An expectation maximization algorithm for high-dimensional model selection for the Ising model with misclassified states

David G. Sinclair and Giles Hooker

Department of Statistical Science, Cornell University, Ithaca, NY, USA

ABSTRACT
We propose the misclassified Ising Model: a framework for analyzing dependent binary data where the binary state is susceptible to error. We extend previous theoretical results of a model selection method based on applying the LASSO to logistic regression at each node and show that the method will still correctly identify edges in the underlying graphical model under suitable misclassification settings. With knowledge of the misclassification process, an expectation maximization algorithm is developed that accounts for misclassification during model selection. We illustrate the increase of performance of the proposed expectation maximization algorithm with simulated data, and using data from a functional magnetic resonance imaging analysis.

1. Introduction
This paper proposes an extension of estimation methods for undirected graphical models to cases where node values are observed with error. In particular, motivated by data from functional magnetic resonance imaging (fMRI), we examine the consequences of misclassification noise in an Ising network model on estimation methods proposed in Ravikumar et al., 2010, [21] and show that the estimated edge set can be improved by accounting for misclassification rates.

Graphical models have proven to be a useful tool in modeling a wide range of data, arising in fields such as neuroscience, genetics, social networks, image restoration, traffic models, and disease case modeling, among many. The graph structure provides a useful mathematical framework for representing complex dependencies among a large collection of objects.

In this paper we focus on undirected graphical models, which are specified by a graph \( G = (V, E) \) for a node set \( V = \{1, 2, \ldots, p\} \) and an edge set \( E \subseteq V \times V \). A random vector with this graph structure is assumed to follow the Markov Property (see [14]): the \( i \)th and \( j \)th element of the vector are dependent conditional on the remaining nodes if and only if \( (i, j) \in E \). Thus, we are concerned with uncovering the structure of the edge set \( E \) and therefore uncovering conditional dependencies within our dataset.

CONTACT Giles Hooker  gjh27@cornell.edu
*Portions of this work appear as a preprint at https://arxiv.org/abs/1704.05995.
In this paper, we assume that our data is binary where the dependencies are entirely captured by pairwise relationships resulting in the Ising Model [12], detailed in Section 2, which corresponds precisely to these assumptions. The Ising Model has proven useful in data analysis settings such as functional magnetic resonance imaging (fMRI) [24], image restoration [8,13], spatial statistics [1], social network analysis [20], and genetics [18].

Structure learning of the edge set in the Ising model is a well-studied problem in the statistics literature. Considerable attention has been given to finding information-theoretic bounds for learning Ising graph structures [22,23,25]. Table 1 in [23] gives a useful summary of the graphical assumptions for which these information-theoretic bounds are known.

Due to the computational intractability of the partition function $Z(\theta^*)$ for the Ising distribution function given in (1) (see [27]), various approaches have been developed in order to perform sound statistical methodology under this practical constraint. Barber et al. [2] show an extended BIC method for uncovering the underlying graph in the Ising data setting with theoretical bounds. Bresler [3] develops a greedy algorithm, which uses a structural property of mutual information associated with Ising models to prove asymptotic exact learning of the underlying graph. Ravikumar et al. [21] show theoretic bounds for a neighborhood-based regularized logistic regression approach for performing model selection analogous to the Meinshausen-Bühlmann approach for Gaussian graphical models [19].

One potential issue with categorical data is the possibility for misclassification which leadsto a coarsening in the observed dataset [10]. This arises in fMRI data where the traditional General Linear Model approach attempts to find areas of the cortex that have been significantly activated corresponding to a thresholding of the BOLD response’s association with the HRF function [16]. When the cortex is reduced to specialized regions via a parcellation [9,24] we can think of this procedure as assigning a latent label to each parcel and may suspect possible misclassification when the BOLD response’s association with the HRF is close to the threshold. If there is a non-zero probability of misclassification, it can be shown that the data no longer follows an Ising distribution, and thus it is not clear if current structure learning methods can still perform adequately. However, we are often able to assess the certainty of the node label – in our example the label denotes a component of an implicit mixture distribution for the continuous-valued BOLD response – and we can make use of this to target specific nodes to be updated.

In this paper we extend the theory behind [21]’s approach to handle misclassification and, conditional on this result, we develop a methodology for further boosting structural learning performance via an expectation maximization (EM) technique [4] that can be used if there is knowledge of the misclassification process. Due to the inherent dependence in our data set, it is difficult to show that the EM method will always increase the marginal log likelihood. However we show that if the learned structural dependence can predict a candidate state with high probability, the EM method can provide gains in efficiency.

In Section 2 of this paper the misclassified Ising model is defined and theoretical guarantees are stated. In Section 3 the algorithm for incorporating misclassification information in an updated edge set estimated is described. Section 4 provides simulations to better understand the performance of this methodology. Section 5 shows how this methodology can be applied in an fMRI setting, and simulations are done to show the method should
still increase structural learning accuracy. Proofs of theoretical results and calculations for the EM updates are provided in appendices.

2. Misclassified Ising model and theoretical guarantees

In this section we develop the Misclassified Ising Model, and discuss theoretical guarantees for estimating the underlying edge set with this added noise assumption.

2.1. Ising model

We focus on the special case of the Ising Model as described in Ravikumar et al. [21], which we refer to as the Ising$(G, \theta^*)$ distribution. Let $X = (x^{(1)}, \ldots, x^{(n)})$ be n i.i.d. observations of $X = (x_1, \ldots, x_p) \sim$ Ising$(G, \theta^*)$ in which $x_s \in \{-1,1\}$, and $\theta^*_{st} \in \mathbb{R}$ for each $s \in V$, with probability mass function

$$P_{\theta^*}(x) = \frac{1}{Z(\theta^*)} \exp \left\{ \sum_{(s,t) \in E} \theta^*_{st}x_sx_t \right\} \quad (1)$$

Here the partition function $Z(\theta^*)$ ensures the distribution sums to one. Recall that $\theta^*_{st} \neq 0 \iff (s,t) \in E$, and therefore our goal is to determine the support of $\theta^*$.

Due to the computational intractability of the partition function [27], a neighborhood-based likelihood method is adopted in Ravikumar et al. [21], a technique akin to the [19] method for Gaussian graphical models [15], where a model selection is undertaken to find the neighborhood of each node separately. The estimated edge set is then consolidated from the neighborhood sets.

