Hub Graphical Lasso

Kean Ming Tan

June 18, 2013

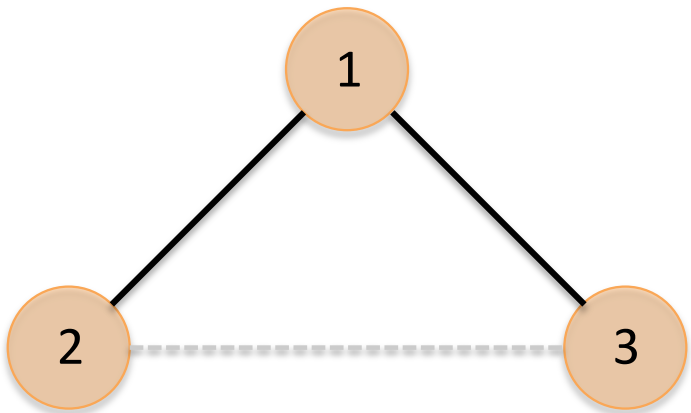
Joint work with Karthik Mohan, Palma London, Maryam Fazel, Su-In Lee, and Daniela Witten

University of Washington

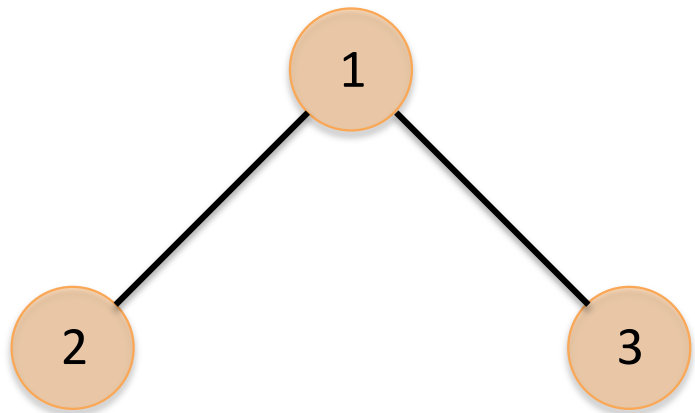
Arabidopsis Thaliana data set

- $n = 118$ samples and $p = 39$ genes
- Two isoprenoid biosynthesis pathways, Mevalonate Acid (MVA) and Methylerythritol Phosphate (MEP)
- Understand the conditional dependence relationship among the genes

Maginal versus conditional



Marginal



Conditional

Gaussian Graphical Model

- Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a $p \times p$ matrix
- We know $\mathbf{\Sigma}^{-1}_{ij} = 0$ implies the ith and jth variables are conditionally independent, given the other variables
- We can construct a conditional independence graph by estimating $\mathbf{\Sigma}^{-1}$
- The MLE for $\mathbf{\Sigma}^{-1}$ is $\mathbf{S}^{-1}$, where $\mathbf{S} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$
- Not invertible and all entries are non-zero

