

# On the sample complexity of graph selection: Practical methods and fundamental limits

Martin Wainwright

UC Berkeley  
Departments of Statistics, and EECS

Based on joint work with:

John Lafferty (CMU)

Pradeep Ravikumar (UT Austin)

Prasad Santhanam (Univ. Hawaii)

# Introduction

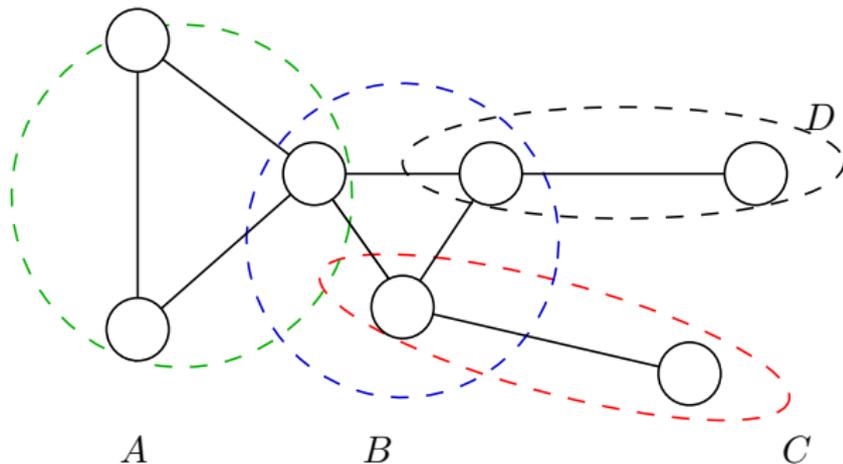
- Markov random fields (undirected graphical models): central to many applications in science and engineering:
  - ▶ communication, coding, information theory, networking
  - ▶ machine learning and statistics
  - ▶ computer vision; image processing
  - ▶ statistical physics
  - ▶ bioinformatics, computational biology ...

# Introduction

- Markov random fields (undirected graphical models): central to many applications in science and engineering:
  - ▶ communication, coding, information theory, networking
  - ▶ machine learning and statistics
  - ▶ computer vision; image processing
  - ▶ statistical physics
  - ▶ bioinformatics, computational biology ...
- some core computational problems
  - ▶ *counting/integrating*: computing marginal distributions and data likelihoods
  - ▶ *optimization*: computing most probable configurations (or top  $M$ -configurations)
  - ▶ *model selection*: fitting and selecting models on the basis of data

# What are graphical models?

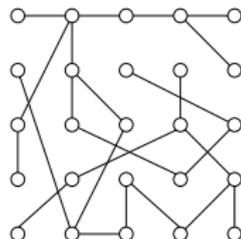
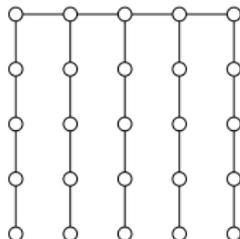
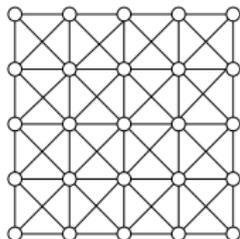
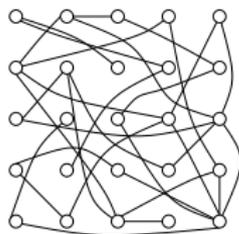
- Markov random field: random vector  $(X_1, \dots, X_p)$  with distribution factoring according to a graph  $G = (V, E)$ :



- Hammersley-Clifford Theorem:  $(X_1, \dots, X_p)$  being Markov w.r.t  $G$  implies factorization over graph cliques
- studied/used in various fields: spatial statistics, language modeling, computational biology, computer vision, statistical physics ....

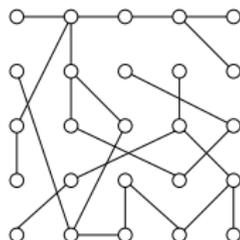
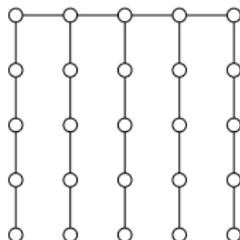
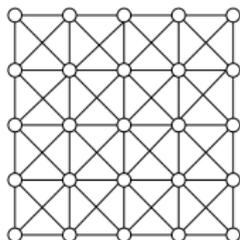
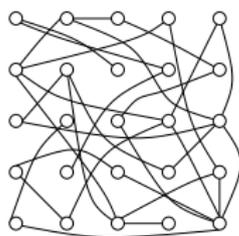
# Graphical model selection

- let  $G = (V, E)$  be an undirected graph on  $p = |V|$  vertices



# Graphical model selection

- let  $G = (V, E)$  be an undirected graph on  $p = |V|$  vertices

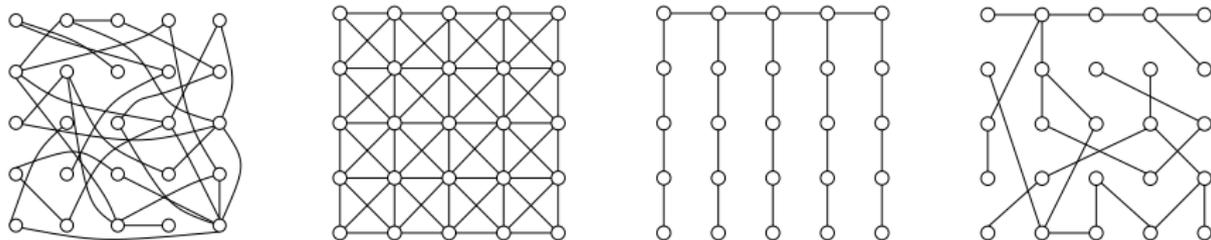


- pairwise Markov random field: family of prob. distributions

$$\mathbb{P}(x_1, \dots, x_p; \theta) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{(s,t) \in E} \langle \theta_{st}, \phi_{st}(x_s, x_t) \rangle \right\}.$$

# Graphical model selection

- let  $G = (V, E)$  be an undirected graph on  $p = |V|$  vertices



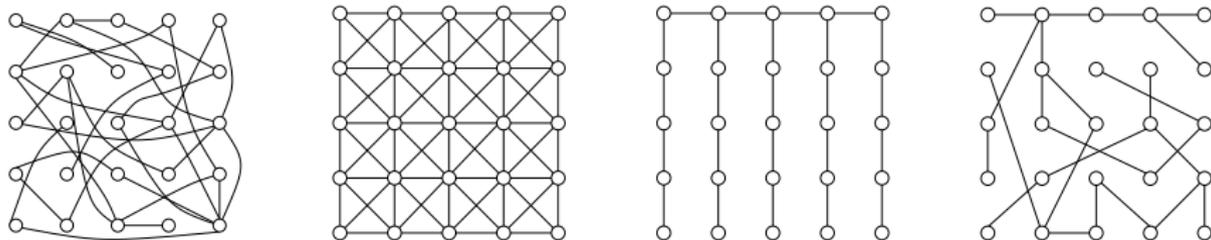
- pairwise Markov random field: family of prob. distributions

$$\mathbb{P}(x_1, \dots, x_p; \theta) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{(s,t) \in E} \langle \theta_{st}, \phi_{st}(x_s, x_t) \rangle \right\}.$$

- Problem of graph selection:** given  $n$  independent and identically distributed (i.i.d.) samples of  $X = (X_1, \dots, X_p)$ , identify the underlying graph structure

