Introduction to graphical models: Lecture III

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

Introduction

- Markov random fields (undirected graphical models): central in many application areas of science/engineering:

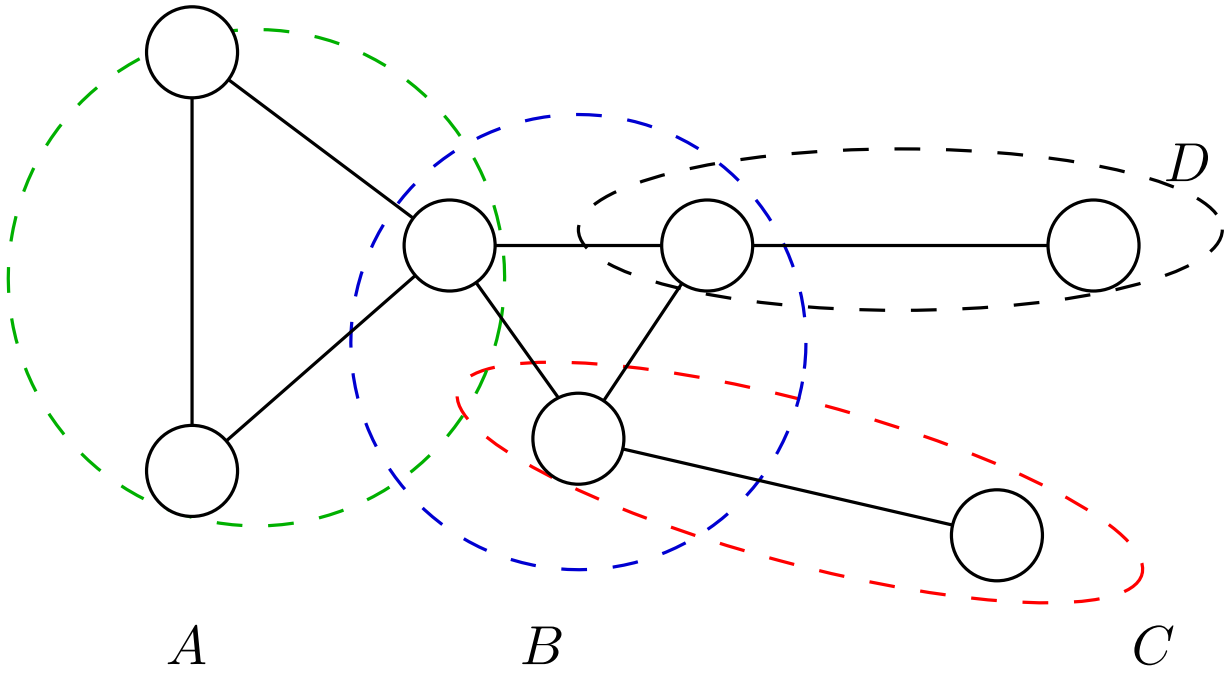
Introduction

- Markov random fields (undirected graphical models): central in many application areas of science/engineering:

- some fundamental problems
  - *counting/integrating:* computing marginal distributions and partition functions
  - *optimization:* computing most probable configurations (or top M-configurations)
  - *model selection:* fitting and selecting models on the basis of data

Graph structure and factorization

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

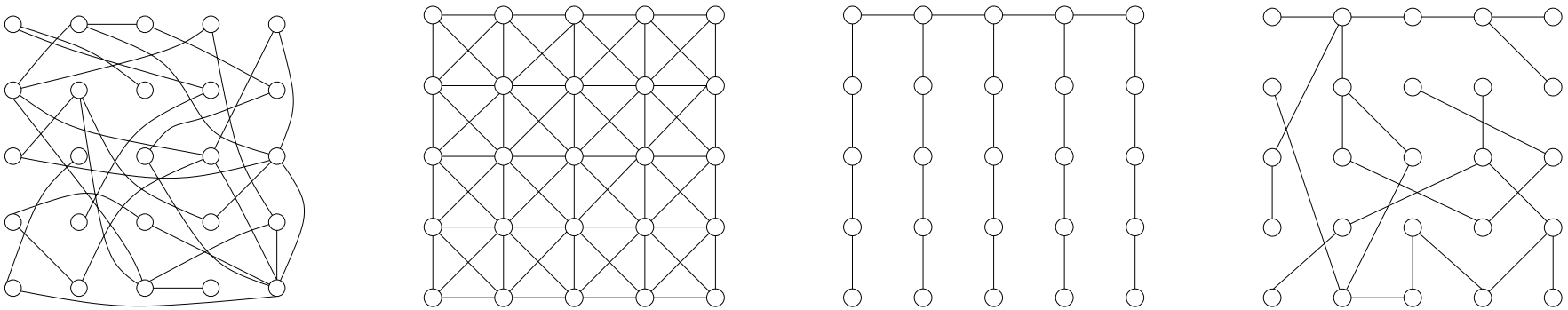


- Hammersley-Clifford theorem: factorization over cliques

$$\mathbb{Q}(x_1, \ldots, x_p; \theta) = \frac{1}{Z(\theta)} \exp\Big\{ \sum_{C \in \mathcal{C}} \theta_C(x_C) \Big\}$$

Graphical model selection

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



- pairwise graphical model factorizes over edges of graph:

$$\mathbb{Q}(x_1, \ldots, x_p; \theta) \quad \propto \quad \exp\Big\{ \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \Big\}.$$

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

Various classes of methods

1. Exact solutions
   - Chow-Liu algorithm for trees (Chow & Liu, 1967)
   - computationally intractable for hypertrees (Srebro & Karger, 2001)

Various classes of methods

1. Exact solutions
   - Chow-Liu algorithm for trees (Chow & Liu, 1967)
   - computationally intractable for hypertrees (Srebro & Karger, 2001)

2. Testing-based approaches
   - PC algorithm (Spirtes et al., 2000; Kalisch & Bühlmann, 2008)
   - thresholding (Bresler et al., 2008; Anandkumar et al., 2010)

Various classes of methods

1. Exact solutions
   - Chow-Liu algorithm for trees (Chow & Liu, 1967)
   - computationally intractable for hypertrees (Srebro & Karger, 2001)

2. Testing-based approaches
   - PC algorithm (Spirtes et al., 2000; Kalisch & Bühlmann, 2008)
   - thresholding (Bresler et al., 2008; Anandkumar et al., 2010)

3. Penalized forms of global likelihood
   - combinatorial penalties (AIC, BIC, GIC etc.)
   - ℓ_1 and related penalties
     - ⋆ classical analysis of penalized Gaussian MLE: Yuan & Lin, 2006
     - ⋆ some fast algorithms: d'Asprémont et al., 2007; Friedman et al, 2008

Various classes of methods

1. Exact solutions
   - Chow-Liu algorithm for trees (Chow & Liu, 1967)
   - computationally intractable for hypertrees (Srebro & Karger, 2001)

2. Testing-based approaches
   - PC algorithm (Spirtes et al., 2000; Kalisch & Bühlmann, 2008)
   - thresholding (Bresler et al., 2008; Anandkumar et al., 2010)

3. Penalized forms of global likelihood
   - combinatorial penalties (AIC, BIC, GIC etc.)
   - ℓ_1 and related penalties
     - ⋆ classical analysis of penalized Gaussian MLE: Yuan & Lin, 2006
     - ⋆ some fast algorithms: d'Asprémont et al., 2007; Friedman et al, 2008

4. Pseudolikelihoods and neighborhood regression
   - pseudolikeliood consistency for Gaussians (Besag, 1977)
   - pseudolikelihood and BIC criterion (Csiszar & Talata, 2006)
   - neighborhood regression for Gaussian MRFs
     (e.g., Meinshausen & Buhlmann, 2005; Wainwright, 2006, Zhao & Yu 2006)
   - logistic regression for Ising models (Ravikumar et al., 2010)

§1. Global maximum likelihood

- given i.i.d. samples $\mathbf{X}_1^n := \{(X_{1\bullet}, \ldots, X_{n\bullet}\}$, might consider methods based on global likelihood $\ell(\theta; \mathbf{X}_1^n) := \frac{1}{n} \sum_{i=1}^n \log \mathbb{Q}(X_{i\bullet}; \theta)$

§1. Global maximum likelihood

- given i.i.d. samples $\mathbf{X}_1^n := \{(X_{1\bullet}, \ldots, X_{n\bullet}\}$, might consider methods based on global likelihood $\ell(\theta; \mathbf{X}_1^n) := \frac{1}{n} \sum_{i=1}^n \log \mathbb{Q}(X_{i\bullet}; \theta)$

- maximum likelihood for graphical model in exponential form

$$\widehat{\theta} \quad = \quad \arg\max_{\theta} \left\{ \sum_{(s,t) \in E} \underbrace{\widehat{\mathbb{E}}[\theta(X_s, X_t)]}_{\text{empirical moments}} - \log Z(\theta) \right\}$$