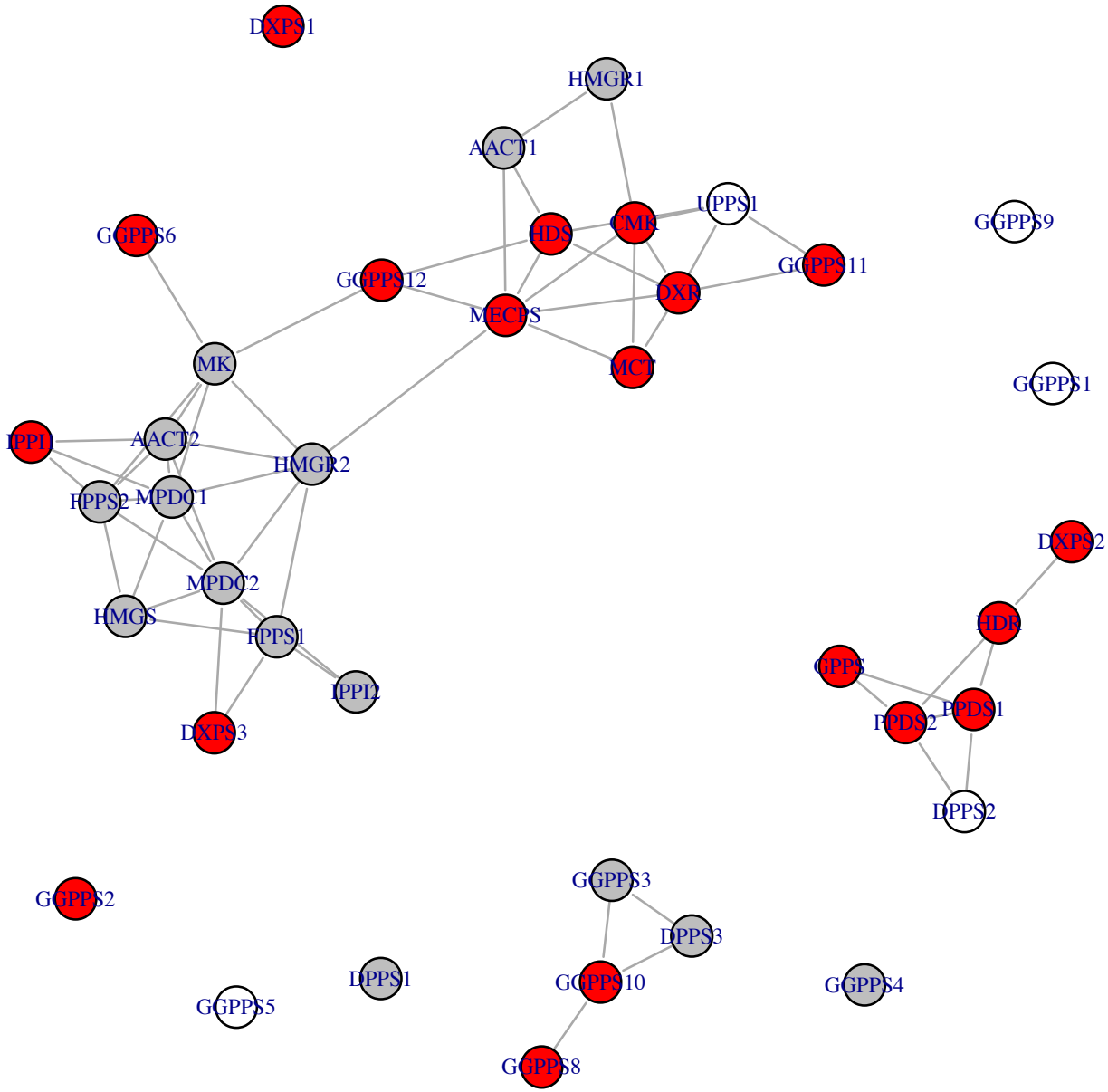
The Graphical Lasso

- ℓ_1 penalized log-likelihood for sparse estimation of $\mathbf{\Sigma}^{-1}$
- Let $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ and $\hat{\mathbf{\Theta}}$ be an estimate of $\mathbf{\Sigma}^{-1}$

$$\underset{\mathbf{\Theta}}{\text{minimize}} \{-\text{logdet } \mathbf{\Theta} + \text{trace}(\mathbf{S\Theta}) + \lambda \|\mathbf{\Theta}\|_1\}$$

