Gene Co-expression Network Reconstruction

With c-level Partial Correlation Graph

Hao Wang

Iowa State University

October 18, 2019

Section 1

Introduction

Backgound & Motivation

- System Biology: to learn the complex functional interactions between all molecules at the level of the cell Barabasi and Oltvai (2004), Boccaletti (2010)
- Network analysis: a popular methods used to study inner structure of a complex system Albert and Barabási (2002)
  - Node: individual
  - Edge: appreciable association between individuals

Figure 1: Social Network

https://www.how2shout.com/tools/top-best-open-source-social-network-platforms.html

Background & Motivation

- Genes and gene products do not work in isolation
- In the context of cellular network
  - Gene-Gene interaction network
  - Gene-protein interaction network
  - Protein-protein interaction network

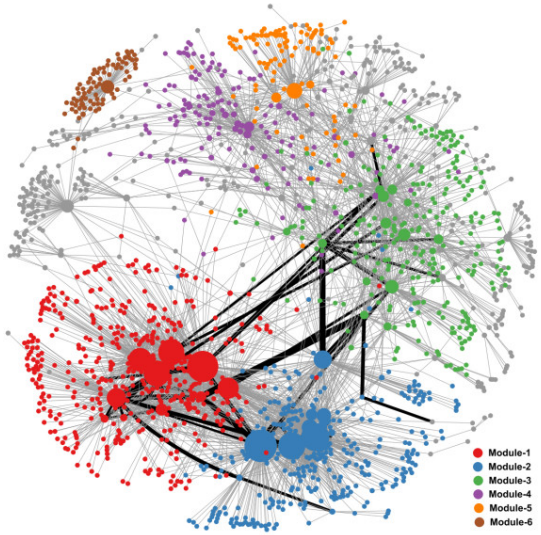

Figure 2: Gene Network

Gu, Zhang, and Wang (2012)

Background & Motivation

- Gene co-expression network
  - Undirected graph
  - Modeling similar gene expression profiles (co-expression relationships)
  - Edges represent pairwise expression similarities
- Gene regulatory network
  - Directed graph
  - Modeling how regulators govern the gene expression levels of mRNA and proteins
  - Edges represent signals/information passed in systems

Background & Motivation

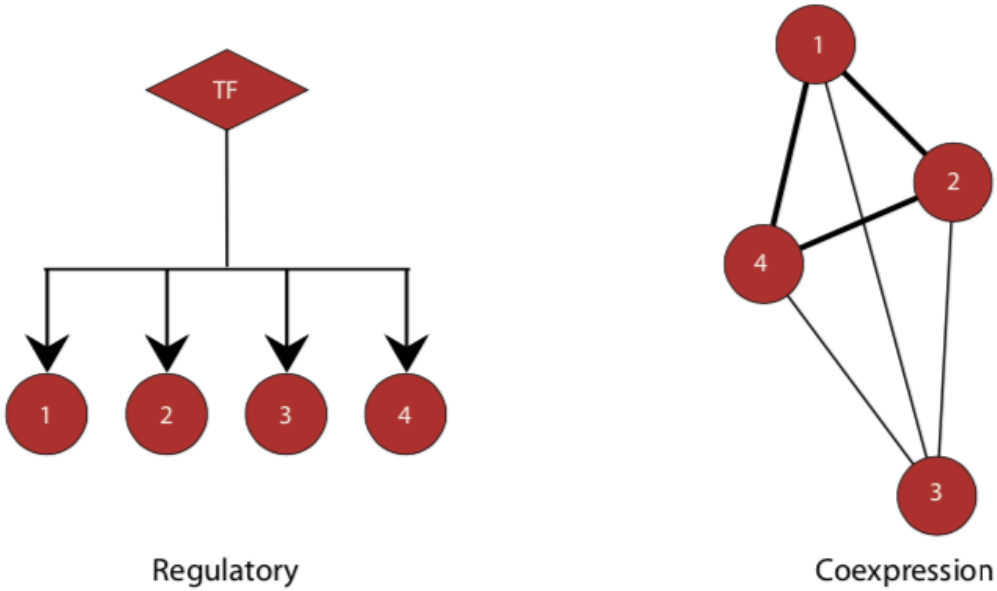

Figure 3: Regulatory and Co-expression Network

Boccaletti (2010)

Background & Motivation

- Gene co-expression network
  - Undirected graph
  - Modeling similar gene expression profiles (co-expression relationships)
  - Edges represent pairwise expression similarities
- Gene regulatory network
  - Directed graph
  - Modeling how regulators govern the gene expression levels of mRNA and proteins
  - Edges represent signals/information passed in systems
- We focused on gene co-expression network

Genomics Technologies

- Integrating biological information to construct networks theoretically or experimentally can be difficult
- Thanks to high-throughput genomics technologies
  - DNA microarray Heller (2002)
  - next-generation sequencing Ansorge (2009)
- Data driven methods can be applied on large-scale and high-quality datasets

Genomics Technologies

- A typical gene expression dataset looks like this

Table 1: Gene Expression Data Example

Genes	Sample 1	Sample 2	...	Sample N
Gene 1	ge_{11}	ge_{12}	...	ge_{1N}
Gene 2	ge_{21}	ge_{22}	...	ge_{2N}
...	...	...	...	...
Gene P	ge_{P1}	ge_{P2}	...	ge_{PN}

Existing Methods

How to investigate pairwise gene-gene associations?

- Correlation based network
  - Using the Pearson correlation as the measurement of indirect linear associations
  - To explain how genes are marginally associated with each other
- Information theory based network
  - Using mutual information to represent indirect non-linear associations
  - A more generalized measure of probabilistic dependency
- Gaussian Graphical Model
  - Using partial correlation to measure direct linear association
  - To assess how genes are directly connected to each other

Wang and Huang (2014), Yu et al. (2013)

Marginal v.s. Conditional Dependency

```
##          A      B      C   D     E
## A    1.000  0.000 -0.177 0.0 0.000
## B    0.000  1.000 -0.729 0.0 0.000
## C   -0.177 -0.729  1.000 0.0 0.393
## D    0.000  0.000  0.000 1.0 0.200
## E    0.000  0.000  0.393 0.2 1.000
```

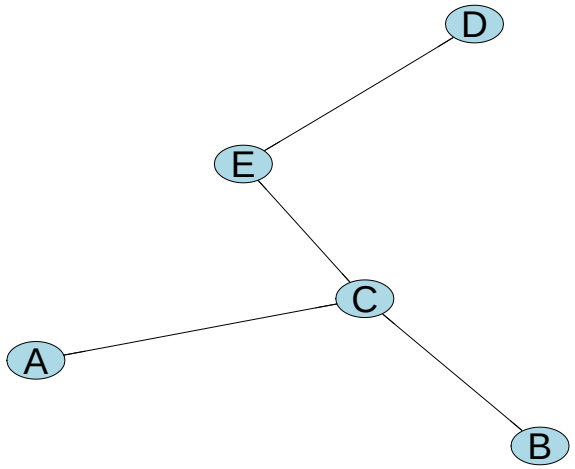

Figure 4: True Network

Marginal v.s. Conditional Dependency

```
##          A      B      C      D      E
## A    1.000  0.265 -0.328 -0.035 -0.162
## B    0.265  1.000 -0.788 -0.117 -0.463
## C   -0.328 -0.788  1.000  0.136  0.587
## D   -0.035 -0.117  0.136  1.000  0.261
## E   -0.162 -0.463  0.587  0.261  1.000
```