§1. Global maximum likelihood

- given i.i.d. samples $\mathbf{X}_1^n := \{(X_{1\bullet}, \ldots, X_{n\bullet}\}$, might consider methods based on global likelihood $\ell(\theta; \mathbf{X}_1^n) := \frac{1}{n} \sum_{i=1}^n \log \mathbb{Q}(X_{i\bullet}; \theta)$

- maximum likelihood for graphical model in exponential form

$$\widehat{\theta} \quad = \quad \arg\max_{\theta} \left\{ \sum_{(s,t) \in E} \underbrace{\widehat{\mathbb{E}}[\theta(X_s, X_t)]}_{\text{empirical moments}} - \log Z(\theta) \right\}$$

- exact likelihood involves log partition function $\log Z(\theta)$:
  - can be computed for Gaussian MRFs (log-determinant)
  - intractable for Ising models (binary pairwise MRFs) (Welsh, 1993)

§1. Global maximum likelihood

- given i.i.d. samples $\mathbf{X}_1^n := \{(X_{1\bullet}, \ldots, X_{n\bullet}\}$, might consider methods based on global likelihood $\ell(\theta; \mathbf{X}_1^n) := \frac{1}{n} \sum_{i=1}^n \log \mathbb{Q}(X_{i\bullet}; \theta)$

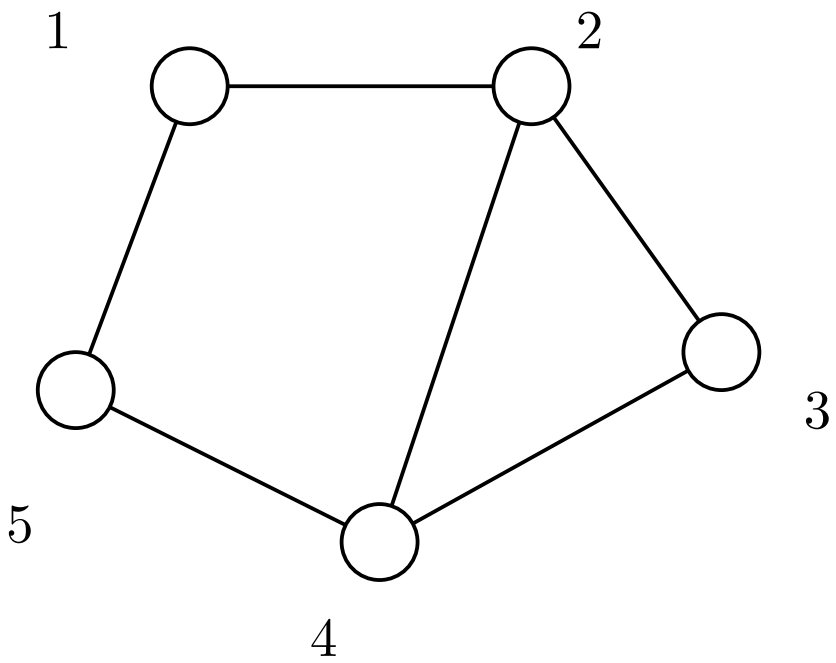
- maximum likelihood for graphical model in exponential form

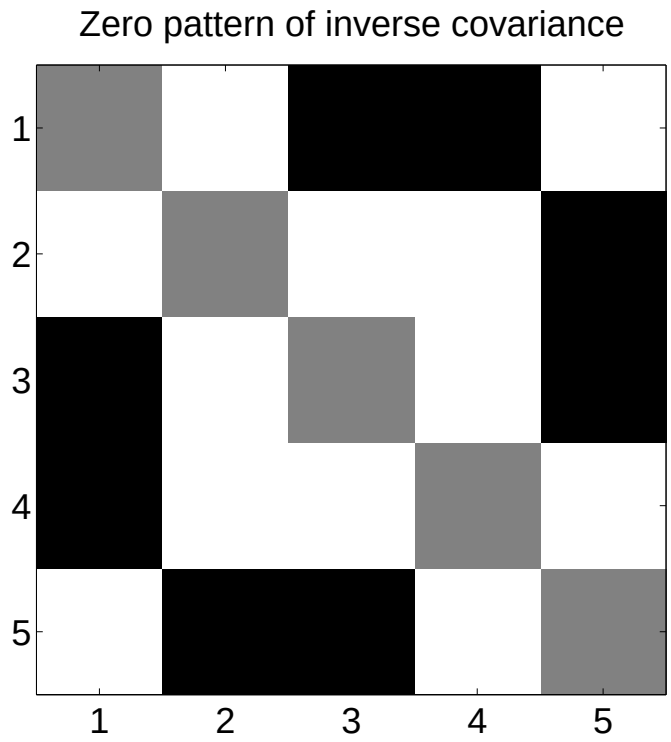
$$\widehat{\theta} \quad = \quad \arg\max_{\theta} \left\{ \sum_{(s,t) \in E} \underbrace{\widehat{\mathbb{E}}[\theta(X_s, X_t)]}_{\text{empirical moments}} - \log Z(\theta) \right\}$$

- exact likelihood involves log partition function $\log Z(\theta)$:
  - can be computed for Gaussian MRFs (log-determinant)
  - intractable for Ising models (binary pairwise MRFs) (Welsh, 1993)

- possible solutions:
  - MCMC methods
  - stochastic approximation methods
  - variational approximations (mean field, Bethe and belief propagation)

Gaussian graphs $\equiv$ Sparse inverse covariances





- Gaussian graphical model specified by *sparse inverse covariance* Θ:

$$\mathbb{Q}(x_1, \ldots, x_p; \Theta) \quad = \quad \frac{\det(\Theta)}{(2\pi)^{p/2}} \exp\Big(-\frac{1}{2} x^T \Theta x \Big).$$

Gaussian ℓ_1-penalized MLE

Estimator: ℓ_1-regularized log-determinant program:

$$\widehat{\Theta} \quad = \quad \arg\min_{\Theta \succ 0} \Big\{ \underbrace{-\log\det\Theta + \langle\langle \widehat{\Sigma}^n, \, \Theta \rangle\rangle}_{\text{Gaussian log likelihood}} \quad + \quad \underbrace{\lambda_n \sum_{i \neq j} |\Theta_{ij}|}_{\text{Regularization}} \Big\} .$$

Gaussian ℓ_1-penalized MLE

Estimator: ℓ_1-regularized log-determinant program:

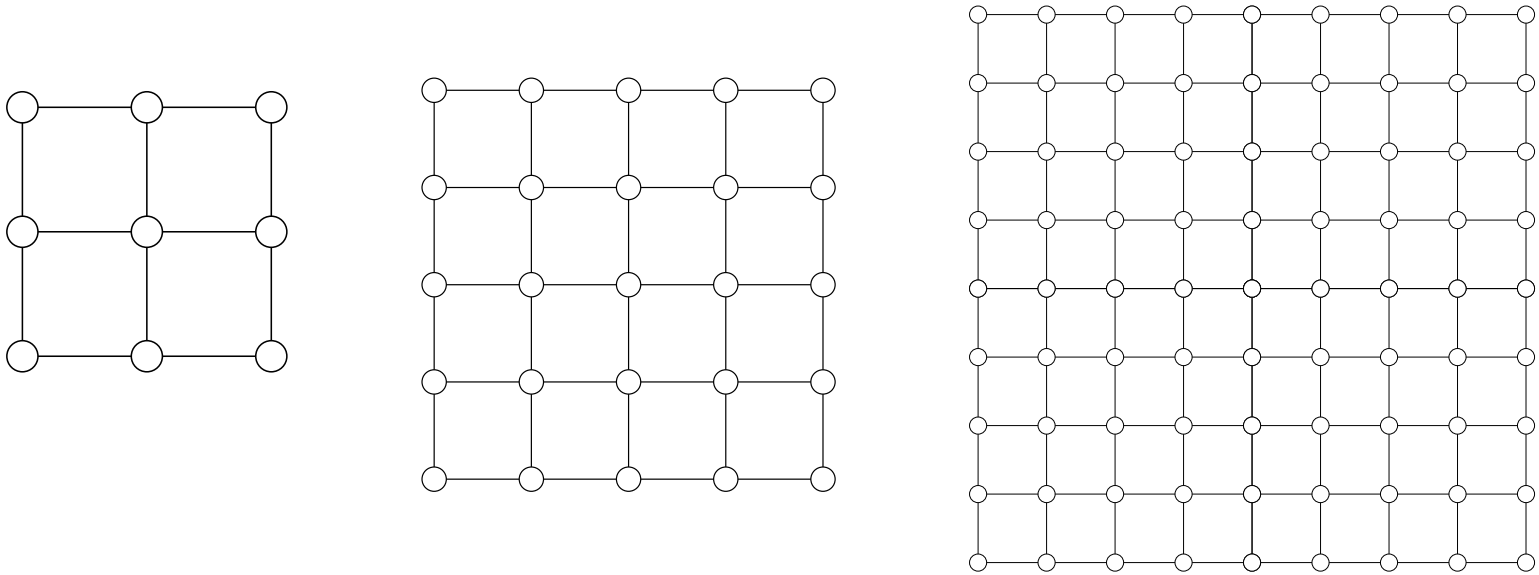
$$\widehat{\Theta} \quad = \quad \arg\min_{\Theta \succ 0} \Big\{ \underbrace{-\log\det\Theta + \langle\langle \widehat{\Sigma}^n, \, \Theta \rangle\rangle}_{\text{Gaussian log likelihood}} \quad + \quad \underbrace{\lambda_n \sum_{i \neq j} |\Theta_{ij}|}_{\text{Regularization}} \Big\} .$$