- λ controls sparsity of $\hat{\mathbf{\Theta}}$

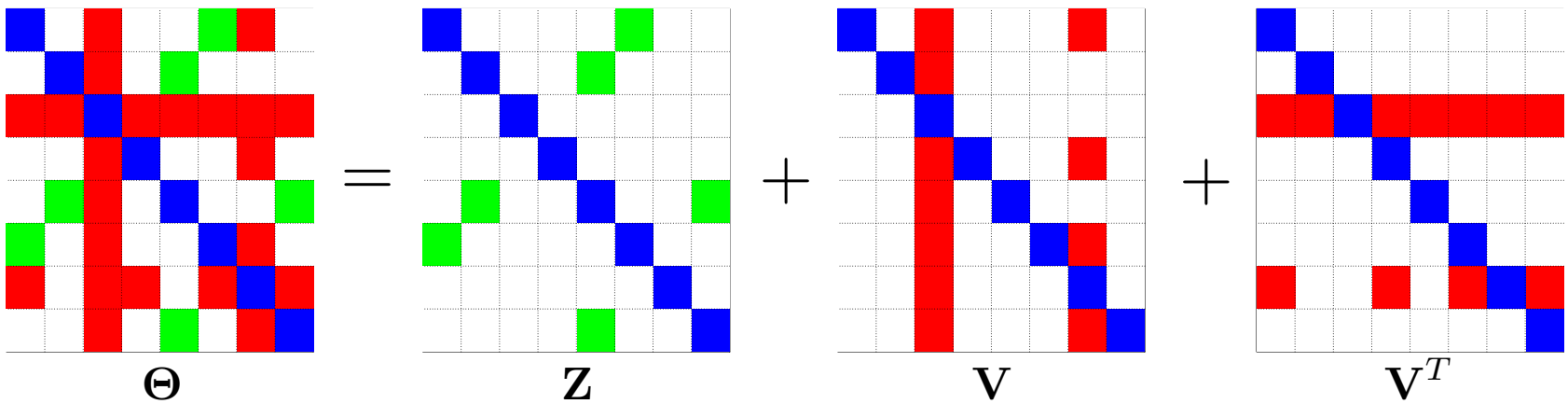
Citation: Friedman et al. 2007, Yuan and Lin 2006, Banerjee et al. 2008


DXPS1
HMGR1
AACT1
UPPS1
GGPPS9
GGPPS6
HDS
CMK
GGPPS11
GGPPS12
DXR
MECPS
MCT
MK
GGPPS1
IPPI1
AACT2
HMGR2
MPDC1
HPPS2
DXPS2
MPDC2
HDR
HMGS
HPPS1
GPPS
IPPI2
PPDS1
PPDS2
DXPS3
DPPS2
GGPPS3
GGPPS2
DPPS3
DPPS1
GGPPS10
GGPPS4
GGPPS5
GGPPS8

Problems with the Graphical Lasso

- Each edge is equally likely to appear, independent of all other edges
- Unrealistic in many settings
  - Gene regulatory network - transcription factors are connected to a number of genes
  - World Wide Web - Google is connected to most webpages
  - Network of collaborations among scientists - Paul Erdos collaborated with 511 scientists
- Want some columns of $\hat{\mathbf{\Theta}}$ to be dense

Motivation of Hub Graphical Lasso (HGL)



- $\mathbf{Z}$ is sparse, something like the graphical lasso solution
- Non-zero columns of $\mathbf{V}$ corresponds to the hub nodes

Optimization problem

$$\underset{\mathbf{\Theta},\mathbf{V},\mathbf{Z}}{\text{minimize}} \quad \Big\{ -\log\det\mathbf{\Theta} + \text{trace}(\mathbf{S\Theta}) + \lambda_1\|\mathbf{Z} - \text{diag}(\mathbf{Z})\|_1 + \lambda_2\|\mathbf{V} - \text{diag}(\mathbf{V})\|_1 + \lambda_3 \sum_{j=1}^{p} \|(\mathbf{V} - \text{diag}(\mathbf{V}))_j\|_2 \Big\}$$

$$\text{subject to} \quad \mathbf{\Theta} = \mathbf{Z} + \mathbf{V} + \mathbf{V}^T$$

