Discovering and exploiting structure in high-dimensional data sets

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

EU-US Frontiers of Engineering Symposium
September 2010

Era of massive data sets

- engineering science in 21st century:
  - rapid technological advances (sensors, storage, computing etc.)
  - tremendous amounts of data being collected

Era of massive data sets

- engineering science in 21st century:
  - rapid technological advances (sensors, storage, computing etc.)
  - tremendous amounts of data being collected
- many examples:
  - biological data: genomics, proteomics, neural recordings etc.
  - astronomy: Sloan digital sky survey, Large synoptic survey telescope etc.
  - consumer preference data: Netflix, Amazon, etc.
  - geosciences: hyperspectral imaging
  - financial data: stocks, bonds, currencies, derivatives etc.

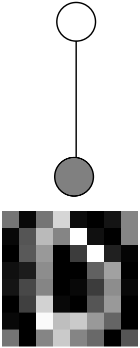
Era of massive data sets

- engineering science in 21st century:
  - rapid technological advances (sensors, storage, computing etc.)
  - tremendous amounts of data being collected
- many examples:
  - biological data: genomics, proteomics, neural recordings etc.
  - astronomy: Sloan digital sky survey, Large synoptic survey telescope etc.
  - consumer preference data: Netflix, Amazon, etc.
  - geosciences: hyperspectral imaging
  - financial data: stocks, bonds, currencies, derivatives etc.
- a wealth of data.....yet a paucity of information

Era of massive data sets

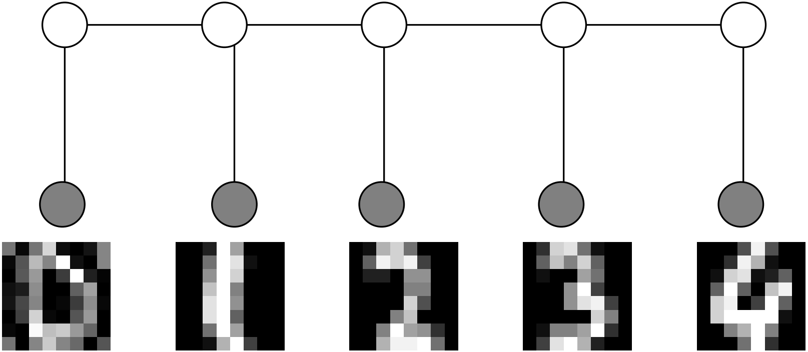
- engineering science in 21st century:
  - rapid technological advances (sensors, storage, computing etc.)
  - tremendous amounts of data being collected
- many examples:
  - biological data: genomics, proteomics, neural recordings etc.
  - astronomy: Sloan digital sky survey, Large synoptic survey telescope etc.
  - consumer preference data: Netflix, Amazon, etc.
  - geosciences: hyperspectral imaging
  - financial data: stocks, bonds, currencies, derivatives etc.
- a wealth of data.....yet a paucity of information
- possible structure in high-dimensional data sets
  - sparsity (data summarized by small number of coefficients)
  - manifolds (data lies on/near curved surfaces)
  - networks (data naturally associated with a graph)

Optical digit/character recognition



- **Goal:** correctly label digits/characters based on “noisy” versions
- E.g., mail sorting; document scanning; handwriting recognition systems

Optical digit/character recognition



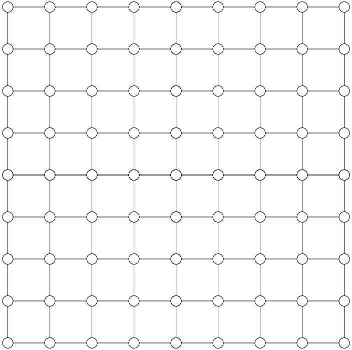
- **Goal:** correctly label digits/characters based on "noisy" versions
- strong sequential dependencies captured by hidden Markov model
- "message-passing" spreads information along chain

(Baum & Petrie, 1966; Viterbi, 1967, and many others)

Digital image processing

- 8-bit digital image: matrix of intensity values $\{0, 1, \ldots 255\}$
- enormous redundancy in "typical" images (useful for denoising, compression, etc.)