Results on this method:

- analysis under classical scaling ($n \to \infty$ with p fixed) (Yuan & Lin, 2006)
- some fast algorithms (d'Asprémont et al., 2007; Friedman et al, 2008)
- high-dimensional analysis of Frobenius norm error (Rothman et al., 2008)
- high-dimensional variable selection and ℓ_∞ bounds (Ravikumar et al., 2011)

High-dimensional analysis

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

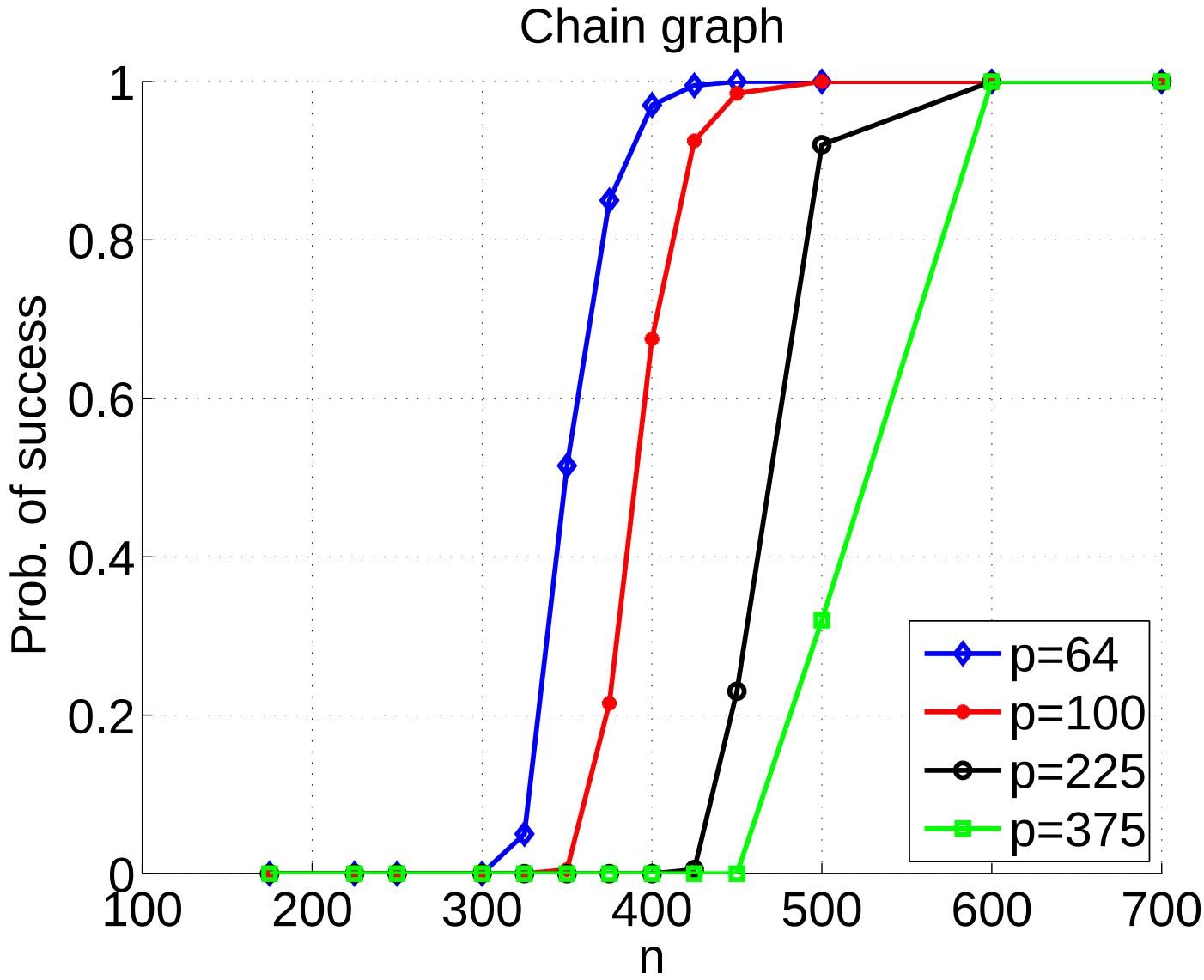


- take n i.i.d. samples from MRF defined by $G_{p,d}$
- study probability of success as a function of three parameters:

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

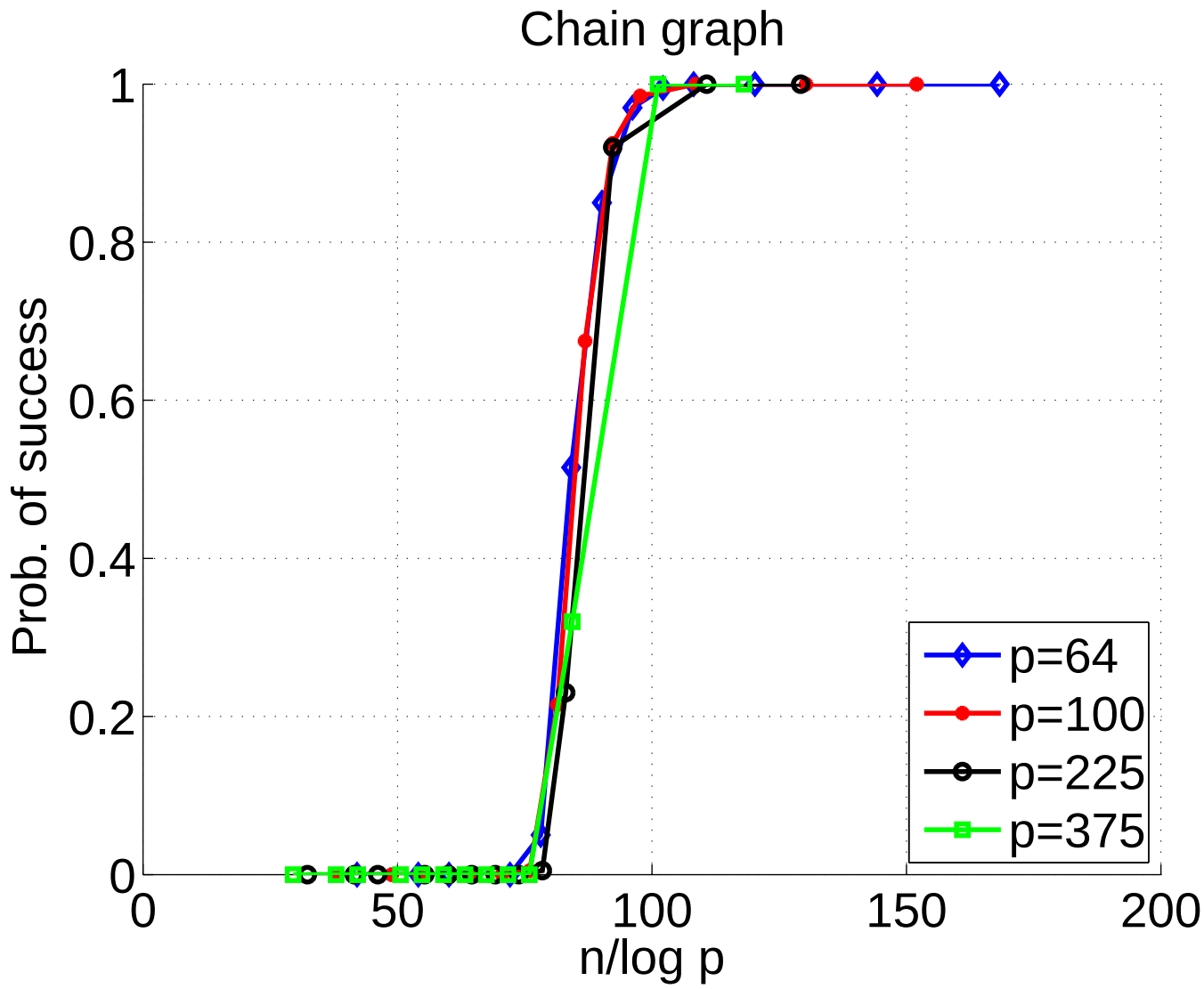
- theory is non-asymptotic: explicit probabilities for finite (n, p, d)

Empirical behavior: Unrescaled plots



Plots of success probability versus raw sample size n.