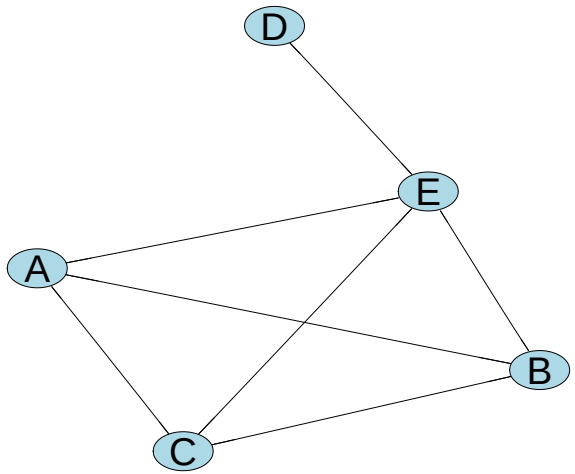

Figure 5: Correlation Network

Marginal v.s. Conditional Dependency

```
##          A      B      C      D     E
## A    1.000  0.011 -0.200  0.001 0.039
## B    0.011  1.000 -0.703 -0.016 0.003
## C   -0.200 -0.703  1.000 -0.026 0.402
## D    0.001 -0.016 -0.026  1.000 0.226
## E    0.039  0.003  0.402  0.226 1.000
```

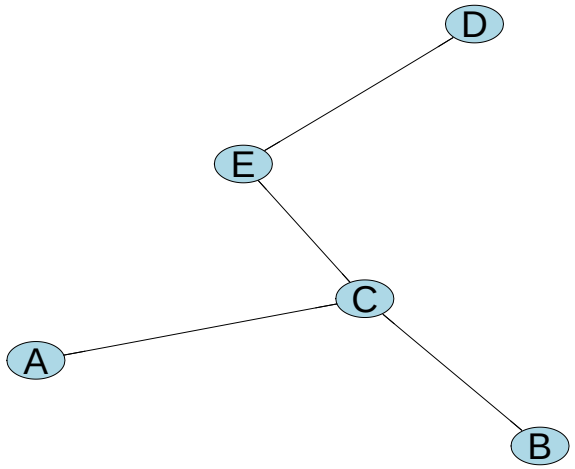

Figure 6: Partial Correlation Network

Marginal v.s. Conditional Dependency

- The Pearson correlation is not a good measure of gene-gene association
- It can not distinguish the direct and indirect association
- Gaussian Graphical Model is preferred

Gaussian Graphical Model

- Starting with an undirected graph $G = (V, E)$
  - Where V is a set of nodes and E is a set of edges
- Let $X = (X_1, X_2, X_3, ..., X_p)$ denote the random vector associated with p genes
- X is assumed to be from $N(0, \Sigma)$
- A Gaussian graphical model is represented by the corresponding partial correlation matrix P.

Gaussian Graphical Model

- Theoretically, there are two approaches to get the partial correlation
  - Using the precision matrix Ω ($\Omega = \Sigma^{-1}$)
    - The partial correlation between variable X_i and variable X_j, denoted as ρ_{ij} is defined by

$$\rho_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} I(i \neq j) + I(i = j)$$

  - The second approach uses regressions
    - The partial correlation between X_i and X_j given the other $p-2$ controlling variables $X_{-(i,j)}$ (the set of variables from X_1 to X_p except X_i and X_j), written ρ_{ij}, is the Pearson correlation between the residuals ϵ_i and ϵ_j resulting from the linear regression of X_i with $X_{-(i,j)}$ and of X_j with $X_{-(i,j)}$

Limitations & Challenges

- High-dimension (lagre P and small N scenario)
  - hard to estimated the measure of interaction (underestimating, singularity, and so on)
  - computationally inefficient
- Much noise & very small sample size
  - low precision
- Lack of statistical inference

Limitations & Challenges

- Several methods have been developed trying to solve mentioned challenges
  - Via inverting sample covariance matrix
    - Empirical Bayes approach (pseudoinverse & bootstrapping) Schäfer and Strimmer (2004)
    - Shrinkage approach with empirical null fitting Schäfer and Strimmer (2005)
  - Via regression
    - GLasso Friedman, Hastie, and Tibshirani (2008)
    - SPACE (Sparse PArtial Correlation Estimation) Peng et al. (2009)
- Lack of statistical inference

Novel Methods

- Covariance matrix based
  - Exact hypothesis testing for shrinkage based Gaussian graphical models (Shrunk MLE) Bernal et al. (2019)
- Regression based
  - c-level Partial Correlation Graph (c-level PCG) Qiu and Zhou (2018)

Section 2

Methods

Shrunk MLE

- Covariance matrix estimator inherits Schäfer and Strimmer (2005)

$$\hat{C}^{\lambda} = (1-\lambda)\hat{C}^{SM} + \lambda T, \quad \hat{\Omega}^{-1} = \hat{C}^{\lambda}, \quad \rho_{ij} = -\frac{\Omega_{ij}}{\sqrt{\Omega_{ii}\Omega_{jj}}}$$

- The distribution across edges $f(\rho)$ is assumed to be a mixture density of the form

$$f(\rho) = \pi_0 f_0(\rho) + (1-\pi_0) f_1(\rho)$$

- Where π_0 is the proportion of the null edges, $f_0(\rho)$ is the probability density for $\rho = 0$, and $f_1(\rho)$ the probability density for the real effects $(\rho \neq 0)$
- $\eta_0 + \eta_A = 1$, $\eta_0 >> \eta_A$

Shrunk MLE

- Schäfer and Strimmer (2005) propose under simulation studies (for small λ) the distribution of the 'shrunk' partial correlation is close to the standard partial correlation (i.e. without shrinkage) as

$$f_0(\rho) = \frac{1}{Beta(\frac{1}{2}, \frac{k-1}{2})}(1-\rho^2)^{(k-3)/2}$$

- To keep it simple, f_1 is assumed to be U(-1,1) density
- k and η_0 are found by maximizing the corresponding likelihood (mixture density)
- Then p-values can be found by f_0
- Alternatively, Prob(null edge$|\rho$)$=\frac{\hat{\eta}_0 f_0 \rho;\hat{k}}{f(\rho;\hat{k})}$ can be computed Efron (2005)
- However, as the shrinkage effects are not included, the P-values are suboptimal

Shrunk MLE

- Bernal et al. (2019) improved this method by proposing the exact distribution of shrunk partial correlations

$$f_0^{\lambda}(r^{\lambda}) = \frac{((1-\lambda)^2 - r^{\lambda^2})^{(k-3)/2}}{Beta(\frac{1}{2}, \frac{k-1}{2})(1-\lambda)^(k-2)}$$