# Graphical model selection

- let  $G = (V, E)$  be an undirected graph on  $p = |V|$  vertices

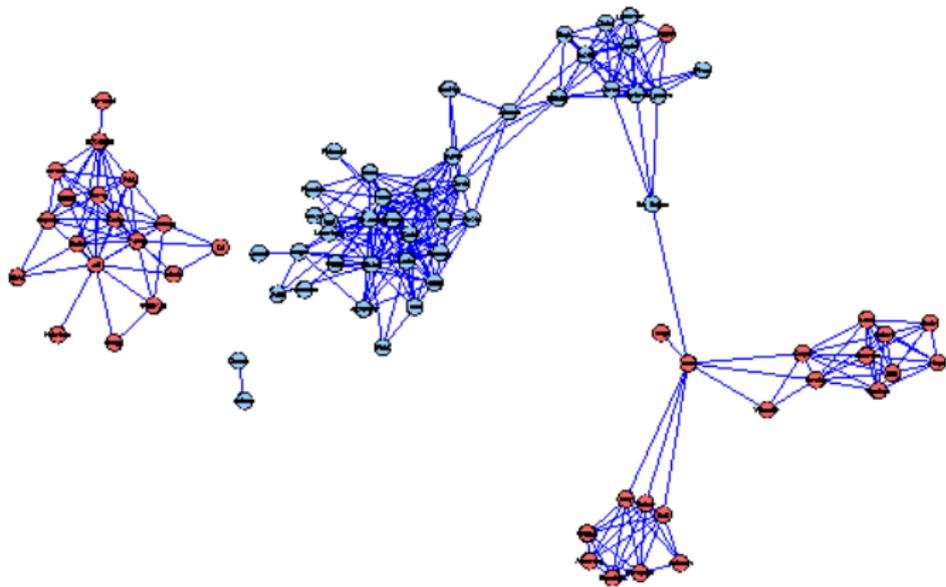


- pairwise Markov random field: family of prob. distributions

$$\mathbb{P}(x_1, \dots, x_p; \theta) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{(s,t) \in E} \langle \theta_{st}, \phi_{st}(x_s, x_t) \rangle \right\}.$$

- Problem of graph selection:** given  $n$  independent and identically distributed (i.i.d.) samples of  $X = (X_1, \dots, X_p)$ , identify the underlying graph structure
- complexity constraint: restrict to subset  $\mathcal{G}_{d,p}$  of graphs with maximum degree  $d$

# Illustration: Voting behavior of US senators



Graphical model fit to voting records of US senators (Bannerjee, El Ghaoui, & d'Aspremont, 2008)

# Outline of remainder of talk

- 1 Background and past work
- 2 A practical scheme for graphical model selection
  - (a)  $\ell_1$ -regularized neighborhood regression
  - (b) High-dimensional analysis and phase transitions
- 3 Fundamental limits of graphical model selection
  - (a) An unorthodox channel coding problem
  - (b) Necessary conditions
  - (c) Sufficient conditions (optimal algorithms)
- 4 Various open questions.....

# Previous/on-going work on graph selection

- methods for Gaussian MRFs
  - ▶  $\ell_1$ -regularized neighborhood regression for Gaussian MRFs (e.g., Meinshausen & Buhlmann, 2005; Wainwright, 2006, Zhao, 2006)
  - ▶  $\ell_1$ -regularized log-determinant (e.g., Yuan & Lin, 2006; d'Aspremont et al., 2007; Friedman, 2008; Ravikumar et al., 2008)

# Previous/on-going work on graph selection

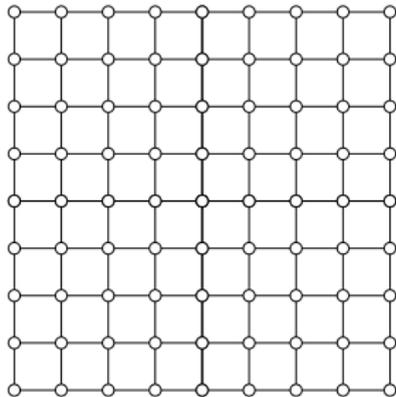
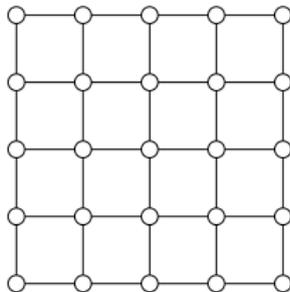
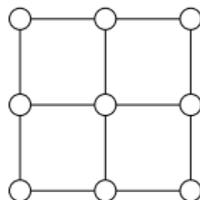
- methods for Gaussian MRFs
  - ▶  $\ell_1$ -regularized neighborhood regression for Gaussian MRFs (e.g., Meinshausen & Buhlmann, 2005; Wainwright, 2006, Zhao, 2006)
  - ▶  $\ell_1$ -regularized log-determinant (e.g., Yuan & Lin, 2006; d'Aspremont et al., 2007; Friedman, 2008; Ravikumar et al., 2008)
- methods for discrete MRFs
  - ▶ exact solution for trees (Chow & Liu, 1967)
  - ▶ local testing (e.g., Spirtes et al, 2000; Kalisch & Buhlmann, 2008)
  - ▶ distribution fits by KL-divergence (Abeel et al., 2005)
  - ▶  $\ell_1$ -regularized logistic regression (Ravikumar, W. & Lafferty et al., 2006, 2008)
  - ▶ approximate max. entropy approach and thinned graphical models (Johnson et al., 2007)
  - ▶ neighborhood-based thresholding method (Bresler, Mossel & Sly, 2008)

# Previous/on-going work on graph selection

- methods for Gaussian MRFs
  - ▶  $\ell_1$ -regularized neighborhood regression for Gaussian MRFs (e.g., Meinshausen & Buhlmann, 2005; Wainwright, 2006, Zhao, 2006)
  - ▶  $\ell_1$ -regularized log-determinant (e.g., Yuan & Lin, 2006; d'Aspremont et al., 2007; Friedman, 2008; Ravikumar et al., 2008)
- methods for discrete MRFs
  - ▶ exact solution for trees (Chow & Liu, 1967)
  - ▶ local testing (e.g., Spirtes et al, 2000; Kalisch & Buhlmann, 2008)
  - ▶ distribution fits by KL-divergence (Abeel et al., 2005)
  - ▶  $\ell_1$ -regularized logistic regression (Ravikumar, W. & Lafferty et al., 2006, 2008)
  - ▶ approximate max. entropy approach and thinned graphical models (Johnson et al., 2007)
  - ▶ neighborhood-based thresholding method (Bresler, Mossel & Sly, 2008)
- information-theoretic analysis
  - ▶ pseudolikelihood and BIC criterion (Csiszar & Talata, 2006)
  - ▶ information-theoretic limitations (Santhanam & W., 2008)

# High-dimensional analysis

- classical analysis: dimension  $p$  fixed, sample size  $n \rightarrow +\infty$
- high-dimensional analysis: allow both dimension  $p$ , sample size  $n$ , and maximum degree  $d$  to increase at arbitrary rates



- take  $n$  i.i.d. samples from MRF defined by  $G_{p,d}$
- study probability of success as a function of three parameters:

$$\text{Success}(n, p, d) = \mathbb{P}[\text{Method recovers graph } G_{p,d} \text{ from } n \text{ samples}]$$

- theory is non-asymptotic: explicit probabilities for finite  $(n, p, d)$

## Some challenges in distinguishing graphs

- clearly, a lower bound on the **minimum edge weight** is required:

$$\min_{(s,t) \in E} |\theta_{st}^*| \geq \theta_{\min},$$

although  $\theta_{\min}(p, d) = o(1)$  is allowed.