Empirical behavior: Appropriately rescaled



Plots of success probability versus rescaled sample size

Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d.
- suitable regularity conditions on Hessian of log-determinant $\Gamma^* := (\Theta^*)^{-1} \otimes (\Theta^*)^{-1}$

Theorem: For multivariate Gaussian and sample size

$$n \quad > \quad c_1 \tau \, d^2 \log p$$

and regularization parameter $\lambda_n \geq c_2 \, \tau \sqrt{\frac{\log p}{n}}$, then with probability greater than $1 - 2\exp\big(-c_3(\tau - 2)\log p\big)$:

Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d.
- suitable regularity conditions on Hessian of log-determinant $\Gamma^* := (\Theta^*)^{-1} \otimes (\Theta^*)^{-1}$

Theorem: For multivariate Gaussian and sample size

$$n \quad > \quad c_1 \tau \, d^2 \log p$$

and regularization parameter $\lambda_n \geq c_2 \, \tau \sqrt{\frac{\log p}{n}}$, then with probability greater than $1 - 2\exp\big(-c_3(\tau - 2)\log p\big)$:

(a) *No false inclusions:* The regularized log-determinant estimate $\widehat{\Theta}$ returns an edge set $\widehat{E} \subseteq E^*$.

Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d.
- suitable regularity conditions on Hessian of log-determinant $\Gamma^* := (\Theta^*)^{-1} \otimes (\Theta^*)^{-1}$

Theorem: For multivariate Gaussian and sample size

$$n \quad > \quad c_1 \tau \, d^2 \log p$$

and regularization parameter $\lambda_n \geq c_2 \, \tau \sqrt{\frac{\log p}{n}}$, then with probability greater than $1 - 2\exp\big(-c_3(\tau - 2)\log p\big)$:

(a) *No false inclusions:* The regularized log-determinant estimate $\widehat{\Theta}$ returns an edge set $\widehat{E} \subseteq E^*$.

(b) ℓ_∞*-control:* Estimate satisfies $\max_{i,j} |\widehat{\Theta}_{ij} - \Theta^*_{ij}| \leq 2\, c_4 \sqrt{\frac{\tau \log p}{n}}$.

Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d.
- suitable regularity conditions on Hessian of log-determinant $\Gamma^* := (\Theta^*)^{-1} \otimes (\Theta^*)^{-1}$

Theorem: For multivariate Gaussian and sample size

$$n \quad > \quad c_1 \tau \, d^2 \log p$$

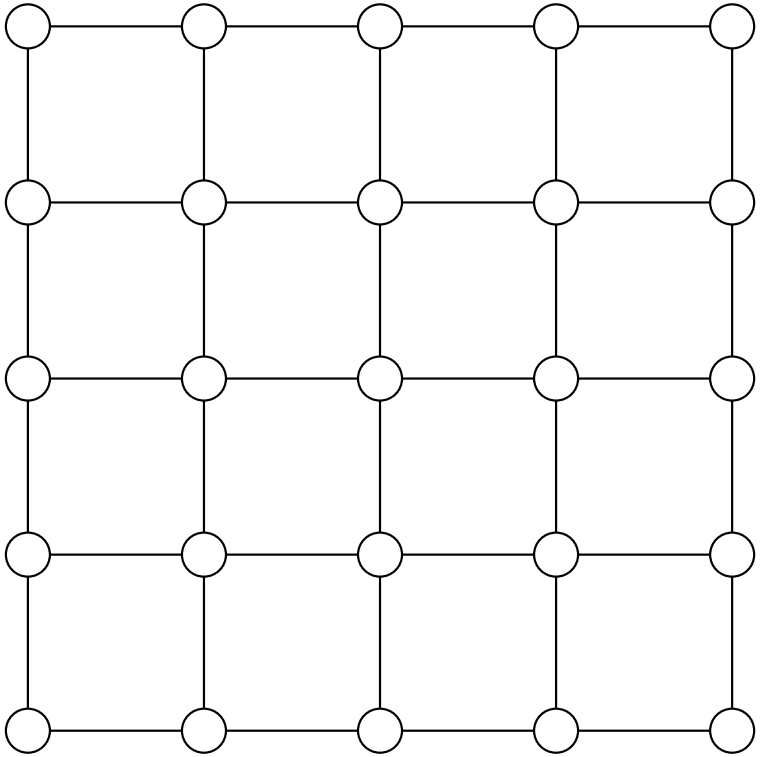
and regularization parameter $\lambda_n \geq c_2 \, \tau \sqrt{\frac{\log p}{n}}$, then with probability greater than $1 - 2\exp\big(-c_3(\tau - 2)\log p\big)$:

(a) *No false inclusions:* The regularized log-determinant estimate $\widehat{\Theta}$ returns an edge set $\widehat{E} \subseteq E^*$.

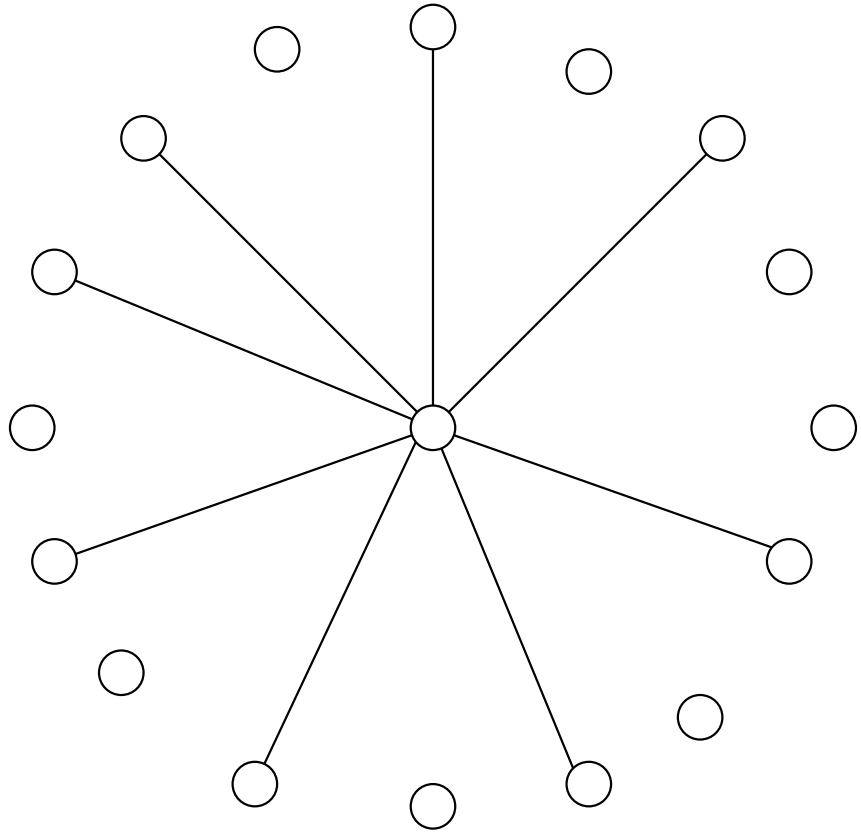
(b) *ℓ_∞-control:* Estimate satisfies $\max_{i,j} |\widehat{\Theta}_{ij} - \Theta^*_{ij}| \leq 2\, c_4 \sqrt{\frac{\tau \log p}{n}}$.

(c) *Model selection consistency:* If $\theta_{\min} \geq c_4 \sqrt{\frac{\tau \log p}{n}}$, then $E = \widehat{E}$.