2.2. $\ell_1$-regularized neighborhood-based model selection

The Ising Model has the useful property that the conditional distribution of a node takes the form of a logistic regression with the canonical link function on all remaining nodes. Therefore, if we let $\theta^*_{\cdot u} = \{\theta^*_{ru}; u \in V \setminus \{r\}\}$ be the edge weights associated with the node $r$, model selection can be done via an $\ell_1$-regularized logistic regression on each node $r$ [6]:

$$\hat{\theta}_{\cdot r} = \arg \min_{\theta_{\cdot r} \in \mathbb{R}^{p-1}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log P_{\theta_{\cdot r}}(x_r^{(i)} | x_{\cdot r}^{(i)}) + \lambda \|\theta_{\cdot r}\|_1 \right\} \quad (2)$$

where $P_{\theta}$ is the logistic regression function with a canonical link with response $1(x_r^{(i)} = 1)$, regression parameters $2\theta_{\cdot r}$, and predictors $x_{\cdot r}^{(i)} = \{x_t | t \in V \setminus \{r\}\}$. Doing this regularized regression over each node can give us an estimate for the edge set $E$ as follows:

$$\hat{E}_{\ell_1} = \{(s,t); (\hat{\theta}_{\cdot s})_t \neq 0 \text{ and } (\hat{\theta}_{\cdot t})_s \neq 0\} \quad (3)$$

In this formulation of the estimated edge set, an edge will be selected between two nodes if the corresponding estimated neighborhood sets both contain these two nodes.

This method is shown in Ravikumar et al. [21] to give a consistent estimate $\hat{E}_{\ell_1}$ in the sense that $P(\hat{E}_{\ell_1} = E) \to 1$ as $n \to \infty$, when $n = \Omega(d^3 \log p)$ for appropriately chosen $\lambda$. 


where $d$ is the maximal degree over nodes in the network. The theoretical results below employ a $\lambda$ that stays constant between nodes $r$; in simulations we find that this gives good results. Improvements may be possible choosing $\lambda$ separately for each node, but this requires large enough sample sizes to select a node-specific $\lambda$ stably. We refer to the method for obtaining this edge set as RWL in recognition of its authors.

### 2.3. Misclassified Ising model

Here we introduce a formalization of the Misclassified Ising Model, which will be defined hierarchically.

We continue to assume $X \sim \text{Ising}(G, \theta^*)$, but define $\tilde{X}$ as the random vector such that $P(\tilde{X} \equiv y | X) = \prod_{s \in V} P(\tilde{x}_s = y_s | x_s) = \prod_{s \in V} (\gamma_s (y_s \neq x_s) (1 - \gamma_s)^{\mathbf{1}(y_s = x_s)})$ for all $y \in \{-1, 1\}^p$. In this sense, each node is misclassified with some probability $\gamma_s$ and the misclassification is independent across nodes. As we only observe the misclassified nodes, $\tilde{X}$, we define their distribution unconditional of $X$ as the Misclassified Ising Model, $\tilde{X} \sim \text{MIsing}_{\gamma}(G, \theta^*)$. The theoretical guarantees for RWL under this distribution shown in Section 2.4 do not directly assume independence of the misclassification probabilities, however this assumption is used when completing the EM update algorithm in Section 3.

As with the Ising Model, let $\tilde{X} = (\tilde{x}^{(1)}, \ldots, \tilde{x}^{(n)})$ be n i.i.d. observations of $\tilde{X}$.

### 2.4. Theoretical guarantees

In this section we show that when the extra noise due to misclassification is small, the estimated edge set $\hat{E}_{\ell_1}$ can still produce a reasonable model selection method. The amount that the added noise hinders our ability to detect edges is captured by the expectation of the score function for each node-conditional distribution for the (not misclassified) Ising Model, where expectation is calculated over the true misclassified Ising Model. Indeed, as misclassification goes to 0, the expectation of the score function goes to 0, which implies that there is no hindrance in obtaining the edges, as expected.

Formally, $W^n_r(\theta) = -\nabla \log P_{\theta_r}(\tilde{x}^{(i)}_r | \tilde{x}^{(j)}_r)$ is the score function for $P_{\theta_r}$ defined in Equation (2). We define the misclassified score and misclassified information as

\[
S_{\text{max}} = \max_{r \in V} |E(W^n_r(\theta^*))| \quad (4)
\]
\[
\hat{Q}^*_r = -E(\nabla W^n_r(\theta^*)) \quad (5)
\]

Note that both of these expectations are over the misclassified distribution. The misclassified score $S_{\text{max}}$ corresponds to the largest deviation of the expected score function over the misclassified distribution from 0.

The first two assumptions we make for our extension, are very similar to those given in Ravikumar et al. [21], however they are made on the misclassified information matrix. These are stated explicitly in Appendix A.1 and are referred to as (A1) and (A2). The third assumption is stated here as:
(A3) Misclassification Condition. For $C_{\text{min}}$ and $D_{\text{max}}$ as defined in (A1), and $\alpha$ as defined in (A2), we assume
\[
S_{\text{max}} \leq \frac{C_{\text{min}}^2 \alpha^2}{400 D_{\text{max}} d (2 - \alpha)^2}
\]  
$C_{\text{min}}$ and $D_{\text{max}}$ provide bounds on the information for the misclassified score and on the covariance of the observations respectively. Note that the expectation in (4) operates over both $\tilde{x}_r^{(i)}$ and $\tilde{x}_r^{(\bar{i})}$ and therefore depends on the underlying network structure, making an explicit expression for $S_{\text{max}}$ difficult to obtain.

If we make the same population assumptions as given in Ravikumar et al. [21] on the underlying Ising Model (stated in Appendix A.1), then for $\alpha$ satisfying (A2) we have the following result that corresponds to Theorem 1 in Ravikumar et al. [21].