- k can be estimated via maximum likelihood estimation with simulated r^{λ} under null
- P-values can be calculated with the null density
- Multiple testing correction (e.g. Benjamini and Hochberg (1995)) is needed to control false discovery rate

c-level PCG

- Recall $\Omega = \{\omega_{j_1,j_2}\}_{p\times p} = \Sigma^{-1}$
- Lemma 1 from Peng et al. (2009) claims the partial correlation can be expressed via only p node-wise regressions
- Let $Y_{-j} = (y_1, ..., y_{j-1}, y_{j+1}, ..., y_p)^T$
- Theoretically

$$y_{j_1} = \alpha_{j_1,0} + \sum_{j_2 \neq j_2} \alpha_{j_1,j_2} y_{j_2} + \epsilon_{j_1},\ j_1 = 1, 2, ..., p.$$

- ϵ_{j_1} is uncorrelated with Y_{-j1} if only if $\alpha_{j_1,j_2} = -\frac{\omega_{j_1,j_2}}{\omega_{j_1,j_1}}$ for any $j_2 \neq j_1$

c-level PCG

- Then it can be shown that

$$Var(\epsilon_{j_1}) = \frac{1}{\omega_{j_1}}, \text{ and } Cov(\epsilon_{j_1}, \epsilon_{j_2}) = \frac{\omega_{j_1,j_2}}{\omega_{j_1,j_1}\omega_{j_2,j_2}} = -\frac{\rho_{j_1,j_2}}{(\omega_{j_1,j_1}\omega_{j_2,j_2})^{1/2}}$$

- Let $\epsilon = (\epsilon_1, ..., \epsilon_p)^T$ and $V = Cov(\epsilon) = \{v_{j_1,j_2}\}_{p\times p}$
- The partial correlation can be expressed as

$$\rho_{j_1,j_2} = -\frac{v_{j_1,j_2}}{\sqrt{\omega_{j_1,j_1}\omega_{j_2,j_2}}}, \ j_1 \neq j_2$$

c-level PCG

- To get estimated partial correlation, node-wide regression is fitted by lasso
- Tunning parameter λ is pre-specified as $\sqrt{2 \times log(p)/n}$
- Let $\hat{\epsilon}_i = (\hat{\epsilon}_{i,1}, ..., \hat{\epsilon}_{i,p})^T$ be the residuals of the i^{th} observation
- Let $\tilde{V} = \{\tilde{v}_{j_1,j_2}\}$ be the sample covariance of the residuals, where $\tilde{v}_{j_1,j_2} = \sum_{i=1}^{n} \frac{\hat{\epsilon}_{i,j_1}\hat{\epsilon}_{i,j_2}}{n}$
- Although $\sum_{i=1}^{n} \frac{\epsilon_{i,j_1}\epsilon_{i,j_2}}{n}$ is an unbiased estimator of v_{j_1,j_2}, replacing $\epsilon_{i,j}$ by $\hat{\epsilon}_{i,j}$ will incur a bias term

c-level PCG

- The authors propose novel estimator of v_{j_1,j_2} with the form

$$\hat{v}_{j_1,j_2} = \begin{cases} -\frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon}_{i,j_1}\hat{\epsilon}_{i,j_2} + \hat{\alpha}_{j_1,j_2}\hat{\epsilon}^2_{i,j_2} + \hat{\alpha}_{j_2,j_1}\hat{\epsilon}^2_{i,j_1}), \ j_1 \neq j_2 \\ \hat{v}_{j_1,j_2} = \sum_{i=1}^{n}\frac{\hat{\epsilon}_{i,j_1}\hat{\epsilon}_{i,j_2}}{n}, \ j_1 = j_2 \end{cases} \tag{1}$$

- Then the partial correlation between gene j_1 and j_2 is estimated by

$$\hat{\rho}_{j_1,j_2} = -\hat{v}_{j_1,j_2} \times \sqrt{\hat{\omega}_{j_1,j_1}\hat{\omega}_{j_2,j_2}} = -\frac{\hat{v}_{j_1,j_2}}{\sqrt{\hat{v}_{j_1,j_1}\hat{v}_{j_2,j_2}}}$$

c-level PCG

- Inference of estimated partial correlation is based on the uncertainty of the estimator and false discovery rate control
- Variance of the estimator is $nVar(\hat{\rho}_{j_1,j_2}) = \kappa(1-\rho^2_{j_1,j_2})^2 1 + o(1)$, where $\kappa = E(\epsilon_j^4)/[3E^2(\epsilon_j^2)]$
- Let $\tilde{\rho}_{j_1,j_2} = \hat{\rho}_{j_1,j_2} I(|\hat{\rho}_{j_1,j_2}| > 2[log(p)/n]^{1/2})$
- $nVar(\hat{\rho}_{j_1,j_2})$ is estimated by $\hat{\kappa}[1-\tilde{\rho}^2_{j_1,j_2}]^2$, where $\hat{\kappa} = \frac{n}{3p}\sum_{j=1}^{p} \frac{\sum_{i=1}^{n}\hat{\epsilon}^4_{i,j}}{(\sum_{i=1}^{n}\hat{\epsilon}^2_{i,j})^2}$

c-level PCG

- Then adaptive thresholding estimator for partial correlation matrix is proposed as

$$\hat{\rho}^{(t)}_{j_1,j_2}(\tau) = \hat{\rho}_{j_1,j_2} I[|\hat{\rho}_{j_1,j_2}| > \tau(1-\tilde{\rho}^2_{j_1,j_2})\{\hat{\kappa} log(p)/n\}^{1/2}]$$

- Particularly, adaptive thresholding estimator for c-level graph has form of

$$\hat{E}_c = [(j_1,j_2) : |\hat{\rho}_{j_1,j_2}| > c + \tau(1-\tilde{\rho}^2_{j_1,j_2})\{\hat{\kappa} log(p)/n\}^{1/2}]$$

c-level PCG

- To choose the threshold parameter τ, the authors control the false discovery proportion at a desired level
  - where FDP is the number of false positives over the number of discovery.
- Let $\#\{A\}$ denote the size of a set A and $\bar{A}$ denote the complementary set of A
- $FDP_c(\tau)$ can be written as $\frac{\#\{\bar{E}_c(\tau)\}FPR_c(\tau)}{max[1,\#\{\hat{E}_c(\tau)\}]}$
- After some complicated theoretical analysis, the authors show that the numerator $\#\{\bar{E}_c(\tau)\}FPR_c(\tau)$ is bounded by a function of τ, denoted as $B(\tau)$
- Then FDP is purely based on known quantities and unknown τ
- For a sequence of candidate τ in (0,2], we could choose

$$\tau_{FDP} = inf\{\tau \in (0,2] : \frac{B(\tau)}{max[1,\#\{\hat{E}_c(\tau)\}]} <= \alpha\}$$

Section 3

Data

Simulated Data

- Simulation setting used in Bernal et al. (2019) Schäfer and Strimmer (2005) by package GeneNet Schaefer, Opgen-Rhein, and Strimmer (2009)
  - Simulate networks with desired proportion of real edges by `ggm.simulate.pcor`
  - Simulate multi-normal data based on simulated networks by `ggm.simulate.data`
- However simulating partial correlation matrix in this way has a limitation
  - When p becomes large, simulated partial correlations can be very small