Some other graphs

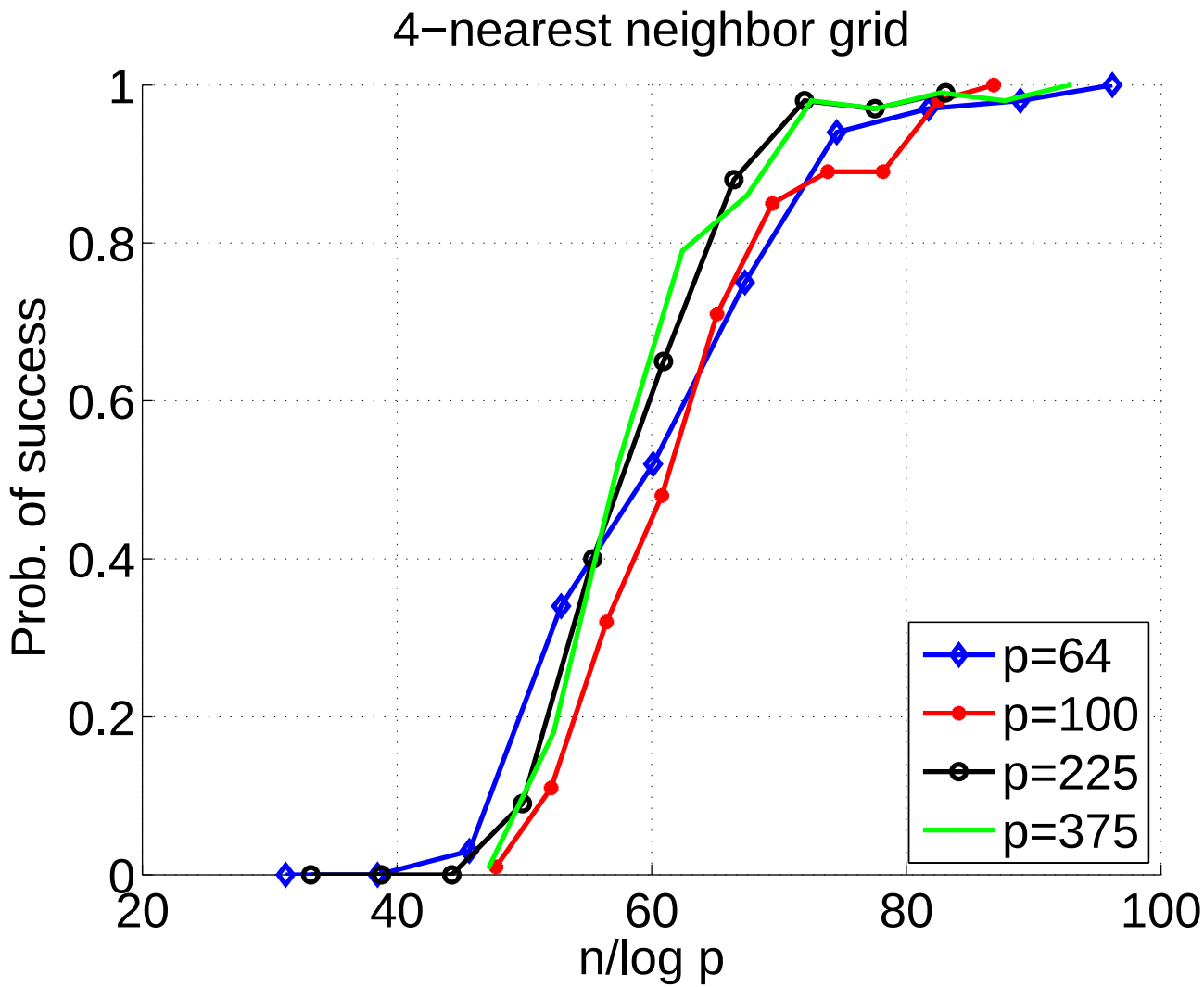


(a) 4-grid
$d = 4$



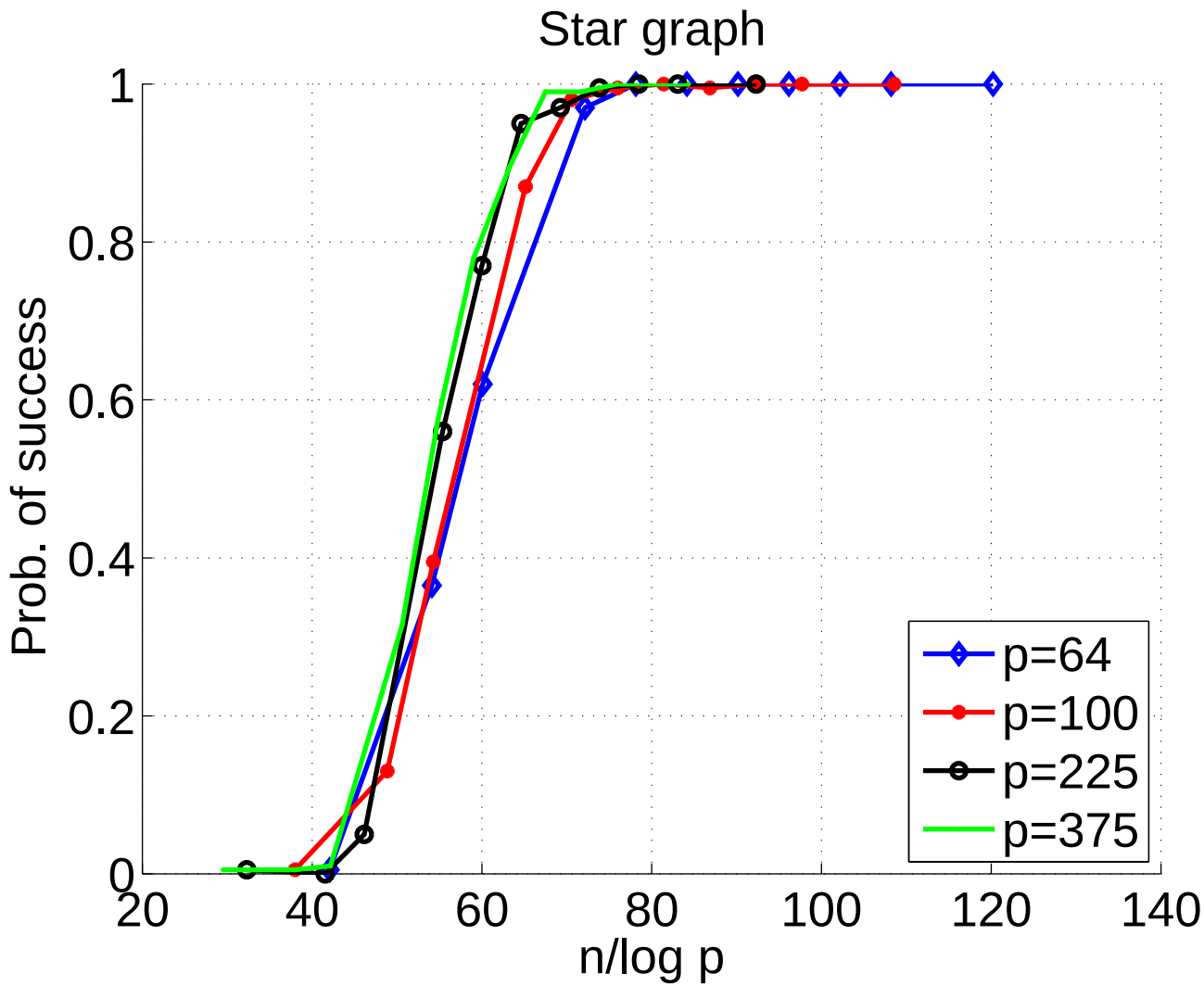
(b) Star
$d \in \{\mathcal{O}(\log p),\ \alpha p\}$

Results for 4-grid graphs



- Vertical axis: success probability $\mathbb{Q}[\widehat{E} = E]$

Results for star graphs



- Vertical axis: success probability $\mathbb{Q}[\widehat{E} = E]$

Proof sketch: Primal-dual certificate

- construct *candidate* primal-dual pair $(\widehat{\theta}, \widehat{z}) \in \mathbb{R}^{p \times p} \times \mathbb{R}^{p \times p}$.
- proof technique—-not a practical algorithm!

(A) Solve the *restricted* log-determinant program

$$\widehat{\theta} = \arg \min_{\Theta \succ 0, \Theta_{S^c} = 0} \Big\{ - \log \det \Theta + \langle\langle \widehat{\Sigma}^n, \Theta \rangle\rangle + \lambda_n \sum_{i \neq j} |\Theta_{ij}| \Big\}$$

thereby obtaining candidate solution $\widehat{\theta} = (\widehat{\theta}_S, 0_{S^c})$.

(B) We choose $\widehat{z}_S \in \mathbb{R}^{|S|}$ as an element of the subdifferential $\partial \|\widehat{\theta}_S\|_1$.

