Graphical models and message-passing: Part III: Learning graphs from data

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

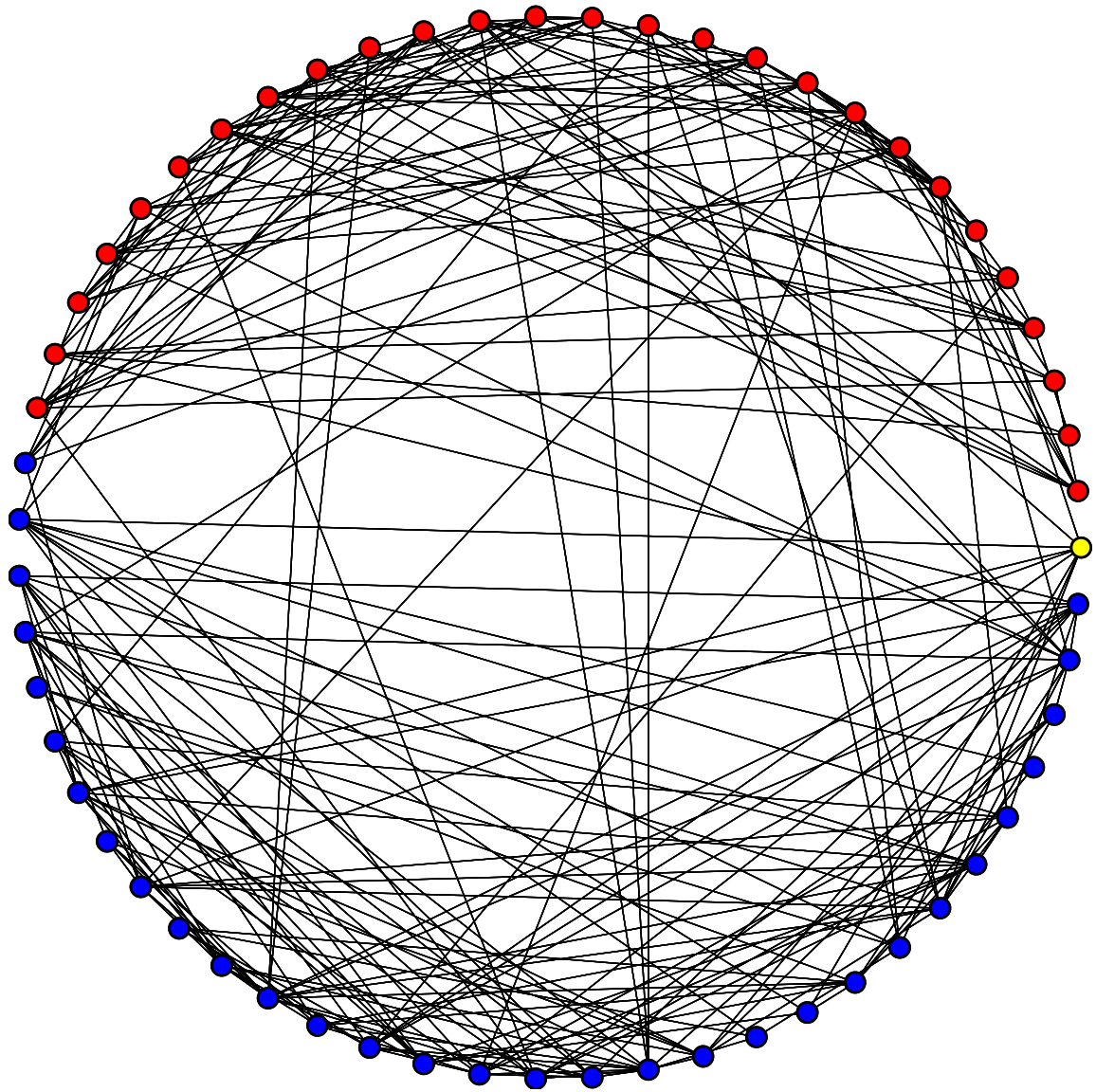
Introduction

- previous lectures on "forward problems": given a graphical model, perform some type of computation
  - Part I: compute most probable (MAP) assignment
  - Part II: compute marginals and likelihoods

- inverse problems concern learning the parameters and structure of graphs from data

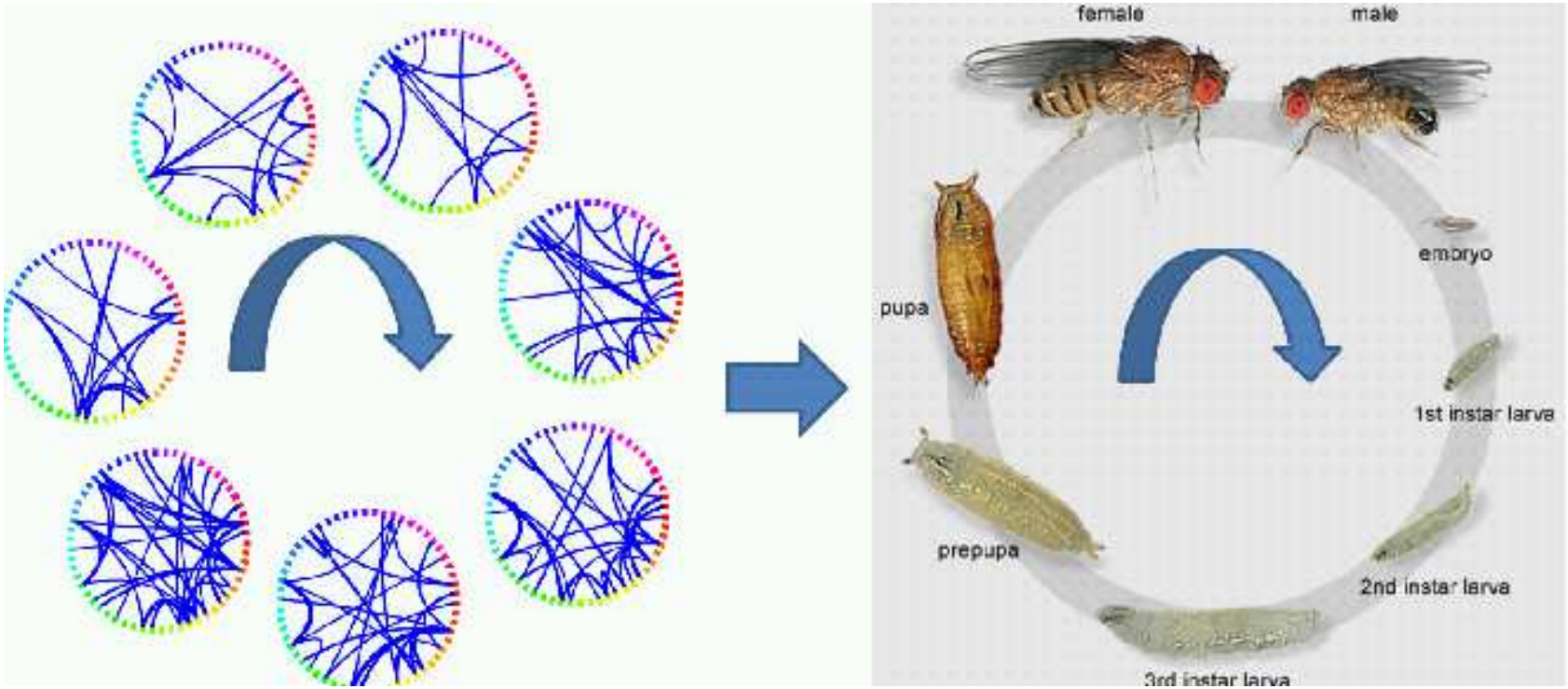
- many instances of such graph learning problems:
  - fitting graphs to politicians' voting behavior
  - modeling diseases with epidemiological networks
  - traffic flow modeling
  - interactions between different genes
  - and so on....

