Learning Scale-Free Networks by Reweighted l_1 Regularization

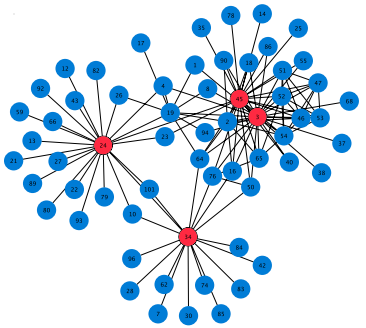
Qiang Liu, Alexander Ihler

UC Irvine

AI & Statistics 2011

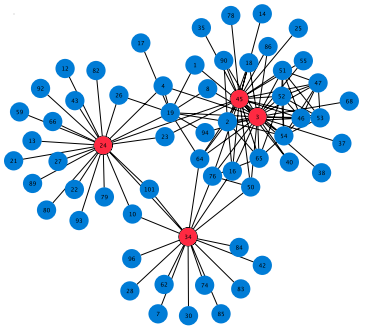
High Dimensional Structure Learning

- Problem: Estimating network structure of MRFs or Bayesian nets

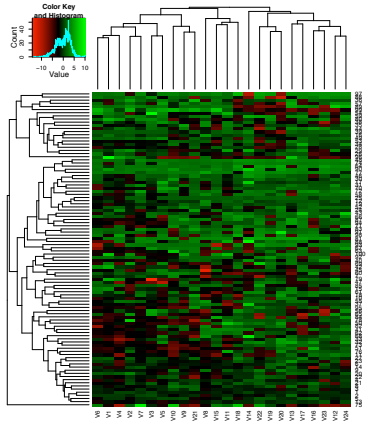


High Dimensional Structure Learning

- Problem: Estimating network structure of MRFs or Bayesian nets

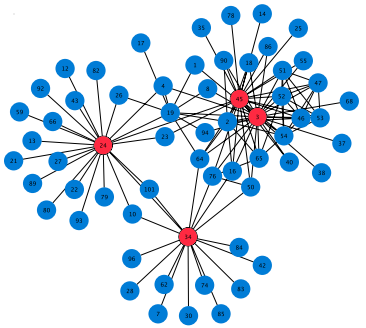


- Applications:
  - Biology (e.g., gene regulatory networks, protein networks)
  - Social networks (e.g., senate networks)

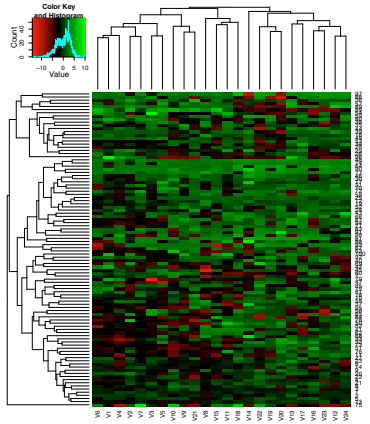


High Dimensional Structure Learning

- Problem: Estimating network structure of MRFs or Bayesian nets



- Applications:
  - Biology (e.g., gene regulatory networks, protein networks)
  - Social networks (e.g., senate networks)



- Challenge: the dimensionality (p) is much larger than the number of data (n), i.e.,

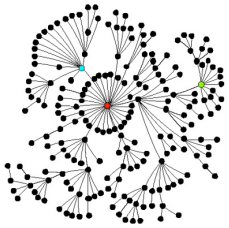
$$n \ll p$$

Prior Information is important

- Prior information is needed to help on learning
- Prior information:
  - System-specific information (expensive): 💲
    - motif map, Chip-Chip
  - Universal information (free lunch): 🙂
    - Sparsity (l_1 -regularization)

Prior Information is important

- Prior information is needed to help on learning
- Prior information:
  - System-specific information (expensive):
    - motif map, Chip-Chip
  - Universal information (free lunch):
    - Sparsity (l_1 -regularization)
  - **Complex network** properties:
    - scale free networks, hubs
    - transitivity



(The rich get richer)

(Friends of friends are friends)

- We incorporate **scale freeness** as prior for learning networks

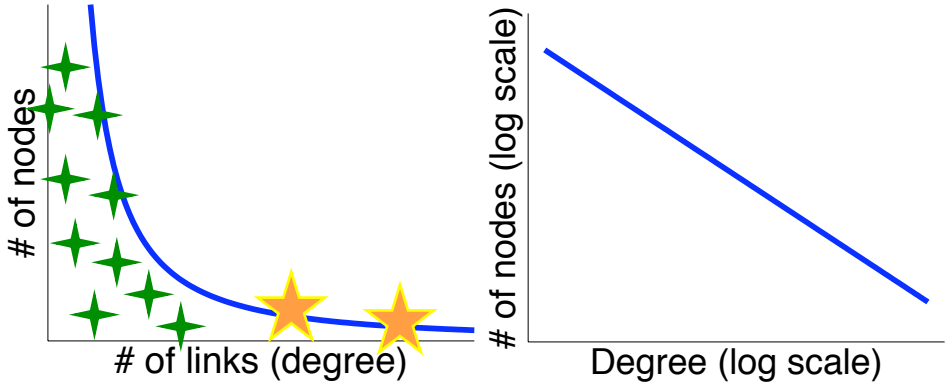
Scale-Free Networks

- The degree distribution follows a power law:

$$p(d) \sim d^{-\gamma}$$

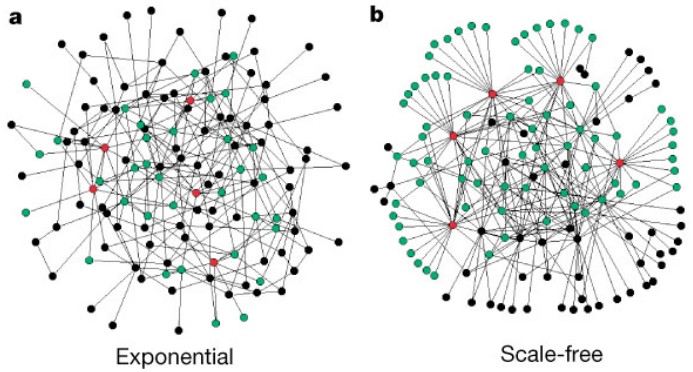
d: degree (number of edges) of the nodes.

- Log-log plot



- Heavy tail, hubs

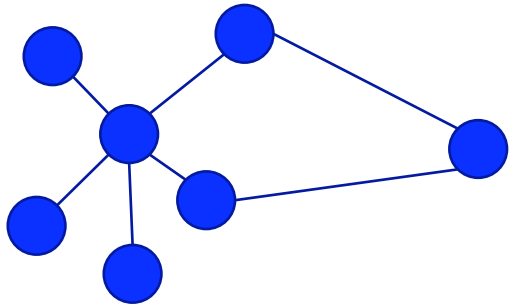
- Applications:
  - Biology networks
  - social networks
  - internet
  - citation networks



http://www.nature.com/nature/journal/v406/n6794/fig_tab/406378a0_F1.html

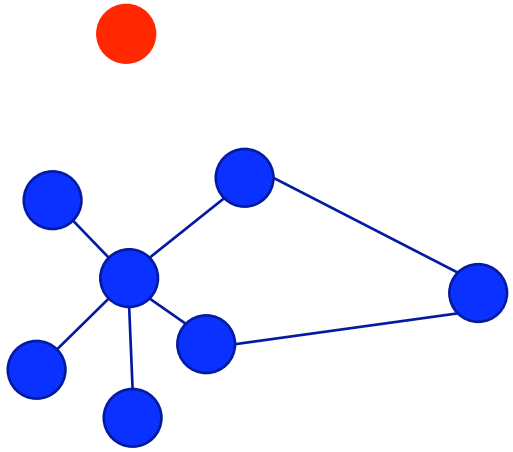
Barabási–Albert (B-A) model

- generates random scale-free networks using a **preferential attachment mechanism**
- Barabási–Albert model:
  - begins with an initial network of m_0 nodes
  - New nodes are connected to existing nodes with a probability that is proportional to the current degree of the existing nodes



Barabási–Albert (B-A) model

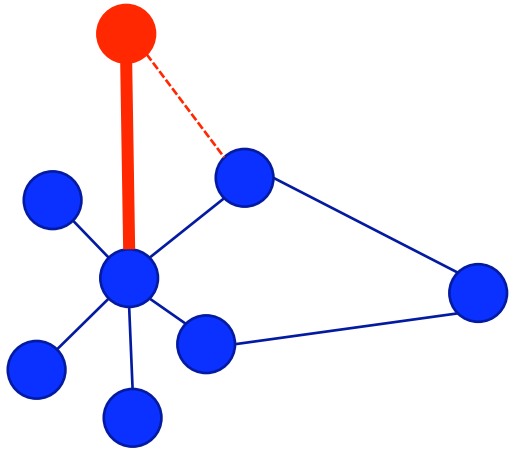
- generates random scale-free networks using a **preferential attachment mechanism**
- Barabási–Albert model:
  - begins with an initial network of m_0 nodes
  - New nodes are connected to existing nodes with a probability that is proportional to the current degree of the existing nodes



Barabási–Albert (B-A) model

- generates random scale-free networks using a **preferential attachment mechanism**
- Barabási–Albert model:
  - begins with an initial network of m_0 nodes
  - New nodes are connected to existing nodes with a probability that is proportional to the current degree of the existing nodes

The rich get richer!



Gaussian Markov Random Field

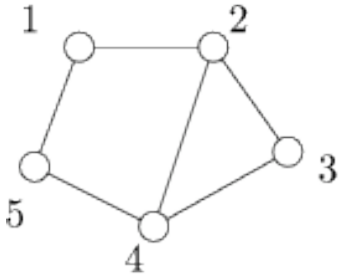
$$p(x) \propto \exp(-\frac{1}{2}x'\Omega x)$$
$$= \prod_{ij} \exp(-\frac{1}{2}w_{ij}x_i x_j)$$

$\Omega = \Sigma^{-1}$

$x = [x_1, x_2, \cdots, x_p]'$ $\Omega = [\, w_{ij}]$

(Factor graph !)