- in contrast to other testing/detection problems, **large  $|\theta_{st}|$  also problematic**

# Some challenges in distinguishing graphs

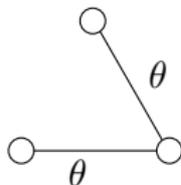
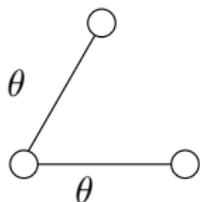
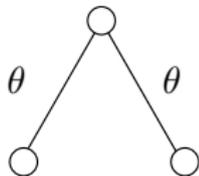
- clearly, a lower bound on the **minimum edge weight** is required:

$$\min_{(s,t) \in E} |\theta_{st}^*| \geq \theta_{\min},$$

although  $\theta_{\min}(p, d) = o(1)$  is allowed.

- in contrast to other testing/detection problems, **large  $|\theta_{st}|$  also problematic**

**Toy example:** Graphs from  $\mathcal{G}_{3,2}$  (i.e.,  $p = 3$ ;  $d = 2$ )



As  $\theta$  increases, all three Markov random fields become arbitrarily close to:

$$\mathbb{P}(x_1, x_2, x_3) = \begin{cases} 1/2 & \text{if } x \in \{(-1)^3, (+1)^3\} \\ 0 & \text{otherwise.} \end{cases}$$

# Markov property and neighborhood structure

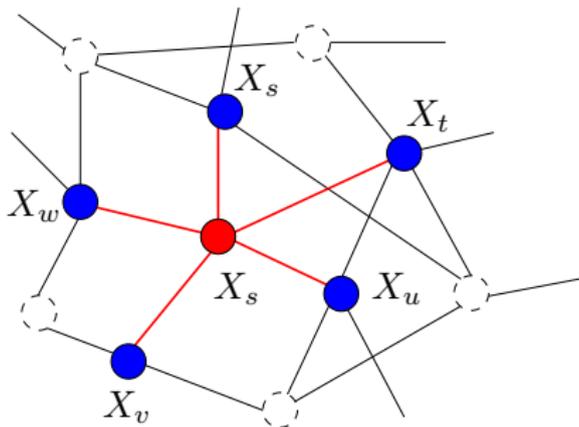
- Markov properties encode neighborhood structure:

$$\underbrace{(X_s \mid X_{V \setminus s})}_\text{Condition on full graph} \stackrel{d}{=} \underbrace{(X_s \mid X_{N(s)})}_\text{Condition on Markov blanket}$$

Condition on full graph

Condition on Markov blanket

$$N(s) = \{s, t, u, v, w\}$$



- basis of pseudolikelihood method
- used for Gaussian model selection

(Besag, 1974)

(Meinshausen & Buhlmann, 2006)

## §2. Practical method via neighborhood regression

**Observation:** Recovering graph  $G$  equivalent to recovering neighborhood set  $N(s)$  for all  $s \in V$ .

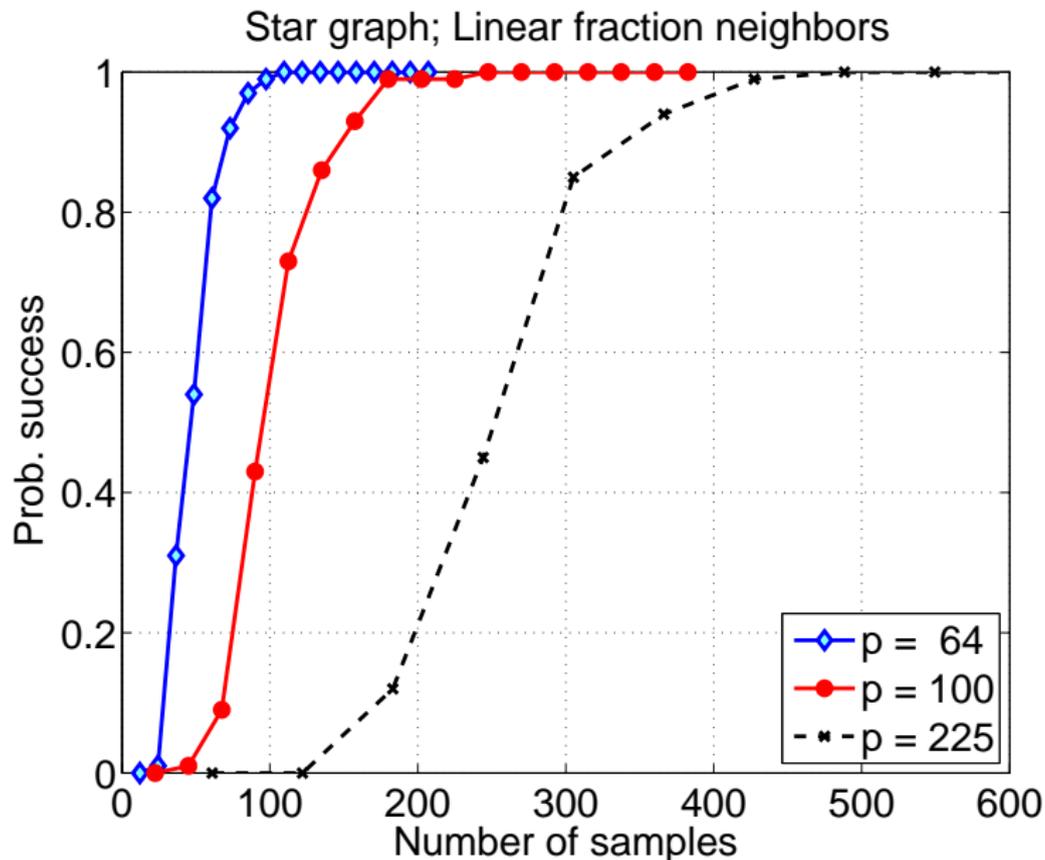
**Method:** Given  $n$  i.i.d. samples  $\{X^{(1)}, \dots, X^{(n)}\}$ , perform logistic regression of each node  $X_s$  on  $X_{\setminus s} := \{X_t, t \neq s\}$  to estimate neighborhood structure  $\hat{N}(s)$ .

- 1 For each node  $s \in V$ , perform  $\ell_1$  regularized logistic regression of  $X_s$  on the remaining variables  $X_{\setminus s}$ :

