Hub discovery in large correlation networks

*Alfred Hero and †Bala Rajaratnam

*University of Michigan, †Stanford University

July 13, 2012

Outline

Random matrix measurement model

	Variable 1	Variable 2	...	Variable p
Sample 1	X_{11}	X_{12}	...	X_{1p}
Sample 2	X_{21}	X_{22}	...	X_{2p}
⋮	⋮	⋮	...	⋮
Sample n	X_{n1}	X_{n2}	...	X_{np}

$n \times p$ measurement matrix $\mathbb{X}$ has i.i.d. elliptically distributed rows

$$\mathbb{X} = \begin{bmatrix} X_{11} & \cdots & \cdots & X_{1p} \\ \vdots & \ddots & \ddots & \vdots \\ X_{n1} & \cdots & \cdots & X_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^1 \\ \vdots \\ \mathbf{X}^n \end{bmatrix} = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$$

Columns of $\mathbb{X}$ index variables while rows index i.i.d. samples

$p \times p$ covariance (dispersion) matrix associated with each row is $\mathrm{cov}(\mathbf{X}^i) = \mathbf{\Sigma}$

Sparse multivariate dependency models

Two types of sparse (ensemble) correlation models:

- Sparse correlation ($\mathbf{\Sigma}$) graphical models:
  - Most correlation are zero, few marginal dependencies
  - Examples: M-dependent processes, moving average (MA) processes
- Sparse inverse-correlation ($\mathbf{K} = \mathbf{\Sigma}^{-1}$) graphical models
  - Most inverse covariance entries are zero, few conditional dependencies
  - Examples: Markov random fields, autoregressive (AR) processes, global latent variables

- Sometimes correlation matrix and its inverse are both sparse
- Often only one of them is sparse

Gaussian graphical models - GGM - (Lauritzen 1996)

Multivariate Gaussian model

$$p(\mathbf{x}) = \frac{|\mathbf{K}|^{1/2}}{(2\pi)^{p/2}} \exp\left(-\tfrac{1}{2}\sum_{i,j=1}^{p} x_i x_j [\mathbf{K}]_{ij}\right)$$

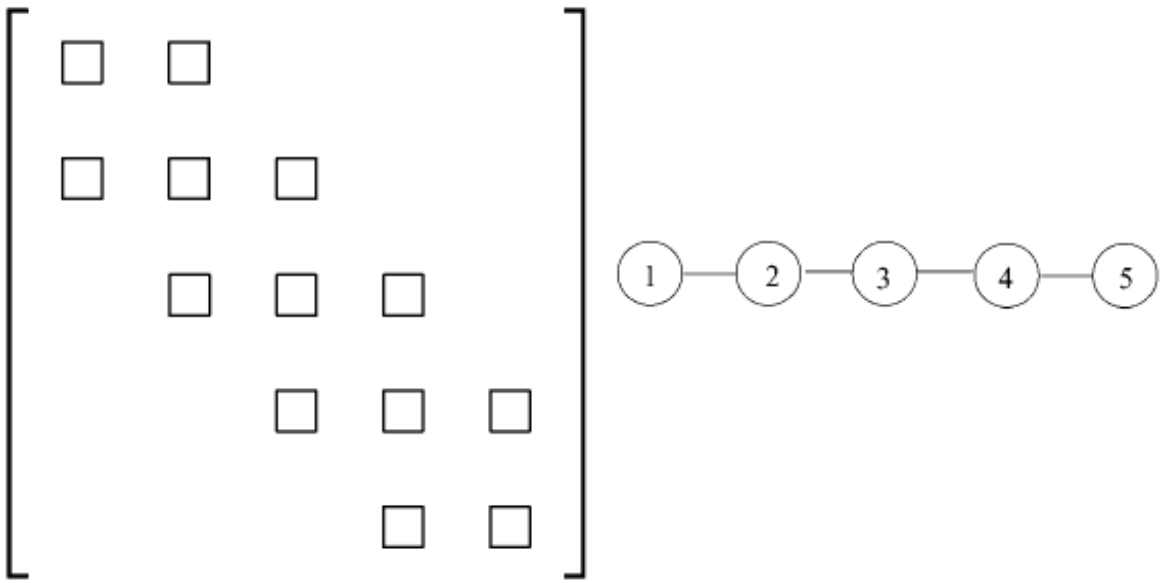
where $\mathbf{K} = [\mathrm{cov}(\mathbf{X})]^{-1}$: $p \times p$ precision matrix

- $\mathcal{G}$ has an edge e_{ij} iff $[\mathbf{K}]_{ij} = 0$
- Adjacency matrix $\mathbf{A}$ of $\mathcal{G}$ obtained by thresholding $\mathbf{K}$

$$\mathbf{A} = h(\mathbf{K}), \quad h(u) = \tfrac{1}{2}(\mathrm{sgn}(|u| - \rho) + 1)$$