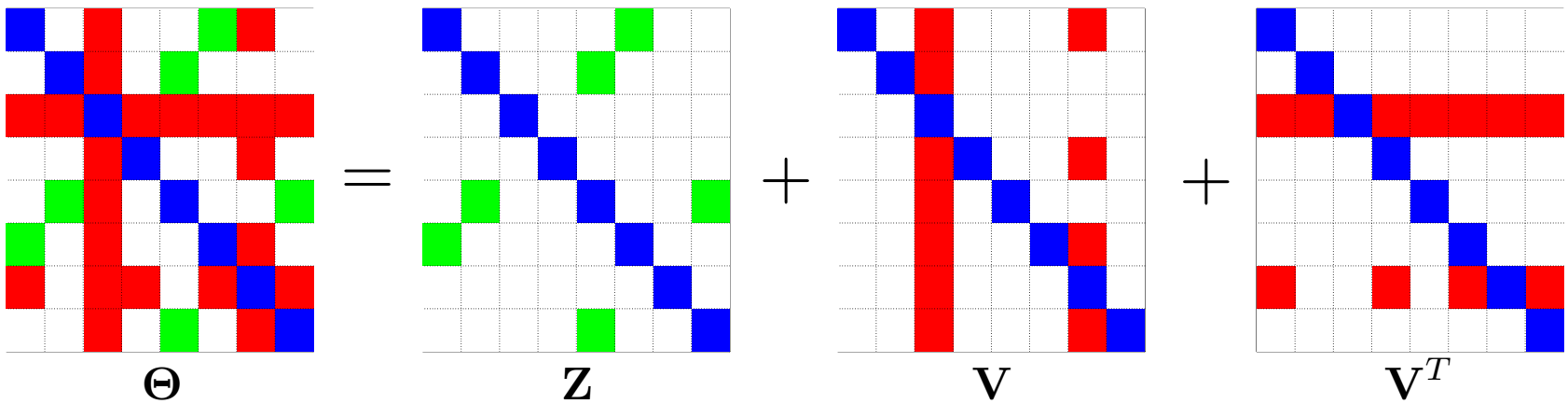
- λ_1 controls the sparsity of the matrix $\mathbf{Z}$
- λ_2 controls the sparsity of the hub nodes
- λ_3 controls the selection of hub nodes

Convex problem - solve using Alternating Direction Method of Multiplier

Measures of performance

- Correctly estimated edges
- Proportion of correctly estimated hub edges
- Proportion of correctly identified hub nodes
  - nodes that have at least r edges and also estimated to be $\mathcal{H}$ most highly-connected nodes