(C) Using optimality conditions from original convex program, solve for $\widehat{z}_{S^c}$ and check whether or not *strict dual feasibility*

$$|\widehat{z}_j| \quad < \quad 1 \qquad \text{for all } j \in S^c \text{ holds.}$$

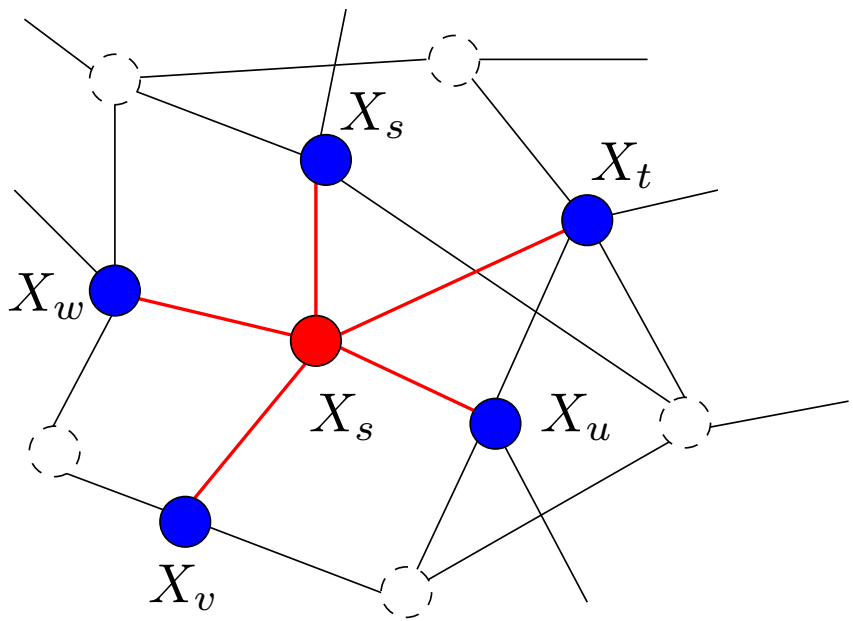
Lemma: Full convex program recovers neighborhood $\Longleftrightarrow$ primal-dual witness succeeds.

§2. Pseudolikelihood and neighborhood approaches

- Markov properties encode neighborhood structure:

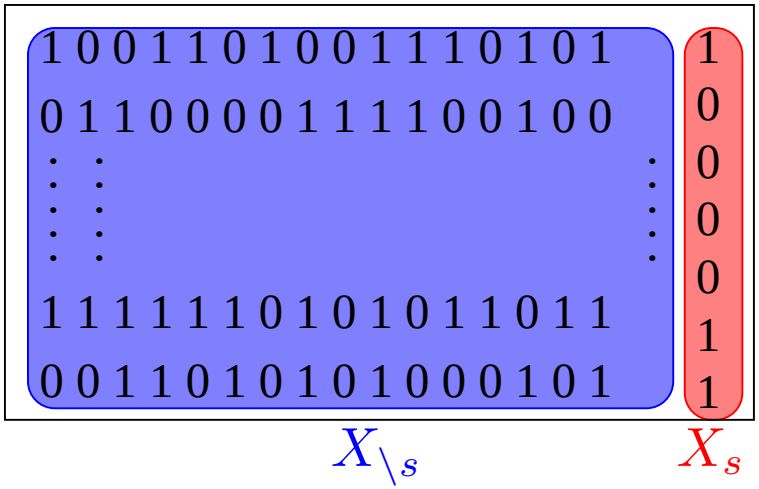
$$\underbrace{(X_s \quad | \quad X_{V\setminus s})}_{\text{Condition on full graph}} \stackrel{d}{=} \underbrace{(X_s \quad | \quad X_{N(s)})}_{\text{Condition on Markov blanket}}$$

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



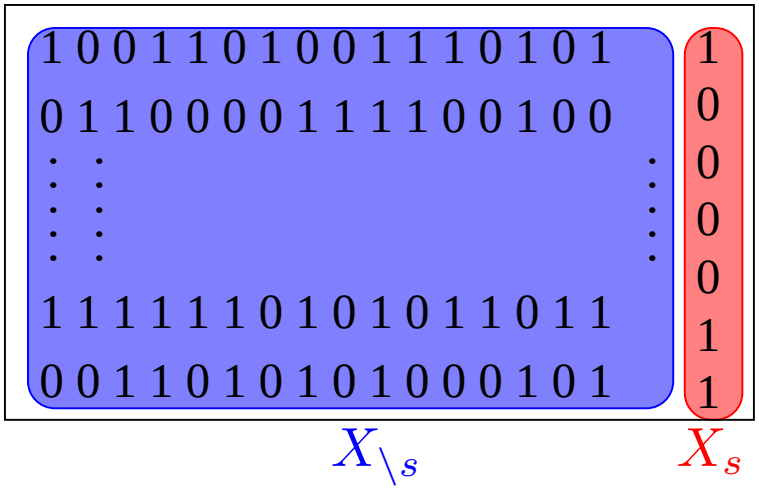
- basis of pseudolikelihood method (Besag, 1974)
- basis of many graph learning algorithm (Friedman et al., 1999; Csiszar & Talata, 2005; Abeel et al., 2006; Meinshausen & Buhlmann, 2006)

Graph selection via neighborhood regression



Predict X_s based on $X_{\setminus s} := \{X_s,\ t \neq s\}$.

Graph selection via neighborhood regression

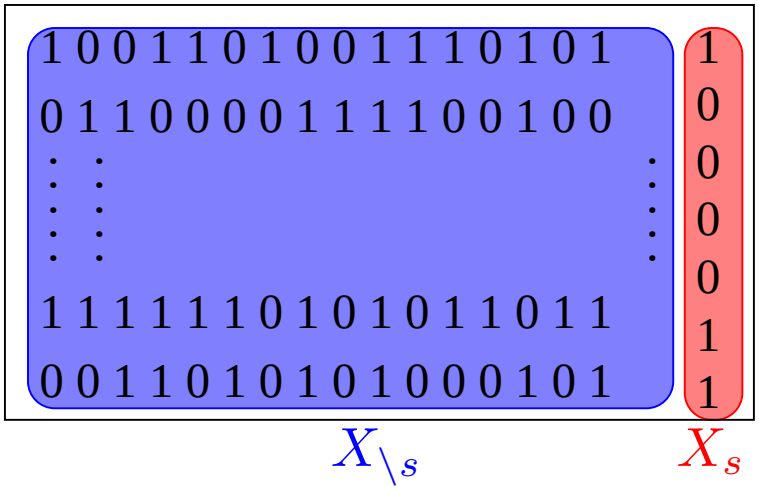


Predict X_s based on $X_{\setminus s} := \{X_s,\ t \neq s\}$.

1. For each node $s \in V$, compute (regularized) max. likelihood estimate:

$$\widehat{\theta}[s] \quad := \quad \arg\min_{\theta \in \mathbb{R}^{p-1}} \left\{ \quad -\frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathcal{L}(\theta; X_{i\,\setminus s})}_{\text{local log. likelihood}} \quad + \quad \lambda_n \underbrace{\|\theta\|_1}_{\text{regularization}} \right\}$$

Graph selection via neighborhood regression



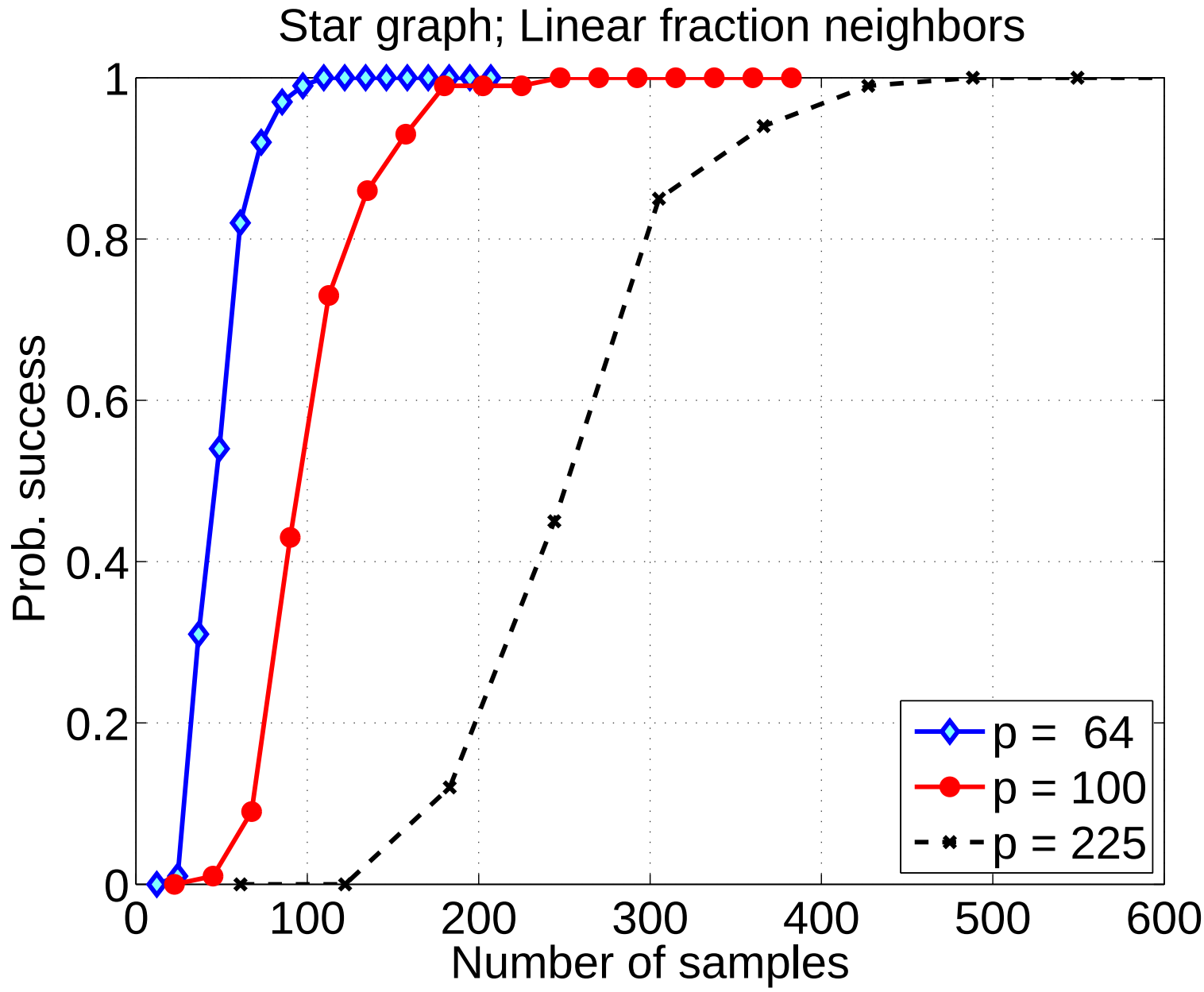
Predict X_s based on $X_{\setminus s} := \{X_s,\ t \neq s\}$.