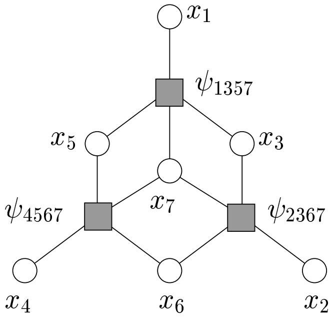
Digital image processing



- 8-bit digital image: matrix of intensity values $\{0, 1, \ldots 255\}$
- enormous redundancy in "typical" images (useful for denoising, compression, etc.)
- simplest graphical model: 2-dimensional grid or lattice

(Ising, 1923; Geman & Geman, 1984, and many others)

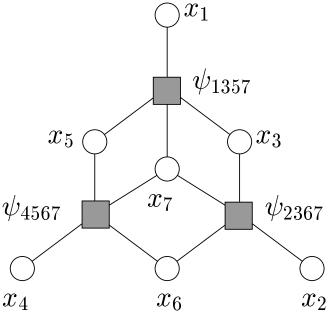
Communication and error-control coding

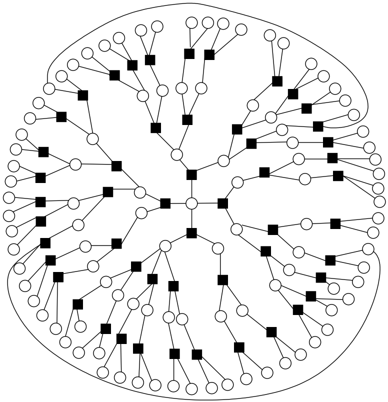


- error-control coding: introduce redundancy via parity checks

$$\psi_{1357}(x_1, x_3, x_5, x_7) = \begin{cases} 1 & \text{if } x_1 \oplus x_3 \oplus x_5 \oplus x_7 = 0, \\ 0 & \text{otherwise.} \end{cases}$$

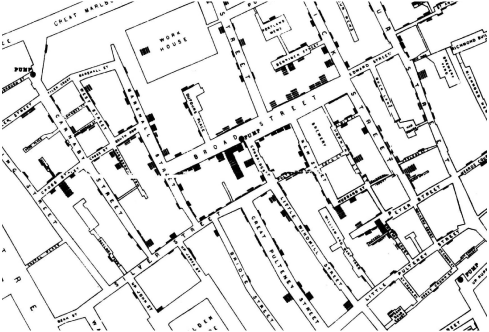
Communication and error-control coding





- error-control coding: introduce redundancy via parity checks
- state-of-the-art codes (turbo, LDPC etc.) based on "tree-like" graphs
  (Gallager, 1963; Berrou et al., 1993; Urbanke & Richardson, 2008, and many others)

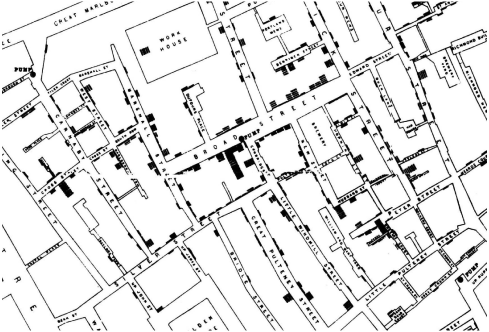
Epidemiological networks



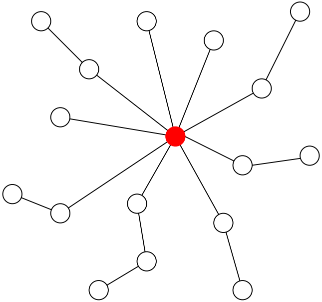
(a) Cholera epidemic (London, 1854)
Snow, 1855