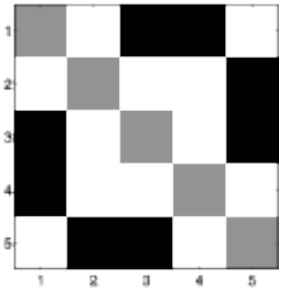
- Ω reflects the structure of GMRF



- Network inference:

  Given data $\mathbf{x} = [x^1, x^2, \cdots, x^n]$

  estimate the non-zero patterns of Ω



- High dimensional: $n \ll p$

Wainwright 08

Useful Properties of Gaussian

- Partial correlation:

$$\rho_{ij} = \text{corr}(x_i, x_j | x_{\neg ij}) = -\frac{w_{ij}}{\sqrt{w_{ii}}\sqrt{w_{jj}}}$$

- Auto-regression:

$$x_i = \sum_{j \neq i} \beta_{ij} x_j + \delta_i$$

$$\text{where } \beta_{ij} = -\frac{w_{ij}}{w_{ii}}, \ \delta_i \sim \mathcal{N}(0, 1/w_{ii}).$$

- w_{ij}, ρ_{ij}, β_{ij} all represents the structure of the network structures.

l_1-based methods

- Neighborhood selection (Meinshausen and Buhlmann, 2006)

$$\min_{\beta} \frac{1}{2}||x_i - \sum_{j \neq i} \beta_{ij} x_j||^2 + \lambda \sum_{j \neq i} |\beta_{ij}|$$

  - Computationally efficient, consistent
  - β_{ij} and β_{ji} maybe inconsistent.
  - AND / OR rules

l_1-based methods

- Neighborhood selection (Meinshausen and Buhlmann, 2006)

$$\min_{\beta} \frac{1}{2}||x_i - \sum_{j \neq i} \beta_{ij} x_j||^2 + \lambda \sum_{j \neq i} |\beta_{ij}|$$

- Joint regression sparse methods (Peng et al., 2009)

$$\min_{\rho, w_{jj}} \frac{1}{2} \sum_{i}^{p} \mu_i ||x_i - \sum_{j \neq i} \rho_{ij} \sqrt{\frac{\omega_{jj}}{\omega_{ii}}} x_j||^2 + \lambda \sum_{j \neq i} |\rho_{ij}| \qquad s.t., \rho_{ij} = \rho_{ji}$$

  - Similar to pseudolikelihood (Rocha et al.08, Friedman et al.10)

l_1-based methods

- Neighborhood selection (Meinshausen and Buhlmann, 2006)

$$\min_{\beta} \frac{1}{2}||x_i - \sum_{j \neq i} \beta_{ij} x_j||^2 + \lambda \sum_{j \neq i} |\beta_{ij}|$$

- Joint regression sparse methods (Peng et al., 2009)

$$\min_{\rho, w_{jj}} \frac{1}{2} \sum_{i}^{p} \mu_i ||x_i - \sum_{j \neq i} \rho_{ij} \sqrt{\frac{\omega_{jj}}{\omega_{ii}}} x_j||^2 + \lambda \sum_{j \neq i} |\rho_{ij}| \qquad s.t., \rho_{ij} = \rho_{ji}$$

  - Similar to pseudolikelihood (Rocha et al.08, Friedman et al.10)

- Graphical lasso (likelihood) (Banerjee et al 2008; Friedman et al., 2008, etc)

$$\min_{\Omega \succ 0} \operatorname{tr}(S\Omega) - \log\det(\Omega) + \lambda \sum_{ij} |\omega_{ij}|$$

S: empirical covariance matrix

l_1-based methods

- They are all in the form of

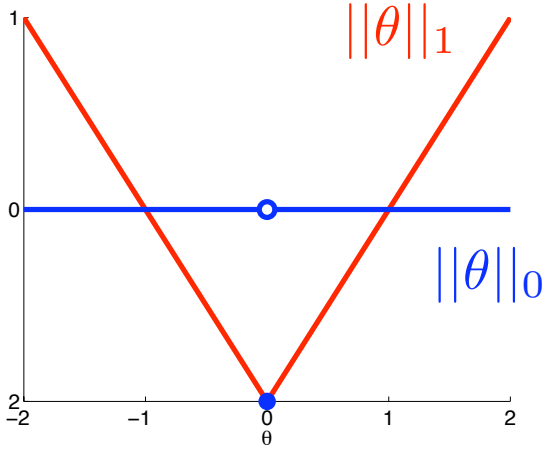
$$\min_{\Theta} S(\mathbf{x}|\Theta) + \alpha||\Theta||_1$$

where θ_{ij} can be w_{ij}, ρ_{ij}, β_{ij}, depending on the methods.

- More about l_1-regularization:
  - Can be treated as a **convex surrogate** of l_0 norm:

$$||\Theta||_0 = \sum_{ij} I(\theta_{ij} = 0)$$

l_0 is NP-hard!



  - Can be understood as a Laplacian prior:

$$p(\theta_{ij}) \propto \exp(-\alpha|\theta_{ij}|)$$

l_1 is not good for scale free networks

- Theoretically
  - Meinshausen & Buhlmann 2006; Wainwright 2009 $n = \Omega(d_{\max} \log p)$
  - Ravikumar et al 2008 $n = \Omega(d_{\max}^2 \log p)$

    $d_{\max}$: maximal degree in the network

- Experimentally
  - Schafer & Strimmer 2005, Ravikumar et al 2008, etc

Power Law Regularization

Power law distribution:

$$p(d_i) \propto d_i^{-\alpha} \quad (\forall i \in V) \qquad d_i\text{: degree of the } i\text{-th node}$$

Degree d_i is l_0 norm of i-th column of Θ:

$$d_i = ||\theta_{\neg i}||_0 = \sum_{j \neq i} I(\theta_{ij} \neq 0)$$

Prior of Θ : $$\log p(\Theta) = -\alpha \sum_i \log(||\theta_{\neg i}||_0) + const$$

Regularized score function:

$$\min_{\Theta} S(\mathbf{x}|\Theta) + \alpha \sum_i \log(||\theta_{\neg i}||_0)$$

Power Law Regularization

Power law distribution:

$$p(d_i) \propto d_i^{-\alpha} \quad (\forall i \in V) \qquad d_i\text{: degree of the } i\text{-th node}$$

Degree d_i is l_0 norm of i-th column of Θ:

$$d_i = ||\theta_{\neg i}||_0 = \sum_{j \neq i} I(\theta_{ij} \neq 0)$$

Prior of Θ : $$\log p(\Theta) = -\alpha \sum_i \log(||\theta_{\neg i}||_0) + const$$

Regularized score function:

l_0 combinatorial problem

Hard!

$$\min_{\Theta} S(\mathbf{x}|\Theta) + \alpha \sum_i \log(||\theta_{\neg i}||_0)$$

l_1 Relaxation

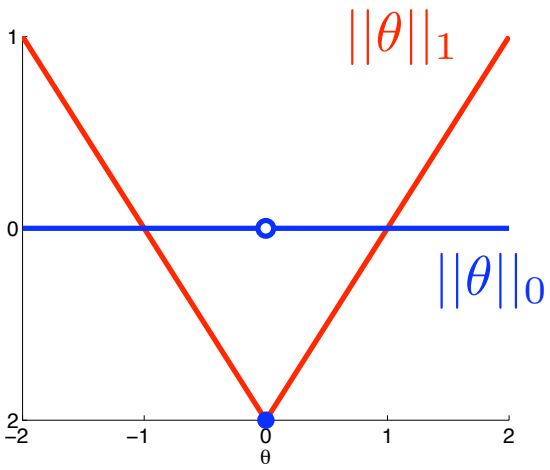
Use l_1 norm as the surrogate of l_0:

$$||\theta_{\neg i}||_0 \rightarrow ||\theta_{\neg i}||_1$$

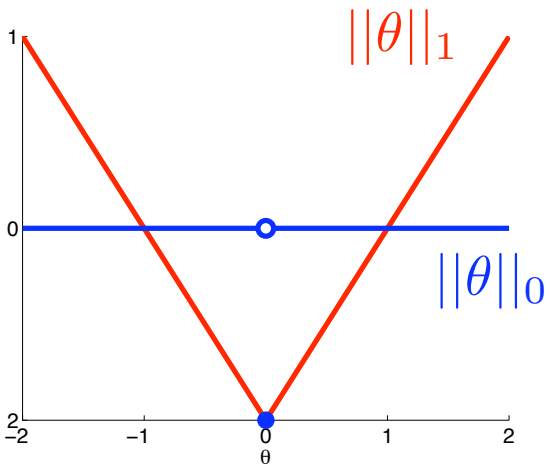
where $||\theta_{\neg i}||_1 = \sum_{j \neq i} |\theta_{ij}|$.

The new regularized score function:

$$\min_{\Theta} S(\mathbf{x}|\Theta) + \alpha \sum_i \log(||\theta_{\neg i}||_1)$$



l_1 Relaxation



Use l_1 norm as the surrogate of l_0:

$$||\theta_{\neg i}||_0 \rightarrow ||\theta_{\neg i}||_1$$

where $||\theta_{\neg i}||_1 = \sum_{j \neq i} |\theta_{ij}|$.

The new regularized score function:

$$\min_{\Theta} S(\mathbf{x}|\Theta) + \alpha \sum_i \log(||\theta_{\neg i}||_1)$$

log-norm distribution: $\log(z) \sim \mathcal{N}(\mu, \sigma^2)$

$$\begin{aligned} \log p(z) &= -\log(z) - (\log(z) - \mu)^2/2\sigma^2 + \text{const} \\ &\approx -\log(z) + \text{const} \qquad (\sigma \rightarrow \infty) \end{aligned}$$

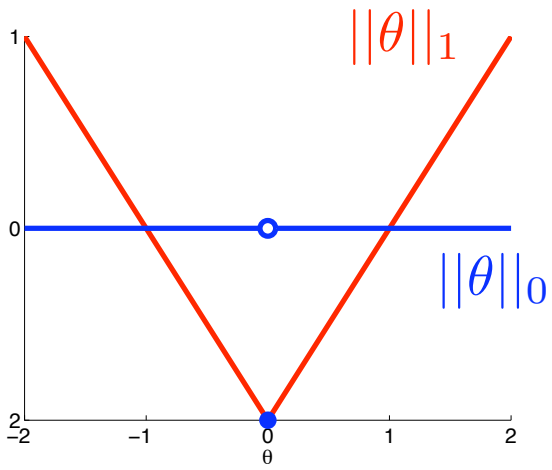
log-norm prior on $||\theta_{\neg i}||_1^{\alpha}$ **!**

l_1 Relaxation

Use l_1 norm as the surrogate of l_0:

$$||\theta_{\neg i}||_0 \rightarrow ||\theta_{\neg i}||_1$$

where $||\theta_{\neg i}||_1 = \sum_{j \neq i} |\theta_{ij}|$.