ρ is arbitrary positive threshold

Banded Gaussian graphical model



Block diagonal Gaussian graphical model

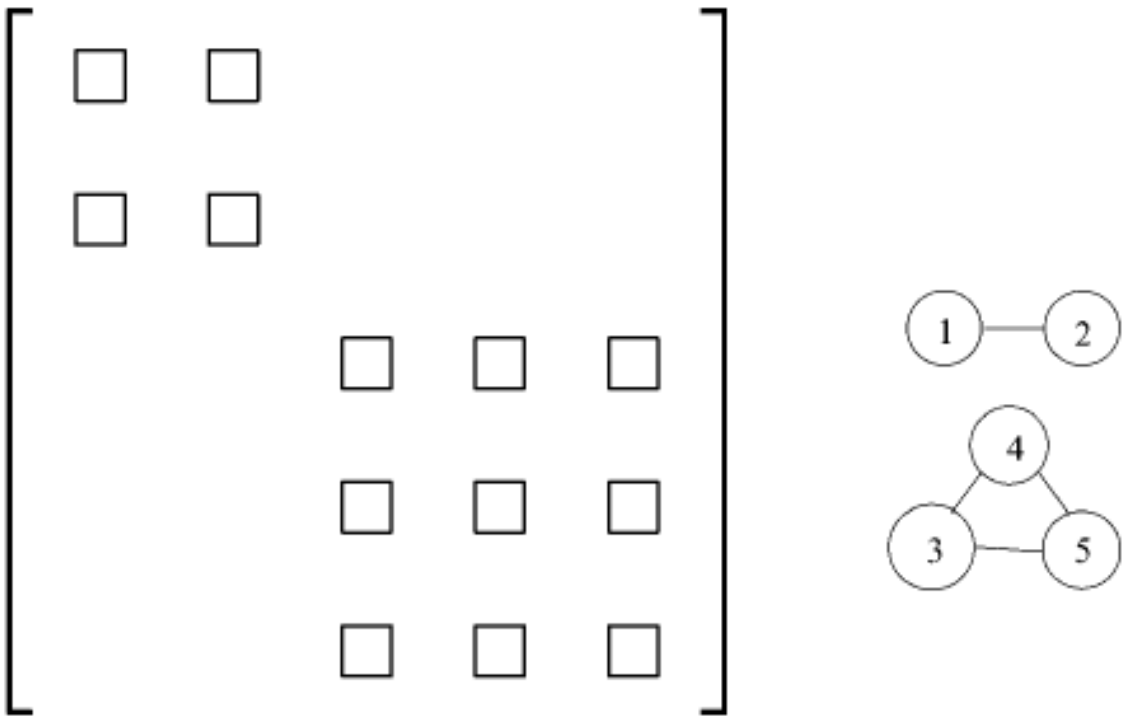
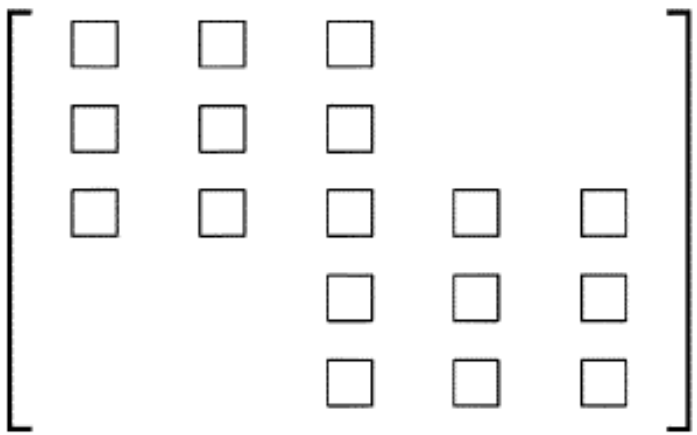
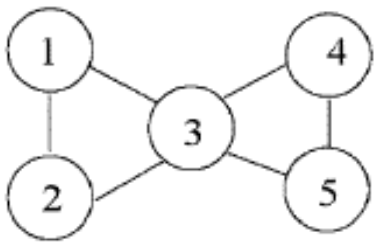


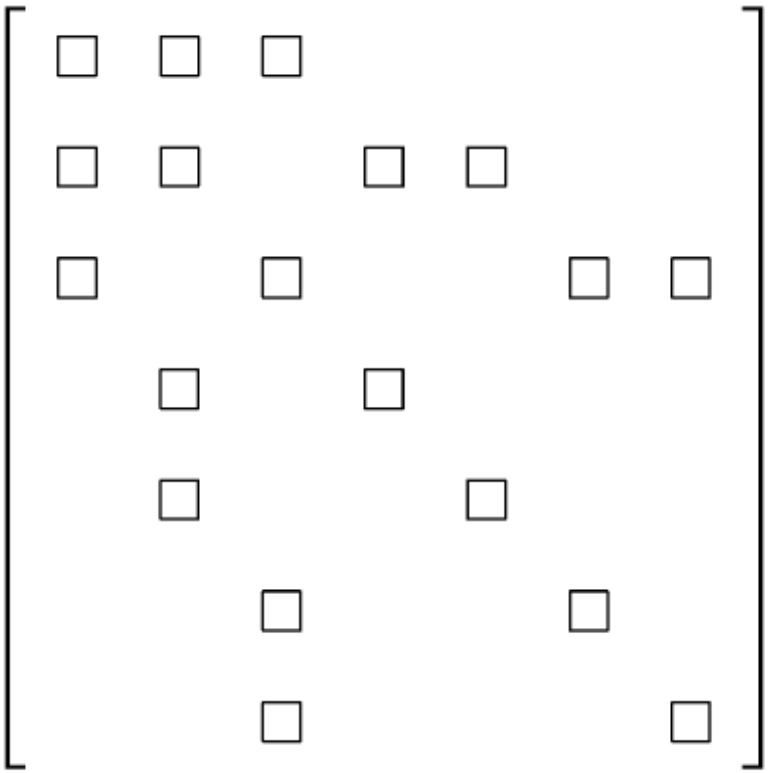
Figure: Left: inverse covariance matrix $\mathbf{K}$. Right: associated graphical model

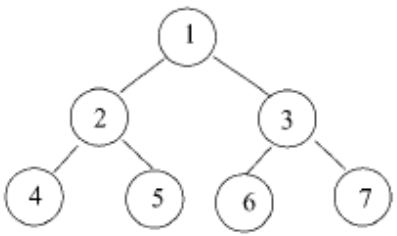
Two-coupled block Gaussian graphical model





Tree (Multiscale) Gaussian graphical model





Outline

Screening for hubs in $\mathcal{G}$

Figure: Star components - hubs of degree $d = 1, \ldots, 5, \ldots$

- In this talk we are interested in star components, i.e., hubs of high degree.
- Define number of hub nodes of degree $>$ d in $\mathcal{G}$

$$N_d = \sum_{i=1}^{p} I(d_i \geq d)$$

- We are interested in determining N_d from measurement matrix $\mathbb{X}$.

Previous work on inferring structure of graphical models

- Regularized l_2 or $l_{\mathcal{F}}$ covariance estimation
  - Banded covariance model: Bickel-Levina (2008)
  - Sparse eigendecomposition model: Johnstone-Lu (2007)
  - Stein shrinkage estimator: Ledoit-Wolf (2005), Chen-Weisel-Eldar-H (2010)
- Gaussian graphical model selection
  - l_1 regularized GGM: Meinshausen-Bühlmann (2006), Wiesel-Eldar-H (2010).
  - Bayesian estimation: Rajaratnam-Massam-Carvalho (2008)
- Independence testing
  - Sphericity test for multivariate Gaussian: Wilks (1935)
  - Maximal correlation test: Moran (1980), Eagleson (1983), Jiang (2004), Zhou (2007)

This work (H, Rajaratnam 2011, 2012): fixed n large p, unrestricted sparsity structure, hubs of partial-correlation.