Example: US Senate network (2004–2006 voting)



(Banerjee et al., 2008; Ravikumar, W. & Lafferty, 2010)

Example: Biological networks



- gene networks during Drosophila life cycle (Ahmed & Xing, PNAS, 2009)
- many other examples:
  - protein networks
  - phylogenetic trees

Learning for pairwise models

- drawn n samples from

$$\mathbb{Q}(x_1, \ldots, x_p; \Theta) \quad = \quad \frac{1}{Z(\Theta)} \exp\Big\{ \sum_{s \in V} \theta_s x_s^2 + \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big\}$$

- graph G and matrix $[\Theta]_{st} = \theta_{st}$ of edge weights are unknown

Learning for pairwise models

- drawn n samples from

$$\mathbb{Q}(x_1, \ldots, x_p; \Theta) \quad = \quad \frac{1}{Z(\Theta)} \exp\Big\{ \sum_{s \in V} \theta_s x_s^2 + \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big\}$$

- graph G and matrix $[\Theta]_{st} = \theta_{st}$ of edge weights are unknown

- data matrix:
  - Ising model (binary variables): $\mathbf{X}_1^n \in \{0,1\}^{n \times p}$
  - Gaussian model: $\mathbf{X}_1^n \in \mathbb{R}^{n \times p}$

- estimator $\mathbf{X}_1^n \mapsto \widehat{\Theta}$

Learning for pairwise models

- drawn n samples from

$$\mathbb{Q}(x_1, \ldots, x_p; \Theta) \quad = \quad \frac{1}{Z(\Theta)} \exp\Big\{ \sum_{s \in V} \theta_s x_s^2 + \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big\}$$

- graph G and matrix $[\Theta]_{st} = \theta_{st}$ of edge weights are unknown

- data matrix:
  - Ising model (binary variables): $\mathbf{X}_1^n \in \{0,1\}^{n \times p}$
  - Gaussian model: $\mathbf{X}_1^n \in \mathbb{R}^{n \times p}$

- estimator $\mathbf{X}_1^n \mapsto \widehat{\Theta}$

- various loss functions are possible:
  - graph selection: $\mathrm{supp}[\widehat{\Theta}] = \mathrm{supp}[\Theta]$?
  - bounds on Kullback-Leibler divergence $D(\mathbb{Q}_{\widehat{\Theta}} \,\|\, \mathbb{Q}_{\Theta})$
  - bounds on $|||\widehat{\Theta} - \Theta|||_{\mathrm{op}}$.

Challenges in graph selection

For pairwise models, negative log-likelihood takes form:

$$\begin{aligned}\ell(\Theta; \mathbf{X}_1^n) &:= -\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{Q}(x_{i1}, \ldots, x_{ip}; \Theta) \\ &= \log Z(\Theta) - \sum_{s \in V} \theta_s \widehat{\mu}_s - \sum_{(s,t)} \theta_{st} \widehat{\mu}_{st}\end{aligned}$$

Challenges in graph selection

For pairwise models, negative log-likelihood takes form:

$$\ell(\Theta; \mathbf{X}_1^n) := -\frac{1}{n} \sum_{i=1}^{n} \log \mathbb{Q}(x_{i1}, \ldots, x_{ip}; \Theta)$$
$$= \log Z(\Theta) - \sum_{s \in V} \theta_s \widehat{\mu}_s - \sum_{(s,t)} \theta_{st} \widehat{\mu}_{st}$$

- maximizing likelihood involves computing $\log Z(\Theta)$ or its derivatives (marginals)
- for Gaussian graphical models, this is a log-determinant program
- for discrete graphical models, various work-arounds are possible:
  - Markov chain Monte Carlo and stochastic gradient
  - variational approximations to likelihood
  - pseudo-likelihoods

Methods for graph selection

- for Gaussian graphical models:
  - ℓ_1-regularized neighborhood regression for Gaussian MRFs
    (e.g., Meinshausen & Buhlmann, 2005; Wainwright, 2006, Zhao & Yu, 2006)
  - ℓ_1-regularized log-determinant (e.g., Yuan & Lin, 2006; d'Asprémont et al., 2007; Friedman, 2008; Rothman et al., 2008; Ravikumar et al., 2008)