The new regularized score function:

$$\min_{\Theta} S(\mathbf{x}|\Theta) + \alpha \sum_i \log(||\theta_{\neg i}||_1 + \epsilon) + \beta \sum_i |\theta_{ii}|$$

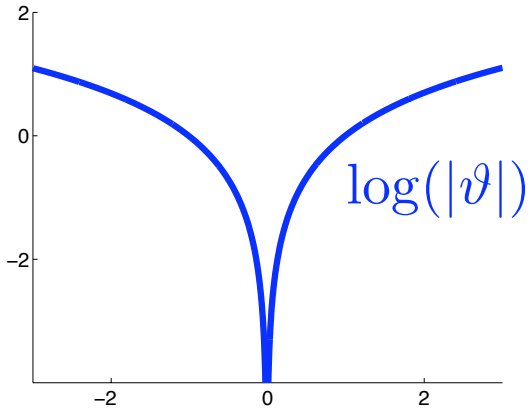
$\epsilon \geq 0$ prevents singularity

diagonal regularization

How to solve the optimization problem?

$$\min_{\Theta} S(\mathbf{x}|\Theta) + \alpha \sum_i \log(||\theta_{\neg i}||_1 + \epsilon) + \beta \sum_i |\theta_{ii}|$$

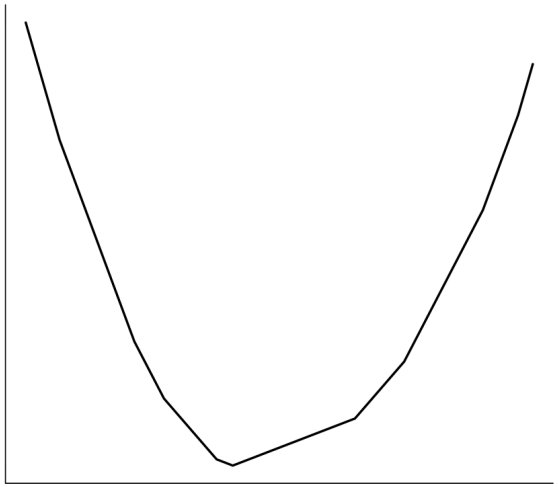
- Difficulty:
  - Non-differentiable
  - Non-convex



- It turns out that we can solve it by
  - a sequence of standard l_1 optimization problems,
  - coefficients updated by a "***preferential attachment***" mechanism

A MM algorithm (Hunter & Lange 2000)

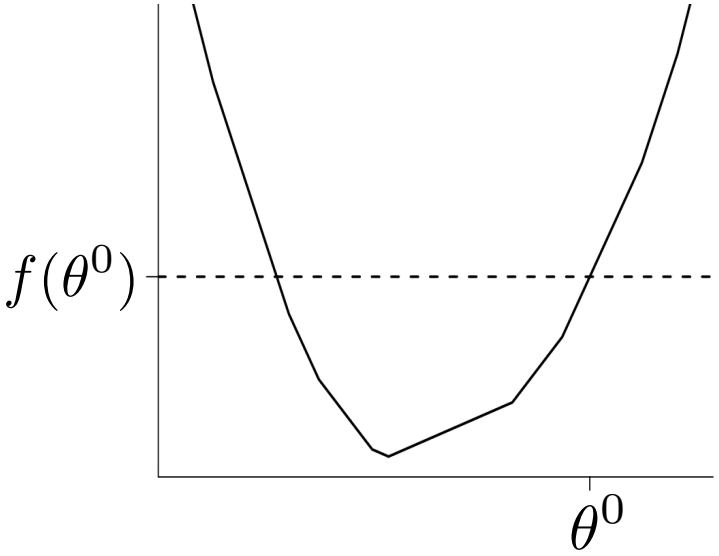
- A generalization of EM
- Approximate the objective using a upper bound that is tangent to it (**Majorizing** the objective)
- **Minimize** the upper bound
- **Monotonically** decreases the objective



Picture from Hunter's slides

A MM algorithm (Hunter & Lange 2000)

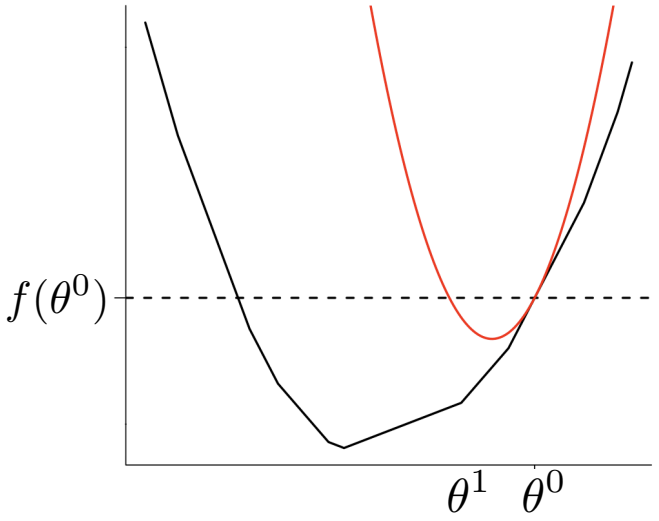
- A generalization of EM
- Approximate the objective using a upper bound that is tangent to it (**Majorizing** the objective)
- **Minimize** the upper bound
- **Monotonically** decreases the objective



Picture from Hunter's slides

A MM algorithm (Hunter & Lange 2000)

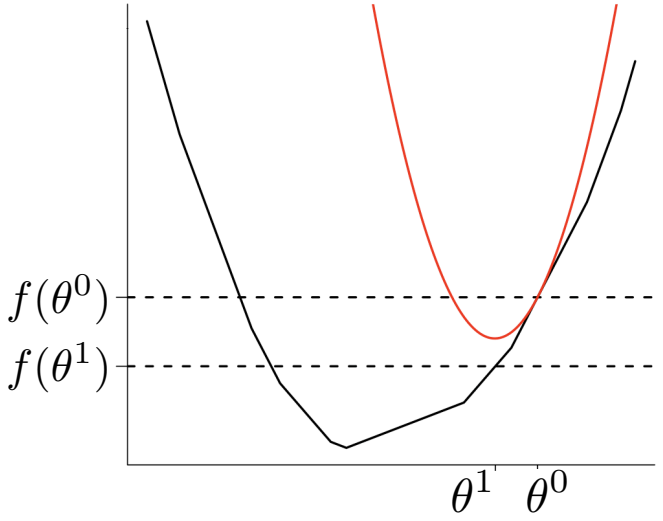
- A generalization of EM
- Approximate the objective using a upper bound that is tangent to it (**Majorizing** the objective)
- **Minimize** the upper bound
- **Monotonically** decreases the objective



Picture from Hunter's slides

A MM algorithm (Hunter & Lange 2000)

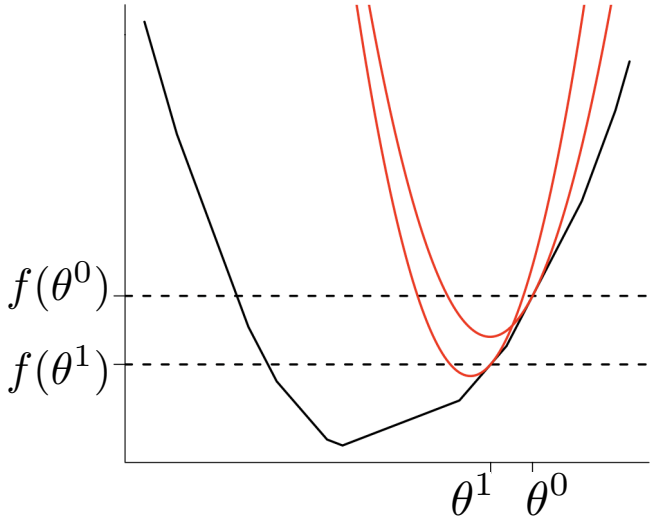
- A generalization of EM
- Approximate the objective using a upper bound that is tangent to it (**Majorizing** the objective)
- **Minimize** the upper bound
- **Monotonically** decreases the objective



Picture from Hunter's slides

A MM algorithm (Hunter & Lange 2000)

- A generalization of EM
- Approximate the objective using a upper bound that is tangent to it (**Majorizing** the objective)
- **Minimize** the upper bound
- **Monotonically** decreases the objective

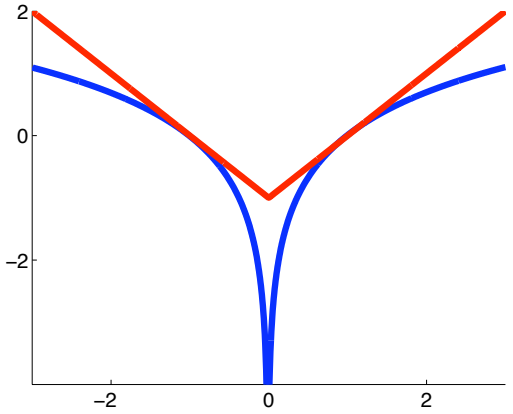


$$f(\theta^1) \leq f(\theta^0)$$

Picture from Hunter's slides

A MM algorithm (Hunter & Lange 2000)

- A generalization of EM
- Approximate the objective using a upper bound that is tangent to it (**Majorizing** the objective)
- **Minimize** the upper bound
- **Monotonically** decreases the objective



$$\alpha \sum_i \log(||\theta_{\neg i}||_1 + \epsilon) \leq \alpha \sum_{i \neq j} \lambda_{ij} |\theta_{ij}| + \text{const}$$

$$\text{where } \lambda_{ij} = \alpha \left(\frac{1}{||\theta_{\neg i}||_1 + \epsilon} + \frac{1}{||\theta_{\neg j}||_1 + \epsilon} \right)$$

λ_{ij}'s value dependents on θ_{ij}

Linear approximation

Reweighted l_1-based Optimization

Initialize $\Theta^0 = I$. (Θ^n: estimation at n-th iteration)

Iterate

$$\lambda_{ij} = \alpha \left(\frac{1}{||\theta^n_{\neg i}||_1 + \epsilon} + \frac{1}{||\theta^n_{\neg j}||_1 + \epsilon} \right)$$

preferential attachment

$$\Theta^{n+1} = \arg\min_{\Theta} S(\mathbf{x}|\Theta) + \sum_{i \neq j} \lambda_{ij} |\theta_{ij}| + \beta \sum_i |\theta_{ii}|$$

until converge

Reweighted l_1-based Optimization

Initialize $\Theta^0 = I$. (Θ^n: estimation at n-th iteration)

Iterate

preferential attachment

$$\lambda_{ij} = \alpha \left(\frac{1}{||\theta^n_{\neg i}||_1 + \epsilon} + \frac{1}{||\theta^n_{\neg j}||_1 + \epsilon} \right)$$

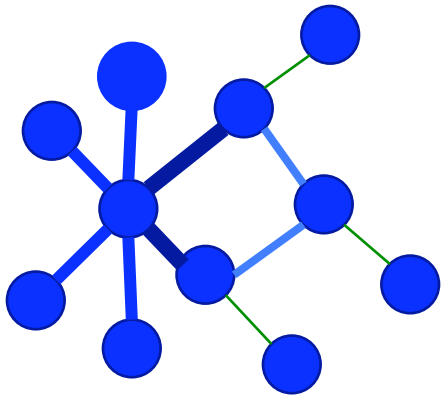
$$\Theta^{n+1} = \arg\min_{\Theta} S(\mathbf{x}|\Theta) + \sum_{i \neq j} \lambda_{ij} |\theta_{ij}| + \beta \sum_i |\theta_{ii}|$$

until converge

- Preferential attachment!

preferential attachment

$d_i \approx ||\theta_{\neg i}||_1 \Uparrow \Rightarrow \lambda_{ij} \Downarrow \Rightarrow d_i \Uparrow$



The rich get richer!

Reweighted l_1 -based Optimization

- Monotonically increasing, work for all score functions

Reweighted l_1-based Optimization

- Monotonically increasing, work for all score functions

- First iteration is regular l_1 regularization with $\lambda_{ij} = \frac{2\alpha}{\epsilon}$

$$\lambda_{ij} = \alpha\left(\frac{1}{||\theta_{\neg i}||_1 + \epsilon} + \frac{1}{||\theta_{\neg j}||_1 + \epsilon}\right) \quad \Rightarrow \quad \lambda_{ij} = \frac{2\alpha}{\epsilon} \text{ when } \Theta = I$$

Reweighted l_1-based Optimization

- Monotonically increasing, work for all score functions
- First iteration is regular l_1 regularization with $\lambda_{ij} = \frac{2\alpha}{\epsilon}$

$$\lambda_{ij} = \alpha\left(\frac{1}{||\theta_{\neg i}||_1 + \epsilon} + \frac{1}{||\theta_{\neg j}||_1 + \epsilon}\right) \quad \Rightarrow \quad \lambda_{ij} = \frac{2\alpha}{\epsilon} \text{ when } \Theta = I$$

- Usually gain most improvement in the first few iterations

Reweighted l_1-based Optimization

- Monotonically increasing, work for all score functions
- First iteration is regular l_1 regularization with $\lambda_{ij} = \frac{2\alpha}{\epsilon}$

$$\lambda_{ij} = \alpha\left(\frac{1}{||\theta_{\neg i}||_1 + \epsilon} + \frac{1}{||\theta_{\neg j}||_1 + \epsilon}\right) \quad \Rightarrow \quad \lambda_{ij} = \frac{2\alpha}{\epsilon} \text{ when } \Theta = I$$

- Usually gain most improvement in the first few iterations
- Similar methods were used for other purposes:
  - Adaptive lasso (Zhou 06) , optimizing SCAD penalty (Fan et al 09 etc)
  - Compressive sensing (Candes et al 08)

Other works on structured Regularization

- Variable selection (group lasso (Yuan & Lin 2006), fused lasso (Tibshirani et al 2005), etc.)

