High-dimensional graphical model selection: Practical and information-theoretic limits

Martin Wainwright

Departments of Statistics, and EECS

UC Berkeley, California, USA

Based on joint work with:

John Lafferty (CMU), Pradeep Ravikumar (UC Berkeley), and
Prasad Santhanam (University of Hawaii)

Supported by grants from National Science Foundation, and
a Sloan Foundation Fellowship

Introduction

- classical asymptotic theory of statistical inference:
  - number of observations $n \to +\infty$
  - model dimension p stays fixed

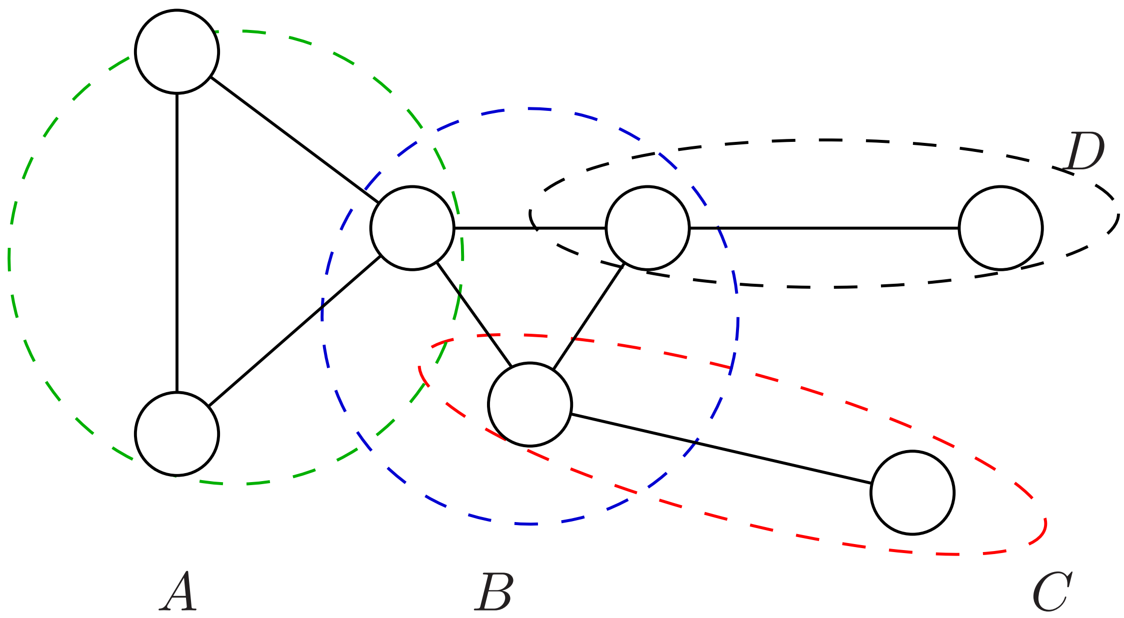
- not suitable for many modern applications:
  - { images, signals, systems, networks } frequently large ($p \approx 10^3 - 10^8$)...
  - function/surface estimation: enforces limit $p \to +\infty$
  - interesting consequences: might have $p = \Theta(n)$ or even $p \gg n$

- curse of dimensionality: frequently impossible to obtain consistent procedures unless $p/n \to 0$

- can be saved by a lower *effective dimensionality*, due to some form of complexity constraint:
  - sparse vectors
  - {sparse, structured, low-rank}-matrices
  - structured regression functions
  - graphical models (Markov random fields)

What are graphical models?

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



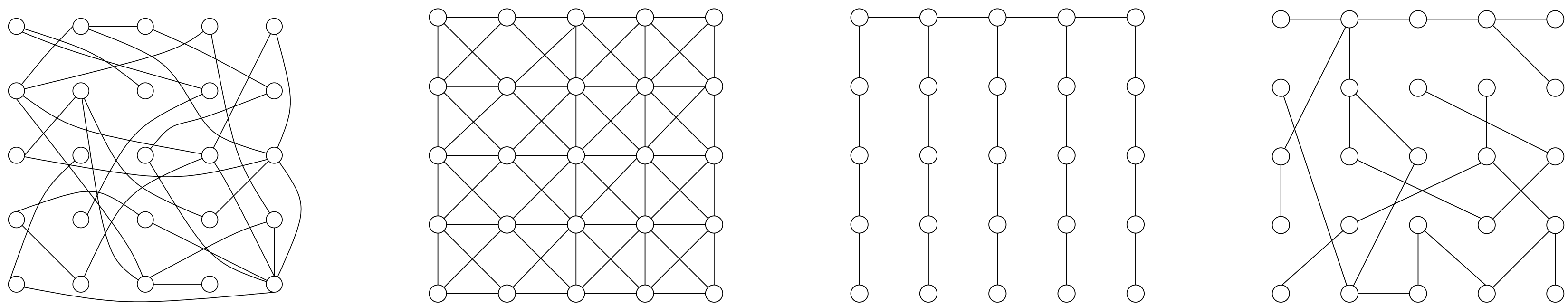
- Hammersley-Clifford Theorem: $(X_1, \ldots, X_p)$ being Markov w.r.t G implies factorization:

$$\mathbb{P}(x_1, \ldots, x_p) \quad \propto \quad \exp\Big\{\theta_A(x_A) + \theta_B(x_B) + \theta_C(x_C) + \theta_D(x_D)\Big\}.$$