Sample covariance and sample correlation

- Sample covariance of columns of $\mathbb{X} = ((X_{lm}))$ $(n \times p)$

$$\hat{\sigma}_{ij} = \frac{1}{n-1} \sum_{m=1}^{n} (X_{mi} - \hat{\mu}_i)(X_{mj} - \hat{\mu}_j), \quad \hat{\mu}_i = \frac{1}{n} \sum_{m=1}^{n} X_{mi}.$$

- Sample covariance matrix $\hat{\mathbf{\Sigma}} = ((\hat{\sigma}_{ij}))$

$$\hat{\mathbf{\Sigma}} = \frac{1}{n-1} \mathbb{X}^T (\mathbf{I} - n^{-1} \mathbf{1}\mathbf{1}^T) \mathbb{X}$$

- Sample correlation matrix $\mathbf{R} = ((r_{ij}))$

$$\mathbf{R} = \mathbf{D}_{\hat{\Sigma}}^{-1/2} \hat{\mathbf{\Sigma}} \mathbf{D}_{\hat{\Sigma}}^{-1/2}, \qquad r_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii} \hat{\sigma}_{jj}}}$$

Z-score representation of sample correlation

- $\mathbf{Z}_i$ is standardized i-th column $\mathbf{X}_i$ of $\mathbb{X}$

$$\mathbf{Z}_i = \frac{\mathbf{X}_i - \hat{\mu}_i \mathbf{1}}{\sqrt{\hat{\sigma}_{ii}}\sqrt{n-1}}, \quad i = 1, \ldots, p$$

- Z-score matrix

$$\mathbb{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_p] = (n-1)^{-1/2}(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T)\mathbb{X}\mathbf{D}^{-1/2}.$$

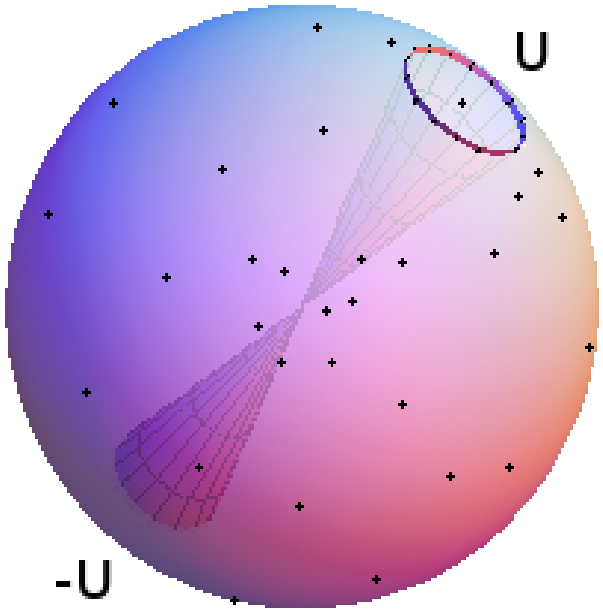
- Z-score representation of $\mathbf{R}$

$$\mathbf{R} = \mathbb{Z}^T\mathbb{Z}$$

- n-dimensional $\mathbf{Z}_i$ lies in $n-2$ dimensional subpsace

$$\mathbf{1}^T\mathbf{Z}_i = 0 \text{ and } \|\mathbf{Z}_i\| = 1$$

Projected Z-scores lie on sphere S_{n-2}



Correlation is related to distance between projected Z-scores

$$\|\mathbf{U}_i - \mathbf{U}_j\| = \sqrt{2(1 - r_{ij})}$$

Relative surface area of spherical cap of radius $\sqrt{2(1-\rho)}$:

$$P_0(\rho, n) = 2B((n-2)/2, 1/2) \int_\rho^1 (1-u^2)^{\frac{n-4}{2}} du$$

Screening hubs in $\mathcal{G}$ from random samples

Problem: Find hubs in $\mathcal{G}$ given n i.i.d. samples $\{\mathbf{X}^j\}_{j=1}^n$
Solution(?): Threshold the sample partial correlation matrix

$$\mathbf{P} = [\mathrm{diag}(\mathbf{R}^{-1})]^{-1/2}\mathbf{R}^{-1}[\mathrm{diag}(\mathbf{R}^{-1})]^{-1/2}$$

Difficulties when $n < p$

- Sample correlation matrix $\mathbf{R}$ not invertible
  Soln: use Moore-Penrose generalized inverse instead.
- False matches can occur at any threshold level $\rho \in [0, 1)$.
  Soln: Derive p-values for the false matches.
- False matches have phase transitions as function of ρ
  Soln: Derive mathematical expressions for critical thresholds.

Key element: projected Z-scores

- Define pseudo-inverse partial correlation:

$$\mathbf{P} = \mathrm{diag}(\mathbf{R}^\dagger)\mathbf{R}^\dagger\mathrm{diag}(\mathbf{R}^\dagger), \quad \mathbf{R}^\dagger = \text{pseudo-inverse } (bR)$$

Lemma

Assume that $n < p$. The Moore-Penrose pseudo-inverse of $\mathbf{R}$ has the Z-score representation

$$\mathbf{R}^\dagger = \mathbb{U}^T[\mathbb{U}\mathbb{U}^T]^{-2}\mathbb{U}.$$

- Sample partial correlation representation $p_{ij} = \mathbf{Y}_i^T\mathbf{Y}_j$

$$\mathbf{P} = \mathbb{Y}^T\mathbb{Y}, \quad \mathbb{Y} = [\mathbb{U}\mathbb{U}^T]^{-1}\mathbb{U}\mathbf{D}^{-1/2}_{\mathbb{U}[\mathbb{U}\mathbb{U}^T]^{-2}\mathbb{U}}$$

Empirical hub discovery

Empirical hub discoveries: For threshold ρ and degree parameter δ define number $N_{\delta,\rho}$ of vertices in sample partial-correlation graph with degree $d_i \geq \delta$

$$N_{\delta,\rho} = \sum_{i=1}^{p} \phi_{\delta,i}$$

$$\phi_{\delta,i} = \begin{cases} 1, & \text{card}\{j : j \neq i, |\mathbf{Y}_i^T \mathbf{Y}_j| \geq \rho\} \geq \delta \\ 0, & o.w. \end{cases}$$