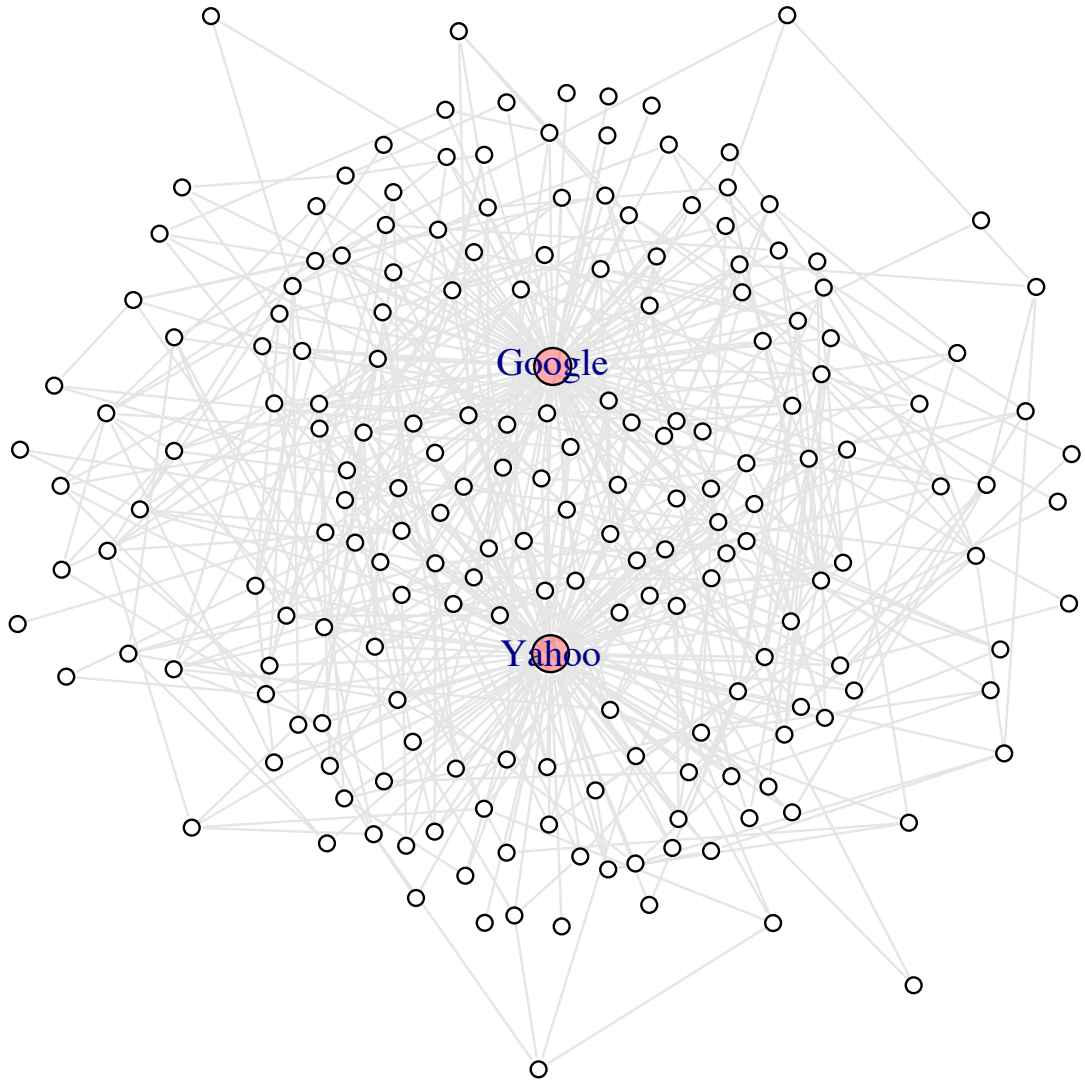
Simulation: Network with several hub nodes