- 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



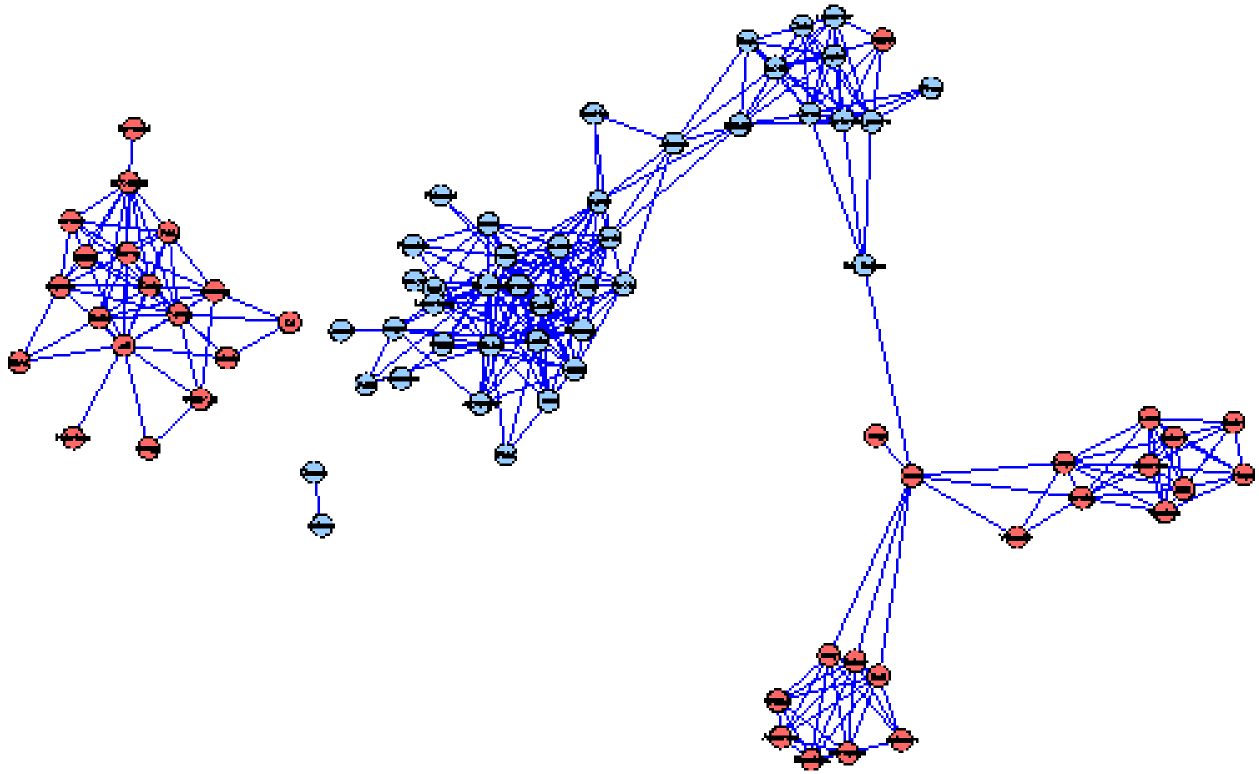
- pairwise Markov random field: family of prob. distributions

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

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

Some issues in high-dimensional inference

Consider some fixed loss function, and a fixed level δ of error.

Limitations of tractable algorithms:

Given particular (polynomial-time) algorithms

- for what sample sizes n do they succeed/fail to achieve error δ?
- given a collection of methods, when does more computation reduce minimum # samples needed?

Information-theoretic limitations:

Data collection as communication from nature $\longrightarrow$ statistician:

- what are fundamental limitations of problem (Shannon capacity)?
- when are known (polynomial-time) methods optimal?
- when are there gaps between poly.-time methods and optimal methods?