Simulated Data

- η_A=0.01, p=100, 300, 500, 1000

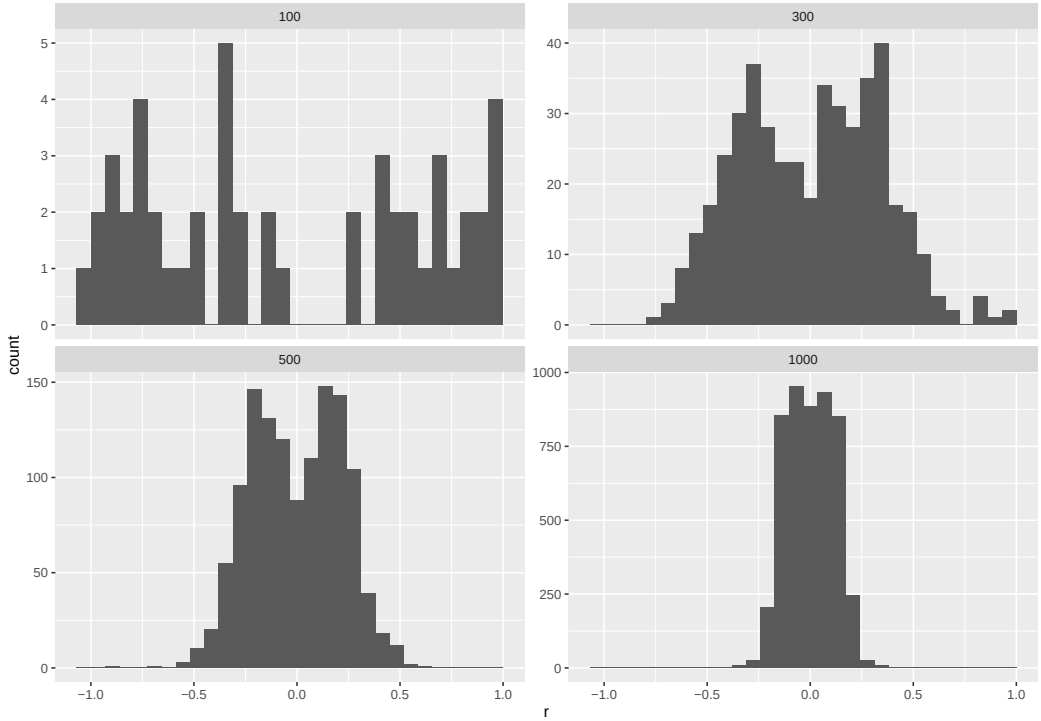

Figure 7: Distribution of Simulated Partial Correlation With different P

Simulated Data

- We simulated partial correlation matrix in a slightly different manner
- Random structure
  - Starting with the p by p empty precision matrix Ω
  - $\omega_{ij} = \omega_{ji}$ for all $i \neq j$ are equal 0 with probability $ 1- _0 $ and equal to a random number from $U(-1, 1)$ with probability ϵ_0, where ϵ_0 is the desired proportion of real edges
  - All diagonal elements are set to be 1.75 to make the matrix positive definite
  - The the corresponding partial correlation matrix is calculated by

$$\rho_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} I(i \neq j) + I(i = j)$$

Simulated Data

- Besides, since it is validated to assume
  - The biological network is very sparse Barabasi and Oltvai (2004)
  - Modules of genes exist in the biological network
- We simulated networks by setting the partial correlation matrix as block diagonal matrix
  - Starting with covariance matrix Σ directly, partial correlation matrix is calculated based on the Σ
  - The covariance matrix contains k by k symmetric sub-block matrix B along the diagonal
  - Where B is simulated by setting all off-diagonal elements to be random number from U(0.3,0.9) and k equals to 4 or 10
  - Random data is simulated from multinormal with mean zero and covariance matrix Σ

E.coli

- The dataset we used consists of E.coli microarray gene-expression from Schmidt-Heck et al. (2004)
- The expression of 4289 protein coding genes of the E.coli in total was measured using microarrays
- 102 genes were selected as they differentially expressed after normalization
- The dataset can be found from the *GeneNet* package
  - It consists of 9 observations (9 time points at 0, 8, 15, 22, 45, 68, 90, 150 and 180 min) and 102 genes.