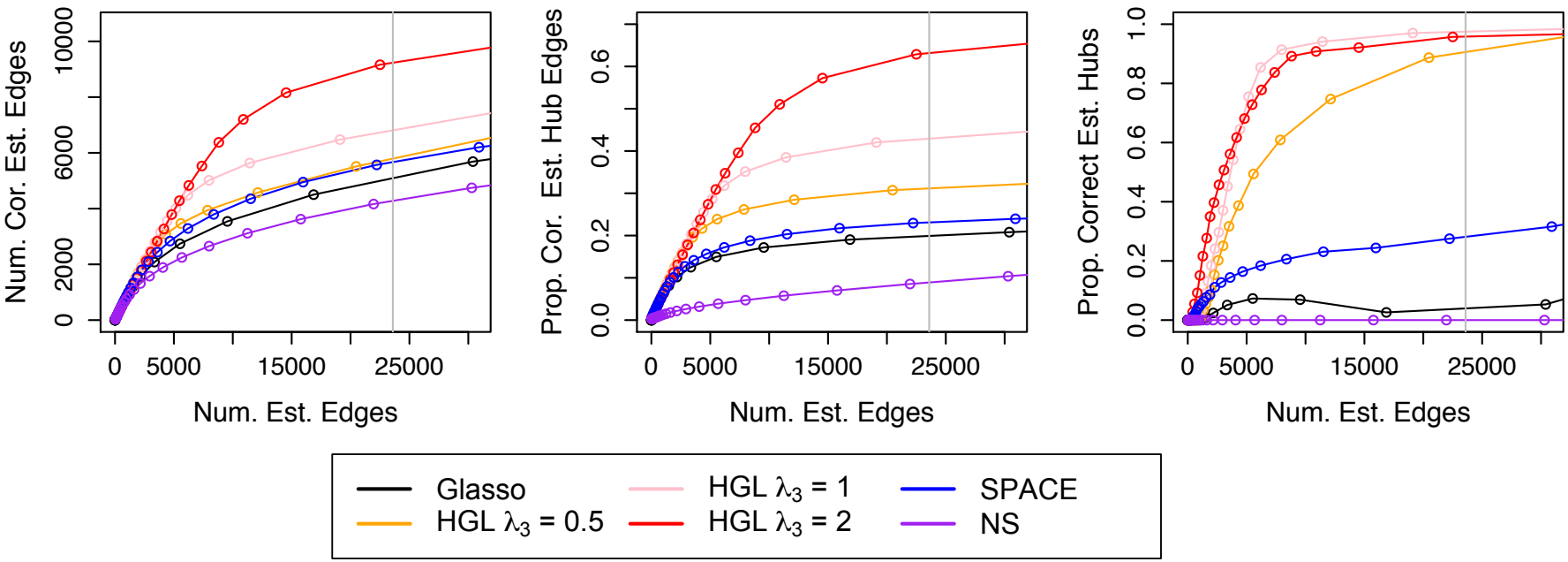
Generate $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ as follows