Previous/on-going work on graph selection

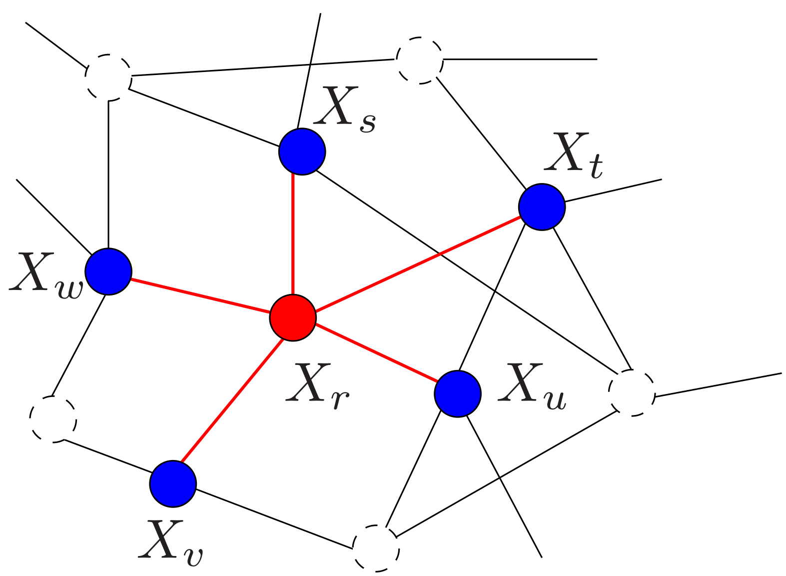
- exact solution for trees (Chow & Liu, 1967)
- local testing-based approaches (e.g., Spirtes et al, 2000; Kalisch & Buhlmann, 2008)
- methods for Gaussian MRFs
  - ℓ_1-regularized neighborhood regression for Gaussian MRFs (e.g., Meinshausen & Buhlmann, 2005; Wainwright, 2006, Zhao, 2006)
  - ℓ_1-regularized log-determinant (e.g., Yuan & Lin, 2006; d'Asprémont et al., 2007; Friedman, 2008; Ravikumar et al., 2008)
- methods for discrete MRFs
  - neighborhood-based search method (Bresler, Mossel & Sly, 2008)
  - ℓ_1-regularized logistic regression (Ravikumar et al., 2006, 2008)
- information-theoretic approaches:
  - pseudolikelihood and BIC criterion (Csiszar & Talata, 2006)
  - information-theoretic limitations (Santhanam & Wainwright, 2008)

Markov property and neighborhood structure

- Markov properties encode neighborhood structure:

$$\underbrace{(X_r \quad | \quad X_{V \backslash r})}_{\text{Condition on full graph}} \stackrel{d}{=} \underbrace{(X_r \quad | \quad X_{N(r)})}_{\text{Condition on Markov blanket}}$$

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



- basis of pseudolikelihood method (Besag, 1974)

Practical method via neighborhood regression

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

Method: Given n i.i.d. samples $\{X^{(1)}, \ldots, X^{(n)}\}$, perform logistic regression of each node X_r on $X_{\backslash r} := \{X_r,\ t \neq r\}$ to estimate neighborhood structure $\widehat{N}(r)$.

1. For each node $r \in V$, perform ℓ_1 regularized logistic regression of X_r on the remaining variables $X_{\backslash r}$:

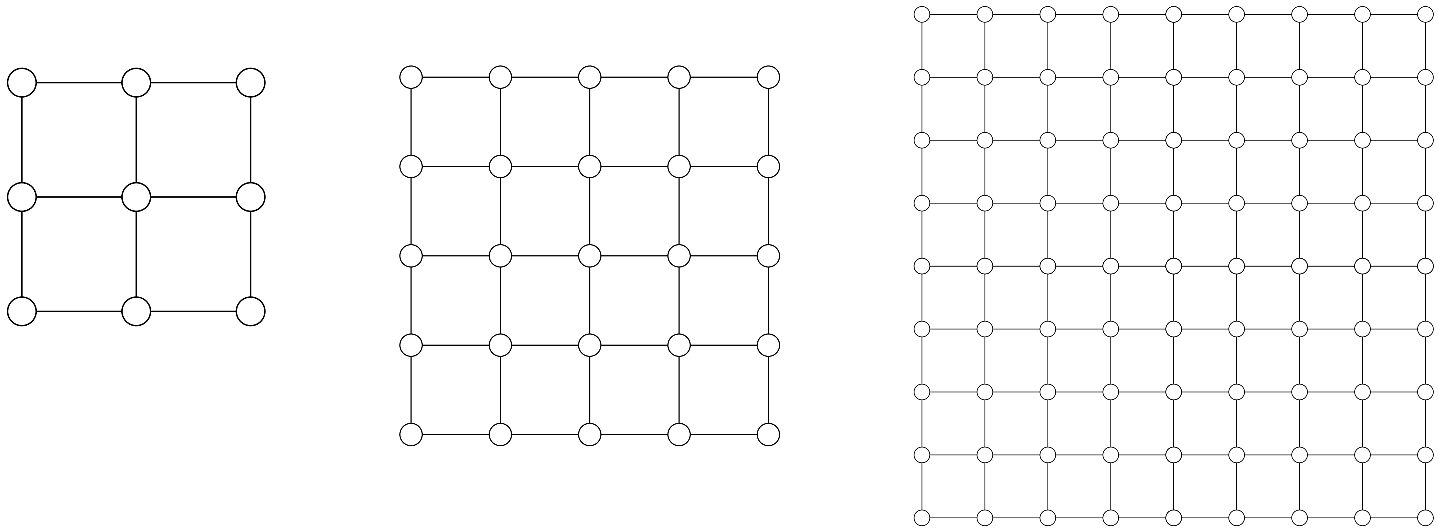
$$\widehat{\theta}[r] \quad := \quad \arg \min_{\theta \in \mathbb{R}^{p-1}} \Big\{ \quad \underbrace{\frac{1}{n} \sum_{i=1}^{n} f(\theta; X_{\backslash r}^{(i)})}_{\text{logistic likelihood}} \quad + \quad \underbrace{\rho_n \|\theta\|_1}_{\text{regularization}} \Big\}$$