Section 4

Results

Simulation Study

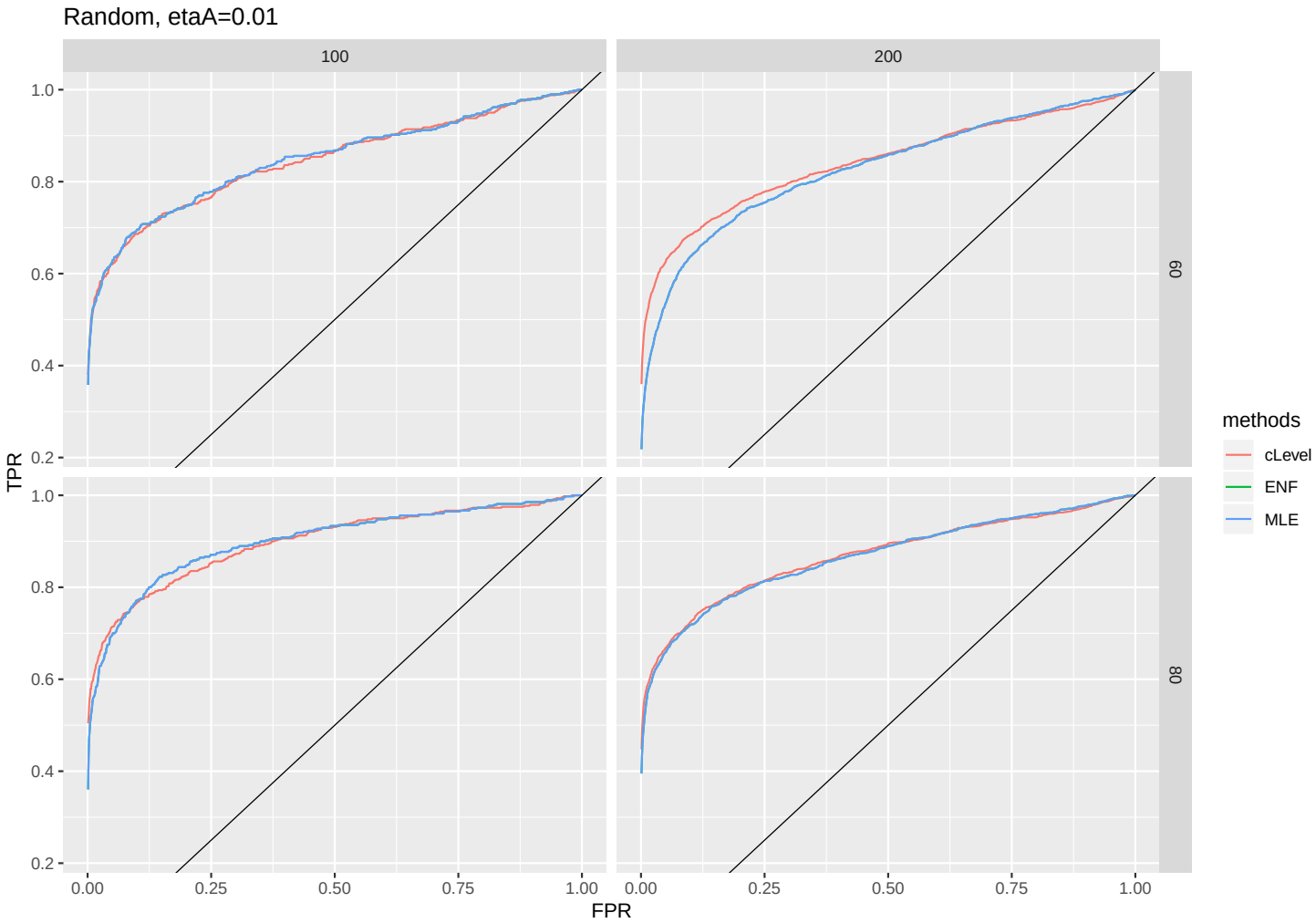

Figure 8: ROC curve, Random 0.01

Simulation Study

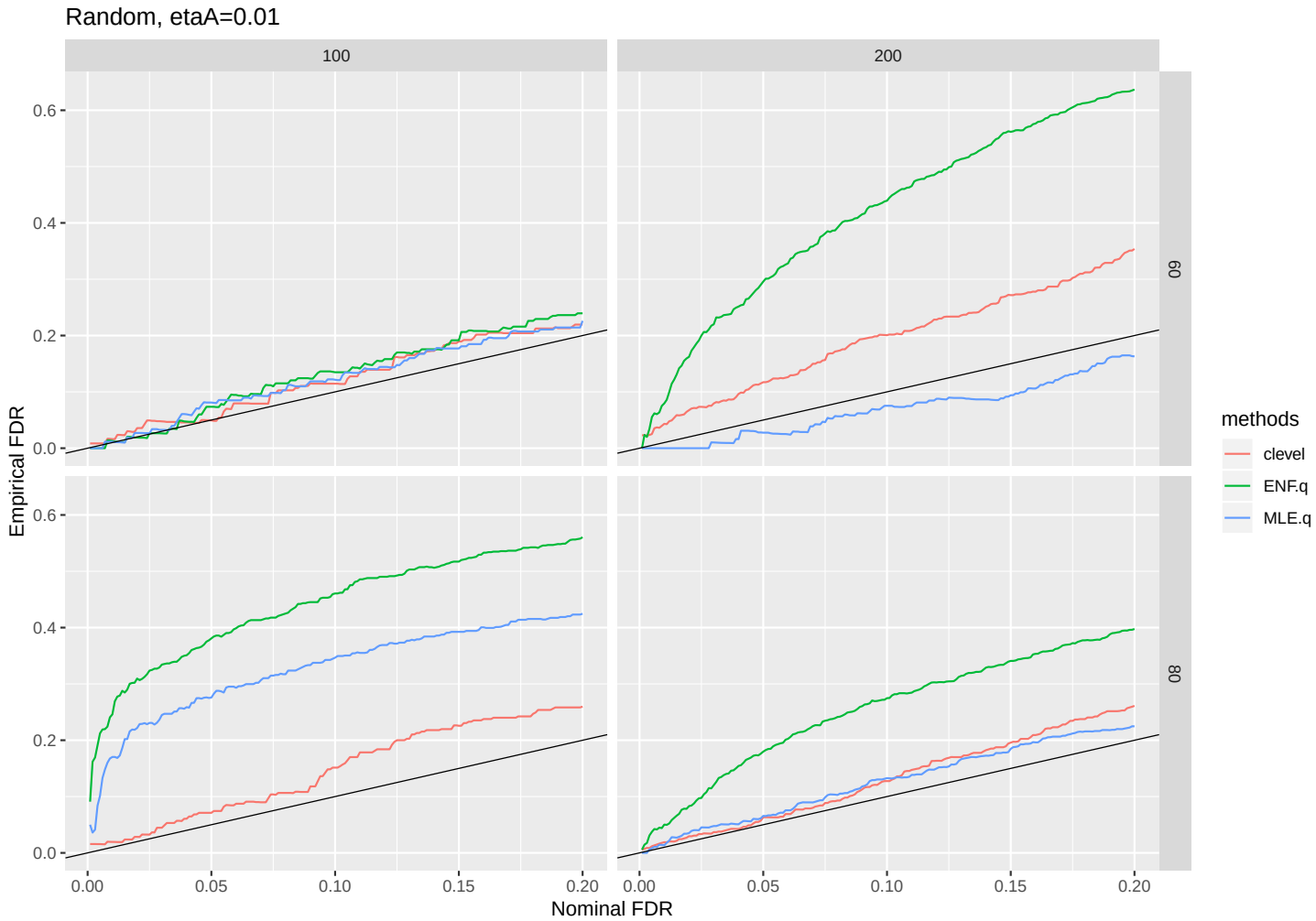

Figure 9: fdr, Random 0.01

Simulation Study