Extended Theorem 1: Consider a Misclassified Ising graphical model, $\text{MIsing}_\gamma(G, \theta^*)$ with parameter vector $\theta^*$ and associated edge set $E^*$ such that conditions (A1) and (A2) are satisfied by the misclassified information matrix $\tilde{Q}^*$ for all $r \in V$. Assume the misclassified score, $S_{\text{max}}$ satisfies (A3) and let $\tilde{X}$ be a set of $n$ i.i.d. samples for the misclassified Ising model. Suppose that the regularization parameter $\lambda_{n,d,p}$ is selected to satisfy
\[
\lambda_{n,d,p} \geq \frac{16(2 - \alpha)}{\alpha} \left( \frac{\log p}{n} + \frac{S_{\text{max}}}{4} \right)
\]  
Then there exist positive constants $L$ and $K$, independent of $(n, d, p)$ such that if
\[
n > D^3 \log p
\]  
then the following properties hold with probability at least $1 - 2 \exp(-K\tilde{\lambda}^2 n)$, where $\tilde{\lambda}_{n,d,p} = \lambda_{n,d,p} - \frac{4(2 - \alpha)}{\alpha} S_{\text{max}}$.

(a) For each node $r \in V$ the $\ell_1$-regularized logistic regression has a unique solution and therefore uniquely specifies a neighborhood $\hat{N}(r)$.
(b) For each node $r \in V$ the estimated neighborhood $\hat{N}(r)$ correctly excludes all edges not in the true neighborhood. Moreover, it correctly includes all edges $(r,t)$ for which $|\theta_{rt}^*| \geq \frac{10}{C_{\text{min}}} \sqrt{d} \lambda_{n,d,p}$.

The proof of this result is located in Appendix 1.

An interesting consequence from this result is that as $n \to \infty$ the tuning parameter does not converge to 0 unless $S_{\text{max}}$ also goes to 0. This means that by part (b) some edges may never be correctly included with high probability due to the conditional independencies of the graphical model being overcome by the misclassification.

3. EM algorithm for updating edges of $\widehat{E}_\ell$

We develop an EM algorithm for obtaining an updated edge set. In Section 2.3, all nodes could potentially have some amount of misclassification probability, however throughout the use of this update we assume that only a subset of nodes can be misclassified. The distinction does not affect the related proofs for the method, although for the method to be
computationally tractable the number of potentially misclassified nodes must be relatively small. This is a fundamental issue in latent network structures and we argue that we can at least improve the log likelihood by targeting a feasible number of the most uncertain nodes.

Conditional on the initial RWL fit, resulting in edge set $\hat{E}_{\ell_1}$ and parameter $\hat{\theta}_{1,r}$, we develop an EM-type algorithm for updating the neighborhood for certain nodes in our graphical model. The method is run on each node individually, similar to RWL. In the usual EM approach the average joint log likelihood of the observed and latent variables is maximized in order to increase the likelihood marginally on the observed data. Due to the complexity of the distribution in the joint case, it is difficult to maximize the log likelihood over all possible latent states.

We instead show in Appendix 2 that maximizing the conditional distributions will still serve to increase the marginal likelihood given that the probability that a node is in the incorrect state is close to 1. By leveraging dependency information from the initial RWL fit, we show in simulations that this condition is satisfied and we are able to increase the marginal likelihood.

In doing our EM update we focus on neighborhoods surrounding nodes that have potentially been misclassified. In order to do this we assume we have some knowledge of the probability of misclassification for each node. This probability can be an average misclassification over all observations for a given node, although the method has better performance when misclassified probabilities are known for each observation. In [24], misclassification probabilities can be derived from the implicit mixture model for continuous signaling in fMRI where the node label indicates a component of the mixture distribution and misclassification probabilities can be obtained from the posterior probabilities of the component labels.

With an appropriate update set of nodes, $\mathcal{U}$, we can then update the edge set to obtain $\hat{E}_{EM}^{\ell_1}$. In the following subsections we go over obtaining the update set $\mathcal{U}$ and completing the $E$ and $M$ steps.

### 3.1. Obtaining update set: $\mathcal{U}$

The update set will be a union of candidate nodes, $\mathcal{C}$, and participant nodes, $\mathcal{P}$. Candidate nodes are those whose estimated neighborhood sets have been potentially affected by misclassification.

If $\hat{\gamma}_s$ is a misclassification estimation for each node, then for a given threshold $q$, a reasonable way to define candidate set is as $\mathcal{C} = \{s \in V : \hat{\gamma}_s > q\}$, although our method is not bound to any procedure for determining the candidate set.

To obtain the participant nodes, first consider the following example. Assume $(r, s) \in E$ and $(s, t) \in E$ but $(r, t) \notin E$. If there were no misclassification in our data then $x_r|x_s \perp \perp x_t|x_s$, but if $x_s$ is a candidate node with some non-zero probability for misclassification, then we have

$$P(x_r = 1, x_t = 1|\hat{x}_s = 1) = P(x_s = \hat{x}_s)P(x_r = 1, x_t = 1|x_s = 1)$$

$$+ P(x_s \neq \hat{x}_s)P(x_r = 1, x_t = 1|x_s = -1)$$

$$= (1 - \gamma_s)P(x_r = 1|x_s = 1)P(x_t = 1|x_s = 1)$$
Thus nodes are no longer independent as long as $\theta_{ri}^* \neq \theta_{si}^*$ and in the fitted network the edge $(r, t)$ may appear. On the other hand, if $x_r$ was a candidate node, then $P(x_t = 1|x_r = 1, x_r = 1) = P(x_t = 1|x_r = 1)$. That is to say that if a node’s shortest path to a candidate node in the true network is greater than or equal to 2, then that node’s neighbors will still be chosen independently from the misclassification. This is not only a useful heuristic for choosing an update set, but will also be a useful property when calculating weights for the EM fit.

Taking this into account, we set the update set to be $U = N(N(C))$, the neighbors of neighbors of the candidate nodes. From here we have the participant nodes as all nodes in $U$ that are not in $C$, i.e. $P = U \setminus C$.

Lastly, let $s$ be the number of disjoint subgraphs induced by $U$ and let $c_{\text{max}}$ be the largest number of candidate nodes in a single subgraph. The computational complexity of the method is $O(sn2^{c_{\text{max}}})$, which can computationally tractable even with up to 20 candidates node in a single subgraph. For the rest of the document, we assume $s = 1$, but for $s > 1$ the $E$ and $M$ steps still hold where a loop is run over each disjoint subgraph.