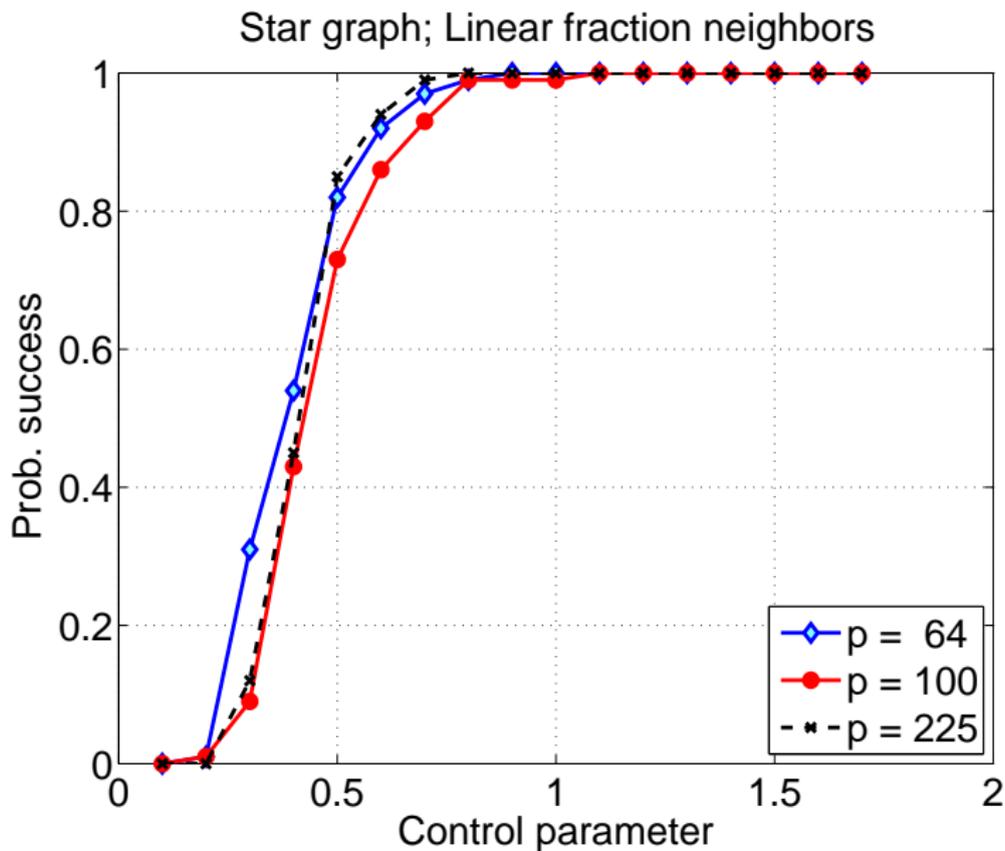
$$\hat{\theta}[s] := \arg \min_{\theta \in \mathbb{R}^{p-1}} \left\{ \underbrace{\frac{1}{n} \sum_{i=1}^n f(\theta; X_{\setminus s}^{(i)})}_{\text{logistic likelihood}} + \underbrace{\rho_n \|\theta\|_1}_{\text{regularization}} \right\}$$

- 2 Estimate the local neighborhood  $\hat{N}(s)$  as the support (non-negative entries) of the regression vector  $\hat{\theta}[s]$ .
- 3 Combine the neighborhood estimates in a consistent manner (AND, or OR rule).

# Empirical behavior: Unrescaled plots



# Empirical behavior: Appropriately rescaled



# Sufficient conditions for consistent model selection

- graph sequences  $G_{p,d} = (V, E)$  with  $p$  vertices, and maximum degree  $d$ .
- edge weights  $|\theta_{st}| \geq \theta_{\min}$  for all  $(s, t) \in E$
- draw  $n$  i.i.d, samples, and analyze prob. success indexed by  $(n, p, d)$

## Theorem

# Sufficient conditions for consistent model selection

- graph sequences  $G_{p,d} = (V, E)$  with  $p$  vertices, and maximum degree  $d$ .
- edge weights  $|\theta_{st}| \geq \theta_{\min}$  for all  $(s, t) \in E$
- draw  $n$  i.i.d. samples, and analyze prob. success indexed by  $(n, p, d)$

## Theorem

Under incoherence conditions, for a rescaled sample size (RavWaiLaf06)

$$\theta_{LR}(n, p, d) := \frac{n}{d^3 \log p} > \theta_{\text{crit}}$$

and regularization parameter  $\rho_n \geq c_1 \tau \sqrt{\frac{\log p}{n}}$ , then with probability greater than  $1 - 2 \exp(-c_2(\tau - 2) \log p) \rightarrow 1$ :

- (a) Uniqueness:** For each node  $s \in V$ , the  $\ell_1$ -regularized logistic convex program has a unique solution. (Non-trivial since  $p \gg n \implies$  not strictly convex).

# Sufficient conditions for consistent model selection

- graph sequences  $G_{p,d} = (V, E)$  with  $p$  vertices, and maximum degree  $d$ .
- edge weights  $|\theta_{st}| \geq \theta_{\min}$  for all  $(s, t) \in E$
- draw  $n$  i.i.d. samples, and analyze prob. success indexed by  $(n, p, d)$

## Theorem

Under incoherence conditions, for a rescaled sample size (RavWaiLaf06)

$$\theta_{LR}(n, p, d) := \frac{n}{d^3 \log p} > \theta_{\text{crit}}$$

and regularization parameter  $\rho_n \geq c_1 \tau \sqrt{\frac{\log p}{n}}$ , then with probability greater than  $1 - 2 \exp(-c_2(\tau - 2) \log p) \rightarrow 1$ :

- (a) Uniqueness:** For each node  $s \in V$ , the  $\ell_1$ -regularized logistic convex program has a unique solution. (Non-trivial since  $p \gg n \implies$  not strictly convex).
- (b) Correct exclusion:** The estimated sign neighborhood  $\hat{N}(s)$  correctly excludes all edges not in the true neighborhood.

# Sufficient conditions for consistent model selection

- graph sequences  $G_{p,d} = (V, E)$  with  $p$  vertices, and maximum degree  $d$ .
- edge weights  $|\theta_{st}| \geq \theta_{\min}$  for all  $(s, t) \in E$
- draw  $n$  i.i.d. samples, and analyze prob. success indexed by  $(n, p, d)$

## Theorem

Under incoherence conditions, for a rescaled sample size (RavWaiLaf06)

$$\theta_{LR}(n, p, d) := \frac{n}{d^3 \log p} > \theta_{\text{crit}}$$

and regularization parameter  $\rho_n \geq c_1 \tau \sqrt{\frac{\log p}{n}}$ , then with probability greater than  $1 - 2 \exp(-c_2(\tau - 2) \log p) \rightarrow 1$ :

- (a) Uniqueness:** For each node  $s \in V$ , the  $\ell_1$ -regularized logistic convex program has a unique solution. (Non-trivial since  $p \gg n \implies$  not strictly convex).
- (b) Correct exclusion:** The estimated sign neighborhood  $\hat{N}(s)$  correctly excludes all edges not in the true neighborhood.
- (c) Correct inclusion:** For  $\theta_{\min} \geq c_3 \tau \sqrt{d} \rho_n$ , the method selects the correct signed neighborhood.

# Sufficient conditions for consistent model selection

- graph sequences  $G_{p,d} = (V, E)$  with  $p$  vertices, and maximum degree  $d$ .
- edge weights  $|\theta_{st}| \geq \theta_{\min}$  for all  $(s, t) \in E$
- draw  $n$  i.i.d. samples, and analyze prob. success indexed by  $(n, p, d)$

## Theorem

Under incoherence conditions, for a rescaled sample size (RavWaiLaf06)

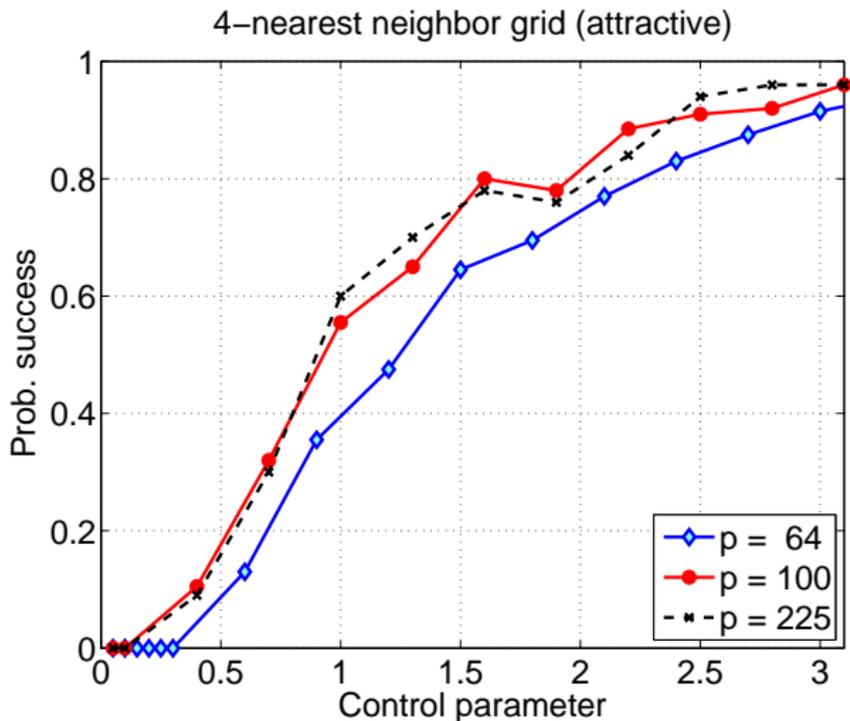
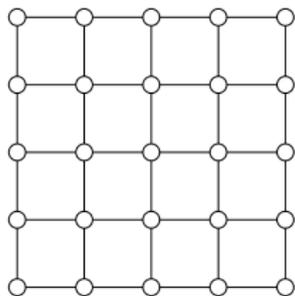
$$\theta_{LR}(n, p, d) := \frac{n}{d^3 \log p} > \theta_{\text{crit}}$$