2. Estimate the local neighborhood $\widehat{N}(r)$ as the support (non-negative entries) of the regression vector $\widehat{\theta}[r]$.

3. Combine the neighborhood estimates in a consistent manner (AND, or OR rule).

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

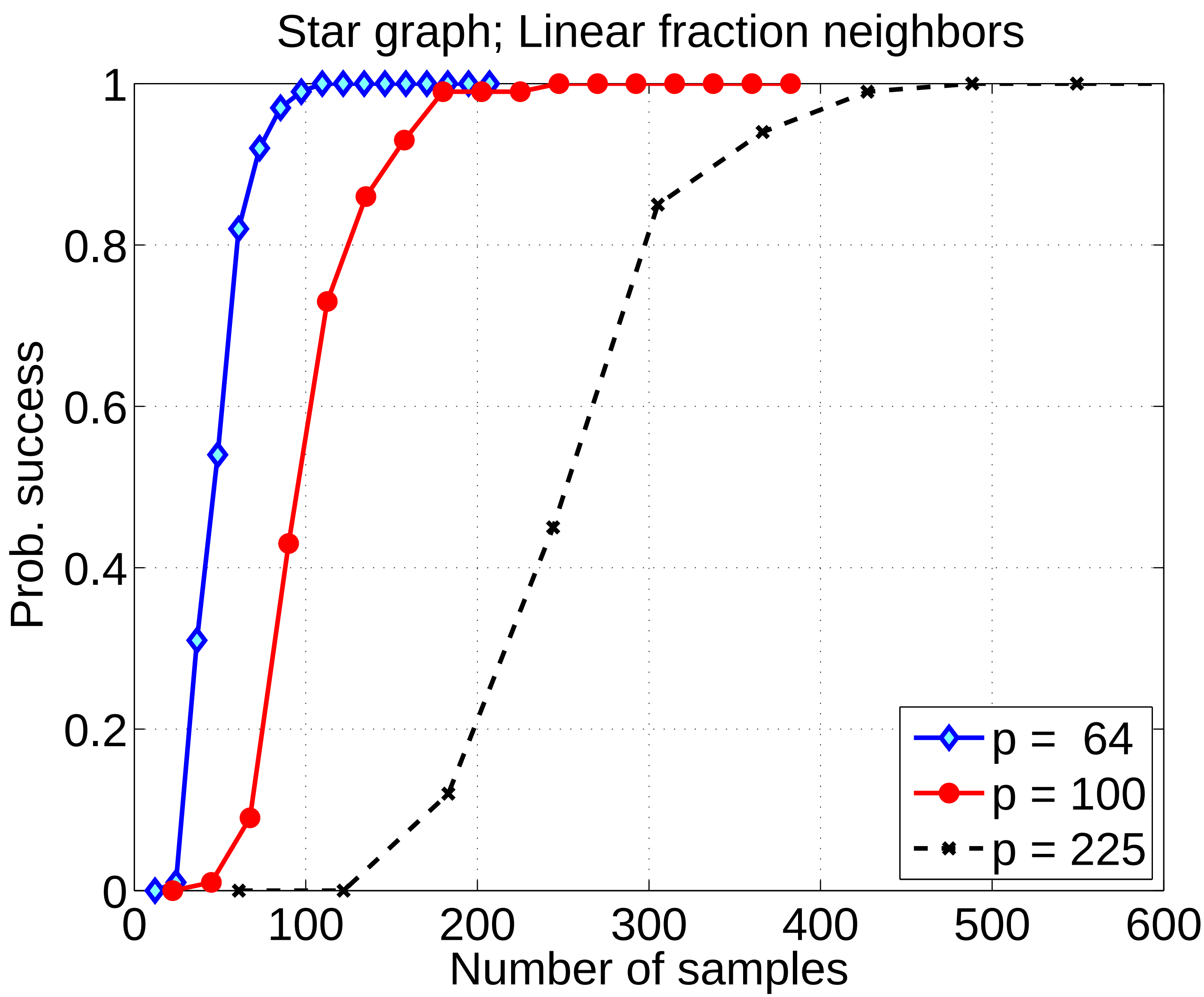


- take n i.i.d. samples from MRF defined by $G_{p,d}$