- network structure associated with spread of disease

Epidemiological networks



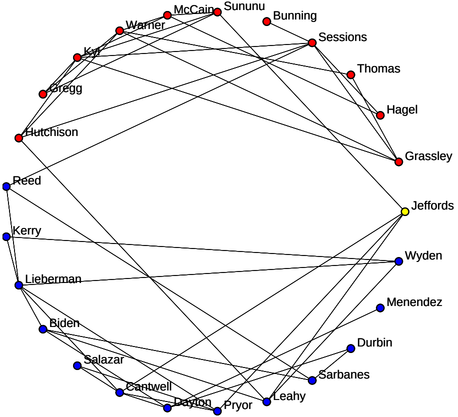
(a) Cholera epidemic (London, 1854)
Snow, 1855



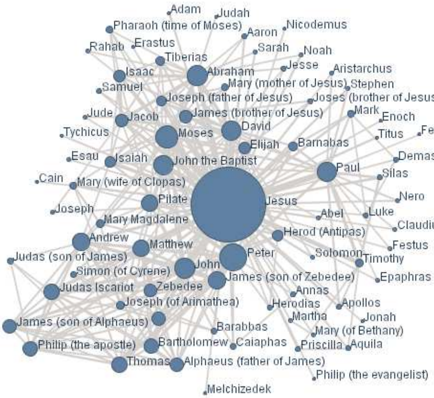
(b) “Spoke-hub” network

- network structure associated with spread of disease
- useful diagnostic information: contaminated water from Broad Street pump

Social networks

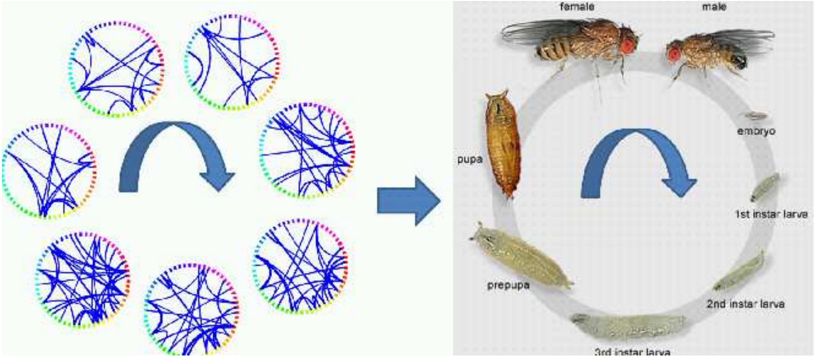


(a) US senators (2004-2006)
(Ravikumar, W. & Lafferty, 2006)



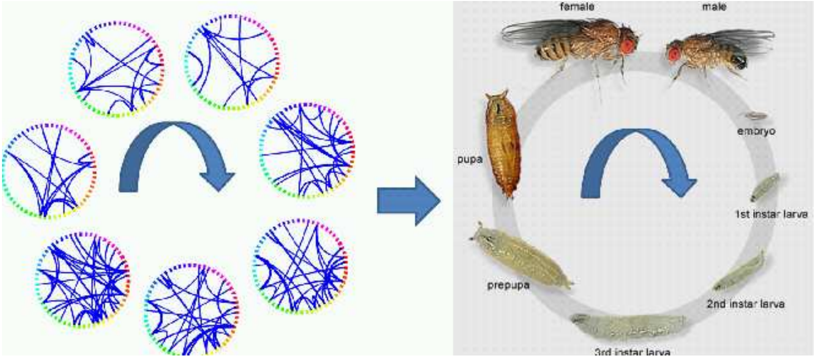
(b) Biblical characters
www.esv.org

Biological networks



- gene networks during Drosophila life cycle (Ahmed & Xing, PNAS, 2009)

Biological networks



- gene networks during Drosophila life cycle (Ahmed & Xing, PNAS, 2009)
- many other examples:
  - protein networks
  - phylogenetic trees
  - neural networks for brain-machine interfaces (e.g., Carmena et al., 2009)

Core challenges

1. Exploiting graphical structure
   - Computing most probable configurations
     - Communication: channel decoding (turbo, LDPC)
     - Image processing: denoising/deblurring
   - Inferring "hidden variables"
     - Computer vision: stereo vision, face recognition
     - Social networks: detecting cliques, party membership etc.