Asymptotic discovery rate

Assume that rows of $\mathbb{X}$ are i.i.d. with bounded elliptically contoured density and sparse graphical model.
Poisson limit: (H and Rajaratnam 2011, 2012):

Theorem

For large p

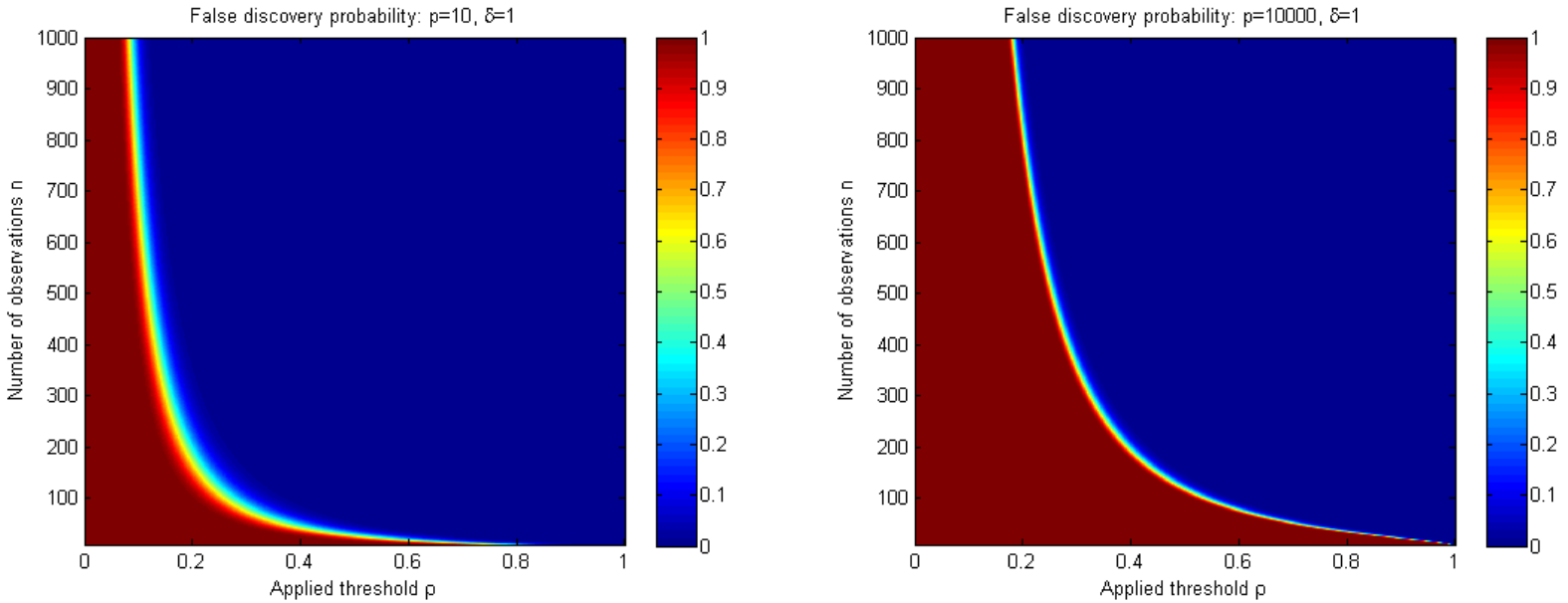
$$P(N_{\delta,\rho} > 0) \approx \begin{cases} 1 - \exp(-\lambda_{\delta,\rho}/2), & \delta = 1 \\ 1 - \exp(-\lambda_{\delta,\rho}), & \delta > 1 \end{cases}.$$

$$\lambda_{\delta,\rho} = p\binom{p-1}{\delta}(P_0(\rho, n))^{\delta}$$

$$P_0(\rho, n) = 2B((n-2)/2, 1/2)\int_{\rho}^{1}(1-u^2)^{\frac{n-4}{2}}\,du$$

False discovery probability heatmaps ($\delta = 1$)

False discovery probability: $P(N_{\delta,\rho} > 0) \approx 1 - \exp(-\lambda_{\delta,\rho})$