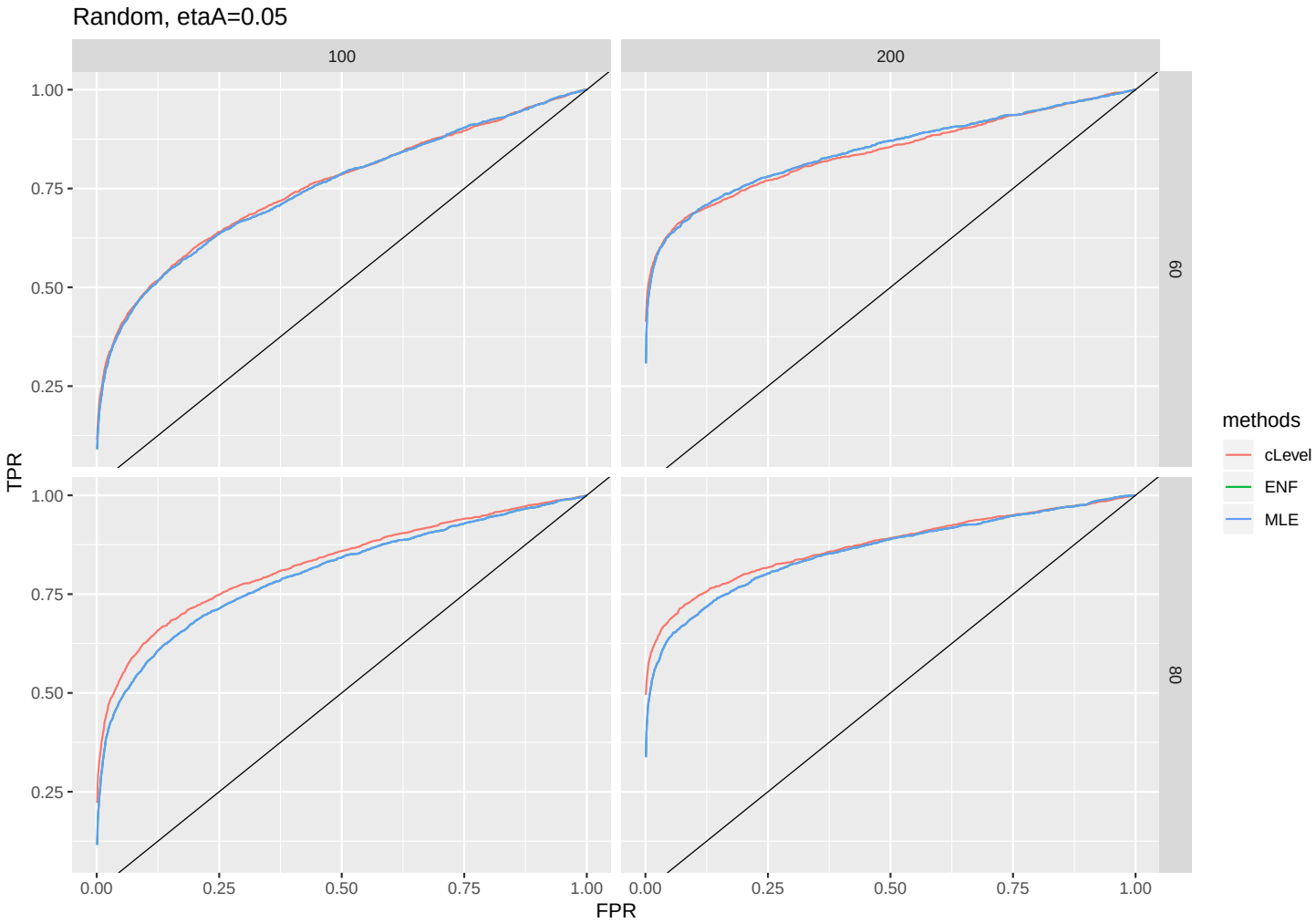

Figure 10: ROC curve, Random 0.05

Simulation Study

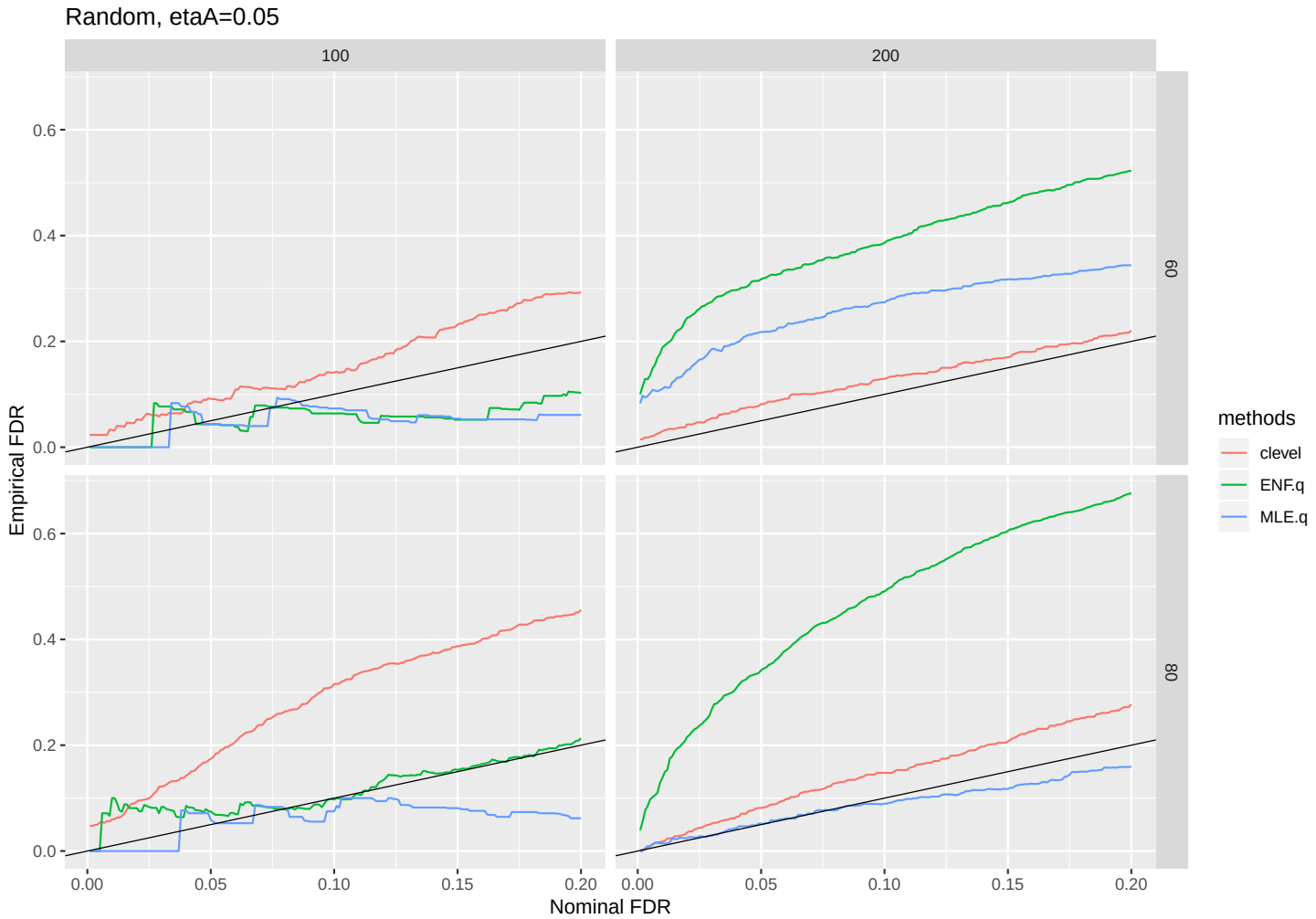

Figure 11: fdr, Random 0.05

Simulation Study