and regularization parameter  $\rho_n \geq c_1 \tau \sqrt{\frac{\log p}{n}}$ , then with probability greater than  $1 - 2 \exp(-c_2(\tau - 2) \log p) \rightarrow 1$ :

- (a) Uniqueness:** For each node  $s \in V$ , the  $\ell_1$ -regularized logistic convex program has a unique solution. (Non-trivial since  $p \gg n \implies$  not strictly convex).
- (b) Correct exclusion:** The estimated sign neighborhood  $\widehat{N}(s)$  correctly excludes all edges not in the true neighborhood.
- (c) Correct inclusion:** For  $\theta_{\min} \geq c_3 \tau \sqrt{d} \rho_n$ , the method selects the correct signed neighborhood.

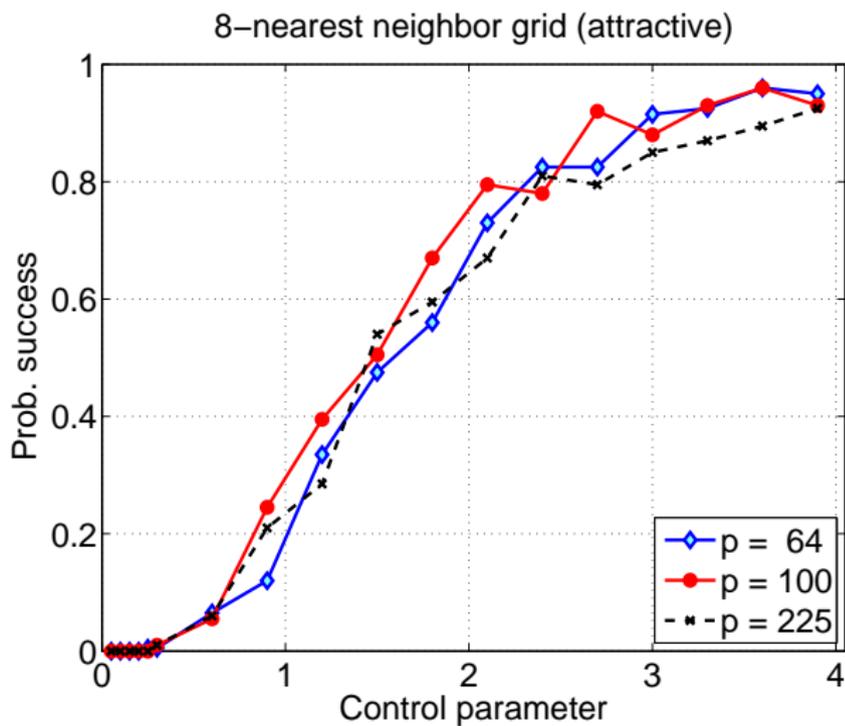
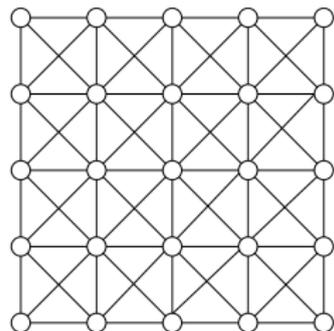
**Consequence:** For  $\theta_{\min} = \Omega(1/d)$ , it suffices to have  $n = \Omega(d^3 \log p)$ .

# Rescaled plots for 4-grid graphs



Prob. of success  $\mathbb{P}[\hat{G} = G]$  versus rescaled sample size  $\theta_{LR}(n, p, d^3) = \frac{n}{d^3 \log p}$

# Results for 8-grid graphs



Prob. of success  $\mathbb{P}[\hat{G} = G]$  versus rescaled sample size  $\theta_{LR}(n, p, d^3) = \frac{n}{d^3 \log p}$

# Assumptions

Define Fisher information matrix of logistic regression:

$$Q^* := \mathbb{E}_{\theta^*} [\nabla^2 f(\theta^*; X)].$$

**A1. Dependency condition:** Bounded eigenspectra:

$$C_{min} \leq \lambda_{min}(Q_{SS}^*), \quad \text{and} \quad \lambda_{max}(Q_{SS}^*) \leq C_{max}.$$
$$\lambda_{max}(\mathbb{E}_{\theta^*} [XX^T]) \leq D_{max}.$$

**A2. Incoherence** There exists an  $\nu \in (0, 1]$  such that

$$\|Q_{S^c S}^* (Q_{SS}^*)^{-1}\|_{\infty, \infty} \leq 1 - \nu.$$

where  $\|A\|_{\infty, \infty} := \max_i \sum_j |A_{ij}|$ .

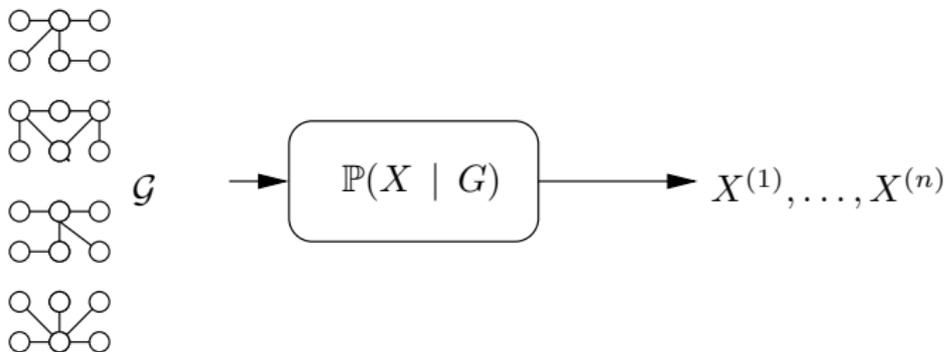
- bounds on eigenvalues are fairly standard
- incoherence condition:
  - ▶ partly necessary (prevention of degenerate models)
  - ▶ partly an artifact of  $\ell_1$ -regularization
- incoherence condition is weaker than correlation decay

## §3. Info. theory: Graph selection as channel coding

- graphical model selection is an *unorthodox* channel coding problem:

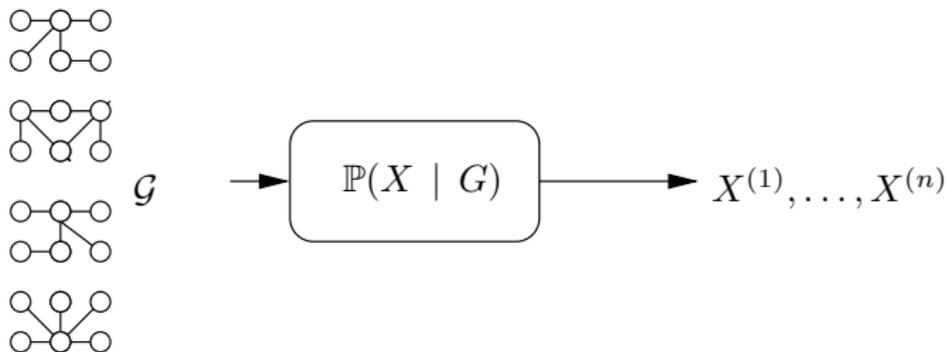
### §3. Info. theory: Graph selection as channel coding

- graphical model selection is an *unorthodox* channel coding problem:
  - codewords/codebook: graph  $G$  in some graph class  $\mathcal{G}$
  - channel use: draw sample  $X^{(i)} = (X_1^{(i)}, \dots, X_p^{(i)})$  from Markov random field  $\mathbb{P}_{\theta(G)}$
  - decoding problem: use  $n$  samples  $\{X^{(1)}, \dots, X^{(n)}\}$  to correctly distinguish the “codeword”



### §3. Info. theory: Graph selection as channel coding

- graphical model selection is an *unorthodox* channel coding problem:
  - codewords/codebook: graph  $G$  in some graph class  $\mathcal{G}$
  - channel use: draw sample  $X^{(i)} = (X_1^{(i)}, \dots, X_p^{(i)})$  from Markov random field  $\mathbb{P}_{\theta(G)}$
  - decoding problem: use  $n$  samples  $\{X^{(1)}, \dots, X^{(n)}\}$  to correctly distinguish the “codeword”



Channel capacity for graph decoding determined by balance between

- log number of models
- relative distinguishability of different models

# Necessary conditions for $\mathcal{G}_{d,p}$

- $G \in \mathcal{G}_{d,p}$ : graphs with  $p$  nodes and max. degree  $d$
- Ising models with:
  - ▶ *Minimum edge weight*:  $|\theta_{st}^*| \geq \theta_{\min}$  for all edges
  - ▶ *Maximum neighborhood weight*:  $\omega(\theta) := \max_{s \in V} \sum_{t \in N(s)} |\theta_{st}^*|$

# Necessary conditions for $\mathcal{G}_{d,p}$

- $G \in \mathcal{G}_{d,p}$ : graphs with  $p$  nodes and max. degree  $d$
- Ising models with:
  - ▶ *Minimum edge weight*:  $|\theta_{st}^*| \geq \theta_{\min}$  for all edges
  - ▶ *Maximum neighborhood weight*:  $\omega(\theta) := \max_{s \in V} \sum_{t \in N(s)} |\theta_{st}^*|$

## Theorem

If the sample size  $n$  is upper bounded by

(Santhanam & W, 2008)

$$n < \max \left\{ \frac{d}{8} \log \frac{p}{8d}, \frac{\exp(\frac{\omega(\theta)}{4}) d \theta_{\min} \log(pd/8)}{128 \exp(\frac{3\theta_{\min}}{2})}, \frac{\log p}{2\theta_{\min} \tanh(\theta_{\min})} \right\}$$

then the probability of error of any algorithm over  $\mathcal{G}_{d,p}$  is at least  $1/2$ .

# Necessary conditions for $\mathcal{G}_{d,p}$

- $G \in \mathcal{G}_{d,p}$ : graphs with  $p$  nodes and max. degree  $d$
- Ising models with:
  - ▶ *Minimum edge weight*:  $|\theta_{st}^*| \geq \theta_{\min}$  for all edges
  - ▶ *Maximum neighborhood weight*:  $\omega(\theta) := \max_{s \in V} \sum_{t \in N(s)} |\theta_{st}^*|$

## Theorem

If the sample size  $n$  is upper bounded by

(Santhanam & W, 2008)

$$n < \max \left\{ \frac{d}{8} \log \frac{p}{8d}, \frac{\exp(\frac{\omega(\theta)}{4}) d \theta_{\min} \log(pd/8)}{128 \exp(\frac{3\theta_{\min}}{2})}, \frac{\log p}{2\theta_{\min} \tanh(\theta_{\min})} \right\}$$

then the probability of error of any algorithm over  $\mathcal{G}_{d,p}$  is at least  $1/2$ .

## Interpretation:

- **Naive bulk effect**: Arises from log cardinality  $\log |\mathcal{G}_{d,p}|$
- **$d$ -clique effect**: Difficulty of separating models that contain a near  $d$ -clique
- **Small weight effect**: Difficult to detect edges with small weights.

# Some consequences

## Corollary

*For asymptotically reliable recovery over  $\mathcal{G}_{d,p}$ , any algorithm requires **at least**  $n = \Omega(d^2 \log p)$  samples.*

# Some consequences

## Corollary

For asymptotically reliable recovery over  $\mathcal{G}_{d,p}$ , any algorithm requires *at least*  $n = \Omega(d^2 \log p)$  samples.

- note that **maximum neighborhood weight**  $\omega(\theta^*) \geq d\theta_{\min} \implies$  require  $\theta_{\min} = \mathcal{O}(1/d)$

# Some consequences

## Corollary

For asymptotically reliable recovery over  $\mathcal{G}_{d,p}$ , any algorithm requires *at least*  $n = \Omega(d^2 \log p)$  samples.

- note that **maximum neighborhood weight**  $\omega(\theta^*) \geq d \theta_{\min} \implies$  require  $\theta_{\min} = \mathcal{O}(1/d)$
- from **small weight effect**

$$n = \Omega\left(\frac{\log p}{\theta_{\min} \tanh(\theta_{\min})}\right) = \Omega\left(\frac{\log p}{\theta_{\min}^2}\right)$$

# Some consequences

## Corollary

For asymptotically reliable recovery over  $\mathcal{G}_{d,p}$ , any algorithm requires *at least*  $n = \Omega(d^2 \log p)$  samples.

- note that **maximum neighborhood weight**  $\omega(\theta^*) \geq d\theta_{\min} \implies$  require  $\theta_{\min} = \mathcal{O}(1/d)$
- from **small weight effect**

$$n = \Omega\left(\frac{\log p}{\theta_{\min} \tanh(\theta_{\min})}\right) = \Omega\left(\frac{\log p}{\theta_{\min}^2}\right)$$

- conclude that  $\ell_1$ -regularized logistic regression (LR) is within  $\Theta(d)$  of optimal for general graphs (Ravikumar., W. & Lafferty, 2006)

# Some consequences

## Corollary

For asymptotically reliable recovery over  $\mathcal{G}_{d,p}$ , any algorithm requires *at least*  $n = \Omega(d^2 \log p)$  samples.