### 3.2. E step

For the $k$th step in the EM update, for node $r \in U$, we take the expectation over the latent variables $x_r$. Define the following three sets of parameters

$$
\theta_{U \setminus r} = \{\theta_{sr}; s \in U\}
$$

$$
\theta_{V \setminus U \setminus r}^{(k)} = \{\theta_{sr}^{(k)}; s \notin U\}
$$

$$
\tilde{\theta}_{r} = \theta_{U \setminus r} \cup \theta_{V \setminus U \setminus r}^{(k)}
$$

$\theta_{U \setminus r}$ corresponds to the neighborhood parameters for node $r$ that will be updated. For $s \notin U$, the corresponding edge parameter $\theta_{sr}^{(k)}$ will not be updated, and thus when running this update, the value $2\theta_{sr}^{(k)}x_rx_s$ is included as an offset in the logistic regression to account for their neighborhood effect.

We are interested in the penalized log likelihood

$$
L_{\lambda}(\theta_{U \setminus r}|\theta_{V \setminus U \setminus r}; \tilde{X}) = \tilde{\ell}_{r}(\theta_{U \setminus r}; \theta_{V \setminus U \setminus r}; \tilde{X}) - \lambda \|\tilde{\theta}_{r}\|_1
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} \log P_{\tilde{\theta}_{r}}(\tilde{x}_{r}^{(i)}|\tilde{X}_{\setminus U \setminus r}) - \lambda \|\tilde{\theta}_{r}\|_1
$$

By including the offset terms in the regularization term, we ensure that the log likelihood will increase over a fixed parameter $\lambda$. Let $\Omega_C = \{-1, +1\}^{|C|}$, and for $z_c \in \Omega_C$, let $\tilde{x}_{c}^{(i)}(z_c)$ be original observation with candidate nodes replaced by $z_c$. An estimate of the expectation of this log likelihood is

$$
\hat{Q}_{r}(\theta_{U \setminus r}|\theta_{V \setminus U \setminus r}, \tilde{X}) = \hat{E}_{\tilde{X}_{c}|X_{V \setminus U \setminus r}; \theta^{(k)}}(\tilde{\ell}_{r}(\theta_{U \setminus r}; \theta_{V \setminus U \setminus r}; \tilde{X})) - \lambda \|\tilde{\theta}_{r}\|_1
$$
\[^{13}n \sum_{i=1}^{n} z_c \in \Omega_c \left[ P_{\theta(k)}(X_C = z_c| \tilde{X}_U = \tilde{x}^{(i)}_U) \log P_{\tilde{\theta}_r}(\tilde{x}^{(i)}_r, z_c| \tilde{x}^{(i)}_{\tilde{U}\setminus r}) \right]
- \lambda \| \tilde{\theta}_r \|_1 (13)\]

However, maximizing over the joint probability $P_{\tilde{\theta}_r}(\tilde{x}^{(i)}_r| \tilde{X}_{\tilde{U}\setminus r}, z_c)$ is computationally intractable unless $|\Omega|$ is very small. We instead look only at conditional distributions, and consider the following estimate of the expectation

\[^{14}n \sum_{i=1}^{n} z_c \in \Omega_c \left[ P_{\theta(k)}(X_C = z_c| \tilde{X}_U = \tilde{x}^{(i)}_U) \log P_{\tilde{\theta}_r}(\tilde{x}^{(i)}_r(z_c)| \tilde{x}^{(i)}_{\tilde{U}\setminus r}(z_c)) \right]
- \lambda \| \tilde{\theta}_r \|_1 (14)\]

In Appendix 2 we show that, for any set of observations $\tilde{X}$ and for any initial fit $\tilde{\theta}$, there exists an open set of misclassification probabilities such that maximizing $\tilde{Q}_r$ will still result in an increase in the penalized likelihood $L_{\lambda}(\theta_{\tilde{U}\setminus r}| \tilde{\theta}_{V\setminus \tilde{U}\setminus r})$.

The function $\tilde{Q}_r$ corresponds to a $\ell_1$-regularized weighted logistic regression. Each $P_{\theta(k)}(X_C = z_c| \tilde{X}_U = \tilde{x}^{(i)}_U)$ can be calculated utilizing factorizations of Ising distribution where the partition function is canceled out due to conditioning the probability. A derivation of these probabilities is located in Appendix 3.

### 3.3. M step

Noting that $\tilde{Q}_r$ corresponds to a weighted penalized logistic regression with an offset, we complete the M step maximization using the `glmnet` package in R [7]. We obtain the updated edge parameter estimates as

$$\theta^{(k+1)}_{\tilde{\lambda}} = \left( \text{arg min}_{\theta_{\tilde{U}\setminus r} \in \mathbb{R}^{(|\tilde{U}\setminus r)|-1}} \tilde{Q}_r(\theta_{\tilde{U}\setminus r}| \theta^{(k)}_{\lambda}, \tilde{\theta}_r, \tilde{X})\right) \cup \hat{\theta}_{V\setminus \tilde{U}\setminus r} (15)$$

With the updated edge set as

$$\tilde{E}_{EM}^{(k+1)} = \{ (s, t); \text{if}(\theta^{(k+1)}_{\tilde{\lambda}})_t \neq 0 \text{and}(\theta^{(k+1)}_{\tilde{\lambda}})_s \neq 0 \} (16)$$

We show through simulations that this methodology tends to increase model selection performance of the underlying graphical model.

### 4. Simulations

The EM method uses information about the misclassification, and also leverages the dependence/structure information to which we have access from the original fit as made formal in Section 2.4.

In the following simulation we demonstrate that candidate nodes will gain spurious connections due to misclassification, which can be overcome using the EM update.
One can also note that given misclassification information, a ‘prior’ weight based solely on misclassification information (i.e. agnostic of any structural dependency information) can be calculated as

\[ P(\tilde{X}_C = z_c) = \prod_{s \in C} \gamma_s \]  

(17)

The EM method updates these state probabilities given dependency information.

Our simulation have been designed with the computational cost \(O(sn^2c_{\text{max}})\) in mind, limiting our ability to update large subgraphs of connected candidate nodes. In practice our methods – including searching over \(\lambda\) – take on the order of an hour per data set when implemented on a standard laptop. For this reason, we initially focus on a small network with a sample size constrained to keep \(p\) a large fraction of \(n\), before investigating a simulation based on real-world data. We note that the \(O(sn^2c_{\text{max}})\) cost applies only to connected sets of candidate nodes; thus our methods are still feasible for a large network with multiple small but disjoint candidate node sets.