- Matrix $\mathbf{Z}$ is 98% sparse
- There are $\mathcal{H} = 20$ hub nodes (30% sparse) in the matrix $\mathbf{V}$
- Simulate data $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} N(0, \mathbf{\Sigma})$

Realization of such network for $p = 200$

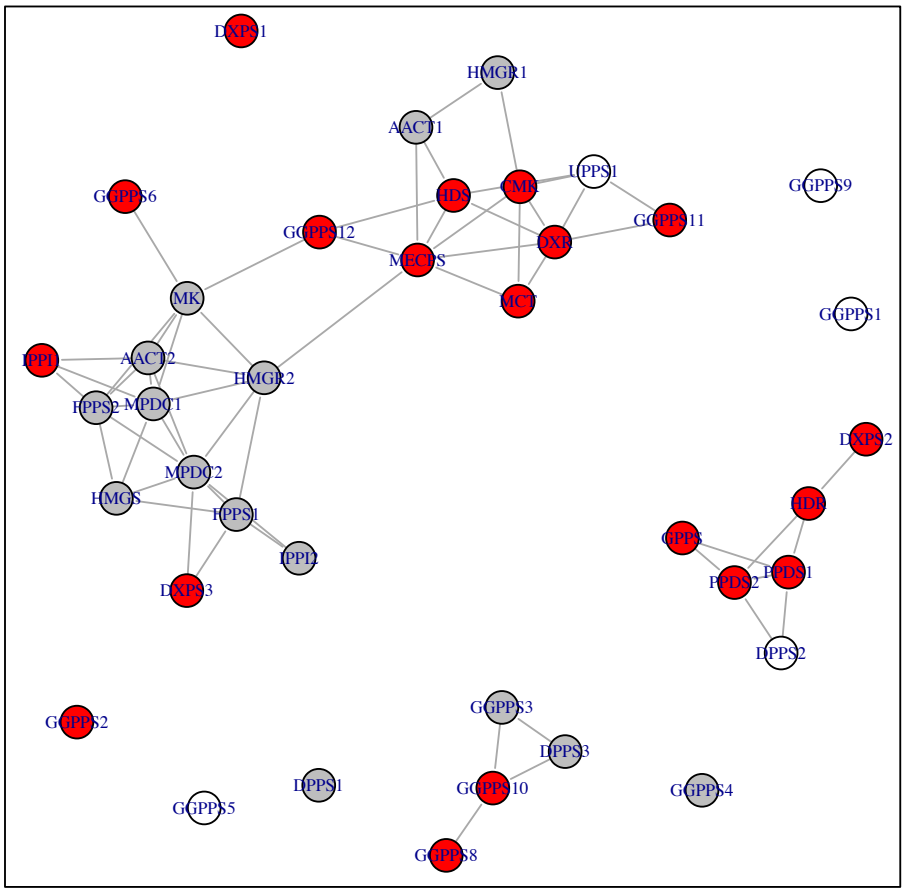


Simulation: $n = 500$ $p = 1000$

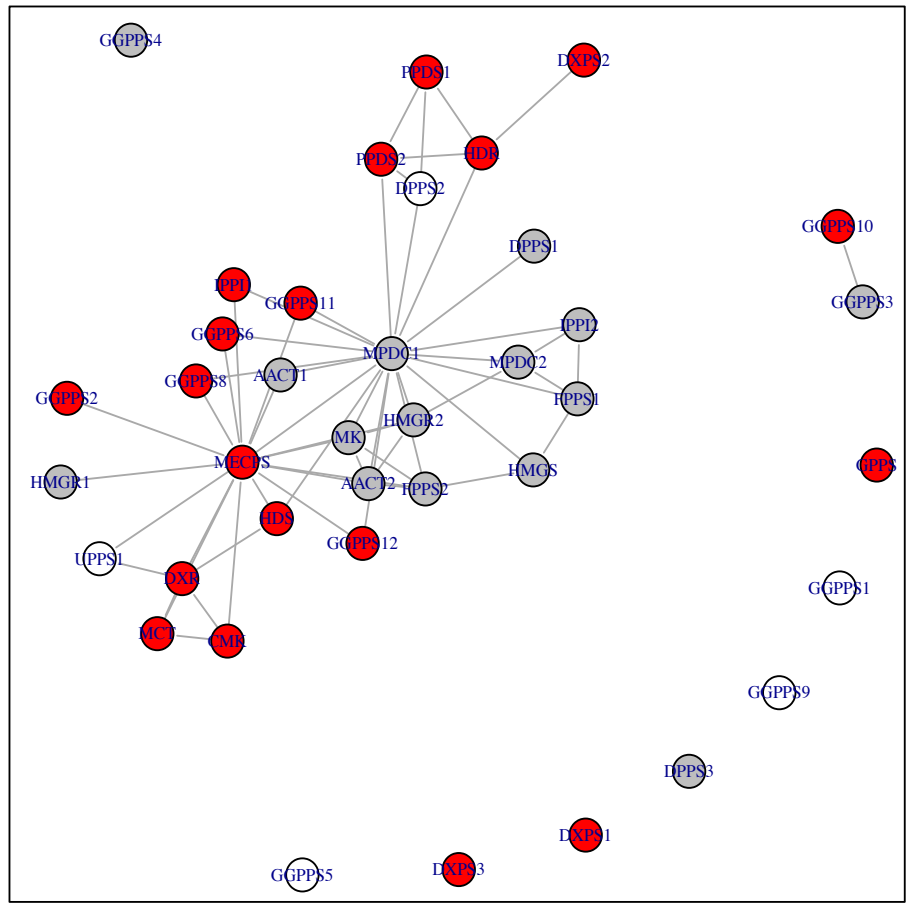


Citation: Meinshausen and Bühlmann 2006, Friedman et al. 2007, Peng et al. 2009

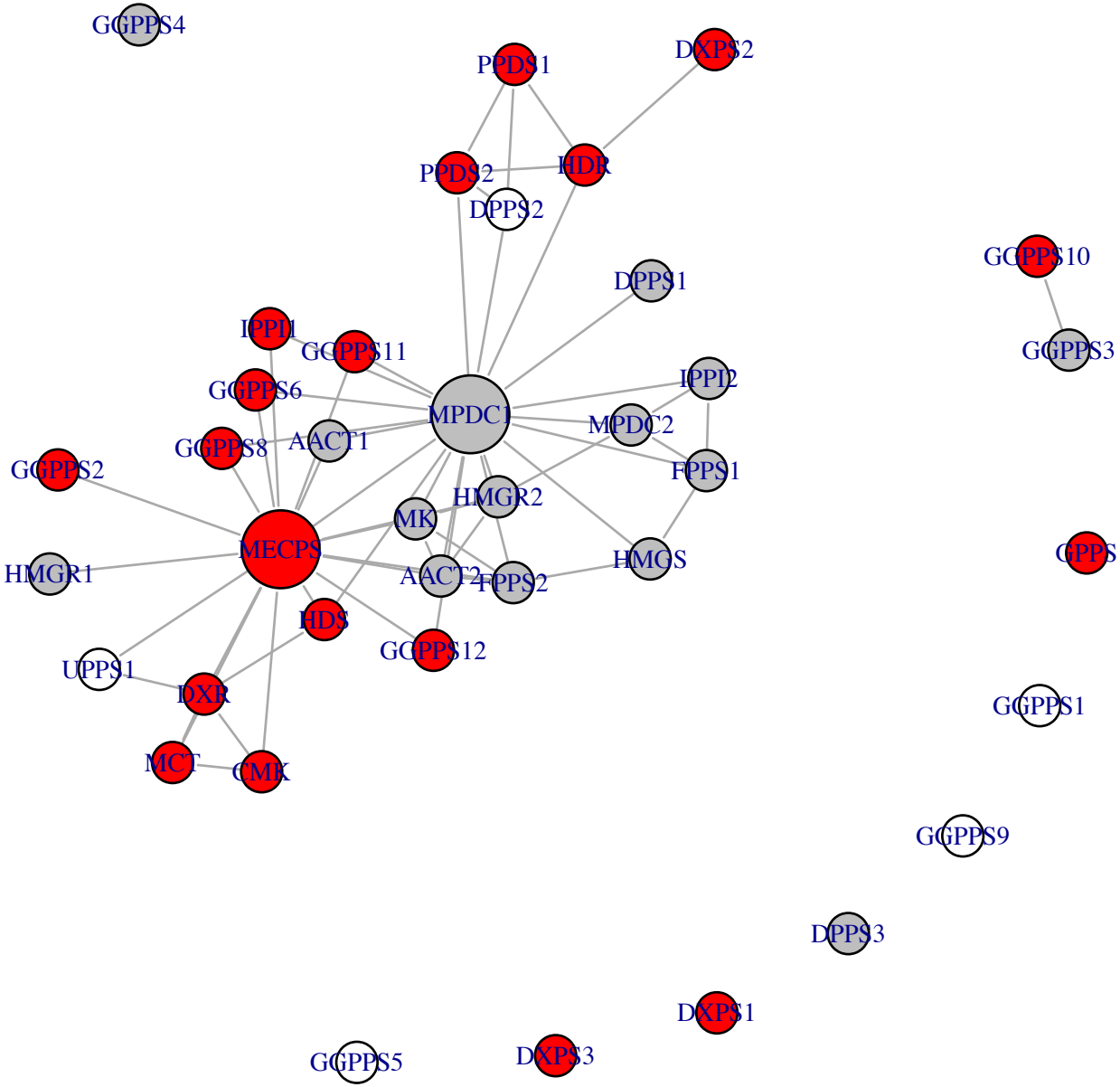
Back to plant data!



The Graphical Lasso



The Hub Graphical Lasso


GGPPS4
PPDS1
DXPS2
PPDS2
HDR
DPPS2
GGPPS10
DPPS1
IPPI1
GGPPS11
GGPPS3
GGPPS6
IPPI2
MPDC1
MPDC2
GGPPS8
AACT1
GGPPS2
FPPS1
HMGR2
MK
MECPS
GPPS
HMGR1
AACT2
HMGS
FPPS2
HDS
GGPPS12
UPPS1
GGPPS1
DXR
MCT
CMK
GGPPS9
DPPS3
DXPS1
GGPPS5
DXPS3

Summary

- Propose HGL to identify hub nodes
- ADMM algorithm guarantees solutions to converge to the global minimum
- Future direction of research
  - Modeling binary network with hub nodes
  - Study the theoretical properties of the estimator

Question?