Methods for graph selection

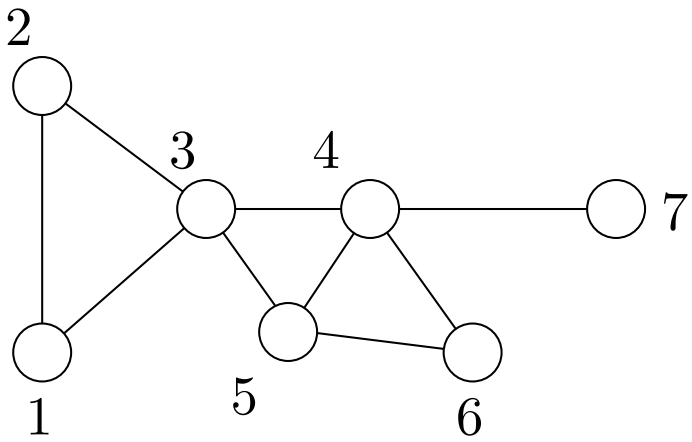
- for Gaussian graphical models:
  - ℓ_1-regularized neighborhood regression for Gaussian MRFs (e.g., Meinshausen & Buhlmann, 2005; Wainwright, 2006, Zhao & Yu, 2006)
  - ℓ_1-regularized log-determinant (e.g., Yuan & Lin, 2006; d'Asprémont et al., 2007; Friedman, 2008; Rothman et al., 2008; Ravikumar 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)
  - various other methods
    - ⋆ distribution fits by KL-divergence (Abeel et al., 2005)
    - ⋆ ℓ_1-regularized log. regression (Ravikumar, W. & Lafferty et al., 2008, 2010)
    - ⋆ approximate max. entropy approach and thinned graphical models (Johnson et al., 2007)
    - ⋆ neighborhood-based thresholding method (Bresler, Mossel & Sly, 2008)

Methods for graph selection

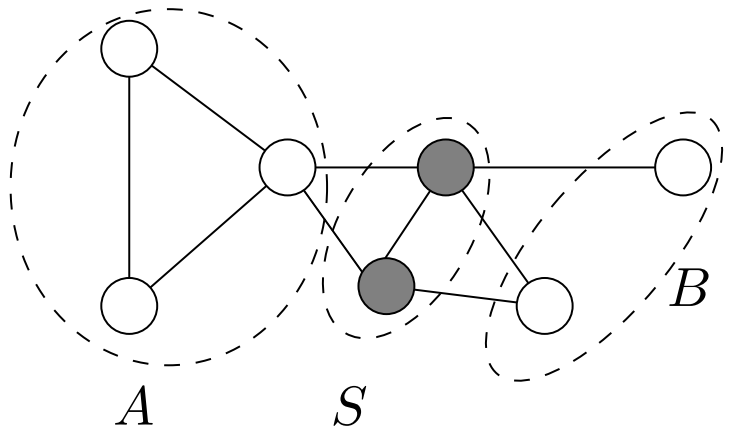
- for Gaussian graphical models:
  - ℓ_1-regularized neighborhood regression for Gaussian MRFs (e.g., Meinshausen & Buhlmann, 2005; Wainwright, 2006, Zhao & Yu, 2006)
  - ℓ_1-regularized log-determinant (e.g., Yuan & Lin, 2006; d'Asprémont et al., 2007; Friedman, 2008; Rothman et al., 2008; Ravikumar 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)
  - various other methods
    - ⋆ distribution fits by KL-divergence (Abeel et al., 2005)
    - ⋆ ℓ_1-regularized log. regression (Ravikumar, W. & Lafferty et al., 2008, 2010)
    - ⋆ 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, 2012)

Graphs and random variables

- associate to each node $s \in V$ a random variable X_s
- for each subset $A \subseteq V$, random vector $X_A := \{X_s, s \in A\}$.



Maximal cliques $(123), (345), (456), (47)$



Vertex cutset S

- a *clique* $C \subseteq V$ is a subset of vertices all joined by edges
- a *vertex cutset* is a subset $S \subset V$ whose removal breaks the graph into two or more pieces

Factorization and Markov properties

The graph G can be used to impose constraints on the random vector $X = X_V$ (or on the distribution $\mathbb{Q}$) in different ways.

Markov property: X is *Markov w.r.t G* if X_A and X_B are conditionally indpt. given X_S whenever S separates A and B.

Factorization: The distribution $\mathbb{Q}$ *factorizes according to G* if it can be expressed as a product over cliques:

$$\mathbb{Q}(x_1, x_2, \ldots, x_p) = \underbrace{\frac{1}{Z}}_{\text{Normalization}} \prod_{C \in \mathcal{C}} \underbrace{\psi_C(x_C)}_{\text{compatibility function on clique } C}$$

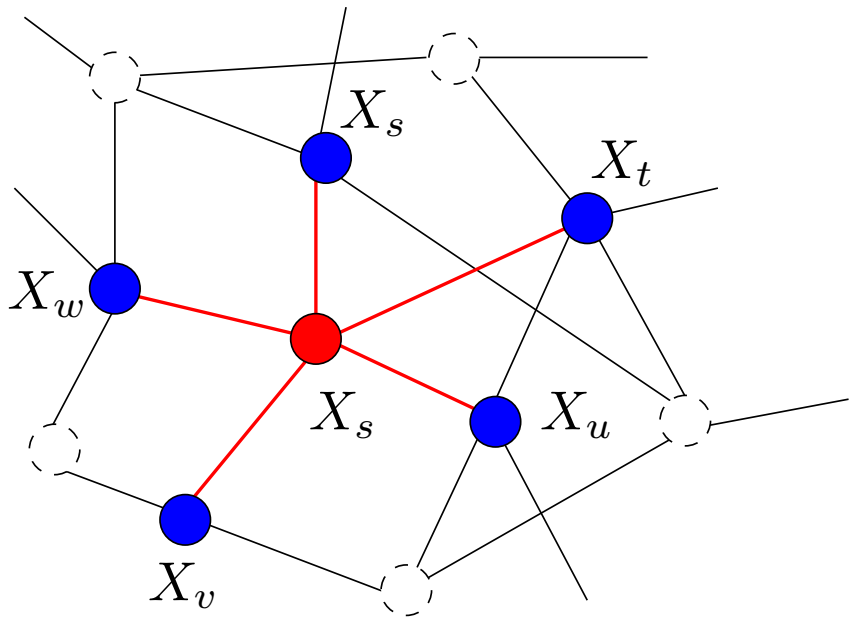
Theorem: (Hammersley & Clifford, 1973) For strictly positive $\mathbb{Q}(\cdot)$, the Markov property and the Factorization property are equivalent.