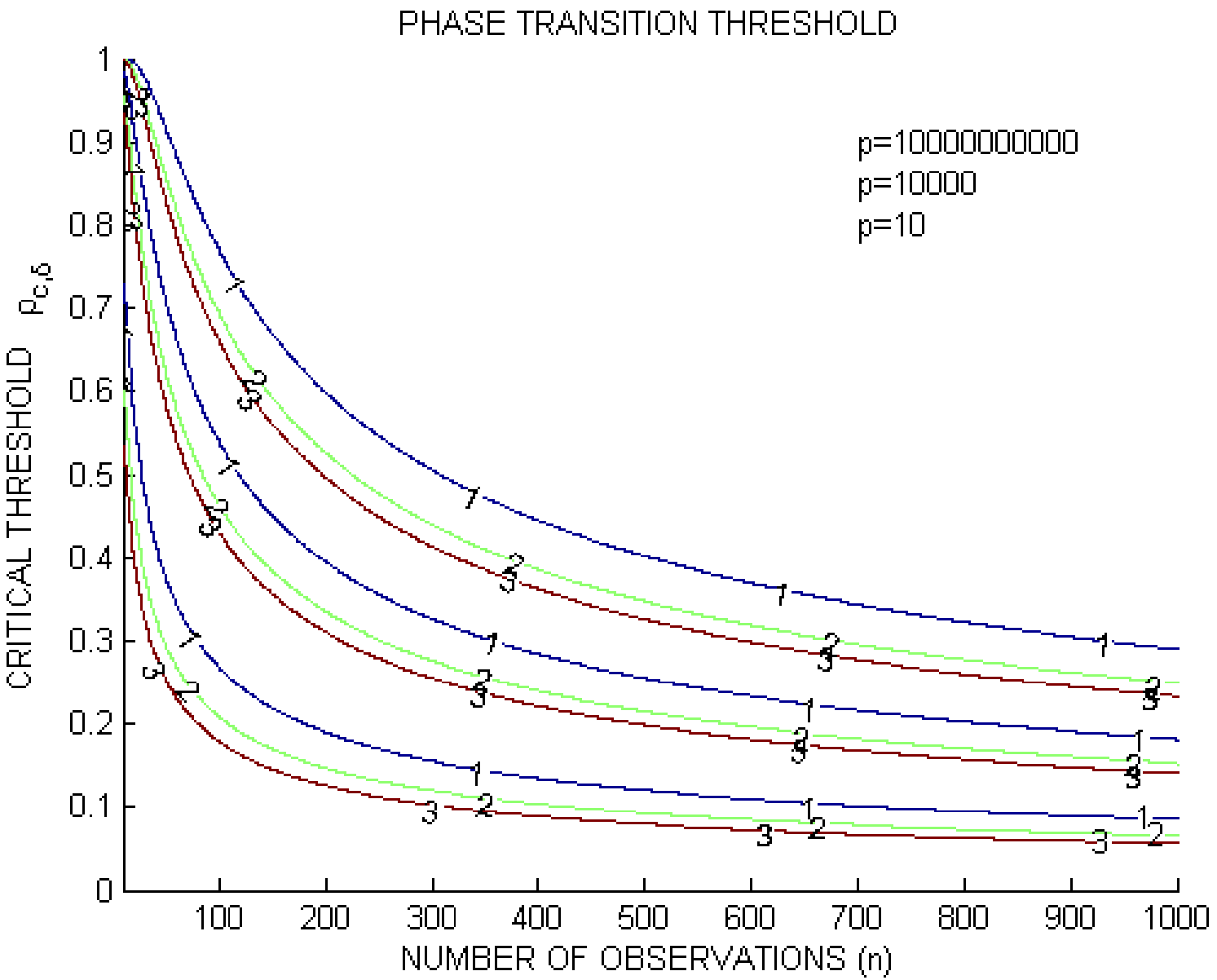


p=10 ($\delta = 1$) p=10000

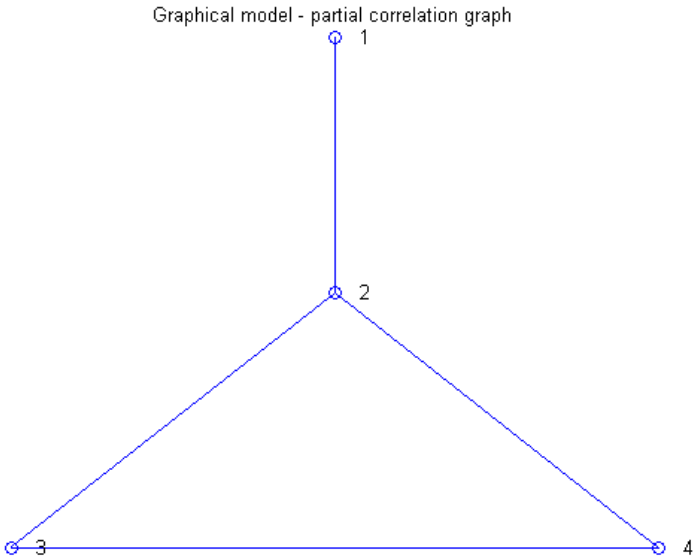
Critical threshold:

$$\rho_c = \sqrt{1 - c_{\delta,n}(p-1)^{-2\delta/\delta(n-2)-2}}$$

Phase transitions as function of δ, p



Example: 4-node Graphical Model



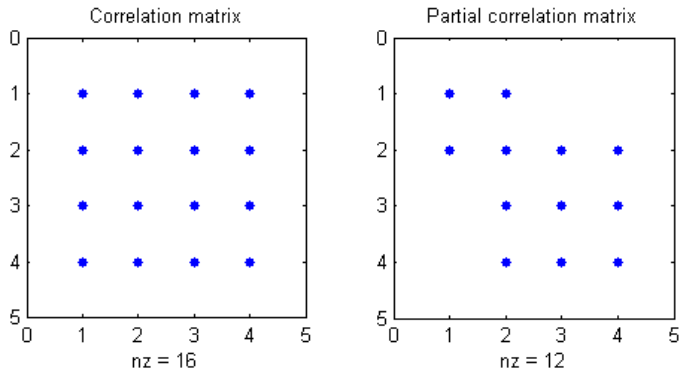
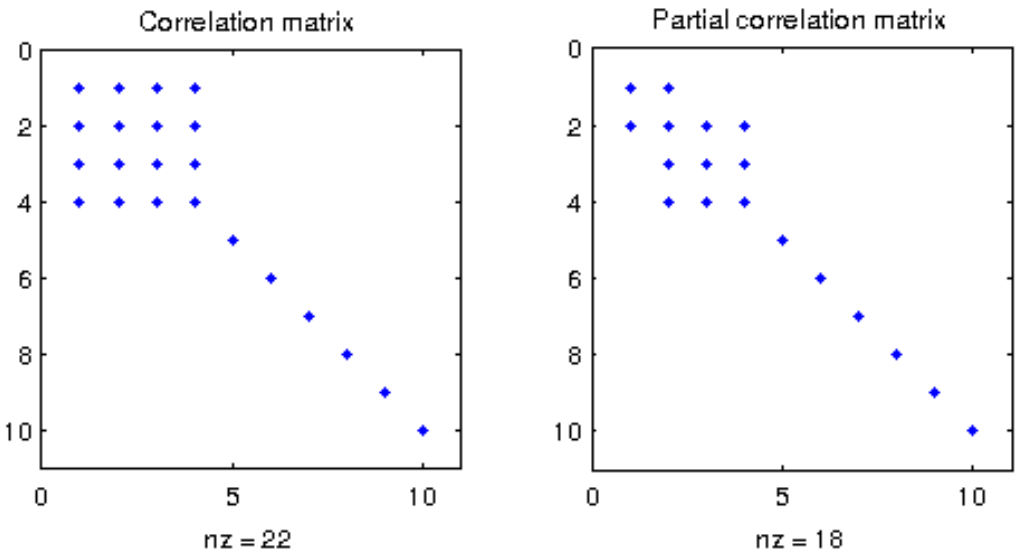


Figure: Graphical model with 4 nodes. Vertex degree distribution: 1 degree 1 node, 2 degree 2 nodes, 1 degree 3 node.