Core challenges

1. Exploiting graphical structure
   - Computing most probable configurations
     - Communication: channel decoding (turbo, LDPC)
     - Image processing: denoising/deblurring
   - Inferring "hidden variables"
     - Computer vision: stereo vision, face recognition
     - Social networks: detecting cliques, party membership etc.
2. Discovering graphical structure in data
   - Appropriate choice of "state variables"
     - Neuroscience: firing rates, spike counts, EEG?
     - Optical character recognition: pixels, Fourier, wavelets?
   - Learning graph structure from data
     - Graph selection: which edges are present/absent?
     - Parameters: what types of interactions?
     - Validation: reliability of fitted models?

Example: Epidemics and graphical models

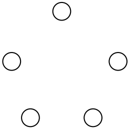
Disease status of person s: $$x_s = \begin{cases} +1 & \text{if individual } s \text{ is infected} \\ -1 & \text{if individual } s \text{ is healthy} \end{cases}$$

Example: Epidemics and graphical models

Disease status of person s: $\quad x_s = \begin{cases} +1 & \text{if individual } s \text{ is infected} \\ -1 & \text{if individual } s \text{ is healthy} \end{cases}$

(1) Independent infection

$$\mathbb{P}(x_1, \ldots, x_5) \propto \prod_{s=1}^{5} \exp(\theta_s x_s)$$

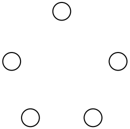


Example: Epidemics and graphical models

Disease status of person s: $$x_s = \begin{cases} +1 & \text{if individual } s \text{ is infected} \\ -1 & \text{if individual } s \text{ is healthy} \end{cases}$$

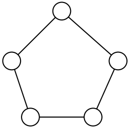
(1) Independent infection

$$\mathbb{P}(x_1, \ldots, x_5) \propto \prod_{s=1}^{5} \exp(\theta_s x_s)$$



(2) Cycle-based infection

$$\mathbb{P}(x_1, \ldots, x_5) \propto \prod_{s=1}^{5} \exp(\theta_s x_s) \prod_{(s,t) \in C} \exp(\theta_{st} x_s \, x_t)$$

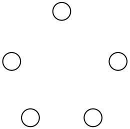


Example: Epidemics and graphical models

Disease status of person s: $\quad x_s = \begin{cases} +1 & \text{if individual } s \text{ is infected} \\ -1 & \text{if individual } s \text{ is healthy} \end{cases}$

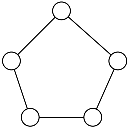
(1) Independent infection

$$\mathbb{P}(x_1, \ldots, x_5) \propto \prod_{s=1}^{5} \exp(\theta_s x_s)$$



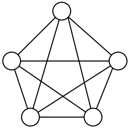
(2) Cycle-based infection

$$\mathbb{P}(x_1, \ldots, x_5) \propto \prod_{s=1}^{5} \exp(\theta_s x_s) \prod_{(s,t) \in C} \exp(\theta_{st} x_s \, x_t)$$