1. For each node $s \in V$, compute (regularized) max. likelihood estimate:

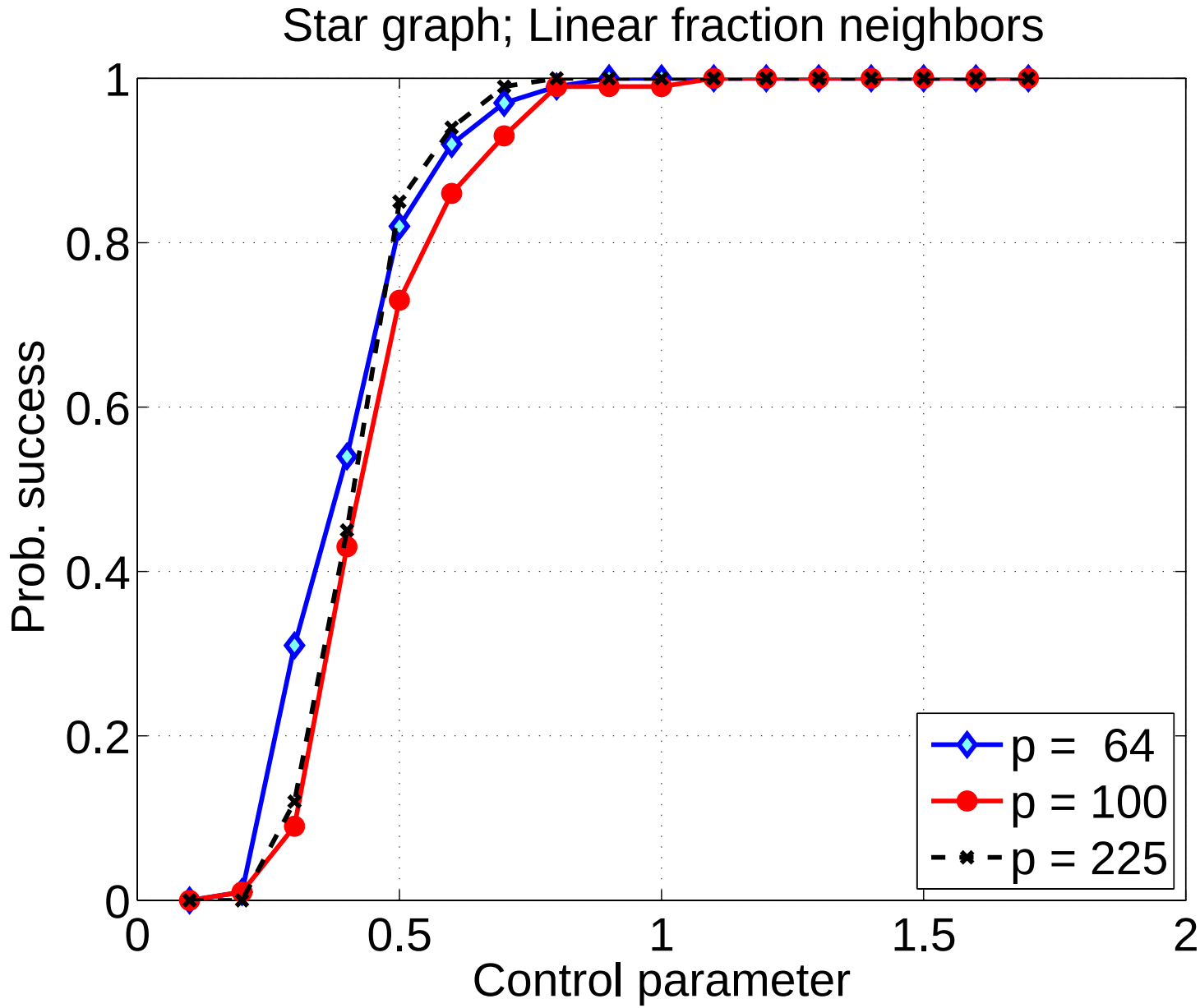
$$\widehat{\theta}[s] \quad := \quad \arg\min_{\theta \in \mathbb{R}^{p-1}} \left\{ \quad -\frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathcal{L}(\theta; X_{i\,\setminus s})}_{\text{local log. likelihood}} \quad + \quad \lambda_n \underbrace{\|\theta\|_1}_{\text{regularization}} \right\}$$

2. Estimate the local neighborhood $\widehat{N}(s)$ as support of regression vector $\widehat{\theta}[s] \in \mathbb{R}^{p-1}$.

Empirical behavior: Unrescaled plots



Empirical behavior: Appropriately rescaled



Plots of success probability versus rescaled sample size

Sufficient conditions for consistent Ising 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 (Ravikumar, W. & Lafferty, 2010)

Sufficient conditions for consistent Ising 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 (Ravikumar, W. & Lafferty, 2010)

Under incoherence conditions, with sample size

$$n \quad > \quad c_1 \, d^3 \log p$$

and regularization parameter $\lambda_n \geq c_2 \sqrt{\frac{\log p}{n}}$, *then with probability greater than* $1 - 2\exp\big(-c_3 \lambda_n^2 n\big)$*:*

(a) Correct exclusion: *The estimated sign neighborhood* $\widehat{N}(s)$ *correctly excludes all edges not in the true neighborhood.*

Sufficient conditions for consistent Ising 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 (Ravikumar, W. & Lafferty, 2010)