### 4.1. Simulation parameters and network specification

We ran the method on a network of 12 nodes \((p = 12)\); Figure 1 shows the topological structure of the network over which we simulate. The intuition for this network topology is that the blue participants nodes will inform the red candidate nodes.

The nodes \(L, D, H\) are each potentially misclassified in 50% of observations, where the probability of misclassification in these observations is 60%. We ran 1000 simulations with \(n = 60\), and true edge parameters \(\theta^*_st = \frac{1}{2}\) for \((s, t) \in E\). All Ising observations were simulated using the IsingSampler package in R [5].

Although nodes \(L, D, H\) are only misclassified in half of observations, the distribution unconditional on knowledge of the misclassification process is still a Misclassified Ising Distribution with non-zero misclassification parameters equal to \(\gamma_L = \gamma_D = \gamma_H = 0.3\).

### 4.2. Fitted models

The models we fit are

1. **RWL**: minimizing (2)
2. **RWL Weighted**: minimizing (2) with a weighted logistic regression using weights defined in (17)
3. **RWL + EM**: running an EM update for edges selected in RWL
4. **Weighted + EM**: running an EM update for edges selected in RWL weighted.

For the initial RWL and RWL Weighted fits, a range of tuning parameters were selected to obtain a ROC curve for candidate and participant nodes. For the EM fits, the selected model was based on the tuning parameter that maximized \(P(\text{True Positive}) + (1 - P(\text{False Positive}))\), and then a range of tuning parameters were employed at the EM stage to create ROC curves. The range of tuning parameters was set to be five times the standard deviation of the optimal smoothing parameters found from a small simulation;
Figure 1. The simulated network topology. Red nodes (L, D, H) are candidate nodes which have a true misclassification probability of 60% in half of observations. Blue nodes correspond to participant nodes. All non-zero edges have an equal weight = 1/2.

this was selected as being both symmetric around an approximate optimum and covering the full range of sensitivity and specificity.

The first set of simulations use only one EM update of our fit. We then investigate the effect of further EM analyses. We look at $RWL + 2EM$ and $RWL + 3EM$, which corresponds to running a second and third EM update on the $RWL$ fitted edge set. When using multiple EM updates we employed the same value of $\lambda$ throughout, rather than selecting an optimum at each stage; we note that the Regularized EM algorithm in Appendix 2 would imply the latter, but comes and significant computational cost.

4.3. Results

In Figure 2 the RWL $+$ EM fit performs at least as well or better than any other method. Even keeping the tuning parameter selected in the first stage when applying the EM update, an increase in classification performance is always observed. Overall, the AUC for candidate nodes increases from 0.6608 to 0.6945, and for participant nodes the AUC increases from 0.8729 to 0.8770.

Interestingly, basing the initial fit off of RWL seems to perform better than the weighted regularized logistic regression ($RWL\text{ Weighted}$). This is consistent with the proof given in Appendix 2, as the misclassification probability for a candidate node will be at most $P(X_r = \tilde{X}_r) = 0.5$ for $RWL\text{ Weighted}$, and therefore this misclassification scenario is far
Figure 2. Output from 1000 simulations with \( n = 60 \) and \( p = 12 \) showing True Positive vs. True Negative and False Positive vs. True Positive relationships for candidate and participant nodes. From the symmetry in the topology of the graph, candidate and participant node results can be aggregated.

Figure 3. Output from 1000 simulations with \( n = 60 \) and \( p = 12 \) showing True Positive vs. True Negative and False Positive vs. True Positive relationships for candidate and participant nodes for running multiple EM updates. Note the decrease in performance for participant nodes for the 3rd EM update.

from the open set \( \Gamma \) defined in Appendix 2. The implication of this result is that misclassification information alone is not enough to provide a gain in model selection performance; dependency information must also be leveraged.

As shown in Section 2.4 some dependency information is obtained in the RWL fit, from which we have that \( P(X_r = \hat{X}_r | \text{RWL}) \approx 0 \) for multiple observations, and therefore the Regularized EM Theorem in Appendix 2 applies. Figure 2 demonstrates this theorized increase in performance, and, as shown in Appendix 2, the increase will occur without needing to change the tuning parameter.

Figure 3 shows the simulations results for running the EM update multiple times. Note that between EM updates it is unlikely the probability that a node is in a given state will change drastically and further iterations may not be helpful. This can be seen in Figure 3,
as by the third EM update, there is a small decrease in participant node detection. After the first EM update the participant node AUC is 0.8770, and it decreases to 0.8593 by the third update.

5. fMRI data example simulations

Ref. [24] documents a method for fitting an Ising model on task-fMRI data. Each node in the graph corresponds to a specialized region of the cortex, and the classification is a discretization of a fit parameter corresponding to blood flow. If the blood flow is above a certain threshold, the area of the cortex is considered active during the task. Due to the inherent noise in the data, misclassification is certainly present.

Figure 4 shows the fit example from [24], using data from the Human Connectome project [26], and the nodes were obtained via the parcellation documented in [9]. An estimate of the node’s state was obtained by investigating the p-values used for the classification procedure. Fourteen out of the 111 regions were found to be closer to the p-value threshold more often, being within 5% of the p-value threshold at over 12% of the time. In Figure 4, these regions are colored in red.

5.1. Choosing update set \( \mathcal{U} \)

A useful consequence of the network that we have fit, is that the update set as defined in Section 3.1 is a disjoint union of \( s = 4 \) disjoint subgraphs. Therefore, we run our

![Figure 4. The fitted connectome network from [24] and the update set \( \mathcal{U} \). Nodes are arranged in a superior (top-down) view of the cortex, where red nodes correspond to candidate nodes and blue nodes are all remaining nodes.](image)
simulations on the largest of the subgraphs denoted as the update set in Figure 4. This corresponds to our $p = 20$ node network topology that we use for simulations.

5.2. Simulation parameters

We ran 500 simulations with $n = 200$, corresponding to the size of the original dataset. Edge parameters in the simulation were selected to correspond to edge parameters from the original fit, however non-zero edges were smoothed towards the average of all edge parameters in order to reduce variance in our estimate of these parameters.

Participant nodes were then misclassified in 50% of observations with a misclassification probability of 75%. Thus, the overall misclassification rate is similar to the observed dataset.

Based on the results from Section 4, we only compare the RWL + EM and RWL, where a range of tuning parameters is selected for each method.