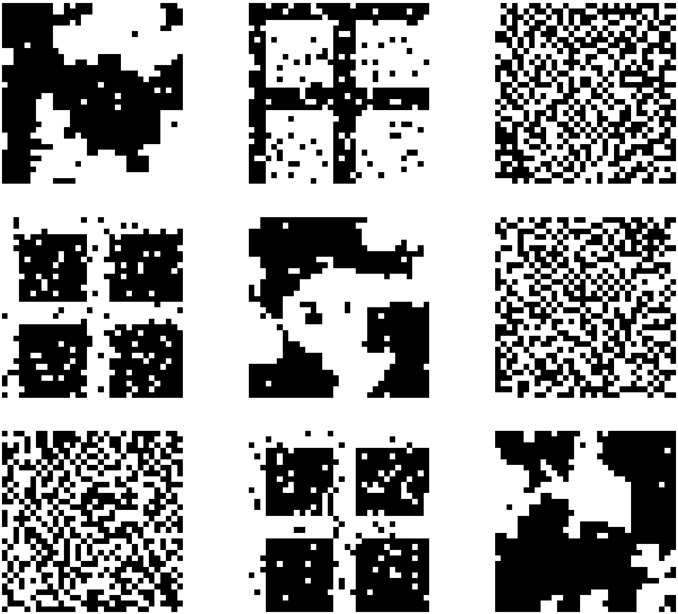


(3) Full clique infection

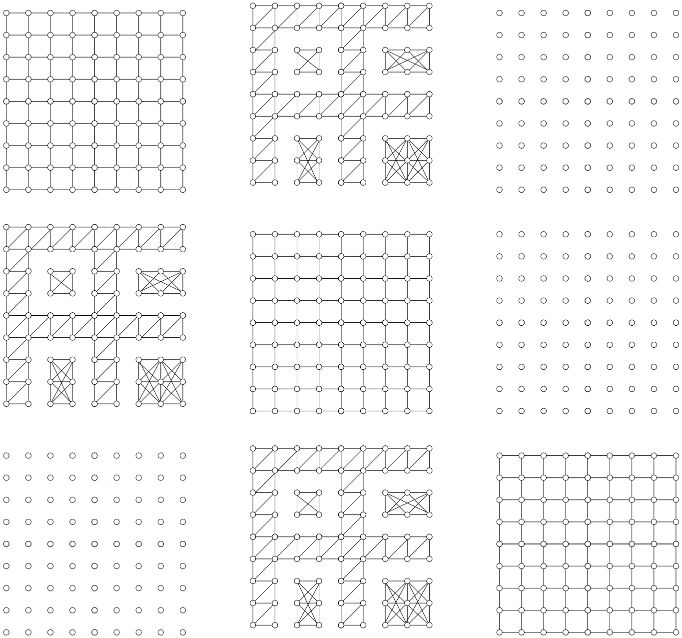
$$\mathbb{P}(x_1, \ldots, x_5) \propto \prod_{s=1}^{5} \exp(\theta_s x_s) \prod_{s \neq t} \exp(\theta_{st} x_s x_t)$$



Possible epidemic patterns



Underlying graphs

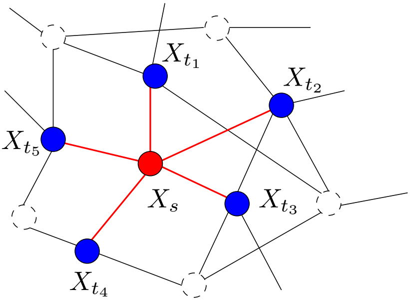


Markov property and neighborhood structure

- Markov properties encode neighborhood structure:

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

$$N(s) = \{t_1, t_2, t_3, t_4, t_5\}$$



- basis of pseudolikelihood method (Besag, 1974)
- used for Gaussian model selection (Meinshausen & Buhlmann, 2006)

Graph selection via neighborhood regression

Key: Graph recovery G equivalent to recovering neighborhood sets $N(s)$.

Method: Based on n samples:

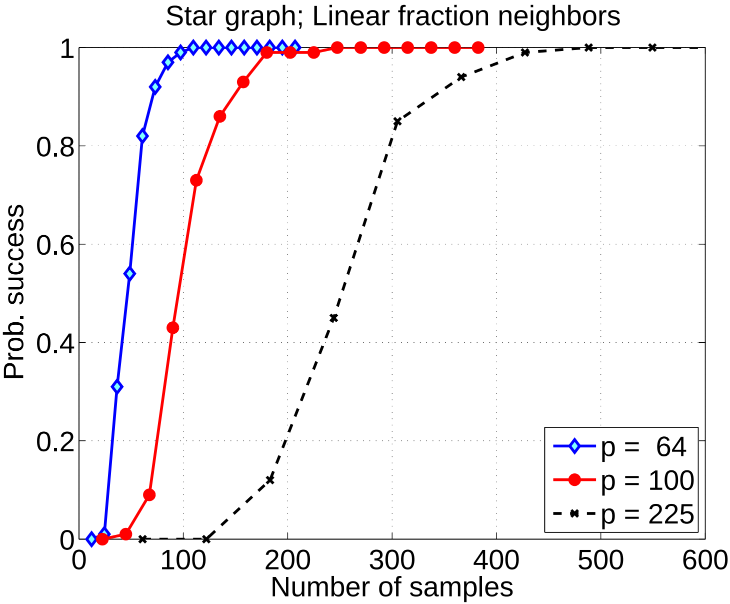
1. For each node s, predict X_s based on other variables $X_{\backslash s}$:

$$\widehat{\theta}[s] \quad := \quad \arg\min_{\theta \in \mathbb{R}^{p-1}} \Big\{ \quad -\frac{1}{n} \sum_{i=1}^{n} \underbrace{\log \mathbb{P}(\theta; X_{\backslash s}^{(i)})}_{\text{negative log likelihood}} \quad + \quad \underbrace{\lambda_n \sum_{t \in V \backslash \{s\}} |\theta_{st}|}_{\ell_1 \text{ regularization}} \Big\}$$

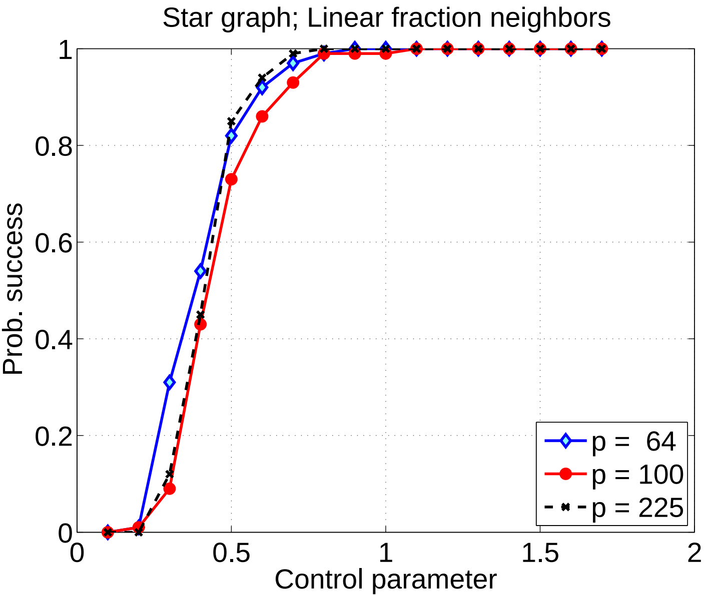
2. Estimate local neighborhood $\widehat{N}(s)$ by extracting non-zero positions within $\widehat{\theta}[s]$.

3. Combine the neighborhood estimates to form a graph estimate $\widehat{G}$.

Empirical behavior: Unrescaled plots



Empirical behavior: Appropriately rescaled



Some theory: Scaling law for graph selection

- graphs $G_{p,d}$ with p nodes and maximum degree d
- minimum absolute weight $\theta_{\min}$ on edges
- how many samples n needed to recover the unknown graph?

Theorem

Some theory: Scaling law for graph selection

- graphs $G_{p,d}$ with p nodes and maximum degree d
- minimum absolute weight $\theta_{\min}$ on edges
- how many samples n needed to recover the unknown graph?

Theorem

Achievable result: *For graph estimate $\widehat{G}$ produced by NR method:*

$$\underbrace{n > c_u \left(d^2 + 1/\theta_{\min}^2\right) \log p}_{\textit{Lower bound on sample size}} \quad \Longrightarrow \quad \underbrace{\mathbb{P}[\widehat{G} \neq G] \to 0}_{\textit{Vanishing probability of error}}$$

Some theory: Scaling law for graph selection

- graphs $G_{p,d}$ with p nodes and maximum degree d
- minimum absolute weight $\theta_{\min}$ on edges
- how many samples n needed to recover the unknown graph?