Markov property and neighborhood structure

- Markov properties encode neighborhood structure:

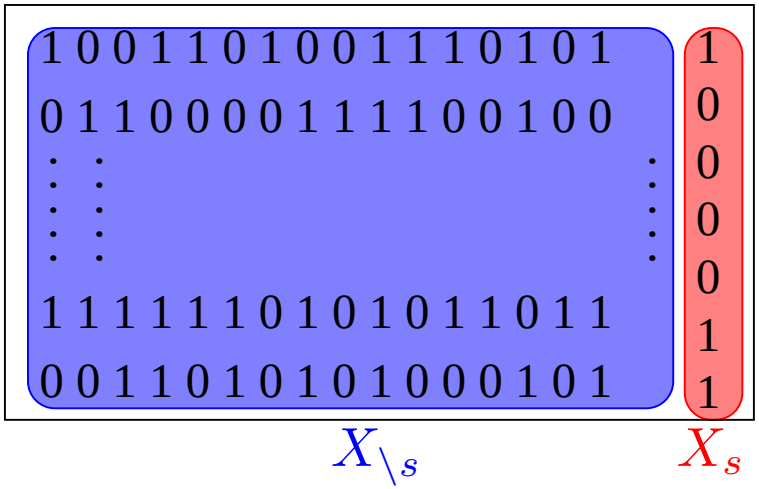
$$\underbrace{(X_s \quad | \quad X_{V \setminus s})}_{\text{Condition on full graph}} \stackrel{d}{=} \underbrace{(X_s \quad | \quad X_{N(s)})}_{\text{Condition on Markov blanket}}$$

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



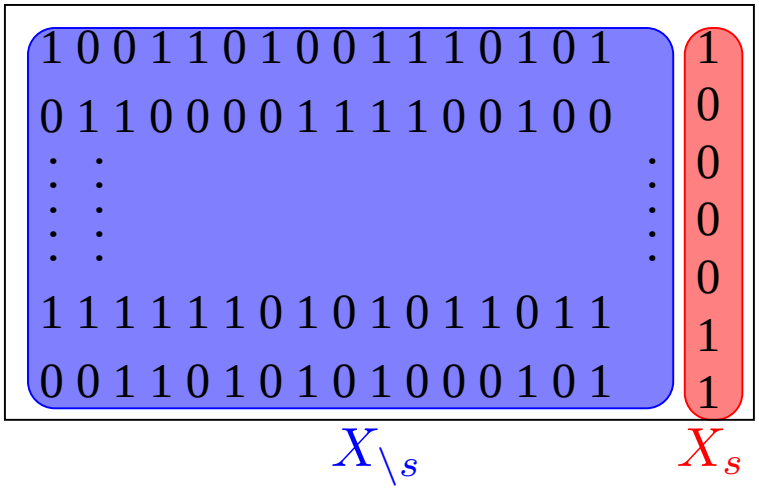
- basis of pseudolikelihood method (Besag, 1974)
- basis of many graph learning algorithms (Friedman et al., 1999; Csiszar & Talata, 2005; Abeel et al., 2006; Meinshausen & Buhlmann, 2006)

Graph selection via neighborhood regression



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

Graph selection via neighborhood regression

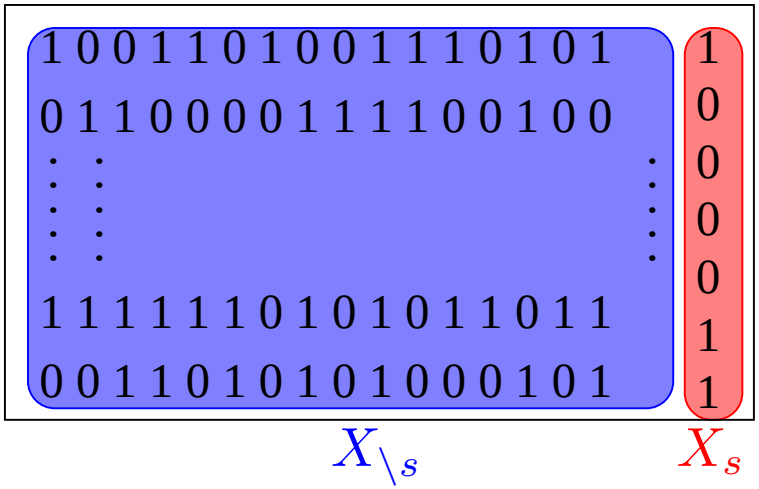


Predict X_s based on $X_{\setminus s} := \{X_s,\ t \neq s\}$.

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

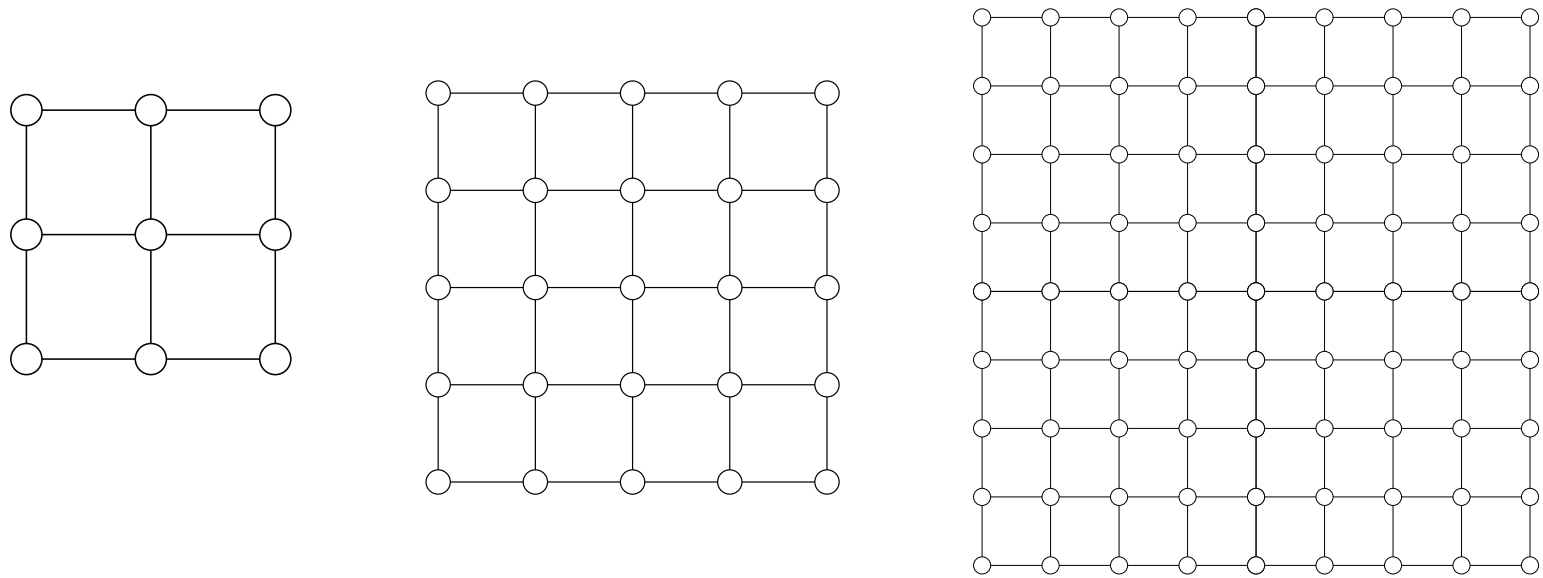
1. For each node $s \in V$, compute (regularized) max. likelihood estimate:

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

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

High-dimensional analysis

- classical analysis: graph size 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

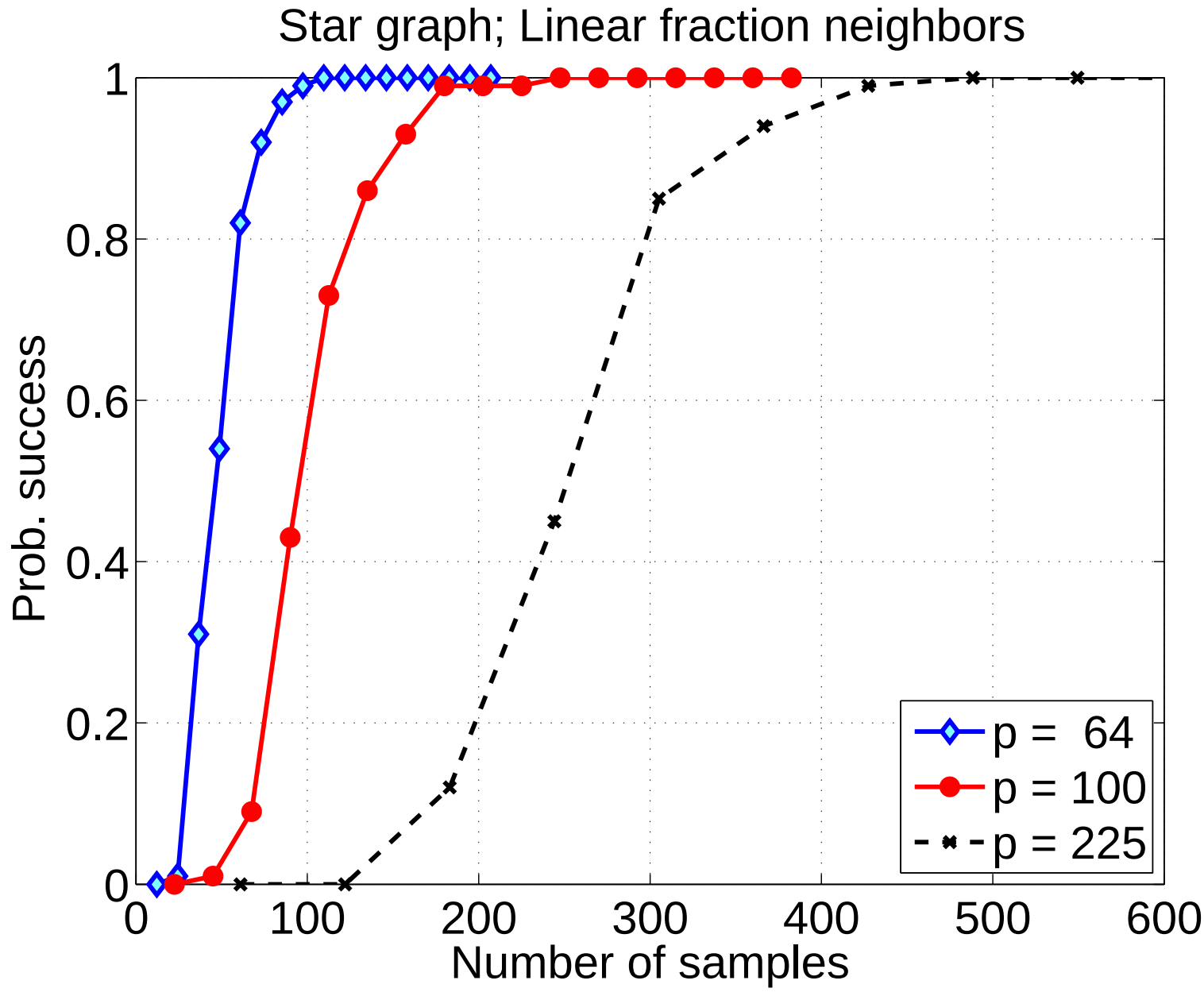


- take n i.i.d. samples from MRF defined by $G_{p,d}$
- study probability of success as a function of three parameters:

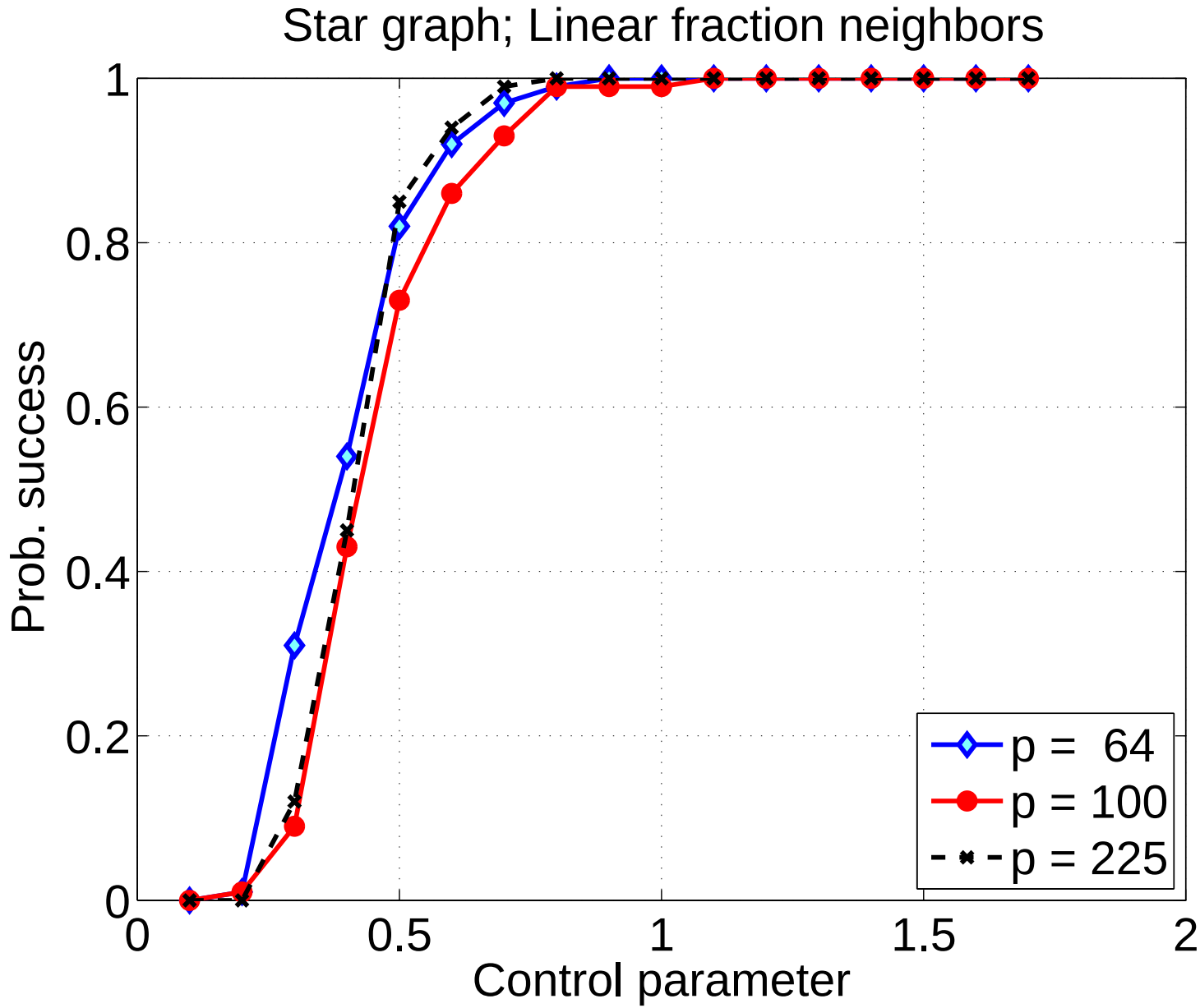
$$\text{Success}(n, p, d) \quad = \quad \mathbb{Q}[\text{Method recovers graph } G_{p,d} \text{ from } n \text{ samples}]$$

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

Empirical behavior: Unrescaled plots

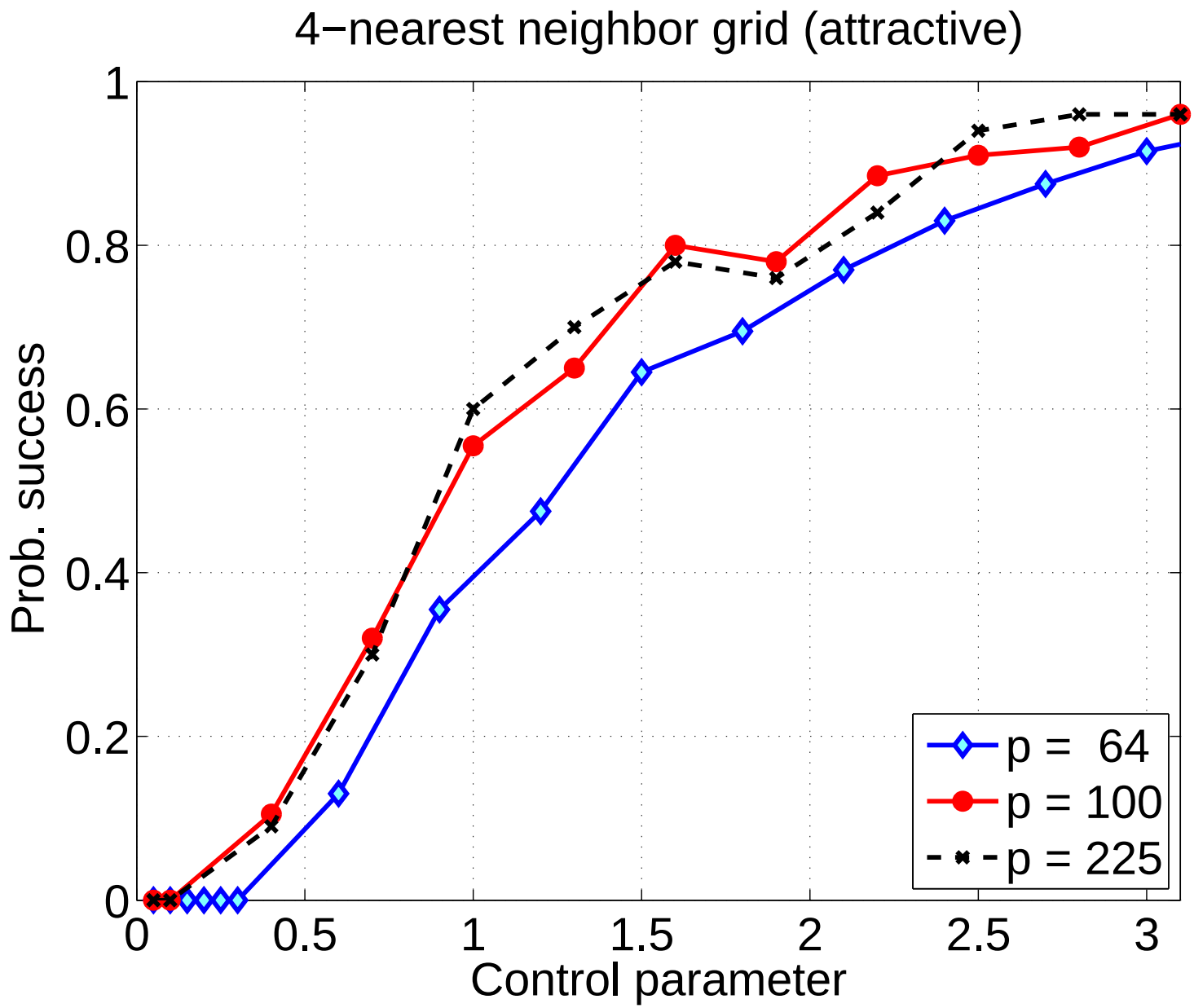


Empirical behavior: Appropriately rescaled



Plots of success probability versus control parameter (α, β)