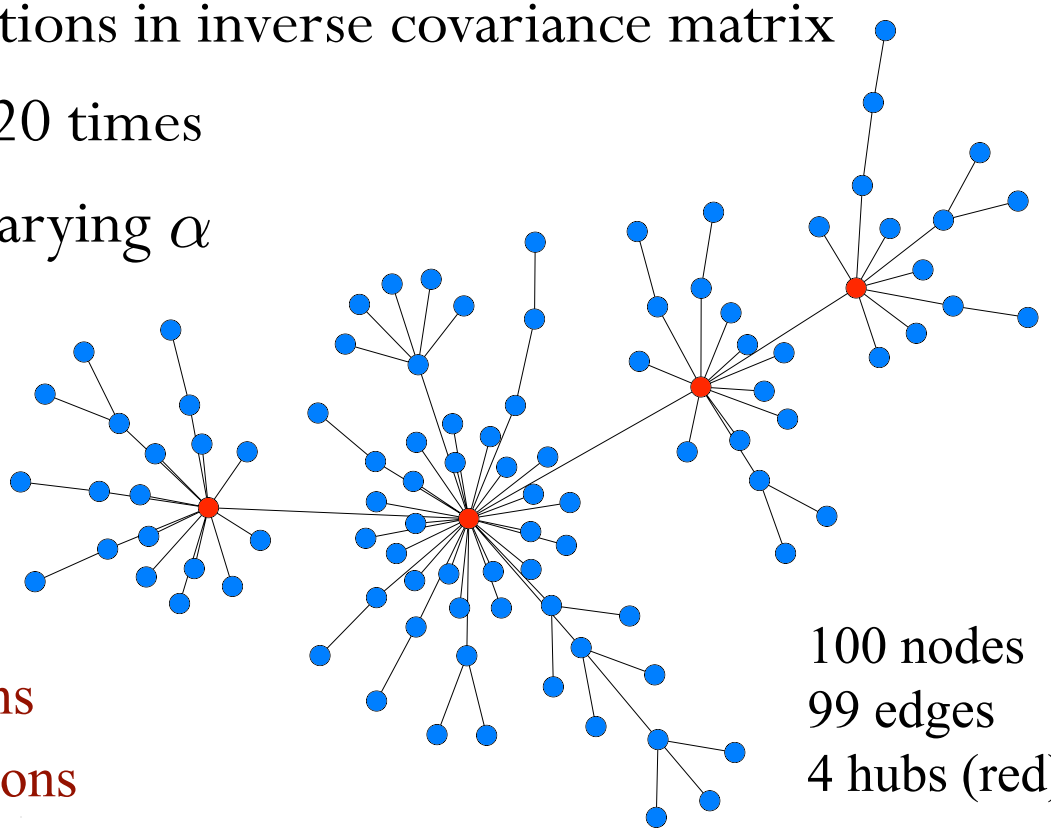
- Structure Learning (Duchi et al 2008; Schmidt et al 2009; Marlin et al 2009; Marlin & Murphy 2009; Jacob et al 2009; Peng et. al 2009, Friedman et.al 2010 …)

Mixed norm ℓ_{pq}: $||\Theta||_{pq} = [\sum_{s \subset \mathcal{S}} (||\Theta_s||_p)^q]^{1/q}$ (usually $\ell_{2,1}$, $\ell_{\infty,1}$)

"Node sparsity" (Peng et al 2009, Friedman et al 2010)

Experiments (simulated scale free network)

- Simulate scale free network using B-A model
- Uniform partial correlations in inverse covariance matrix
- n=100, p=100, repeat 20 times
- Regularization path by varying α

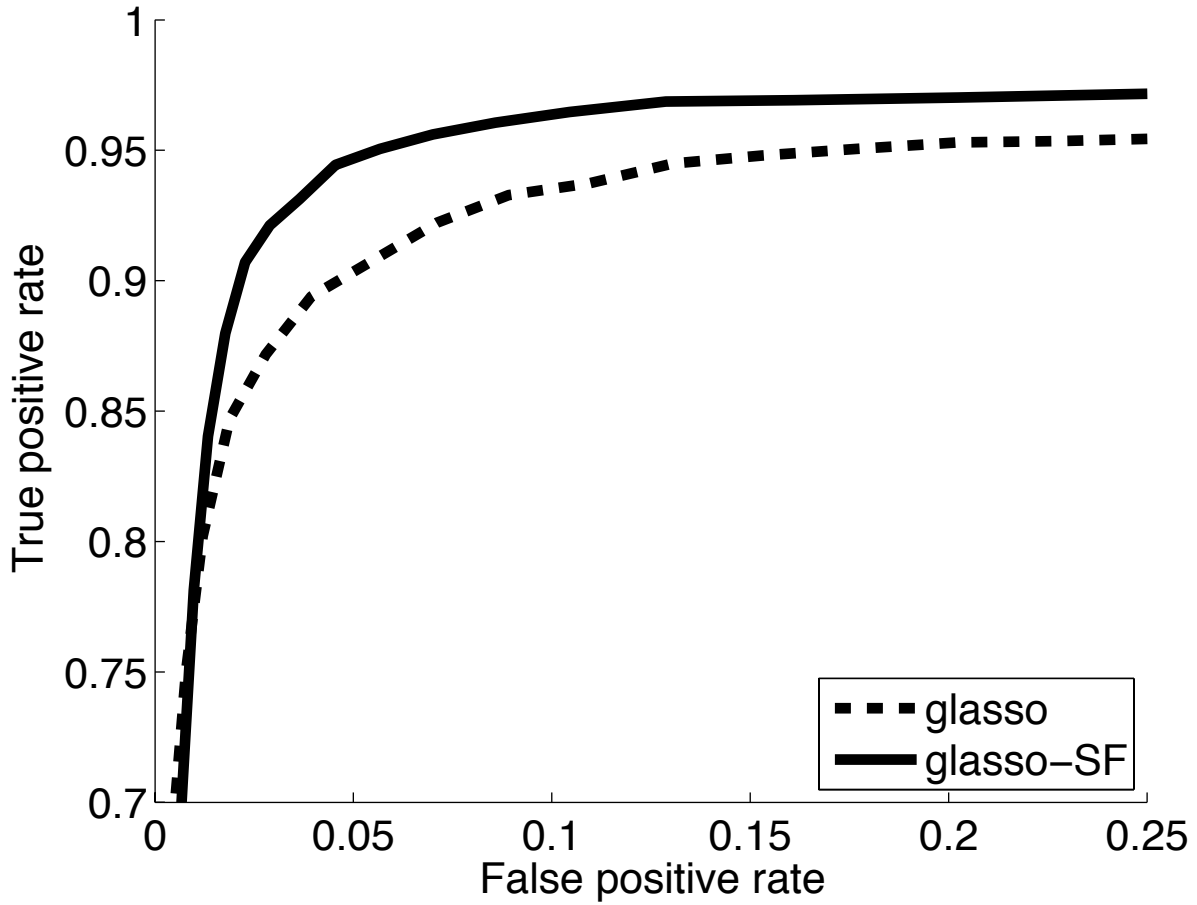


100 nodes
99 edges
4 hubs (red)