P =

1.0000	0.4069	0	0
0.4069	1.0000	-0.5179	-0.8138
0	-0.5179	1.0000	0.7071
0	-0.8138	0.7071	1.0000

Example: First 10 nodes of 1000-node Graphical Model



- 4 node Gaussian graphical model embedded into 1000 node network with 996 i.i.d. "nuisance" nodes
- Simulate 40 observations from these 1000 variables.
- Critical threshold is $\rho_{c,1} = 0.593$. 10% level threshold is $\rho = 0.7156$.

Hub screening p-value computation

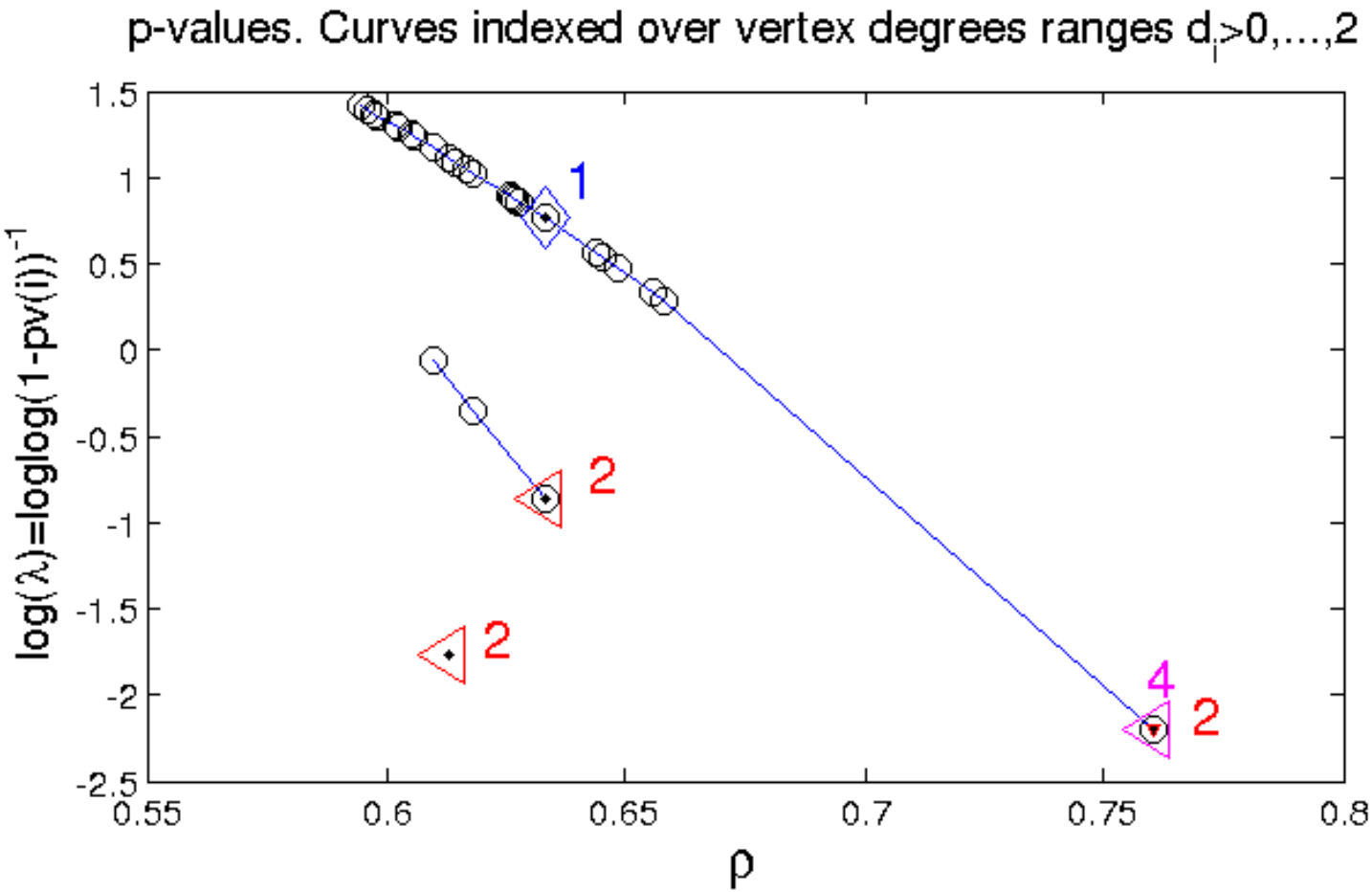
- Hub screening p-value algorithm:
  - Step 1: Compute critical phase transition threshold $\rho_{c,1}$ for discovery of connected vertices
  - Step 2: Generate partial correlation graph with threshold $\rho^* > \rho_{c,1}$
  - Step 3: Compute p-values for each vertex of degree $\delta = k$ found

$$pv_k(i) = P(N_{k,\rho(i)} > 0) = 1 - \exp(-\lambda_{k,\rho(i,k)})$$

    where $\rho(i,k)$ is sample correlation between $\mathbf{X}_i$ and its k-th NN.
  - Step 4: Render these p-value trajectories as a "waterfallplot'

$$\log(\lambda)_{k,\rho(i,k)} \text{ vs. } \rho(i,k) \text{ for } k = 1, 2, \ldots$$

Example: 1000-node Graphical Model



Note: $\log(\lambda) = -2$ is equivalent to pv$= 1 - e^{-e^{\log\lambda}} = 0.127$.

Outline

Example: NKI gene expression dataset

Netherlands Cancer Institute (NKI) early stage breast cancer

- $p = 24,481$ gene probes on Affymetrix HU133 GeneChip
- 295 samples (subjects)
- Peng *et al* used 266 of these samples to perform covariance selection
  - They preprocessed (Cox regression) to reduce number of variables to $1,217$ genes
  - They applied sparse partial correlation estimation (SPACE)
- Here we apply hub screening directly to all $24,481$ gene probes
- Theory predicts phase transition threshold $\rho_{c,1} = 0.296$