Rescaled plots (2-D lattice graphs)



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, 2006, 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, 2006, 2010)

Under incoherence conditions, for a rescaled sample

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

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

(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, 2006, 2010)

Under incoherence conditions, for a rescaled sample

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

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

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

(b) Correct inclusion: *For* $\theta_{\min} \geq c_3 \lambda_n$, *the method selects the correct signed neighborhood.*

Some related work

- thresholding estimator (poly-time for bounded degree) works with $n \gtrsim 2^d \log p$ samples (Bresler et al., 2008)

Some related work

- thresholding estimator (poly-time for bounded degree) works with $n \gtrsim 2^d \log p$ samples (Bresler et al., 2008)

- information-theoretic lower bound over family $\mathcal{G}_{p,d}$: any method requires at least $n = \Omega(d^2 \log p)$ samples (Santhanam & W., 2008)

Some related work

- thresholding estimator (poly-time for bounded degree) works with $n \gtrsim 2^d \log p$ samples (Bresler et al., 2008)

- information-theoretic lower bound over family $\mathcal{G}_{p,d}$: any method requires at least $n = \Omega(d^2 \log p)$ samples (Santhanam & W., 2008)

- ℓ_1-based method: sharper achievable rates, also failure for θ large enough to violate incoherence (Bento & Montanari, 2009)

Some related work

- thresholding estimator (poly-time for bounded degree) works with $n \gtrsim 2^d \log p$ samples (Bresler et al., 2008)

- information-theoretic lower bound over family $\mathcal{G}_{p,d}$: any method requires at least $n = \Omega(d^2 \log p)$ samples (Santhanam & W., 2008)

- ℓ_1-based method: sharper achievable rates, also failure for θ large enough to violate incoherence (Bento & Montanari, 2009)

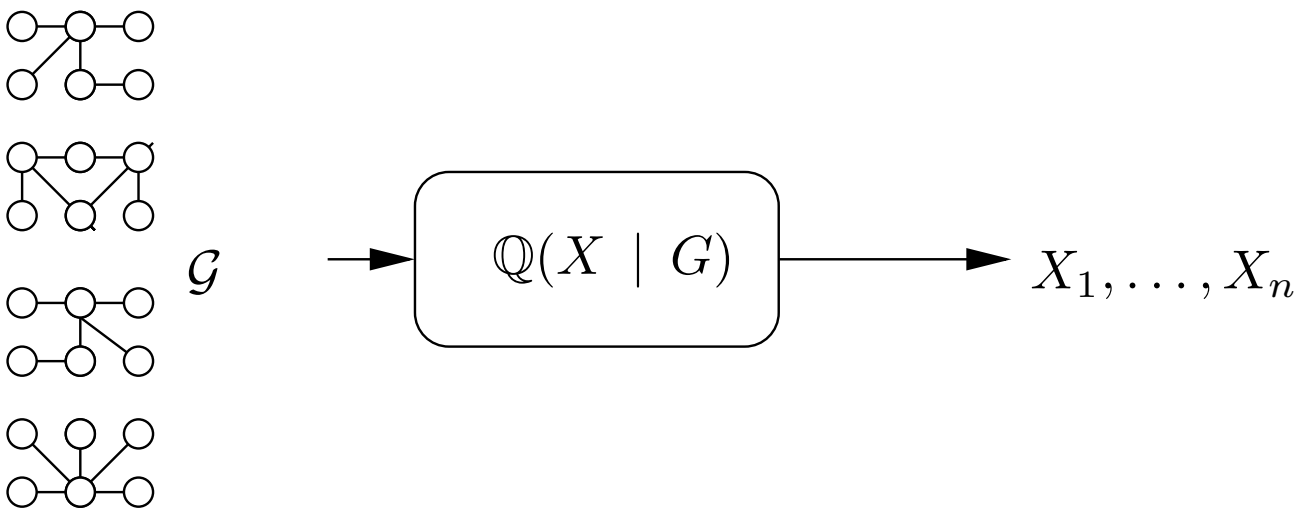
- empirical study: ℓ_1-based method can succeed beyond phase transition on Ising model (Aurell & Ekeberg, 2011)

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

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

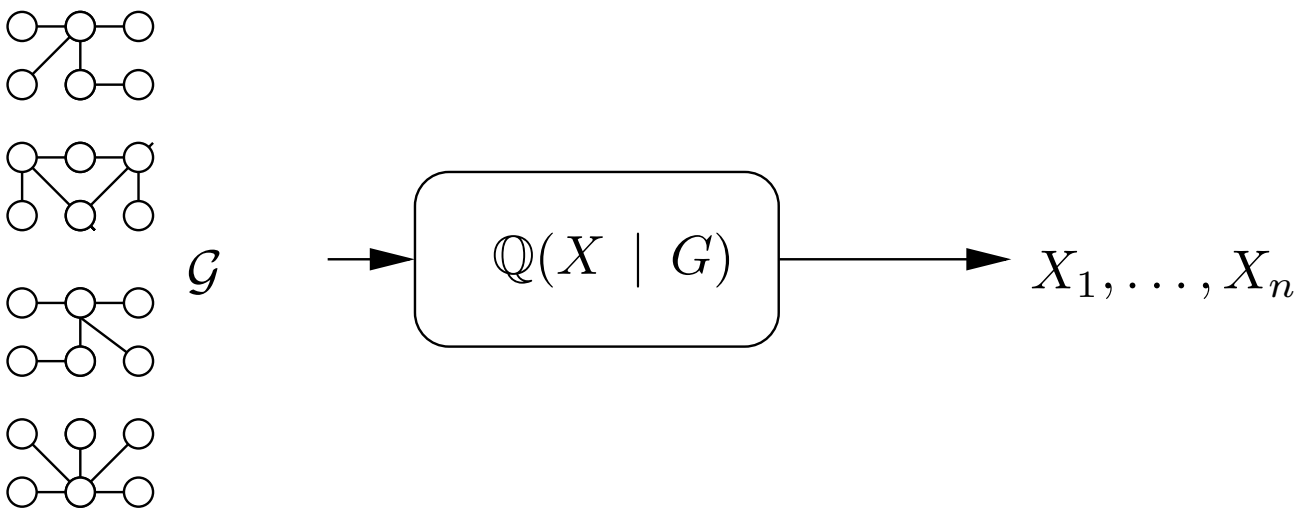
§3. Info. theory: Graph selection as channel coding