Test on all the three score functions
with l_1 and power law regularizations

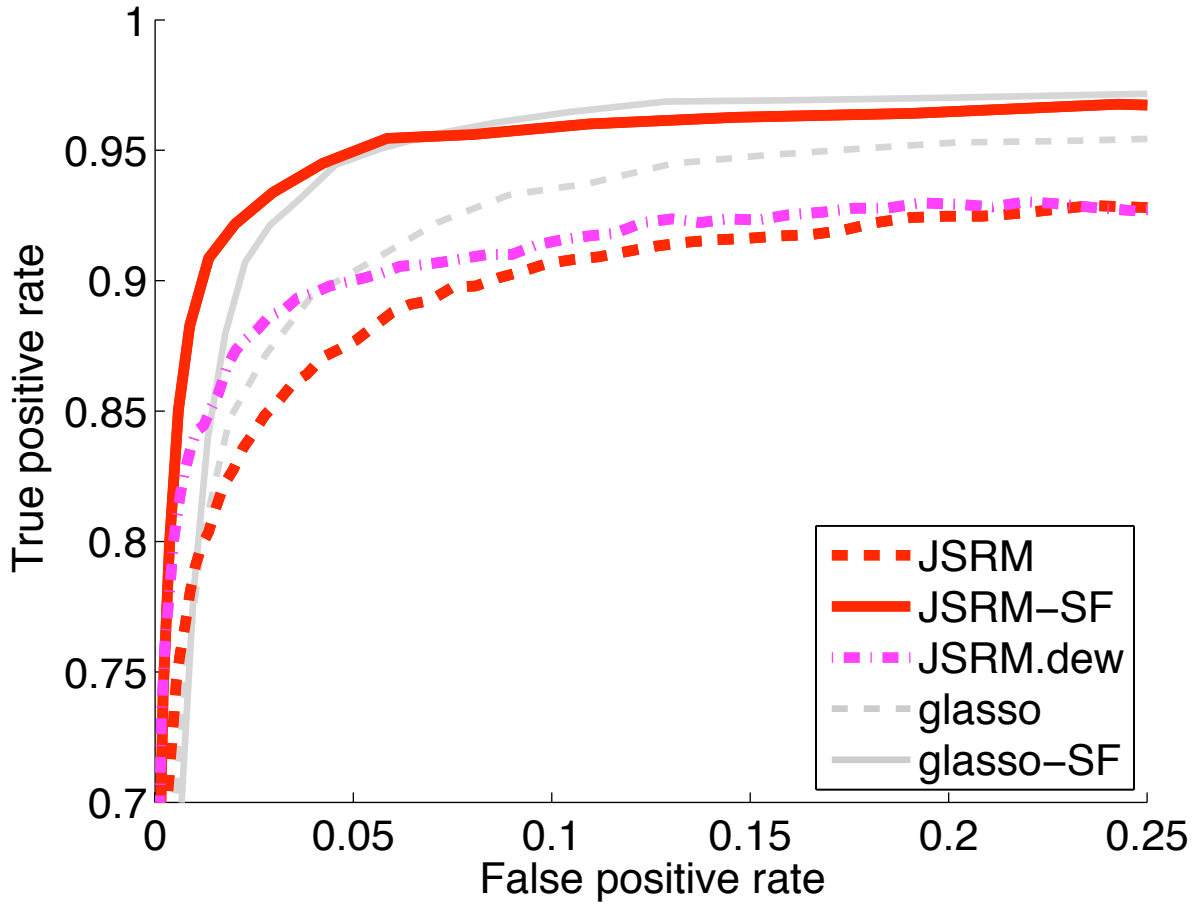
ROC curves

Score function = graphical lasso



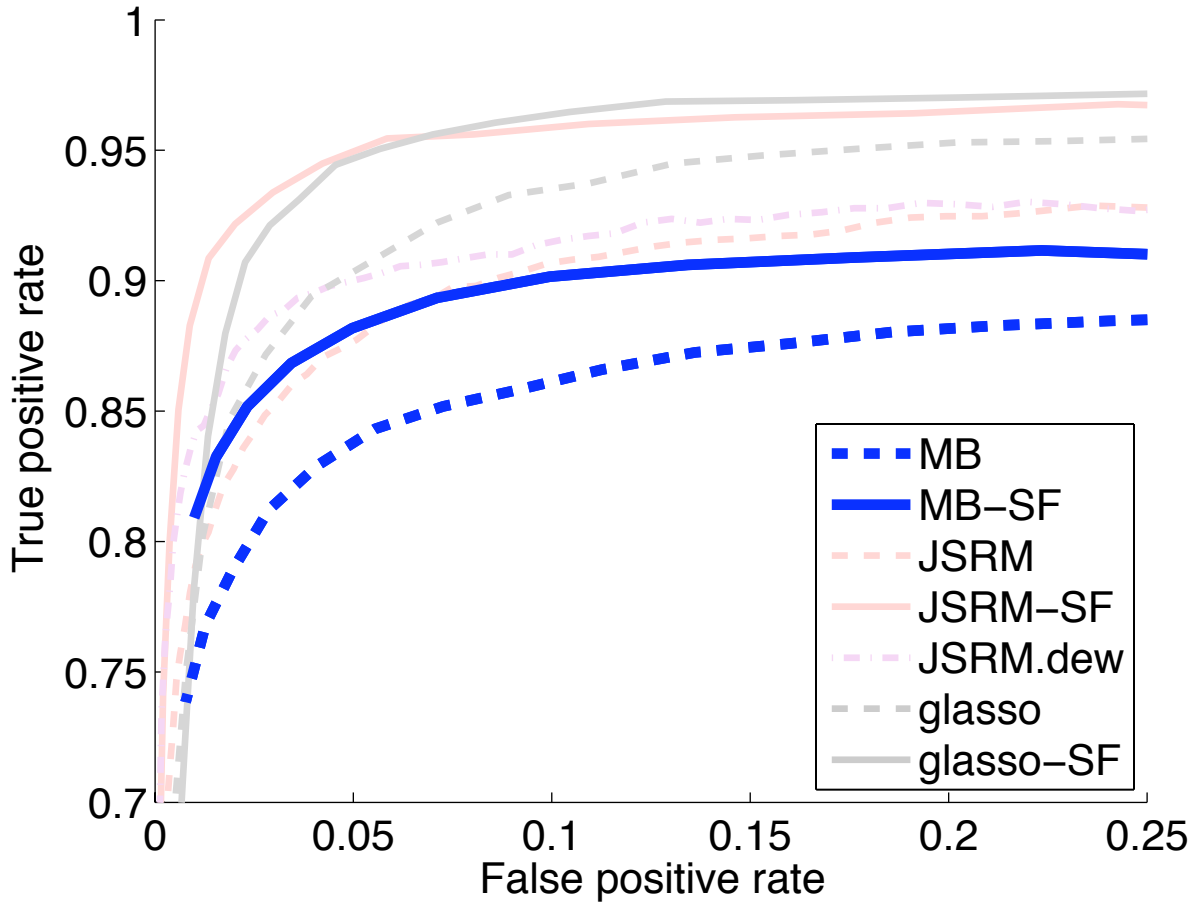
ROC curves

Score function = JSRM



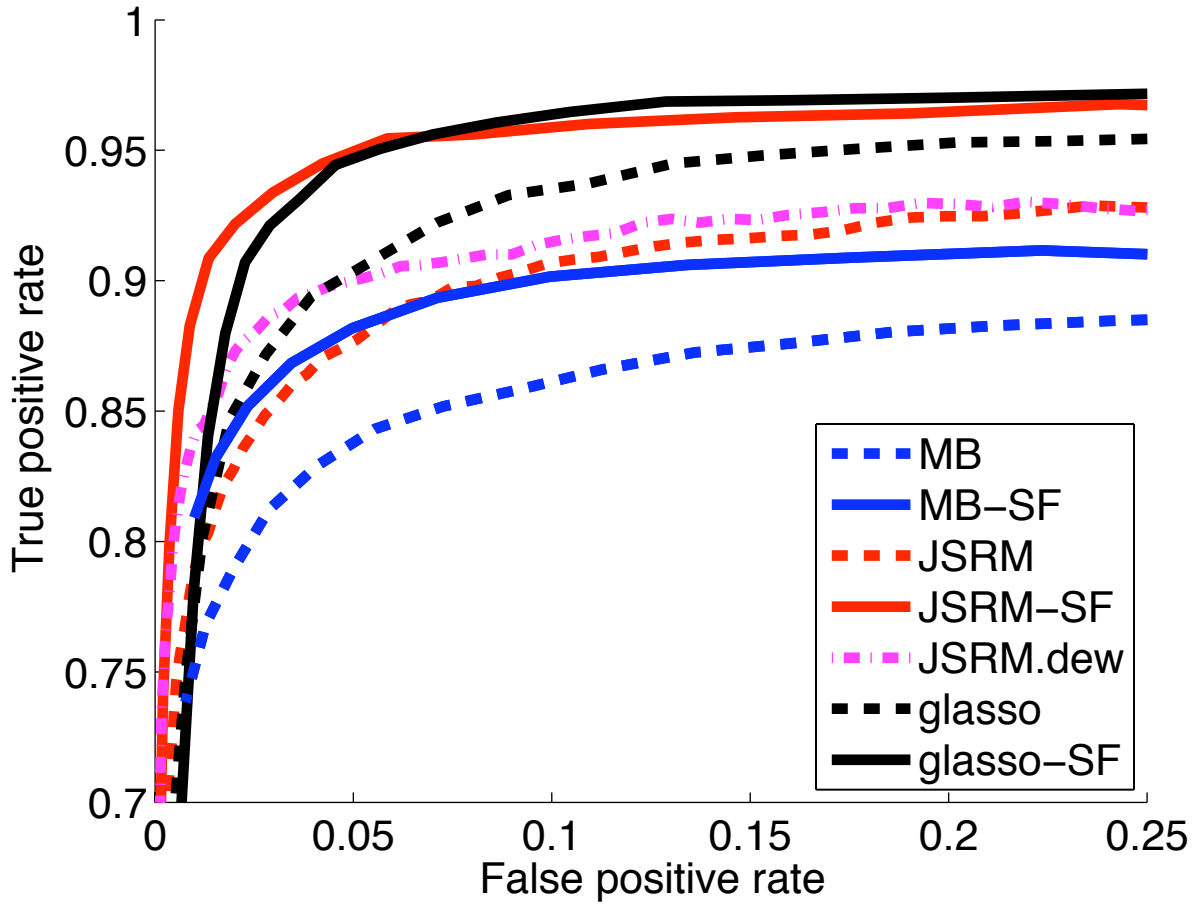
ROC curves

Score function = Neighborhood selection