5.3. Results

Figure 5 shows the True Positive vs False Positive relationship. A consistent increase in classification performance is observed for the first 13 nodes. The overall error rate decreases for the neighborhood of candidate nodes drops from 21.1% to 10.0% when choosing the

![Figure 5](image-url)
optimal tuning parameter for the EM fit. If the tuning parameter is not changed for the EM fit, we still see a decrease in the error from 21.1% to 14.6%. There does appear to be a small decrease in performance for participant nodes that were not a direct neighbor with a candidate (e.g. nodes 75, 32 and 76) however this difference contributed to a 2.8% increase in error rates, compared to the 11.1% decrease in error rates for candidate nodes.

Figure 6 orders the nodes by overall error rate across simulations for the two different methods. The decrease in error rate is consistently better after running the EM fit.

Figure 7 plots the adjacency matrix for \( \mathcal{U} \). This plot has a few interesting characteristics. The red areas, which correspond to false edges that were selected often for the RWL fit tend to correspond to edges between participant nodes that are highly connected to candidate nodes. The error rate is particularly high for nodes 104, 84, and 55. Figure 8 looks only at error rate, and focusses on nodes that had at least one neighbor with a candidate node.

Overall a consistent increase in performance is observed across candidate nodes and their neighbors. The small decrease in performance for participant nodes 75, 32, and 76 may potentially be remedied in practice when including the offset from the rest of the
network discussed in Section 3.2, as more dependency information will be available for determining the posterior state probabilities.

6. Conclusion

In this paper we introduce the misclassified Ising model. We show that under suitable misclassification assumptions RWL can still be used as a model selection technique. We then show that RWL can be extended in order to account for misclassification. Sections 4 and 5 show simulation results for a symmetric network and for a network obtained from fMRI data.

The fMRI node states correspond to discretizations of a continuous variable and therefore provide a useful setting for discussing misclassification. Depending on the discretization method used to determine the latent state, acquiring an estimate for the probability of misclassification is potentially straightforward.

In both cases, the EM-based algorithm is shown to provide significant performance gains in model selection. Given a binary network data set with an estimated misclassification probability, one can therefore obtain more reliable connections between nodes within the update set $U$ by performing this update.

The method is computationally constrained by the greatest number of candidate nodes within the largest disjoint subnetwork of the update set $U$. However, this computational complexity depends only linearly on the number of remaining nodes in the update set. Therefore even with a high degree dataset, if there are few candidate nodes, this method can still be tractable.

The analysis in this paper can be extended easily to the signed edge selection as discussed in Ravikumar et al. [21]. The EM approach can also be extended to the Potts model corresponding to multiple states per node, although this would serve to further increase the computational complexity. Future work within the misclassified Ising framework could be to understand the effect of dependent misclassification across nodes on the misclassified score and information functions.
Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] S. Banerjee, B.P. Carlin and A.E. Gelfand, Hierarchical Modeling and Analysis for Spatial Data, CRC Press, Boca Raton, 2014.
[2] R.F. Barber and M. Drton, High-dimensional Ising model selection with Bayesian information criteria, Electron. J. Stat. 9 (2015), pp. 567–607.
[3] G. Bresler, Efficiently learning Ising models on arbitrary graphs, in Proceedings of the 47th Annual ACM on Symposium on Theory of Computing, ACM, 39 (2015), pp. 771–782.
[4] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B 1977), pp. 1–38.
[5] S. Epskamp, Isingsampler: Sampling methods and distribution functions for the Ising model, R package version 0.1 1, 2014.
[6] J. Friedman, T. Hastie and R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (2010), pp. 1–22. Available at http://www.jstatsoft.org/v33/i01/.
[7] J. Friedman, T. Hastie and R. Tibshirani, glmnet: Lasso and elastic-net regularized generalized linear models, R package version 1, 2009.
[8] S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE. Trans. Pattern. Anal. Mach. Intell. 6 (1984), pp. 721–741.
[9] E.M. Gordon, T.O. Laumann, B. Adeyemo, J.F. Huckins, W.M. Kelley and S.E. Petersen, Generation and evaluation of a cortical area parcellation from resting-state correlations, Cerebral Cortex 26 (2016), pp. 288–303.
[10] D.F. Heitjan and D.B. Rubin, Ignorability and coarse data, Ann. Stat. 19 (1991), pp. 2244–2253. Available at https://doi.org/10.1214/aos/1176348396.
[11] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Am. Stat. Assoc. 58 (1963), pp. 13–30.
[12] E. Ising, Beitrag zur theorie des ferromagnetismus, Zeitsch. Phys. Hadrons Nucl. 31 (1925), pp. 253–258.
[13] M.C. Kandes, Statistical image restoration via the Ising model, Final Project, 2008.
[14] R. Kindermann and J.L. Snell, Markov Random Fields and Their Applications, Vol. 1, American Mathematical Society, Providence, RI, 1980.
[15] S.L. Lauritzen, Graphical Models, Clarendon Press, Oxford, 1996.
[16] M.A. Lindquist, The statistical analysis of fMRI data, Stat. Sci. 23 (2008), pp. 439–464.
[17] R. Little and D. Rubin, Statistical Analysis with Missing Data, Wiley, New York, 2002.
[18] J. Majewski, H. Li and J. Ott, The Ising model in physics and statistical genetics, Am. J. Human Genet. 69 (2001), pp. 853–862.
[19] N. Meinshausen and P. Bühlmann, High-dimensional graphs and variable selection with the Lasso, Ann. Stat. 34 (2006), pp. 1436–1462.
[20] A. Montanari and A. Saberi, The spread of innovations in social networks, Proc. Natl. Acad. Sci. 107 (2010), pp. 20196–20201.
[21] P. Ravikumar, M.J. Wainwright and J.D. Lafferty, High-dimensional Ising model selection using l1-regularized logistic regression, Ann. Stat. 38 (2010), pp. 1287–1319.
[22] N.P. Santhanam and M.J. Wainwright, Information-theoretic limits of selecting binary graphical models in high dimensions, IEEE Trans. Inf. Theory 58 (2012), pp. 4117–4134.
[23] J. Scarlett and V. Cevher, On the difficulty of selecting Ising models with approximate recovery, IEEE Trans. Signal Inf. Process. Over Netw. 2 (2016), pp. 625–638.
[24] David Sinclair. Model Selection Results for High-Dimensional Graphical Models on Binary and Count Data with Applications to FMRI and Genomics, PhD thesis, Cornell University, 2017.
Appendices

Appendix 1. Proof of Extended Theorem 1

In this appendix we state the assumptions for Extended Theorem 1 and complete the proof.