- study probability of success as a function of three parameters:

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

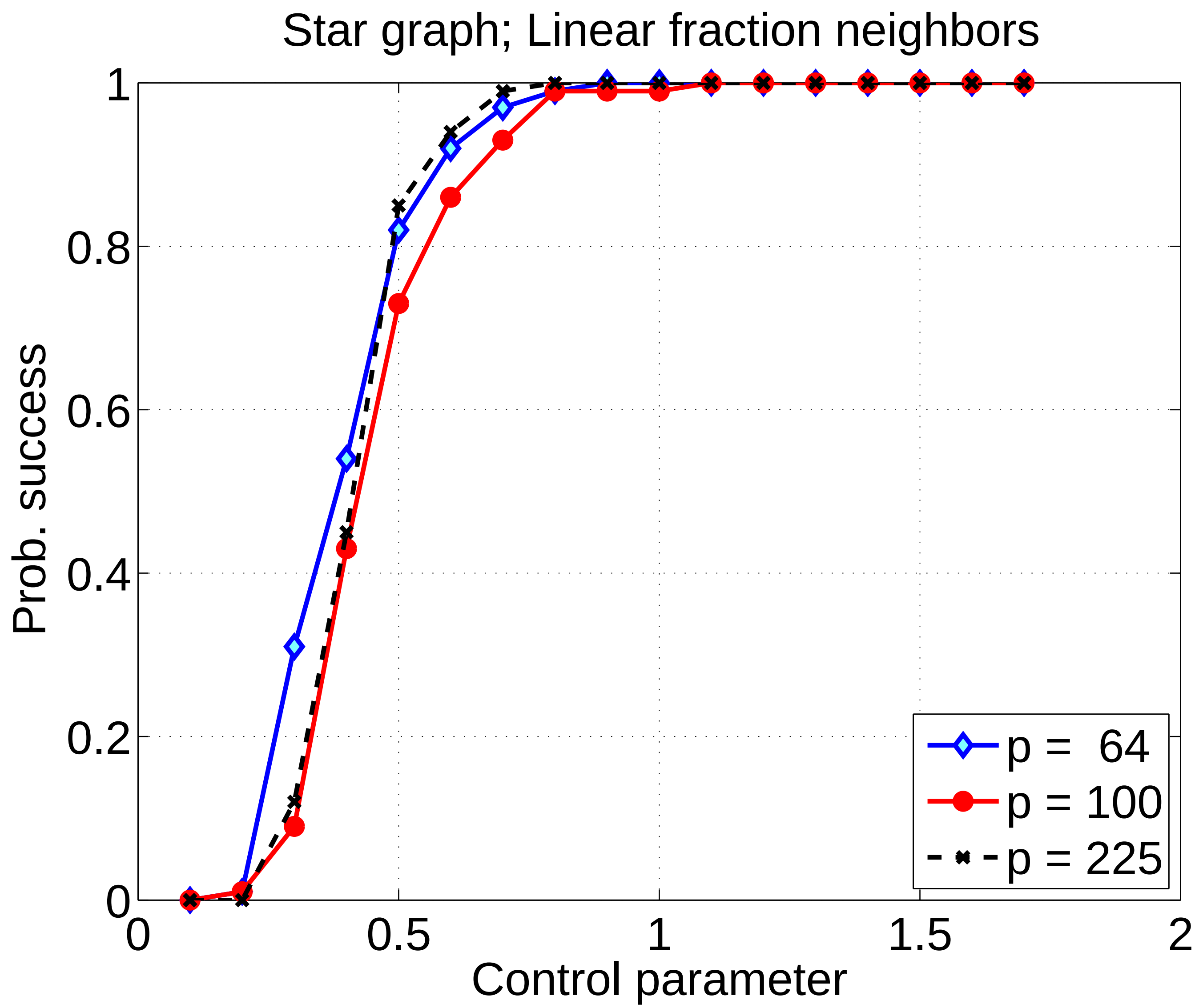
- theory is non-asymptotic: explicit probabilities for finite (n, p, d)

Empirical behavior: Unrescaled plots



Plots of success probability versus raw sample size n.