Under incoherence conditions, with sample size

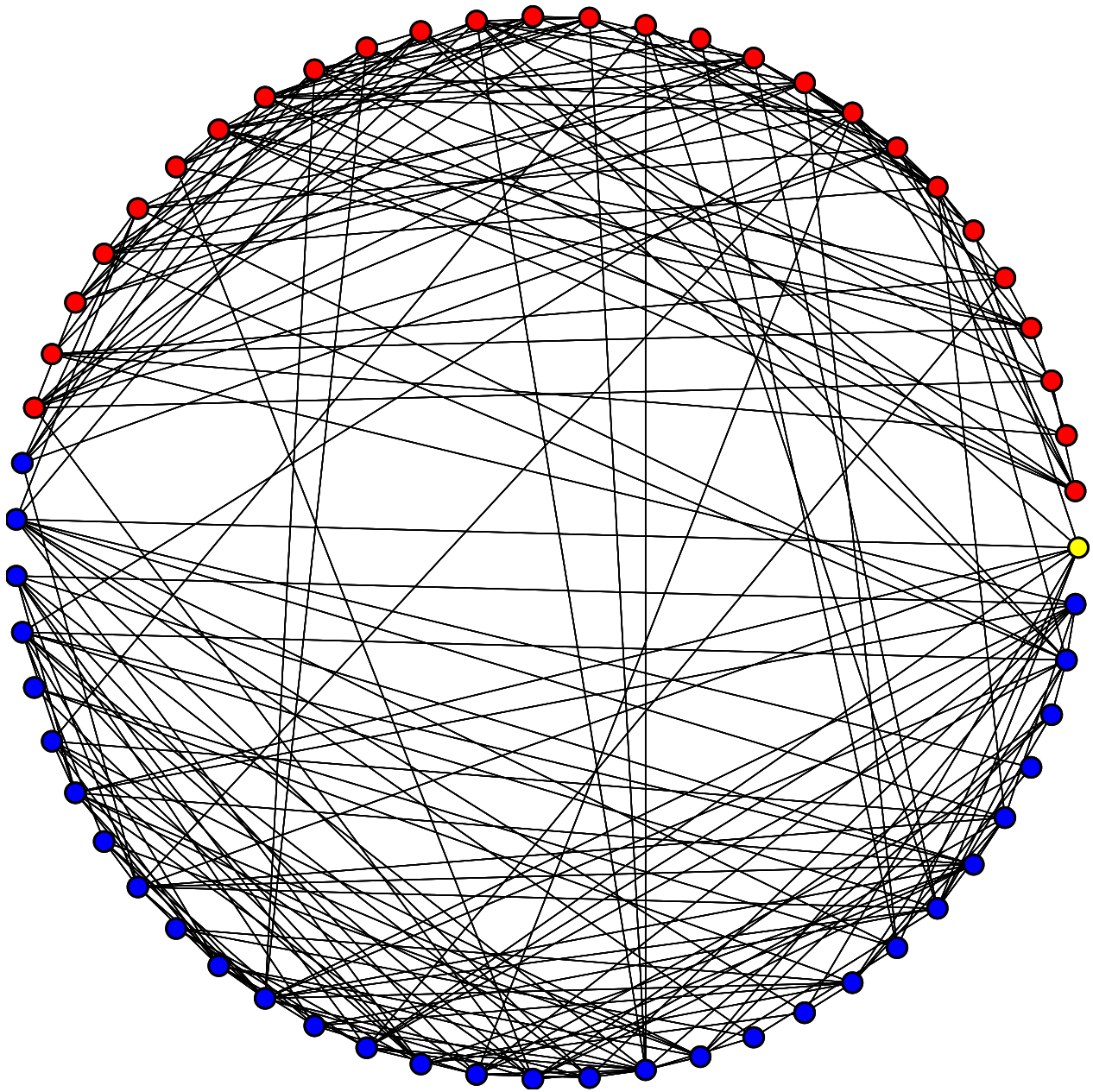
$$n \quad > \quad c_1 \, d^3 \log p$$

and regularization parameter $\lambda_n \geq c_2 \sqrt{\frac{\log p}{n}}$, *then with probability greater than* $1 - 2\exp\big(-c_3 \lambda_n^2 n\big)$*:*

(a) **Correct exclusion:** *The estimated sign neighborhood* $\widehat{N}(s)$ *correctly excludes all edges not in the true neighborhood.*

(b) **Correct inclusion:** *For* $\theta_{\min} \geq c_4 \sqrt{d} \lambda_n$, *the method selects the correct signed neighborhood.*

US Senate network (2004–2006 voting)

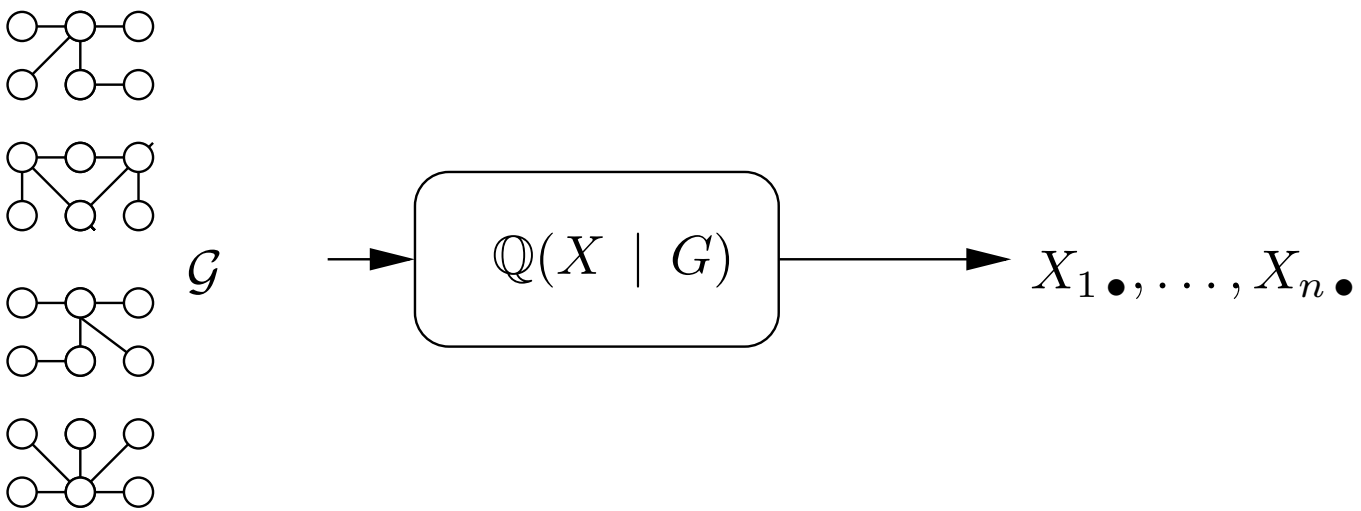


§3. Info. theory: Graph selection as channel coding

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

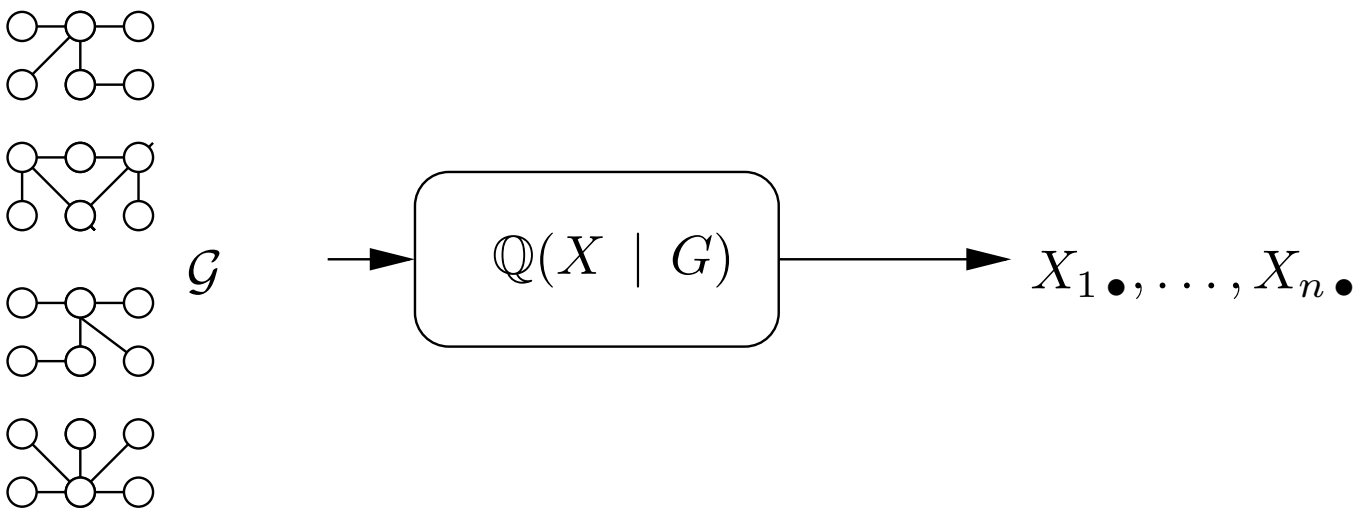
§3. Info. theory: Graph selection as channel coding

- graphical model selection is an *unorthodox* channel coding problem:
  - codewords/codebook: graph G in some graph class $\mathcal{G}$
  - channel use: draw sample $X_{i\bullet} = (X_{i1}, \ldots, X_{ip})$ from Markov random field $\mathbb{Q}_{\theta(G)}$
  - decoding problem: use n samples $\{X_{1\bullet}, \ldots, X_{n\bullet}\}$ 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 $X_{i\bullet} = (X_{i1}, \ldots, X_{ip})$ from Markov random field $\mathbb{Q}_{\theta(G)}$
  - decoding problem: use n samples $\{X_{1\bullet}, \ldots, X_{n\bullet}\}$ 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, 2012)*

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

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, 2012)*

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

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 $\omega(\theta^*) \geq d\,\theta_{\min} \Longrightarrow$ 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 $\omega(\theta^*) \geq d\,\theta_{\min} \Longrightarrow$ require $\theta_{\min} = \mathcal{O}(1/d)$
- from small weight effect

$$n = \Omega\Big(\frac{\log p}{\theta_{\min}\tanh(\theta_{\min})}\Big) \;=\; \Omega\Big(\frac{\log p}{\theta^2_{\min}}\Big)$$

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 $\omega(\theta^*) \geq d\,\theta_{\min} \Longrightarrow$ require $\theta_{\min} = \mathcal{O}(1/d)$
- from small weight effect

$$n = \Omega\Big(\frac{\log p}{\theta_{\min} \tanh(\theta_{\min})}\Big) \;=\; \Omega\Big(\frac{\log p}{\theta^2_{\min}}\Big)$$

- conclude that ℓ_1-regularized logistic regression (LR) is within $\Theta(d)$ of optimal for general graphs