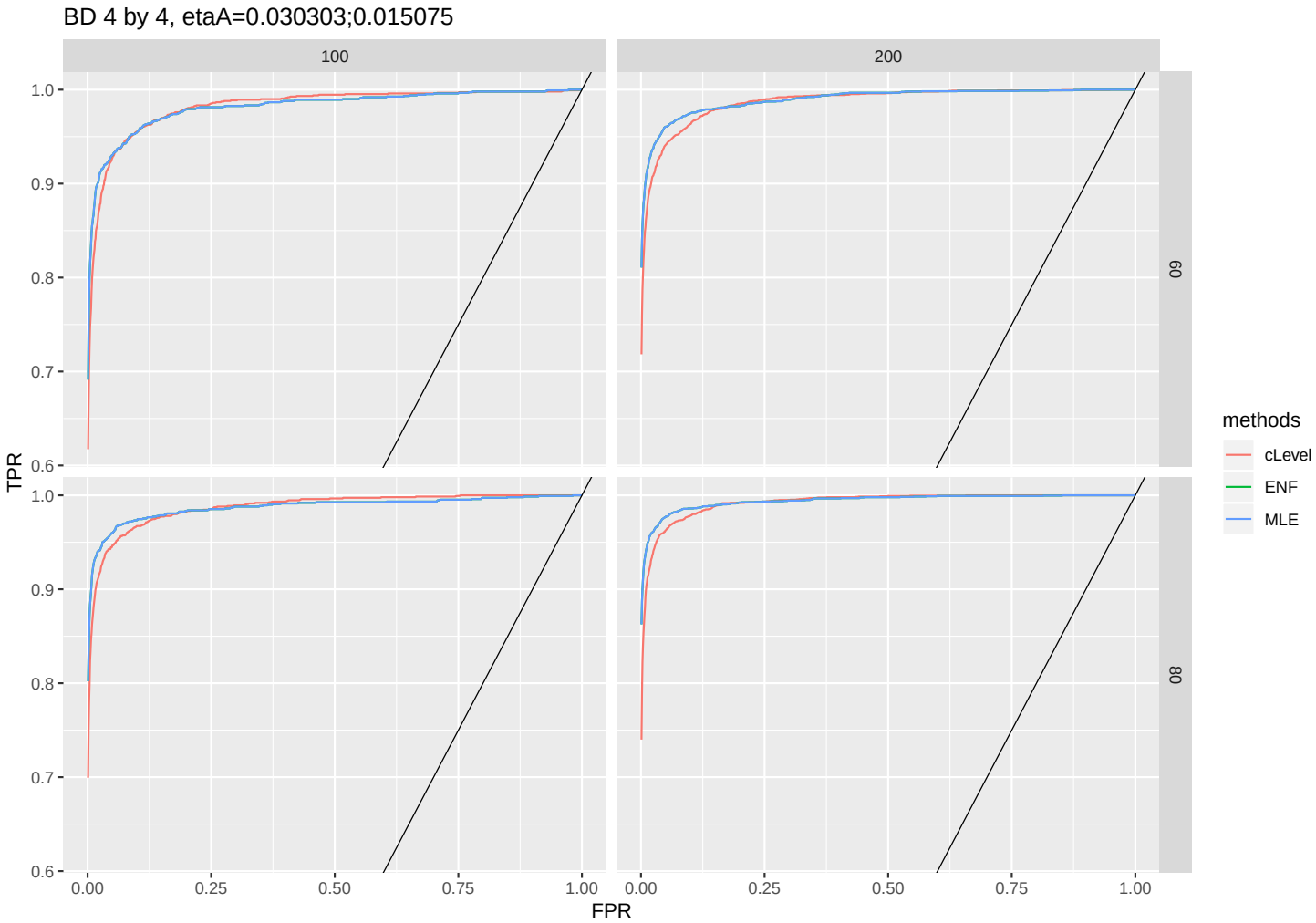

Figure 12: ROC curve, BD 4 by 4

Simulation Study

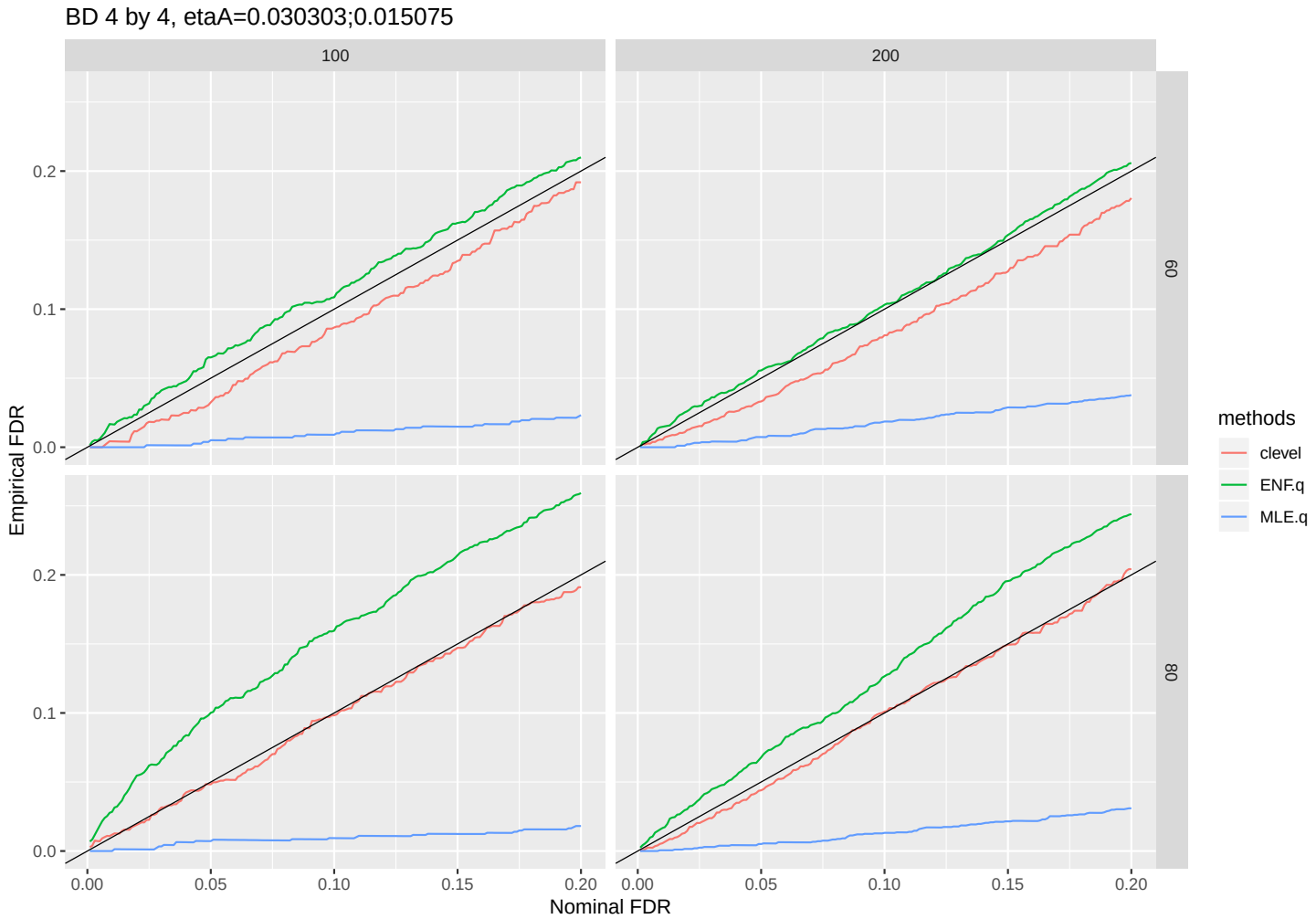

Figure 13: fdr, BD 4 by 4

Simulation Study

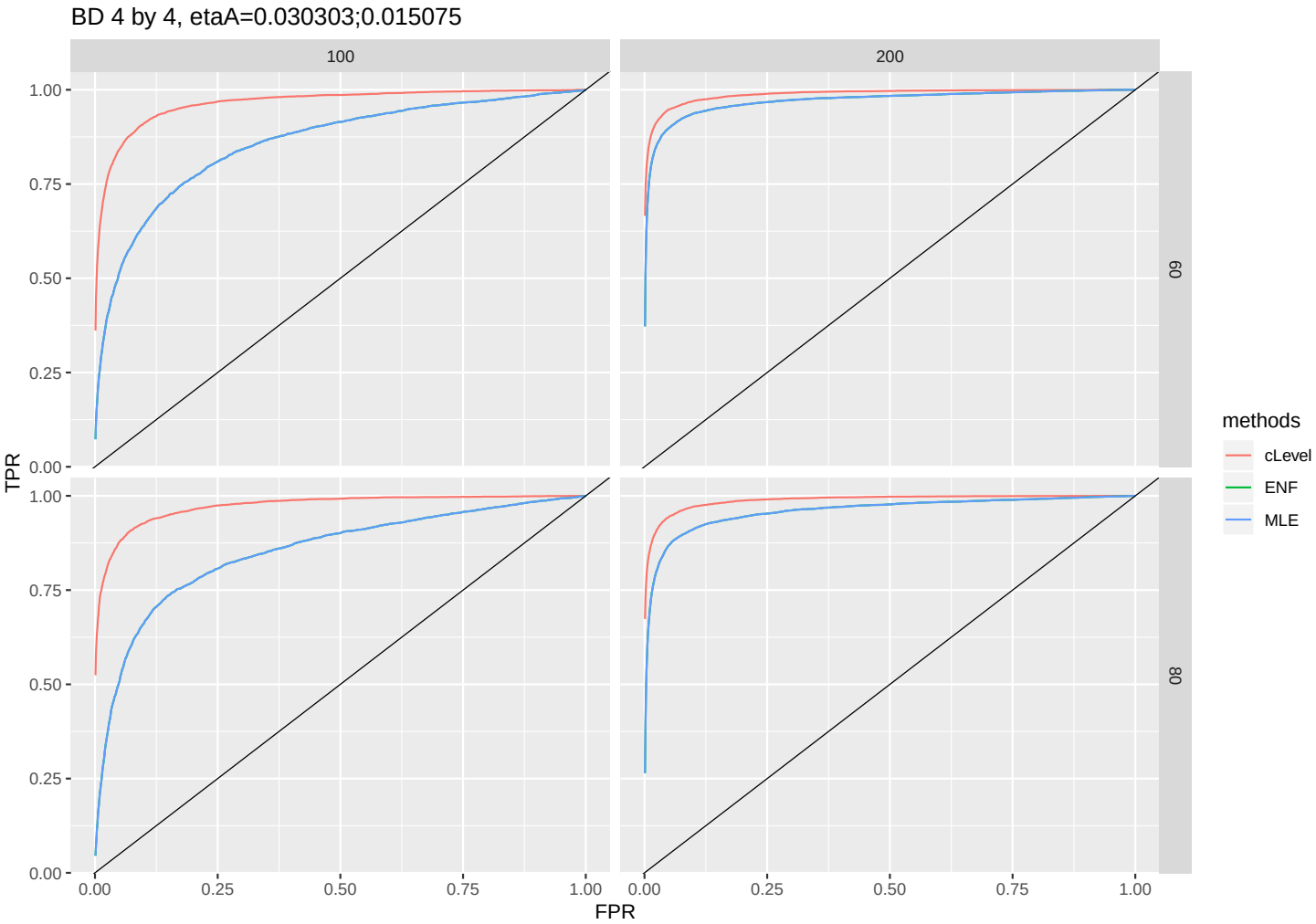