Empirical behavior: Appropriately rescaled



Plots of success probability versus control parameter $T_{\mathrm{LR}}(n, p, d)$.

Sufficient conditions for consistent model selection

- graph sequences $G_{p,d} = (V, E)$ with p vertices, and maximum degree d.
- drawn n i.i.d, samples, and analyze prob. success indexed by (n, p, d)

Theorem: For a rescaled sample size (RavWaiLaf06, RavWaiLaf08)

$$T_{\mathrm{LR}}(n,p,d) \quad := \quad \frac{n}{d^3 \log p} \;>\; T^*_{\text{crit}}$$

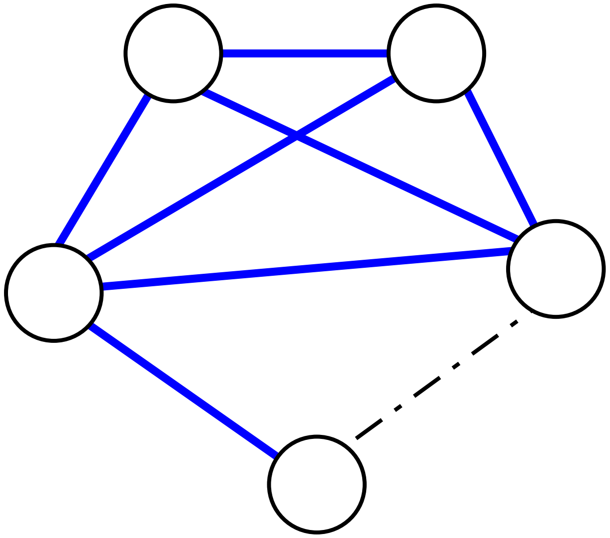
and regularization parameter $\rho_n \geq c_1\, \tau \sqrt{\frac{\log p}{n}}$, then with probability greater than $1 - 2\exp\big(-c_2(\tau - 2)\log p\big) \rightarrow 1$:

(a) For each node $r \in V$, the ℓ_1-regularized logistic convex program has a unique solution. (Non-trivial since $p \gg n \implies$ not strictly convex).

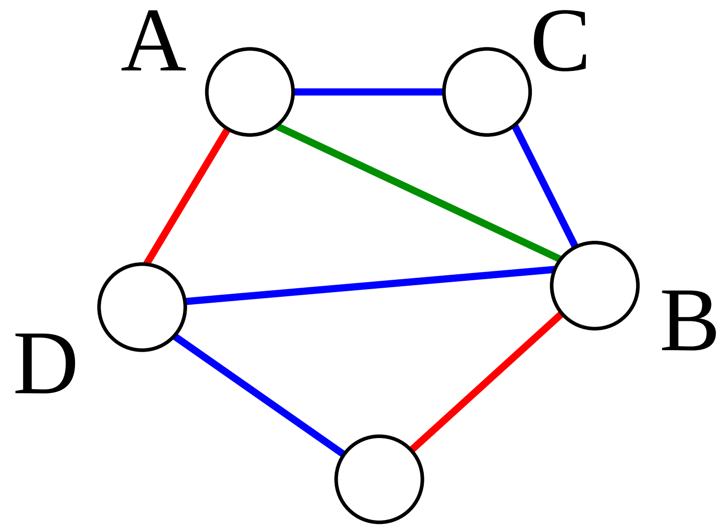
(b) The estimated sign neighborhood $\widehat{N}_{\pm}(r)$ correctly excludes all edges *not* in the true neighborhood.

(c) For $\theta_{min} \geq \; c_3 \tau \sqrt{\frac{d^2 \log p}{n}}$, the method selects the correct signed neighborhood.

Some challenges in distinguishing graphs



Guilt by association



Hidden interactions

Conditions on Fisher information matrix $Q^* = \mathbb{E}[\nabla^2 f(\theta^*; X)]$

A1. Bounded eigenspectra: $\lambda(Q^*_{SS}) \in [C_{min}, C_{max}]$.

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

$$\|Q^*_{S^cS}(Q^*_{SS})^{-1}\|_{\infty,\infty} \quad \leq \quad 1 - \nu.$$

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

Proof sketch: Primal-dual certificate

- construct *candidate* primal-dual pair $(\widehat{\theta}, \widehat{z}) \in \mathbb{R}^{p-1} \times \mathbb{R}^{p-1}$.
- proof technique—-not a practical algorithm!

(A) For a fixed node r with $S = N(r)$, we solve the restricted program

$$\widehat{\theta} = \arg \min_{\theta \in \mathbb{R}^{p-1}, \theta_{S^c}=0} \Big\{ \frac{1}{n} \sum_{i=1}^{n} f(\theta; X_{\backslash r}^{(i)}) \; + \; \rho_n \|\theta\|_1 \Big\},$$

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

(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.