- note that **maximum neighborhood weight**  $\omega(\theta^*) \geq d\theta_{\min} \implies$  require  $\theta_{\min} = \mathcal{O}(1/d)$
- from **small weight effect**

$$n = \Omega\left(\frac{\log p}{\theta_{\min} \tanh(\theta_{\min})}\right) = \Omega\left(\frac{\log p}{\theta_{\min}^2}\right)$$

- conclude that  $\ell_1$ -regularized logistic regression (LR) is within  $\Theta(d)$  of optimal for general graphs (Ravikumar., W. & Lafferty, 2006)
- for bounded degree graphs:
  - ▶  $\ell_1$ -LR order-optimal under incoherence conditions with cost  $\mathcal{O}(p^4)$
  - ▶ thresholding procedure order-optimal under correlation decay, also with polynomial complexity (Bresler, Sly & Mossel, 2008)

# Proof sketch: Main ideas for necessary conditions

- based on assessing difficulty of graph selection over various sub-ensembles  
 $\mathcal{G} \subseteq \mathcal{G}_{p,d}$

# Proof sketch: Main ideas for necessary conditions

- based on assessing difficulty of graph selection over various sub-ensembles  $\mathcal{G} \subseteq \mathcal{G}_{p,d}$
- choose  $G \in \mathcal{G}$  u.a.r., and consider multi-way hypothesis testing problem based on the data  $\mathbf{X}_1^n = \{X^{(1)}, \dots, X^{(n)}\}$

## Proof sketch: Main ideas for necessary conditions

- based on assessing difficulty of graph selection over various sub-ensembles  $\mathcal{G} \subseteq \mathcal{G}_{p,d}$
- choose  $G \in \mathcal{G}$  u.a.r., and consider multi-way hypothesis testing problem based on the data  $\mathbf{X}_1^n = \{X^{(1)}, \dots, X^{(n)}\}$
- for any graph estimator  $\psi : \mathcal{X}^n \rightarrow \mathcal{G}$ , Fano's inequality implies that

$$\mathbb{P}[\psi(\mathbf{X}_1^n) \neq G] \geq 1 - \frac{I(\mathbf{X}_1^n; G)}{\log |\mathcal{G}|} - o(1)$$

where  $I(\mathbf{X}_1^n; G)$  is mutual information between observations  $\mathbf{X}_1^n$  and randomly chosen graph  $G$

# Proof sketch: Main ideas for necessary conditions

- based on assessing difficulty of graph selection over various sub-ensembles  $\mathcal{G} \subseteq \mathcal{G}_{p,d}$
- choose  $G \in \mathcal{G}$  u.a.r., and consider multi-way hypothesis testing problem based on the data  $\mathbf{X}_1^n = \{X^{(1)}, \dots, X^{(n)}\}$
- for any graph estimator  $\psi : \mathcal{X}^n \rightarrow \mathcal{G}$ , Fano's inequality implies that

$$\mathbb{P}[\psi(\mathbf{X}_1^n) \neq G] \geq 1 - \frac{I(\mathbf{X}_1^n; G)}{\log |\mathcal{G}|} - o(1)$$

where  $I(\mathbf{X}_1^n; G)$  is mutual information between observations  $\mathbf{X}_1^n$  and randomly chosen graph  $G$

- remaining steps:
  - 1 Construct “difficult” sub-ensembles  $\mathcal{G} \subseteq \mathcal{G}_{p,d}$
  - 2 Compute or lower bound the log cardinality  $\log |\mathcal{G}|$ .
  - 3 Upper bound the mutual information  $I(\mathbf{X}_1^n; G)$ .

# Two straightforward ensembles

## Two straightforward ensembles

- 1 Naive bulk ensemble: All graphs on  $p$  vertices with max. degree  $d$  (i.e.,  $\mathcal{G} = \mathcal{G}_{p,d}$ )

## Two straightforward ensembles

- 1 **Naive bulk ensemble:** All graphs on  $p$  vertices with max. degree  $d$  (i.e.,  $\mathcal{G} = \mathcal{G}_{p,d}$ )
- ▶ simple counting argument:  $\log |\mathcal{G}_{p,d}| = \Theta(pd \log(p/d))$
  - ▶ trivial upper bound:  $I(\mathbf{X}_1^n; G) \leq H(\mathbf{X}_1^n) \leq np$ .
  - ▶ substituting into Fano yields necessary condition  $n = \Omega(d \log(p/d))$
  - ▶ this bound independently derived by different approach by Bresler et al. (2008)

# Two straightforward ensembles

- 1 **Naive bulk ensemble:** All graphs on  $p$  vertices with max. degree  $d$  (i.e.,  $\mathcal{G} = \mathcal{G}_{p,d}$ )
  - ▶ simple counting argument:  $\log |\mathcal{G}_{p,d}| = \Theta(pd \log(p/d))$
  - ▶ trivial upper bound:  $I(\mathbf{X}_1^n; G) \leq H(\mathbf{X}_1^n) \leq np$ .
  - ▶ substituting into Fano yields necessary condition  $n = \Omega(d \log(p/d))$
  - ▶ this bound independently derived by different approach by Bresler et al. (2008)
  
- 2 **Small weight effect:** Ensemble  $\mathcal{G}$  consisting of graphs with a single edge with weight  $\theta = \theta_{\min}$

# Two straightforward ensembles

- ① **Naive bulk ensemble:** All graphs on  $p$  vertices with max. degree  $d$  (i.e.,  $\mathcal{G} = \mathcal{G}_{p,d}$ )
- ▶ simple counting argument:  $\log |\mathcal{G}_{p,d}| = \Theta(pd \log(p/d))$
  - ▶ trivial upper bound:  $I(\mathbf{X}_1^n; G) \leq H(\mathbf{X}_1^n) \leq np$ .
  - ▶ substituting into Fano yields necessary condition  $n = \Omega(d \log(p/d))$
  - ▶ this bound independently derived by different approach by Bresler et al. (2008)

- ② **Small weight effect:** Ensemble  $\mathcal{G}$  consisting of graphs with a single edge with weight  $\theta = \theta_{\min}$
- ▶ simple counting:  $\log |\mathcal{G}| = \log \binom{p}{2}$
  - ▶ upper bound on mutual information:

$$I(\mathbf{X}_1^n; G) \leq \frac{1}{\binom{p}{2}} \sum_{(i,j),(k,\ell) \in E} D(\theta(G^{ij}) \parallel \theta(G^{k\ell})).$$

- ▶ upper bound on symmetrized Kullback-Leibler divergences:

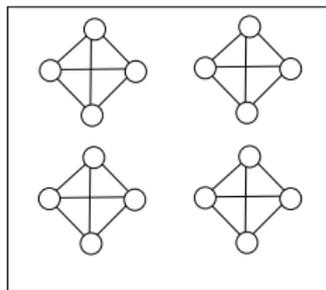
$$D(\theta(G^{ij}) \parallel \theta(G^{k\ell})) + D(\theta(G^{k\ell}) \parallel \theta(G^{ij})) \leq 2\theta_{\min} \tanh(\theta_{\min}/2)$$

- ▶ substituting into Fano yields necessary condition  $n = \Omega\left(\frac{\log p}{\theta_{\min} \tanh(\theta_{\min}/2)}\right)$

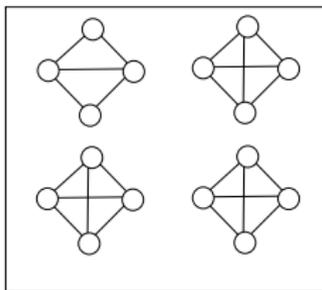
# A harder $d$ -clique ensemble

Constructive procedure:

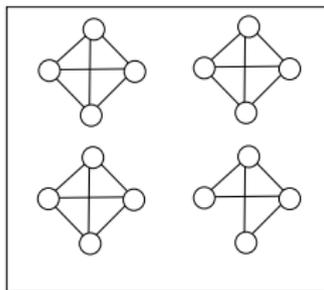
- 1 Divide the vertex set  $V$  into  $\lfloor \frac{p}{d+1} \rfloor$  groups of size  $d+1$ .
- 2 Form the base graph  $\bar{G}$  by making a  $(d+1)$ -clique within each group.
- 3 Form graph  $G^{uv}$  by deleting edge  $(u, v)$  from  $\bar{G}$ .
- 4 Form Markov random field  $\mathbb{P}_{\theta(G^{uv})}$  by setting  $\theta_{st} = \theta_{\min}$  for all edges.



