Chatterjee, B. Yu *et al.*, "Spectral clustering and the high-dimensional stochastic blockmodel," *The Annals of Statistics*, vol. 39, no. 4, pp. 1878–1915, 2011.

[28] P. Langfelder and S. Horvath, "Wgcna: an r package for weighted correlation network analysis," *BMC bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[29] E. Belilovsky, K. Kastner, G. Varoquaux, and M. B. Blaschko, "Learning to discover sparse graphical models," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 440–448.

[30] J. Fan, Y. Feng, and L. Xia, "A projection based conditional dependence measure with applications to high-dimensional undirected graphical models," *arXiv preprint arXiv:1501.01617*, 2015.

[31] Z. Ren, T. Sun, C.-H. Zhang, H. H. Zhou *et al.*, "Asymptotic normality and optimalities in estimation of large gaussian graphical models," *The Annals of Statistics*, vol. 43, no. 3, pp. 991–1026, 2015.

[32] T. Sun and C.-H. Zhang, "Scaled sparse linear regression," *Biometrika*, vol. 99, no. 4, pp. 879–898, 2012.

[33] T. Wang, Z. Ren, Y. Ding, Z. Fang, Z. Sun, M. L. MacDonald, R. A. Sweet, J. Wang, and W. Chen, "Fastggm: an efficient algorithm for the inference of gaussian graphical model in biological networks," *PLoS computational biology*, vol. 12, no. 2, p. e1004755, 2016.

[34] Wikipedia. (2019) Wikipedia page of David R. Obey. [Online]. Available: https://en.wikipedia.org/wiki/Dave$_{Obey}$

[35] ——. (2019) Wikipedia page of Gresham Barrett. [Online]. Available: https://en.wikipedia.org/wiki/Gresham$_{Barrett}$