- graphical model selection is an *unorthodox* channel coding problem:
  - codewords/codebook: graph G in some graph class $\mathcal{G}$
  - channel use: draw sample $X_i = (X_{i1}, \ldots, X_{ip}$ from Markov random field $\mathbb{Q}_{\theta(G)}$
  - decoding problem: use n samples $\{X_1, \ldots, X_n\}$ 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 $X_i = (X_{i1}, \ldots, X_{ip}$ from Markov random field $\mathbb{Q}_{\theta(G)}$
  - decoding problem: use n samples $\{X_1, \ldots, X_n\}$ 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)*

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

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, 2008)*

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

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 optimal up to a factor $\mathcal{O}(d)$ (Ravikumar., W. & Lafferty, 2010)

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 $\mathbf{X}_1^n = \{X_1, \ldots, X_n\}$

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 $\mathbf{X}_1^n = \{X_1, \ldots, X_n\}$

- for any graph estimator $\psi : \mathcal{X}^n \to \mathcal{G}$, Fano's inequality implies that

$$\mathbb{Q}[\psi(\mathbf{X}_1^n) \neq G] \geq 1 - \frac{I(\mathbf{X}_1^n; G) + \log 2}{\log |\mathcal{G}|}$$

  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 $\mathbf{X}_1^n = \{X_1, \ldots, X_n\}$
- for any graph estimator $\psi : \mathcal{X}^n \to \mathcal{G}$, Fano's inequality implies that

$$\mathbb{Q}[\psi(\mathbf{X}_1^n) \neq G] \geq 1 - \frac{I(\mathbf{X}_1^n; G) + \log 2}{\log |\mathcal{G}|}$$

  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 $I(\mathbf{X}_1^n; G)$.

Summary

- simple ℓ_1-regularized neighborhood selection:
  - polynomial-time method for learning neighborhood structure
  - natural extensions (using block regularization) to higher order models

- information-theoretic limits of graph learning

Some papers:

- Ravikumar, W. & Lafferty (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*.
- Santhanam & W (2012). Information-theoretic limits of selecting binary graphical models in high dimensions, *IEEE Transactions on Information Theory*.

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: $\log |\mathcal{G}_{p,d}| = \Theta\big(pd \log(p/d)\big)$
   - 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: $\log |\mathcal{G}_{p,d}| = \Theta\big(pd \log(p/d)\big)$
   - 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: $\log|\mathcal{G}_{p,d}| = \Theta\big(pd\log(p/d)\big)$
   - trivial upper bound: $I(\mathbf{X}_1^n; G) \le H(\mathbf{X}_1^n) \le np$.
   - 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:

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

   - upper bound on symmetrized Kullback-Leibler divergences:

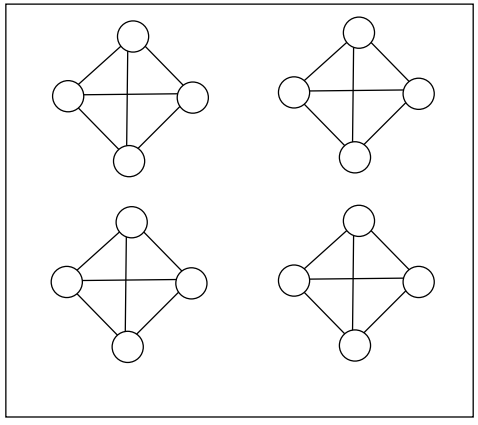
$$D\big(\theta(G^{ij})\|\theta(G^{k\ell})\big) + D\big(\theta(G^{k\ell})\|\theta(G^{ij})\big) \le 2\theta_{\min}\tanh(\theta_{\min}/2)$$

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

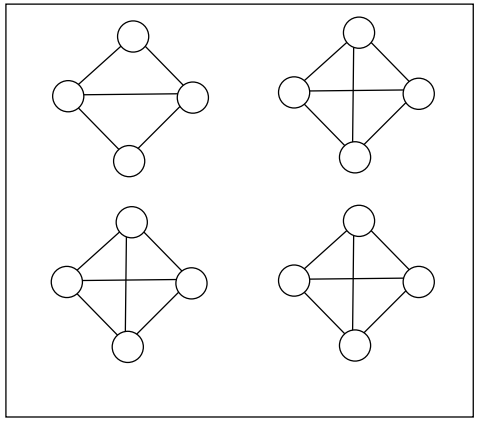
A harder d-clique ensemble

Constructive procedure:

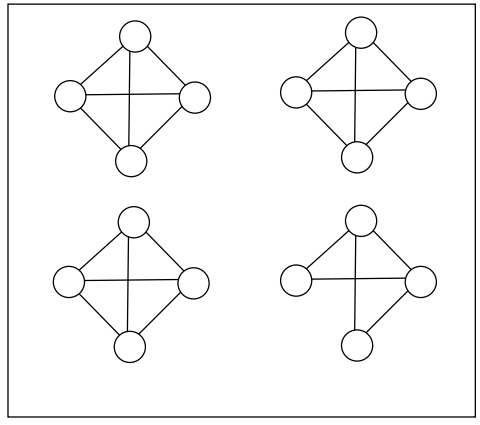
1. Divide the vertex set V into $\lfloor \frac{p}{d+1} \rfloor$ groups of size $d+1$.
2. Form the base graph $\overline{G}$ by making a $(d+1)$-clique within each group.
3. Form graph G^{uv} by deleting edge (u, v) from $\overline{G}$.
4. Form Markov random field $\mathbb{Q}_{\theta(G^{uv})}$ by setting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



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



(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.