Waterfall plot of p-values for sham NKI dataset

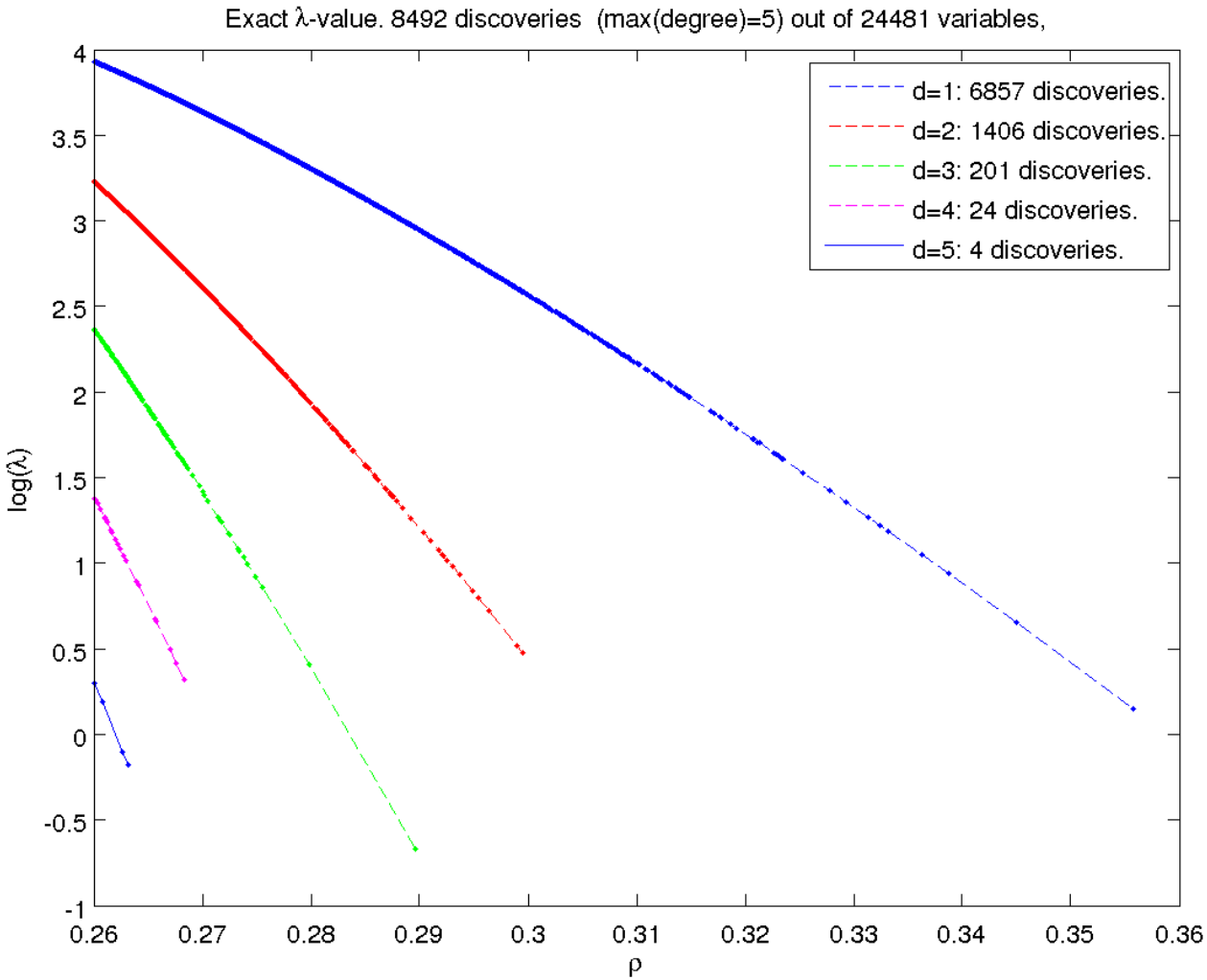


Figure: Waterfall plot of log p-values for concentration hub screening of a sham version of the NKI dataset.

Mean discovery rate validation for sham NKI dataset

observed degree	# predicted ($E[N_{\delta,\rho^*}]$)	# actual (N_{δ,ρ^*})
$d_i \geq \delta = 1$	8531	8492
$d_i \geq \delta = 2$	1697	1635
$d_i \geq \delta = 3$	234	229
$d_i \geq \delta = 4$	24	28
$d_i \geq \delta = 5$	2	4

Table: Fidelity of the predicted (mean) number of false positives and the observed number of false positives in the realization of the sham NKI dataset experiment shown in Fig. 6

.