A.1 Assumptions

In order to prove Extended Theorem 1 we need to make assumptions $\tilde{A}_1$, $\tilde{A}_2$, $\tilde{A}_3$. Assumptions $\tilde{A}_1$ and $\tilde{A}_2$ are analogous to [21] except under the misclassified information matrix. Assumption $\tilde{A}_3$ bounds the amount of misclassification in our data.

Define $S = \{(r, t) \in V \times V | t \in \mathcal{N}(r)\}$.

Assume the following assumptions hold uniformly for all $r \in V$:

(A1) Dependency Condition. For the misclassified information matrix and for the sample covariance matrix, there exists a constant $C_{\min}, D_{\max} > 0$ such that

$$\Lambda_{\min}((\tilde{Q}^*_r)^{SS}) \geq C_{\min}$$

$$\Lambda_{\max}(E_{Y, \theta^*}[X_{\gamma}X_{\gamma}^T]) \leq D_{\max}$$

(A2) Incoherence Condition. There exists $\alpha \in (0, 1]$ such that

$$\|\tilde{Q}^*_S^{-1}Q^*_S\|_{\infty} \leq 1 - \alpha$$

(A3) Misclassification Condition. For $C_{\min}$ and $D_{\max}$ as defined in (A1), and $\alpha$ as defined in (A2), we assume

$$S_{\max} \leq \frac{C_{\min}^3 \alpha^2}{400D_{\max}d(2 - \alpha)^2}$$

A.2 Proof

Within this proof we drop the node-specific subscript $r$. The proof is done within node, and a union bound is applied to obtain the result across nodes.

Define the sample misclassified information as

$$\tilde{Q}^n = -\tilde{E}(\nabla W^n(\theta^*))$$

In [21], Lemma 5, 6, and 7 can be applied to show that if $\tilde{X}$ is such that $\tilde{A}_1$ and $\tilde{A}_2$ hold for $\tilde{Q}^n$, then the assumptions will hold for with high probability for $\tilde{Q}^*$ for $n = \Omega(d^3 \log p)$. These lemmas directly apply to the misclassified case since their only dependence on the Ising distribution is that $\tilde{Q}^n - \tilde{Q}^*$ can be written as an iid mean of bounded observations, which still holds.

Therefore, to complete the proof it suffices to show that Extended Theorem 1 is true only for observations where the event $M = \{\tilde{X} : \tilde{A}_1 \text{and} \tilde{A}_2 \text{hold for} \tilde{Q}^n\}$ occurs. This corresponds to Proposition 1 of [21].

Define $\lambda_{n,d,p} = \lambda_{n,d,p} - \frac{4(2 - \alpha)}{\alpha}S_{\max}$. We can use Lemma 3, and Lemma 4 from [21] to show Extended Theorem 1 holds when $M$ occurs. In order to utilize these lemmas we need to establish an upper bound for the misclassified score function with high probability, and we need to establish
an upper bound for the quantity \( \lambda_{n,d,p,d} \). The following lemma proven in Appendix A.3. established an upper bound on the misclassified score function.

**Lemma:** For the specified incoherence parameter \( \alpha \in (0, 1] \), we have

\[
P\left( \|W^n\|_\infty \geq \frac{\lambda_{n,d,p,d}}{4} \right) = O(\exp(-K\tilde{\lambda}^2 \alpha n))
\]

for \( K \) independent of \( (n, d, p) \) and for \( \lambda_{n,d,p,d} \geq \frac{16(2-\alpha)}{\alpha} \left( \frac{\log p}{n} + \frac{S_{\max}}{4} \right)^{\frac{d}{2}} \),

In order to establish bounds for \( \lambda_{n,d,p,d} \), set \( n > \frac{400c_{\min}^2 d_{\max}}{C_{\min}} \left( \frac{\log p}{n} + \frac{S_{\max}}{4} \right)^{\frac{d}{2}} \). Then by assumption (A3) on \( S_{\max} \), and since \( \frac{\alpha}{2-\alpha} \leq 1 \) we have

\[
\lambda_{n,d,p,d} = \frac{16(2-\alpha)}{\alpha} \left( \frac{\log p}{n} + \frac{S_{\max}}{4} \right)^{\frac{d}{2}}
\]

We can bound (A11) from below by setting \( \delta = \frac{\alpha}{2-\alpha} \left( \frac{\log p}{n} + \frac{S_{\max}}{4} \right)^{\frac{d}{2}} \), and noting that \( \frac{\alpha}{2-\alpha} \leq 1 \). We get

\[
P\left( \frac{\|W^n\|_\infty}{\lambda_{n,d,p,d}} > \frac{n\delta^2}{8} \right) \leq 2 \exp\left( -\frac{n\delta^2}{8} \right)
\]

We bound (A11) from above as follows

\[
2 \exp\left( -\frac{n\delta^2}{8} \right) = 2 \exp\left( -\frac{n}{8} \left[ \frac{\alpha \lambda_{n,d,p,d}}{4(2-\alpha)} - |E(W^n)| \right]^2 \right)
\]

Combining (A11)–(A13) finishes the proof of the lemma.
Appendix 2. Proof of Regularized EM approach

In this appendix we show the following.

For data $\tilde{X}$, for $\hat{\theta}$ the parameter estimate from the RWL fit, and for $\theta^*$ the parameter estimate from the first EM update, there exists an open set of misclassification laws $\Gamma$ such that for the marginal penalized likelihood of our data as defined in Equation (11) we have that

$$L_\lambda(\theta^{*\Gamma}_U|\hat{\theta}_V\setminus U\Gamma, \tilde{X}) \geq L_\lambda(\hat{\theta}_U|\hat{\theta}_V\setminus U\Gamma, \tilde{X}) \quad (A16)$$

For notational convenience, we suppress the parameters $\hat{\theta}_V\setminus U\Gamma$, and we refer to our parameters of interest simply as $\theta$ as they do not change throughout the proof.