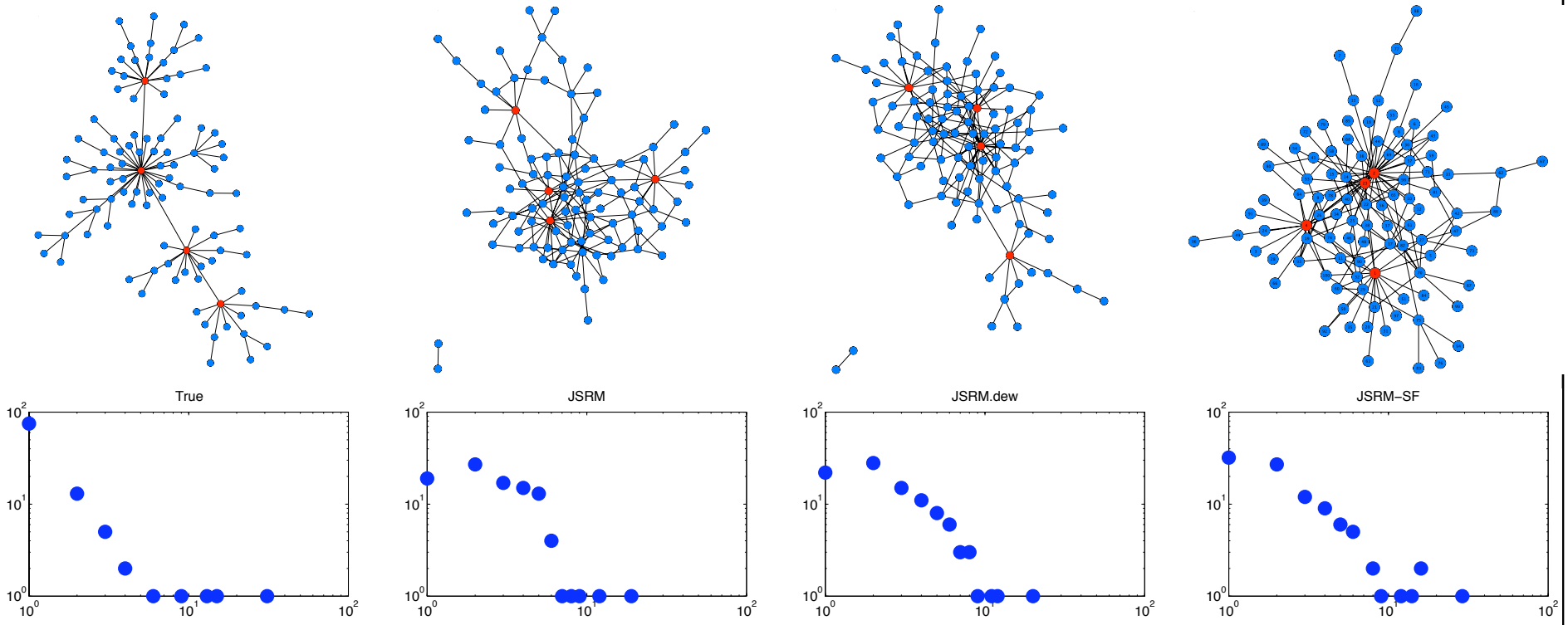


ROC curves

All score function

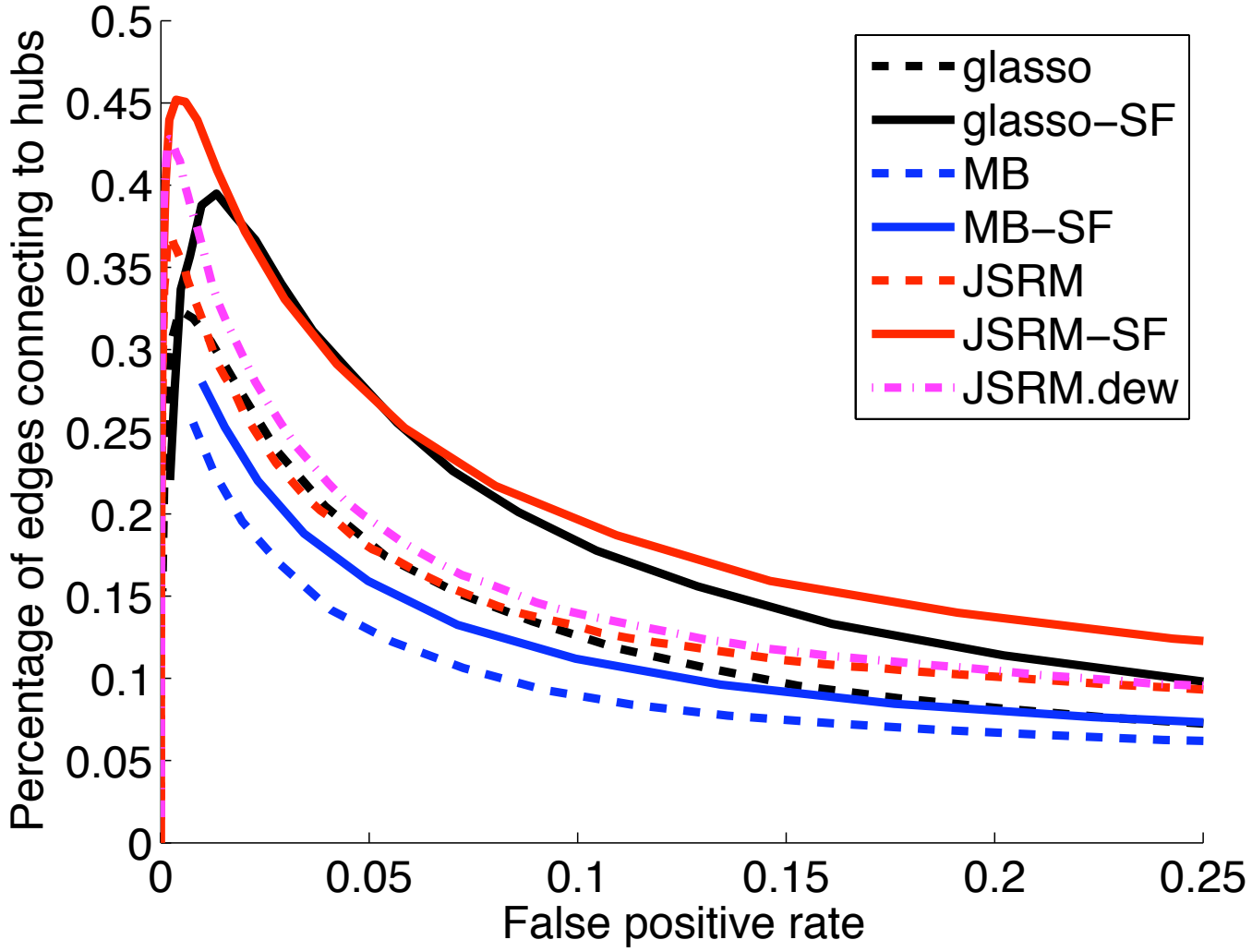


Degree distributions

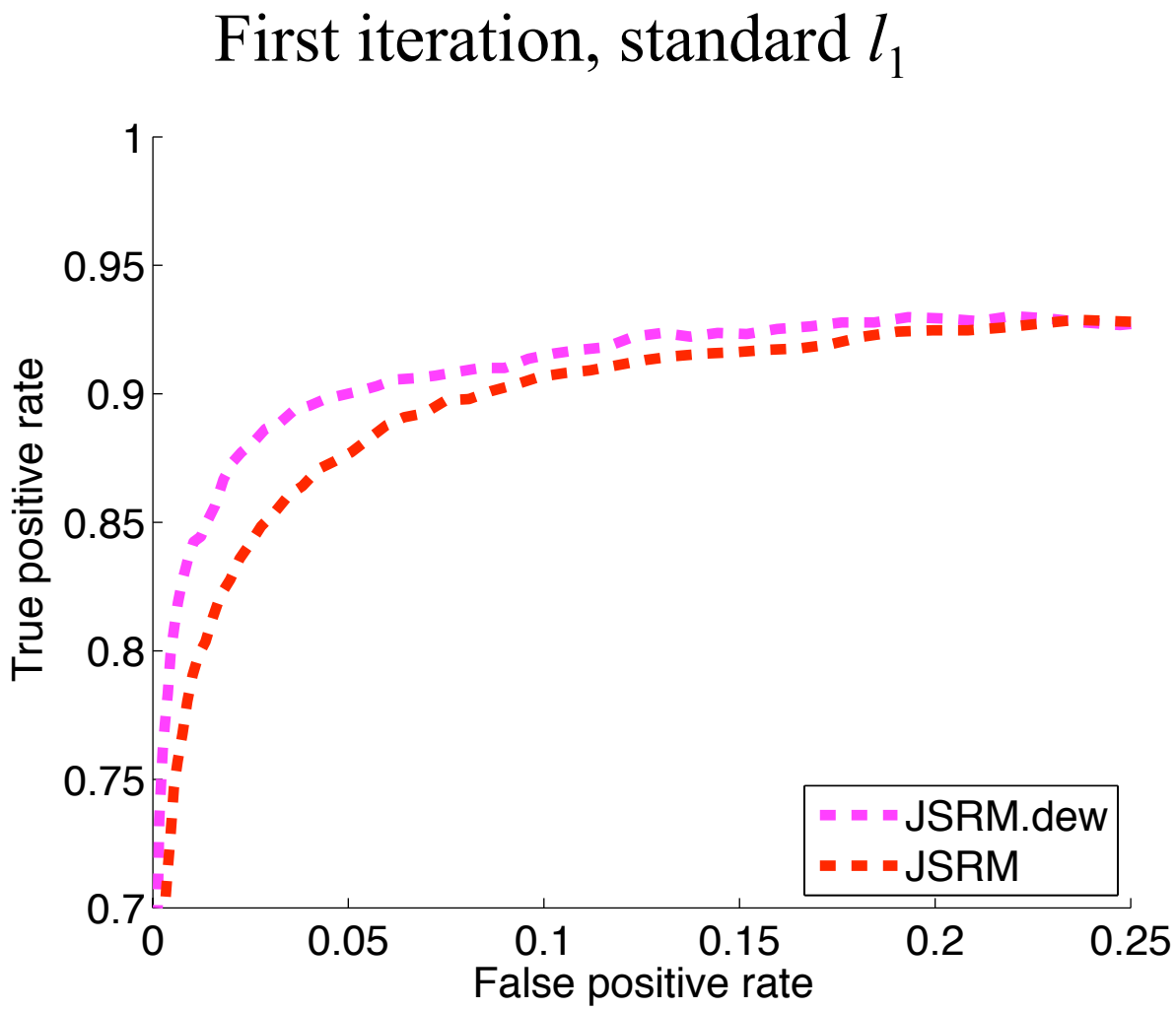


(a) True (b) JSRM (c) JSRM.dew (d) JSRM-SF

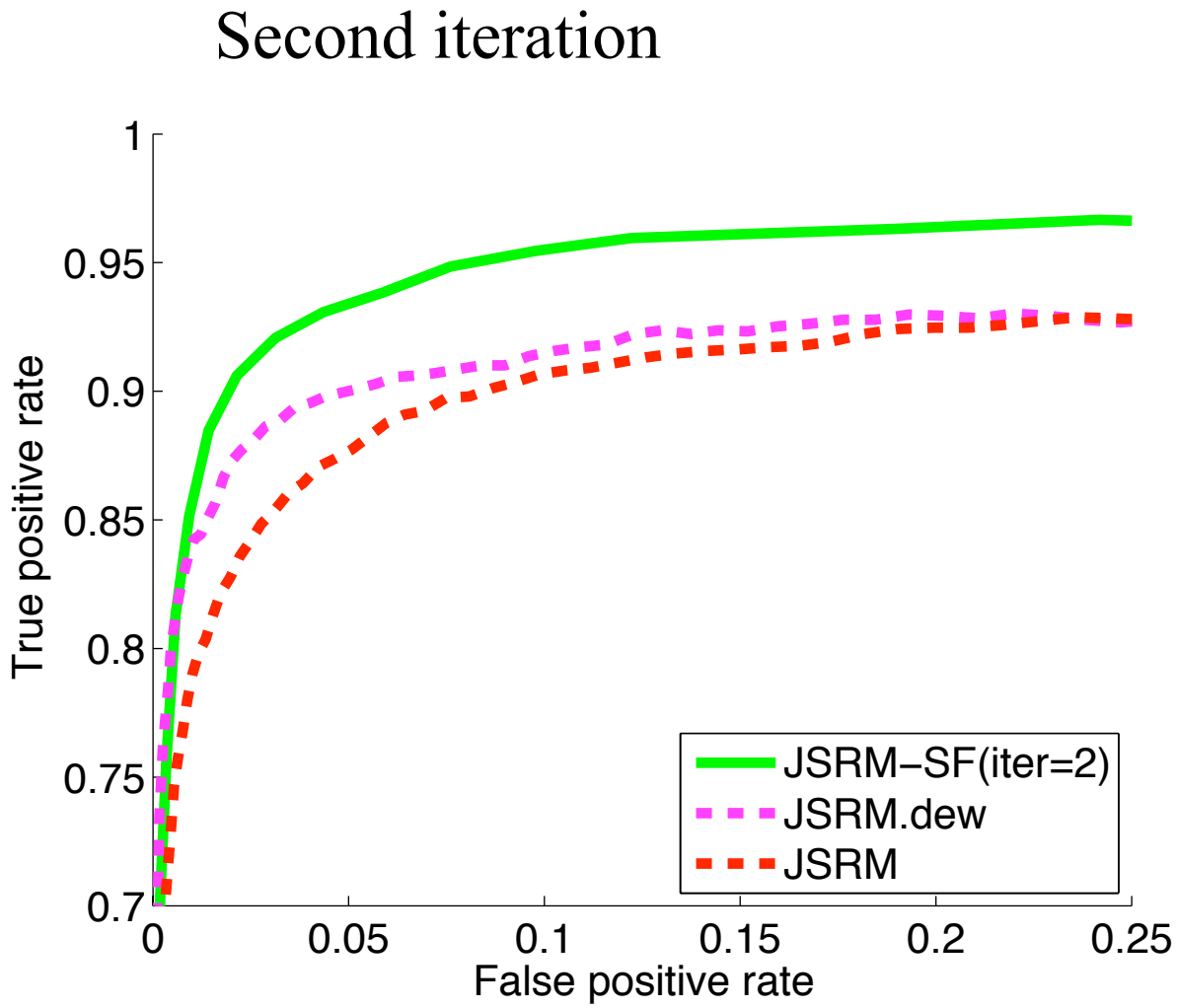
Percentage of edges connecting to hubs



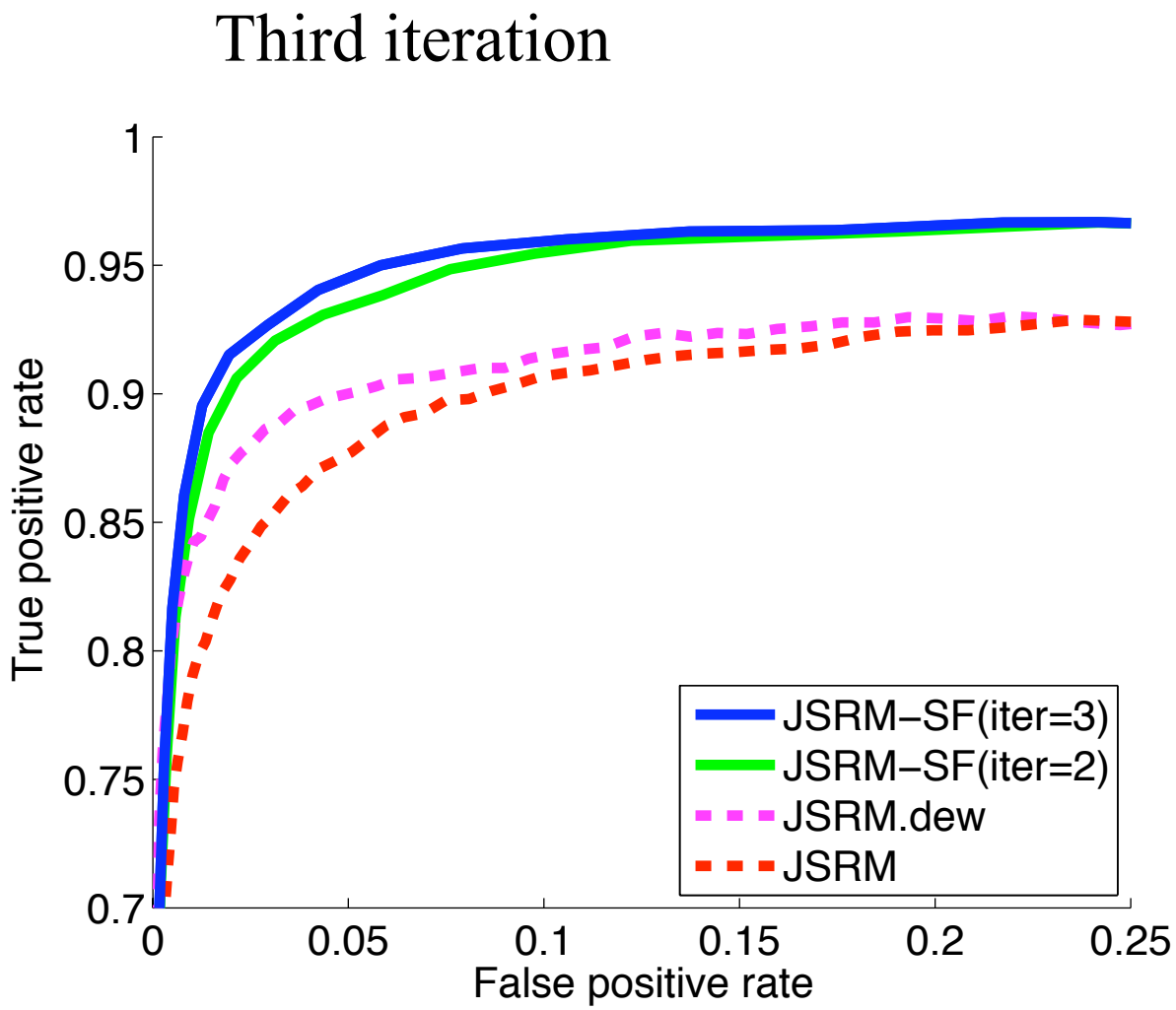