Information-theoretic limits on graph selection

- thus far: have exhibited a a particular polynomial-time method can recover structure if

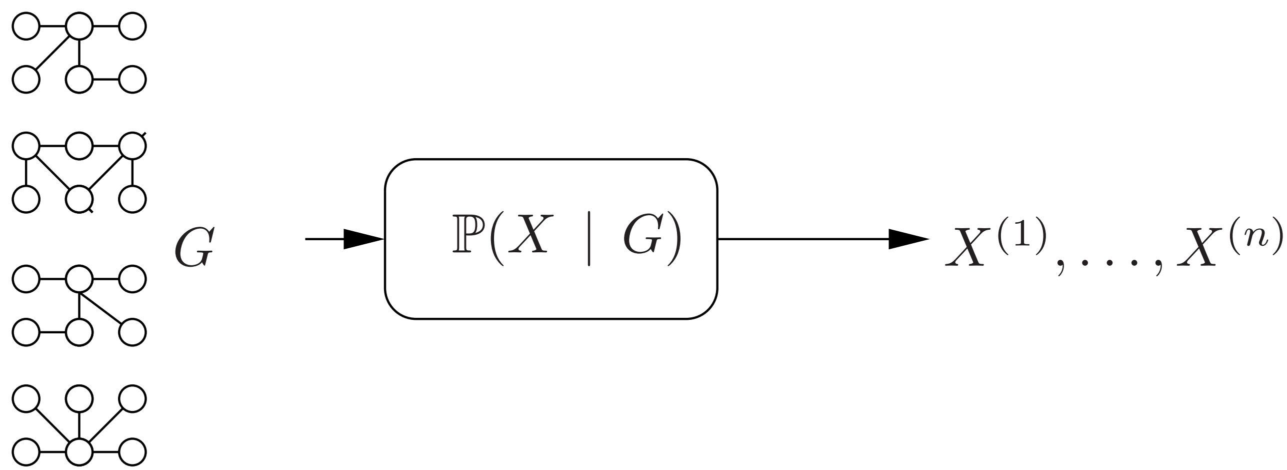
$$n > \Omega(d^3 \log(p - d))$$

- but....is this a "good" result?
- are there polynomial-time methods that can do better?
- information theory can answer the question: is there an exponential-time method that can do better?

(Santhanam & Wainwright, 2008)

Graph selection as channel coding

- graphical model selection is an *unorthodox* channel coding problem:
- nature sends $G \in \mathcal{G}_{d,p} := \{$ graphs on p vertices, max. degree d $\}$



- decoding problem: use observations $\{X^{(1)}, \ldots, X^{(n)}\}$ to correctly distinguish the "codeword"
- channel capacity for graph decoding: balance between
  - log number of models: $\log |M(p,d)| = \Theta\left(pd \log \frac{p}{d}\right)$.
  - relative distinguishability of different models

Necessary conditions for graph recovery

- take Ising models $\mathbb{P}_{\theta(G)}$ from $\mathcal{G}_{d,p}(\lambda, \omega)$:
  - graphs with p nodes and max. degree d
  - parameters $|\theta_{st}| \geq \lambda$ for all edges (s,t)
  - maximum neighborhood weight $\omega = \max_{s \in V} \sum_{t \in N(s)} |\theta_{st}|$.
- take n i.i.d. observations, and study probability of success in terms of (n, p, d)

Theorem: Necessary conditions: For sample size n

$$n \quad \leq \quad \max\left\{ \frac{\log p}{2\lambda \tanh(\lambda)}, \; \frac{\exp(\omega/2)\,\lambda}{16 \sinh(\lambda)} \; d \log(pd), \; \frac{d}{8} \log \frac{p}{8d}, \right\},$$

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

(Santhanam & W., 2008)