Theorem

Achievable result: *For graph estimate $\widehat{G}$ produced by NR method:*

$$\underbrace{n > c_u \left(d^2 + 1/\theta_{\min}^2\right) \log p}_{\textit{Lower bound on sample size}} \quad \Longrightarrow \quad \underbrace{\mathbb{P}[\widehat{G} \neq G] \to 0}_{\textit{Vanishing probability of error}}$$

Necessary condition: *For graph estimate $\widetilde{G}$ produced by any algorithm.*

$$\underbrace{n < c_\ell \left(d^2 + 1/\theta_{\min}^2\right) \log p}_{\textit{Upper bound on sample size}} \quad \Longrightarrow \quad \underbrace{\mathbb{P}[\widetilde{G} \neq G] \geq 1/2}_{\textit{Constant probability of error}}$$

Illustration: Social network of US senators

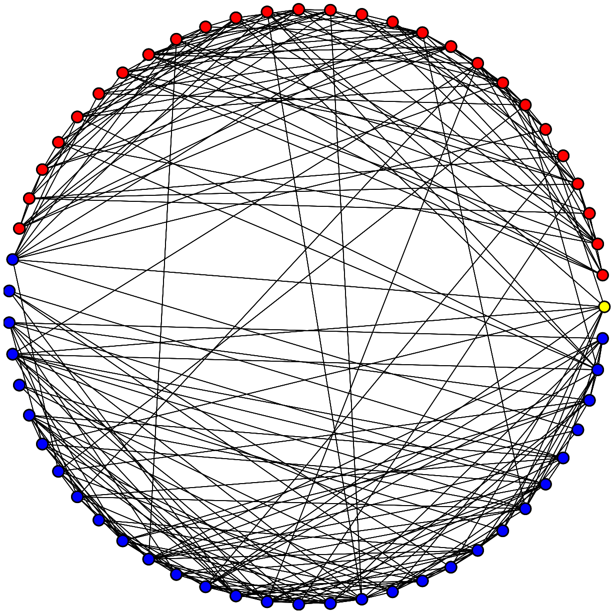
- originally studied by Bannerjee, Aspremont and El Ghaoui (2008)
- discrete data set of voting records for $p = 100$ senators:

$$X_{ij} = \begin{cases} +1 & \text{if senator } i \text{ voted yes on bill } j \\ -1 & \text{otherwise.} \end{cases}$$

- full data matrix $X \in \mathbb{R}^{n \times p}$ with $n = 542$:

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ X_{31} & X_{32} & \cdots & X_{3p} \\ \vdots & \ldots & \ldots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

Estimated senator network (subgraph of 55)



Summary

- graphical models: one framework for capturing structure in high-dimensional data sets
  - classical history....
  - increasingly relevant in era of massive data sets

Summary

- graphical models: one framework for capturing structure in high-dimensional data sets
  - classical history....
  - increasingly relevant in era of massive data sets

- various areas to be further explored:
  - other priors over graph spaces
  - dynamic graph models
  - mixed modality graphs (e.g., switching Markov models)
  - inferring causality
  - theory for message-passing on “non-tree-like” graphs

Summary

- graphical models: one framework for capturing structure in high-dimensional data sets
  - classical history....
  - increasingly relevant in era of massive data sets

- various areas to be further explored:
  - other priors over graph spaces
  - dynamic graph models
  - mixed modality graphs (e.g., switching Markov models)
  - inferring causality
  - theory for message-passing on "non-tree-like" graphs

- interactions between graphs and other signal structures
  - graphs and sparse signals: e.g., Cevher, Hegde, Duarte & Baraniuk, 2009
  - graphs and manifolds: e.g., Belkin et al., 2009