Improvement over iterations



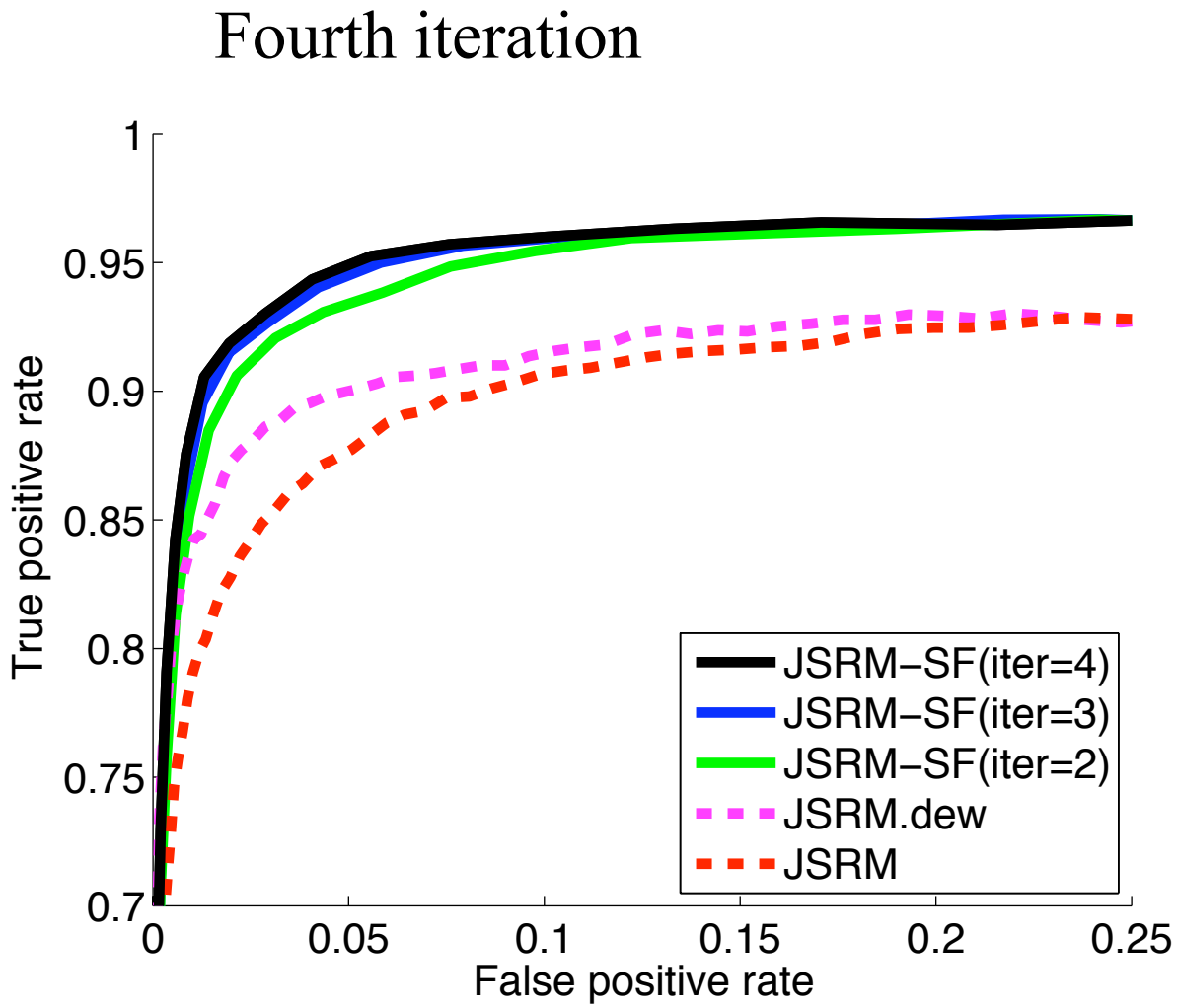
Improvement over iterations



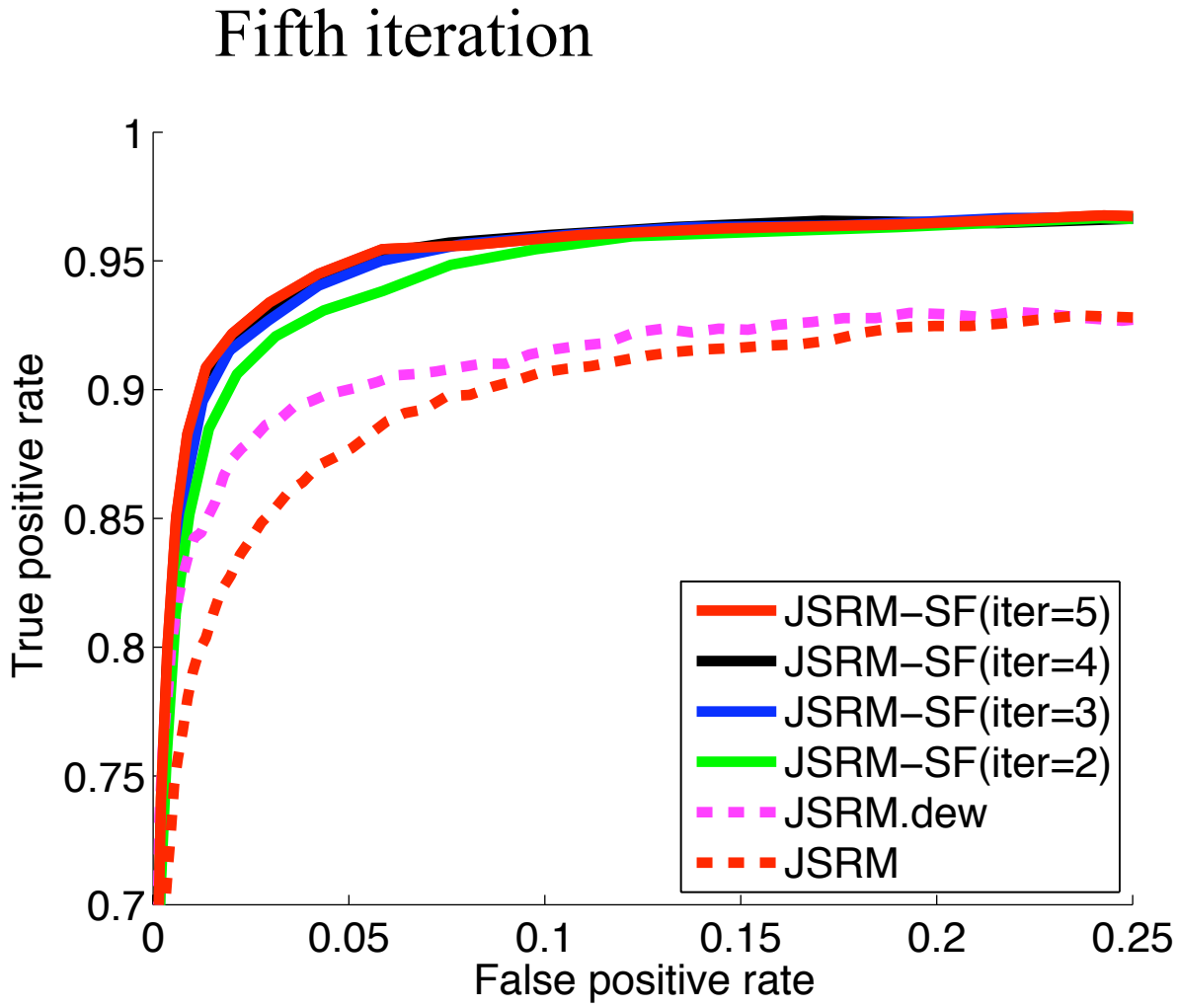
Improvement over iterations



Improvement over iterations

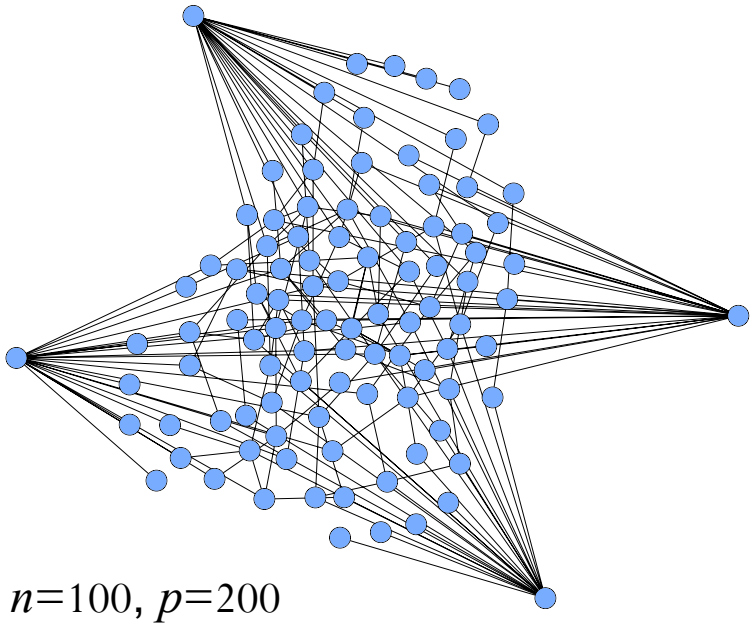


Improvement over iterations



Most improvement is obtained in the 2-nd iteration!