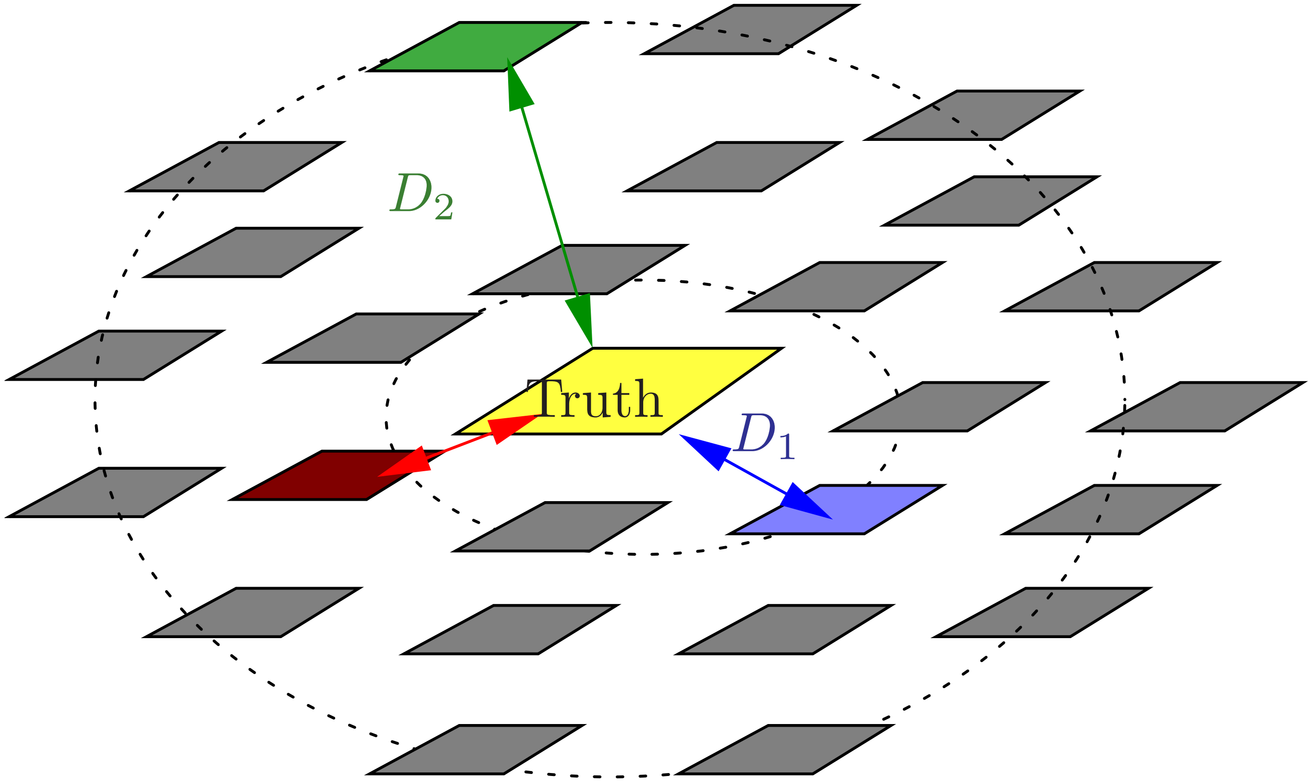
Some consequences

- note neighborhood weight $\omega = \max_{s \in V} \sum_{t \in N(s)} |\theta_{st}|$ is at least $d\lambda$
- hence, need at least

$$n \quad > \quad \frac{\exp(\frac{d\lambda}{2})\,\lambda}{16\sinh(\lambda)} \; d\log(pd)$$

- if $\lambda = \mathcal{O}(1/d)$, then need at least $n > \frac{\log p}{\lambda^2} = \Omega(d^2 \log p)$ samples
- ℓ_1-regularized log. regression (LR) order-optimal for constant degrees
- for d tending to infinity, gap between optimal methods and ℓ_1
  - any method requires $n = \Omega(d^2 \log p)$ samples
  - LR method: guaranteed to work with $n = \Omega(d^3 \log p)$ samples

Geometric intuition underlying proofs



Error probability controlled by two competing quantities:

Model type	Log # models	Distance scaling
Near-by	$\log p$	c_2/θ^2
Intermediate	$d \log p$	$\frac{\sinh(\theta)}{\theta \exp(\theta d)}$
Far-away	$pd \log \frac{p}{d}$	$c_2 p$

Summary and open questions

- ℓ_1-regularized regression to select neighborhoods: succeeds with sample size

$$n \quad > \quad \Big(\frac{c_1}{\theta_{min}^2} + c_2 d^3\Big) \log p.$$

- any method (including those with exponential complexity) fails for

$$n \quad < \quad \Big(\frac{c_3}{\theta_{min}^2} + c_4 d^2\Big) \log p$$

- some extensions....
  - non-binary MRFs via block-structured regularization schemes
  - non-i.i.d. sampling models
  - other performance metrics (e.g, $(1-\delta)$ edges correct)
- broader issue: optimal trade-offs between statistical/computational efficiency?

Some papers

- Ravikumar, P., Wainwright, M. J. and Lafferty, J. (2008). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. Appeared at NIPS Conference (2006); To appear in *Annals of Statistics*.
- Santhanam, P. and Wainwright, M. J. (2008). Information-theoretic limitations of high-dimensional graphical model selection. Presented at Int. Symposium on Information Theory.
- Wainwright, M. J. (2006). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming. To appear in *IEEE Trans. on Information Theory*.
- Wainwright, M. J. (2007). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. UC Berkeley, Department of Statistics, Technical Report, January 2007.