Waterfall plot of p-values for NKI dataset

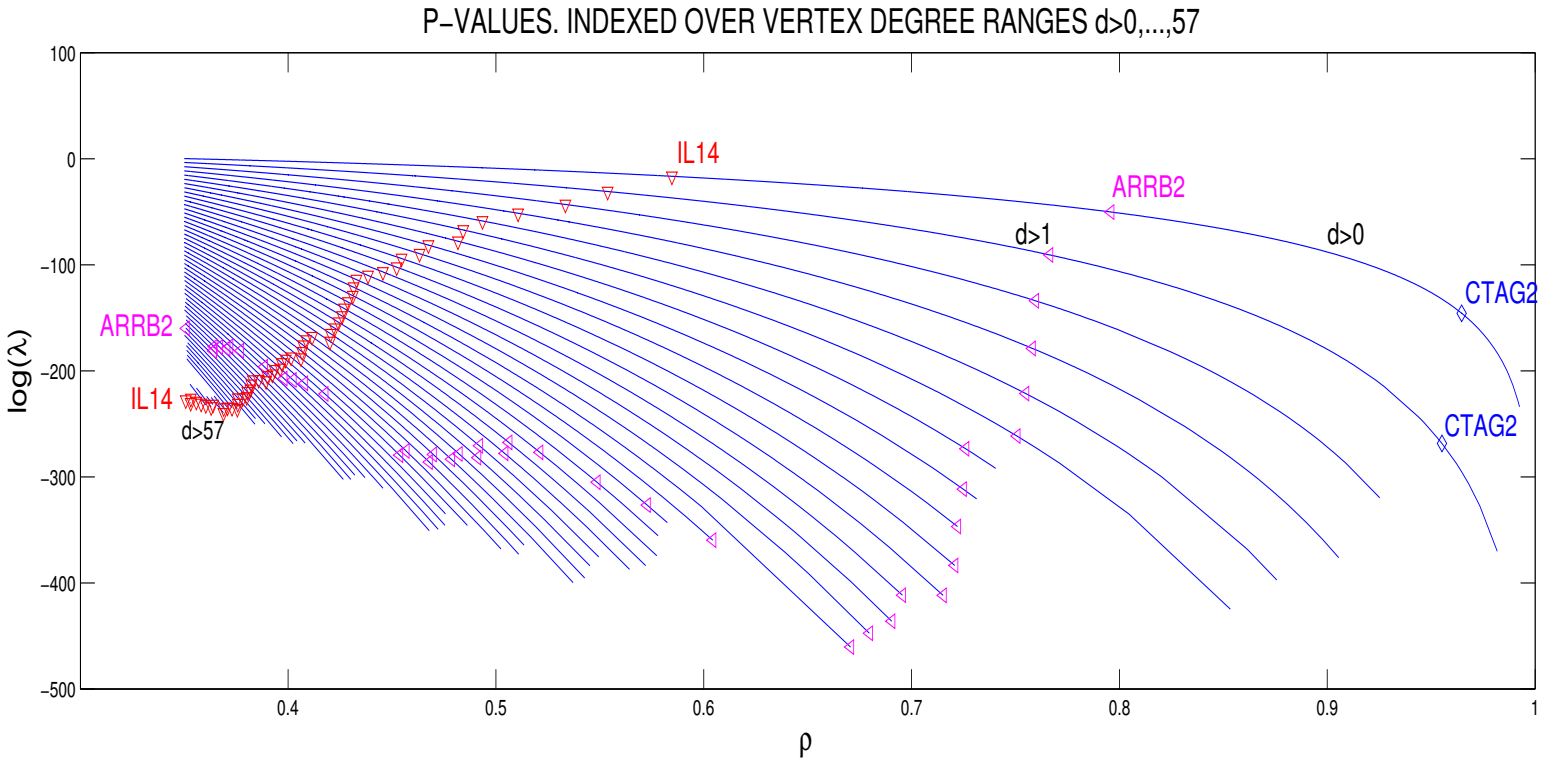


Figure: Waterfall plot of p-values for concentration hub screening of a of NKI dataset. Selected vertex discoveries.

Waterfall plot of p-values of NKI dataset

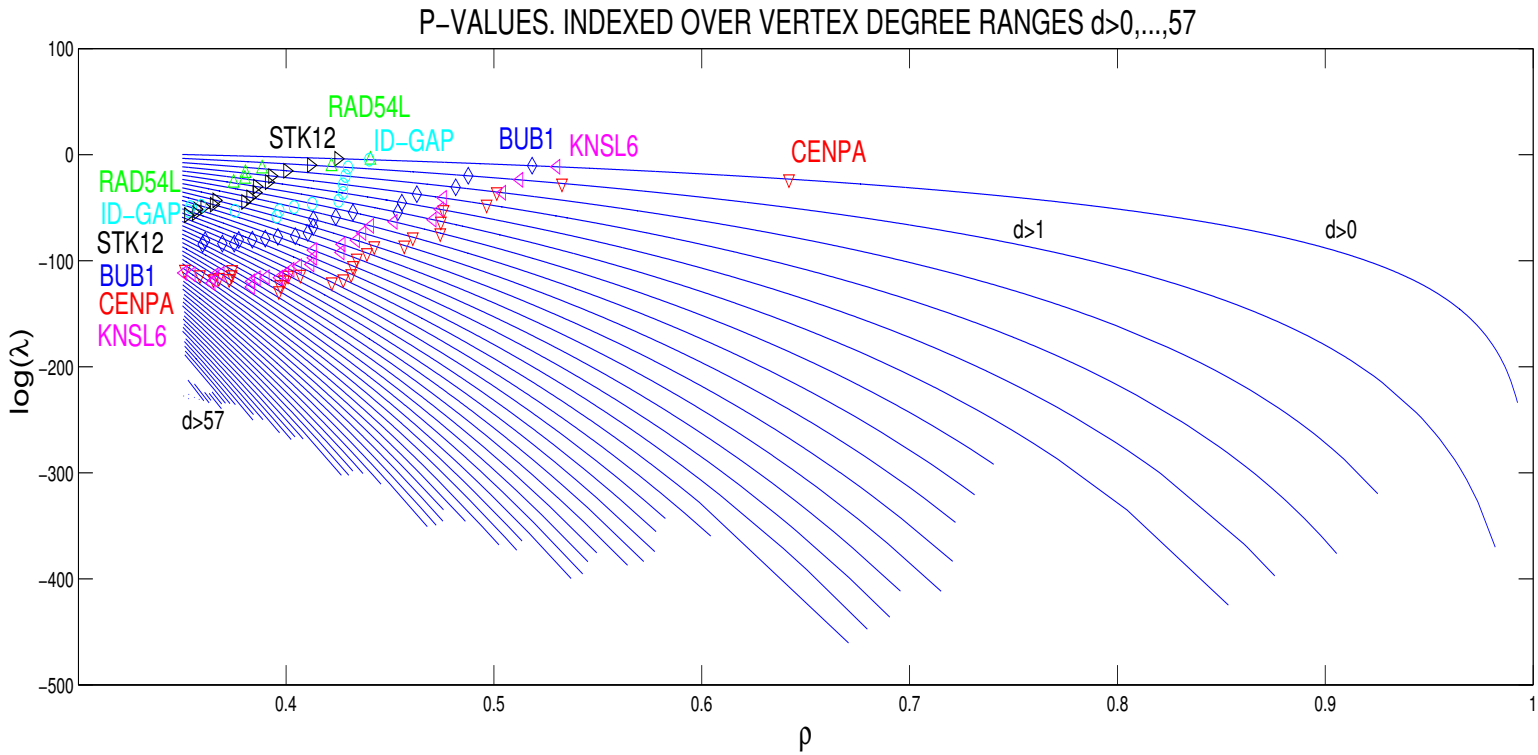


Figure: Waterfall plots of p-values for concentration hub screening of a of NKI dataset. Vertex discoveries using SPACE (Peng *et al*).

Outline

Final remarks

Large scale hub screening of correlation graphs

- Number of variables $= p \gg n =$ number of samples
- Mean number of discoveries: exhibit sharp phase transition
- Critical phase transition threshold exists
- Poisson-type limits hold on the number of discoveries
- P-value waterfall plots facilitate large scale hub discovery

References:

• H and Rajaratnam (2011), "Large scale correlation screening," JASA 2011 and arXiv 2011.

• H and Rajaratnam (2012), "Hub discovery in partial correlation graphical models," IEEE IT 2012 and arXiv 2011.