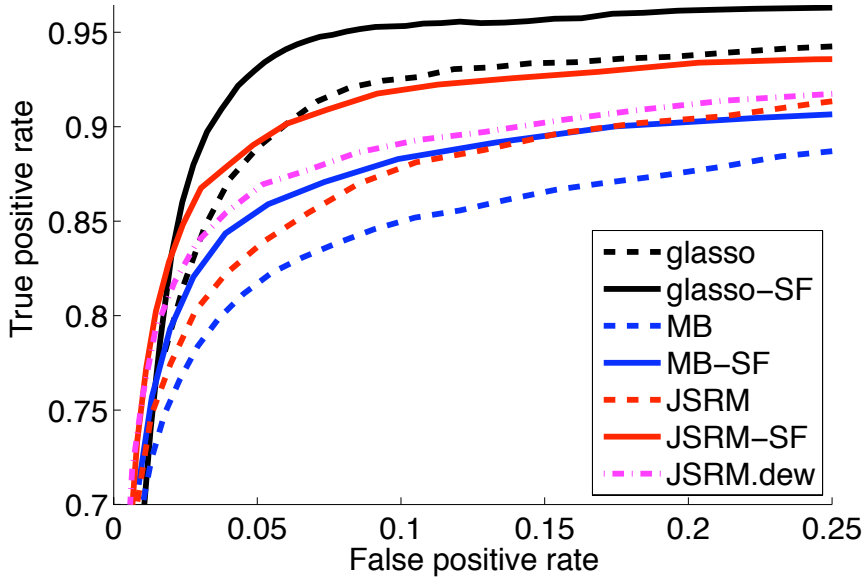
Experiments (simulated hub network)

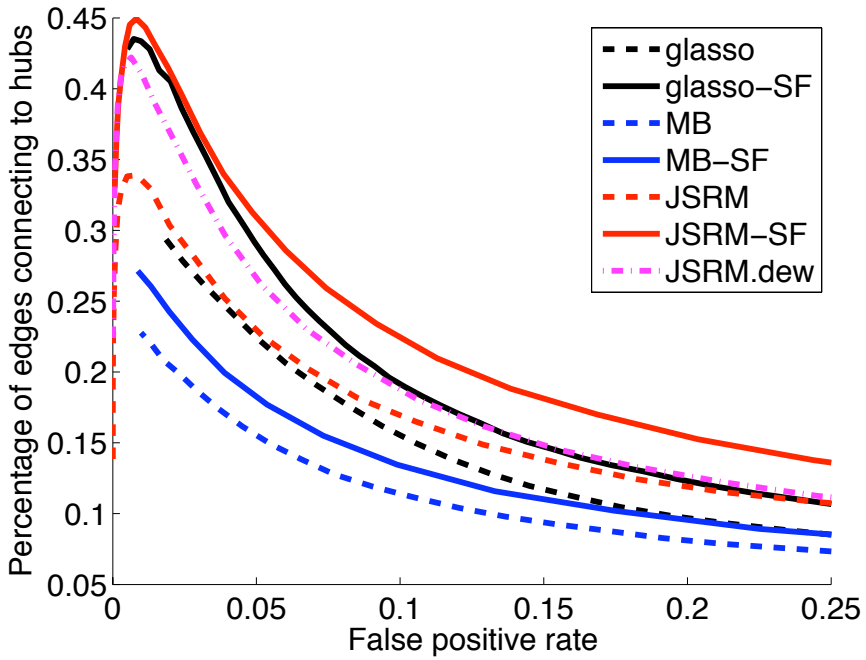


n=100, p=200

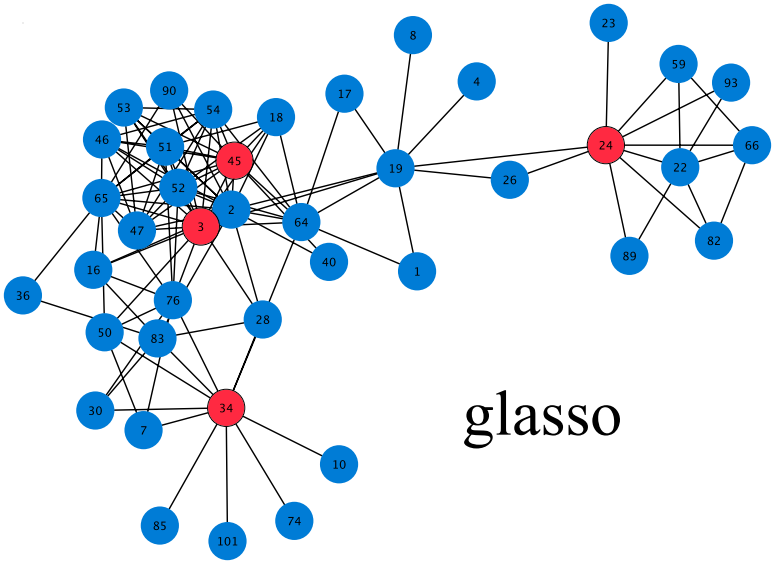
edges = 190

4 hubs (degree=24)

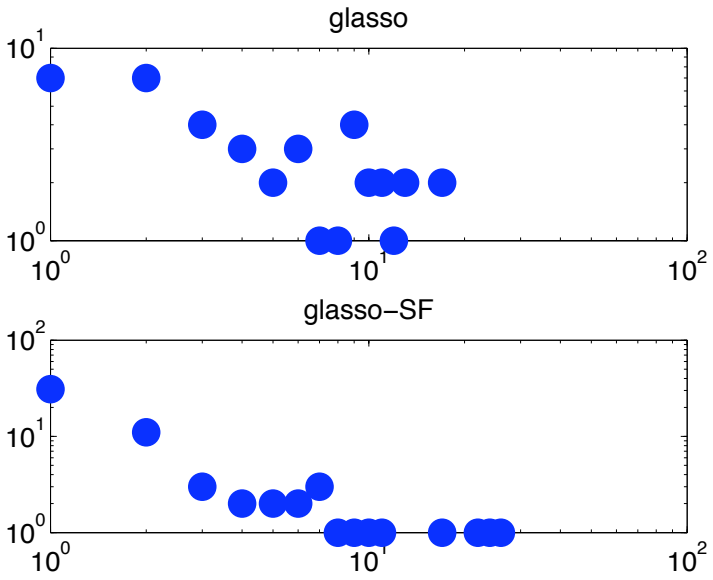


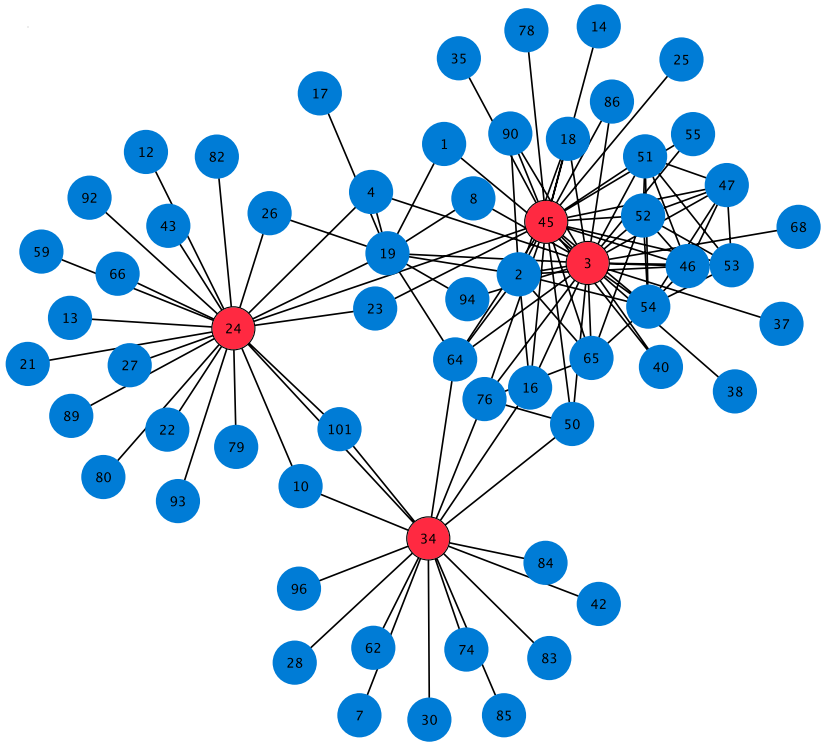


Experiments (Microarray data)



glasso





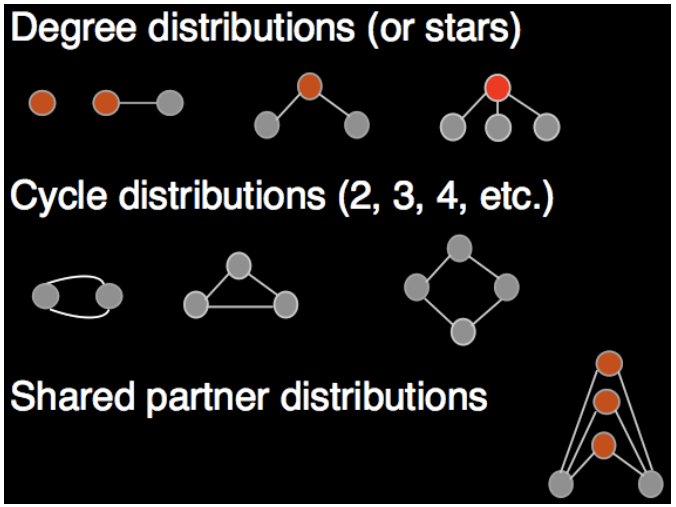
glasso-SF

Saccharomyces cerevisiae

P. T. Spellman et.al. *Mol. Biol. Cell*, 9(12):3273–3297, 1998.

Future works

- Verifying on real biological datasets
- Theoretical analysis
- More structured priors, Priors by Exponential-family random graph models (**ERGM)**



Thank You!