For $z_c$ as the latent states, by following the proof of the EM given in [17] we have the following relationship for the marginal likelihoods, which still holds when the regularization parameter is added

$$L_\lambda(\theta|\tilde{X}) = \sum_{i=1}^n \sum_{z_i \in \Omega_C} \mathbb{P}(z_i|\tilde{X}(i)) \log(\mathbb{P}(\tilde{X}(i)|z_i|\tilde{X}(i)))$$

$$- \sum_{i=1}^n \sum_{z_i \in \Omega_C} \mathbb{P}(z_i|\tilde{X}(i)) \log(\mathbb{P}(\mathbb{P}(z_i|\tilde{X}(i))) + \lambda \|\theta\|_1$$

$$= A_{\Gamma}(\theta) + B_{\Gamma}(\theta) + \lambda \|\theta\|_1 \quad (A17)$$

Where $A_{\Gamma}(\theta)$ and $B_{\Gamma}(\theta)$ correspond the two large summations in Equation (A17). $\Gamma$ is included in the notation for these functions to emphasize their dependence on the misclassification scheme.

For $B_{\Gamma}(\theta)$ we have that by Gibb’s inequality, $B_{\Gamma}(\theta) \geq B_{\Gamma}(\hat{\theta})$ for all $\theta$, and for all $\Gamma$. Therefore $B_{\Gamma}(\theta)$ will increase at $\theta^*$. Our goal is thus to show that $A(\theta) + \lambda \|\theta\|_1$ will increase.

Choose the misclassification setting $\Gamma'$ such that $\prod_{z \in C} P(z_i \neq \tilde{x}^{(i)}_c) = 1$. Define $\tilde{x}^{(i)}_c$ component-wise as $(z_c^{(i)})_s = -\tilde{x}^{(i)}_s$. Under this $\Gamma'$, we have the following representation for $A_{\Gamma'}(\theta)$

$$A_{\Gamma'}(\theta) = \sum_{i=1}^n \sum_{z_i \in \Omega_C} \mathbb{P}(z_i|\tilde{X}(i)) \log(\mathbb{P}((\tilde{x}^{(i)}_c)(z_i)|\tilde{x}^{(i)}(c)(\tilde{X}(i)))\mathbb{P}(z_i|\tilde{x}^{(i)}_c)) \quad (A18)$$

$$= \sum_{i=1}^n \mathbb{P}((\tilde{z}^{(i)}_c)|\tilde{X}(i)) \log(\mathbb{P}(\tilde{x}^{(i)}_c(z^{(i)}_c)|\tilde{x}^{(i)}(c))) \quad (A19)$$

For this selection of $\Gamma'$ we have that $\theta^*$ is chosen to maximize $A_{\Gamma'}(\theta) + \lambda \|\theta\|_1$, and therefore $A_{\Gamma'}(\theta^*) + B_{\Gamma'}(\theta^*) + \lambda \|\theta^*\|_1 \geq A_{\Gamma'}(\hat{\theta}) + B_{\Gamma'}(\hat{\theta}) + \lambda \|\hat{\theta}\|$. Since $A_{\Gamma'}(\hat{\theta}) + B_{\Gamma'}(\hat{\theta}) + \lambda \|\hat{\theta}\|_1$ is continuous in $\Gamma$, there exists an open set $\Gamma$ such that if $\Gamma \in \Gamma'$ then $L_\lambda(\theta^{*\Gamma}_U|\hat{\theta}_V\setminus U\Gamma, \tilde{X}) \geq L_\lambda(\hat{\theta}_U|\hat{\theta}_V\setminus U\Gamma, \tilde{X})$ as needed.

Appendix 3. Calculating weights for E-step

Here we calculate the weights $P_{\theta^{(k)}}(X_c = z_c|\tilde{X}_U = \tilde{x}_U)$ from equation (13). In these calculations we assume we have $y_s^i$ corresponding to the misclassification probability for node $s$ at observation $i$.

We remove the subscript for estimate $\theta^{(k)}$, and superscript for observation $(i)$ for notational convenience. Rearranging conditional and joint probabilities give us

$$P(X_c = z_c|\tilde{X}_U = \tilde{x}_U) = \frac{P(X_c = z_c, \tilde{X}_c = \tilde{x}_c, \tilde{X}_P = \tilde{x}_P)}{P(\tilde{X}_C = \tilde{x}_c, \tilde{X}_P = \tilde{x}_P)} \quad (A20)$$

$$= \frac{P(X_c = z_c, \tilde{X}_P = \tilde{x}_P)}{P(\tilde{X}_C = \tilde{x}_c, \tilde{X}_P = \tilde{x}_P)} P(\tilde{X}_C = \tilde{x}_c|X_c = z_c, \tilde{X}_P = \tilde{x}_P) \quad (A21)$$
The conditional probability in (A21) gives the proportion of the weight associated with the observed misclassification probability. This is calculated as

\[
P(\tilde{X}_C = \tilde{x}_C | X_C = z_c, \tilde{X}_P = \tilde{x}_P) = P(\tilde{X}_C = \tilde{x}_C | X_C = z_c)
\]

\[
= \prod_{s \in C} (\gamma_s 1(\tilde{x}_s \neq (z_c)_s) + (1 - \gamma_s) 1(\tilde{x}_s = (z_c)_s))
\]

\[
= c(z_c, \tilde{x}_U)
\]

The ratio of probabilities gives the weight of the observation associated with the estimated dependency structure. Define \(A(x_C, x_P) = \sum_{(s,t) \in E_U} \theta^{(f)}_{st} x_s x_t\); this corresponds to the association between nodes in \(\mathcal{U}\) as it relates to the full distribution given in (1). From the selection of \(\mathcal{U}\) the ratio of probabilities factors allowing this calculation to ignore nodes outside of \(\mathcal{U}\).

\[
\frac{P(X_C = z_c, \tilde{X}_P = \tilde{x}_P)}{P(\tilde{X}_C = \tilde{x}_C, \tilde{X}_P = \tilde{x}_P)} = \frac{B \exp(A(x_C, x_P))}{B \sum_{z'_c \in \Omega_C} c(z'_c, \tilde{x}_U)}
\]

\[
= \frac{\exp(A(x_C, x_P))}{\sum_{z'_c \in \Omega_C} c(z'_c, \tilde{x}_U)}
\]

Where \(B\) in the above equation corresponds to the potential from all nodes outside of \(\mathcal{U}\).