(a) Base graph  $\bar{G}$



(b) Graph  $G^{uv}$



(c) Graph  $G^{st}$

- For  $d \leq p/4$ , we can form

$$|\mathcal{G}| \geq \lfloor \frac{p}{d+1} \rfloor \binom{d+1}{2} = \Omega(dp)$$

such graphs.

## A key separation lemma

**Strategy:** Upper bound the mutual information by controlling the *symmetrized Kullback-Leibler divergence*:

$$S(\theta(G^{st})\|\theta(G^{uv})) = D(\theta(G^{st})\|\theta(G^{uv})) + D(\theta(G^{uv})\|\theta(G^{st}))$$

## A key separation lemma

**Strategy:** Upper bound the mutual information by controlling the *symmetrized Kullback-Leibler divergence*:

$$S(\theta(G^{st})\|\theta(G^{uv})) = D(\theta(G^{st})\|\theta(G^{uv})) + D(\theta(G^{uv})\|\theta(G^{st}))$$

### Lemma

For the given ensemble, the symmetrized KL divergence is upper bounded as

$$S(\theta(G^{st})\|\theta(G^{uv})) \leq \frac{8d\theta_{\min} \exp(3\theta_{\min}/2)}{\exp(d\theta_{\min}/2)}$$

## A key separation lemma

**Strategy:** Upper bound the mutual information by controlling the *symmetrized Kullback-Leibler divergence*:

$$S(\theta(G^{st})\|\theta(G^{uv})) = D(\theta(G^{st})\|\theta(G^{uv})) + D(\theta(G^{uv})\|\theta(G^{st}))$$

### Lemma

For the given ensemble, the symmetrized KL divergence is upper bounded as

$$S(\theta(G^{st})\|\theta(G^{uv})) \leq \frac{8d\theta_{\min} \exp(3\theta_{\min}/2)}{\exp(d\theta_{\min}/2)}$$

**Key consequences:**

- complexity controls exponentially in **maximum neighborhood weight**

$$\omega(\theta^*) := \max_{s \in V} \sum_{t \in N(s)} |\theta_{st}|.$$

## A key separation lemma

**Strategy:** Upper bound the mutual information by controlling the *symmetrized Kullback-Leibler divergence*:

$$S(\theta(G^{st})\|\theta(G^{uv})) = D(\theta(G^{st})\|\theta(G^{uv})) + D(\theta(G^{uv})\|\theta(G^{st}))$$

### Lemma

For the given ensemble, the symmetrized KL divergence is upper bounded as

$$S(\theta(G^{st})\|\theta(G^{uv})) \leq \frac{8d\theta_{\min} \exp(3\theta_{\min}/2)}{\exp(d\theta_{\min}/2)}$$

**Key consequences:**

- complexity controls exponentially in **maximum neighborhood weight**

$$\omega(\theta^*) := \max_{s \in V} \sum_{t \in N(s)} |\theta_{st}|.$$

- combining with Fano's inequality yields the necessary condition

$$n > \frac{\exp(\frac{\omega(\theta)}{4}) d\theta_{\min} \log(pd/8)}{128 \exp(\frac{3\theta_{\min}}{2})}$$

# Sufficient conditions for $G_{d,p}$

- $G \in \mathcal{G}_{d,p}$ : graphs with  $p$  nodes and max. degree  $d$
- Ising models with:
  - ▶ *Minimum edge weight*:  $|\theta_{st}^*| \geq \theta_{\min}$  for all edges
  - ▶ *Maximum neighborhood weight*:  $\omega(\theta) := \max_{s \in V} \sum_{t \in N(s)} |\theta_{st}^*|$

# Sufficient conditions for $G_{d,p}$

- $G \in \mathcal{G}_{d,p}$ : graphs with  $p$  nodes and max. degree  $d$
- Ising models with:
  - ▶ *Minimum edge weight*:  $|\theta_{st}^*| \geq \theta_{\min}$  for all edges
  - ▶ *Maximum neighborhood weight*:  $\omega(\theta) := \max_{s \in V} \sum_{t \in N(s)} |\theta_{st}^*|$

## Theorem

There is an (exponential-time) method that succeeds if

$$n > \max \left\{ d \log p, \frac{6 \exp(2\omega(\theta))}{\sinh^2\left(\frac{|\theta|}{2}\right)} d \log p, \frac{8 \log p}{\theta_{\min}^2} \right\}.$$

# Sufficient conditions for $G_{d,p}$

- $G \in \mathcal{G}_{d,p}$ : graphs with  $p$  nodes and max. degree  $d$
- Ising models with:
  - ▶ *Minimum edge weight*:  $|\theta_{st}^*| \geq \theta_{\min}$  for all edges
  - ▶ *Maximum neighborhood weight*:  $\omega(\theta) := \max_{s \in V} \sum_{t \in N(s)} |\theta_{st}^*|$

## Theorem

There is an (exponential-time) method that succeeds if

$$n > \max \left\{ d \log p, \frac{6 \exp(2\omega(\theta))}{\sinh^2(\frac{|\theta|}{2})} d \log p, \frac{8 \log p}{\theta_{\min}^2} \right\}.$$

## Comments:

- to avoid exponential penalty via **maximum neighborhood term**, require that  $\theta_{\min} = \mathcal{O}(1/d)$
- leads to simplified lower bound  $n = \Omega\left(\max\left\{\frac{\log p}{\theta_{\min}^2}, d^3 \log p\right\}\right)$

# Summary and open questions

- *Practical method:*  $\ell_1$ -regularized regression succeeds with sample size

$$n > c_1 \max\left\{\frac{d}{\theta_{\min}^2}, d^3\right\} \log p.$$

# Summary and open questions

- *Practical method:*  $\ell_1$ -regularized regression succeeds with sample size

$$n > c_1 \max\left\{\frac{d}{\theta_{\min}^2}, d^3\right\} \log p.$$

- *Fundamental limit:* any algorithm fails for sample size

$$n < c_2 \max\left\{\frac{1}{\theta_{\min}^2}, d^2\right\} \log p$$

- various open questions:

- ▶ determine exact capacity of problem (including  $d^2$  versus  $d^3$  and control of constants)

# Summary and open questions

- *Practical method*:  $\ell_1$ -regularized regression succeeds with sample size

$$n > c_1 \max\left\{\frac{d}{\theta_{\min}^2}, d^3\right\} \log p.$$

- *Fundamental limit*: any algorithm fails for sample size

$$n < c_2 \max\left\{\frac{1}{\theta_{\min}^2}, d^2\right\} \log p$$

- various open questions:

- ▶ determine exact capacity of problem (including  $d^2$  versus  $d^3$  and control of constants)
- ▶ some extensions....
  - ★ non-binary MRFs via block-structured regularization schemes
  - ★ other performance metrics (e.g.  $(1 - \delta)$  edges correct)
- ▶ broader issue: optimal trade-offs between statistical/computational efficiency?

## Some papers on graph selection

- Ravikumar, P., Wainwright, M. J. and Lafferty, J. (2008). High-dimensional Ising model selection using  $\ell_1$ -regularized logistic regression. Appeared at NIPS Conference (2006); To appear in *Annals of Statistics*.
- Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. High-dimensional covariance estimation: Convergence rates of  $\ell_1$ -regularized log-determinant divergence. Appeared at *NIPS Conference 2008*.
- Santhanam, P. and Wainwright, M. J. (2008). Information-theoretic limitations of high-dimensional graphical model selection. Presented at *International Symposium on Information Theory*, 2008.
- Wainwright, M. J. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using  $\ell_1$ -constrained quadratic programming. *IEEE Trans. on Information Theory*, May 2009.