Figure 14: ROC curve, BD 10 by 10

Simulation Study

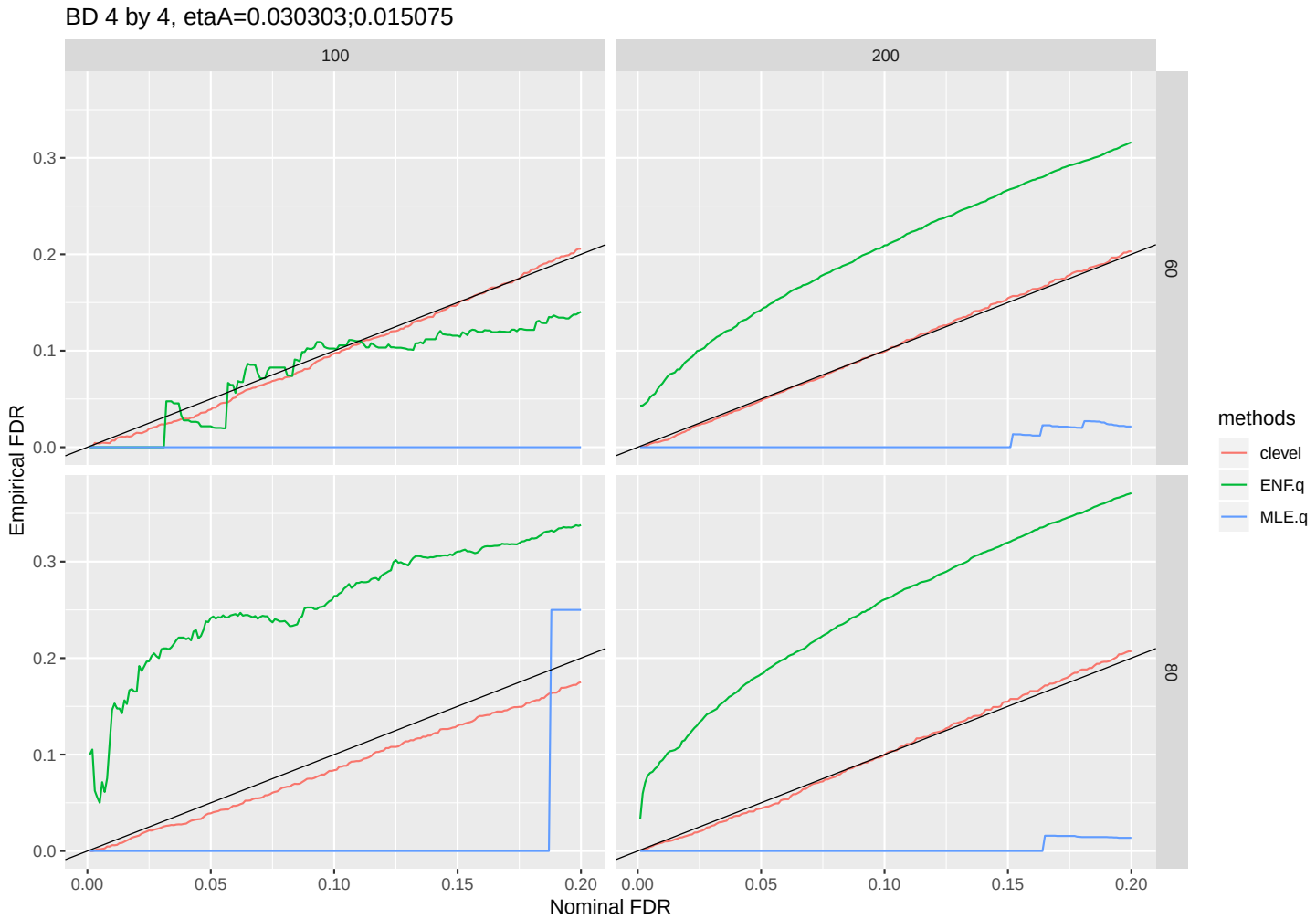

Figure 15: fdr, BD 10 by 10

Simulation Study

- n=60, $\eta_A = 0.03$
- For each network, 20 random datasets are generated

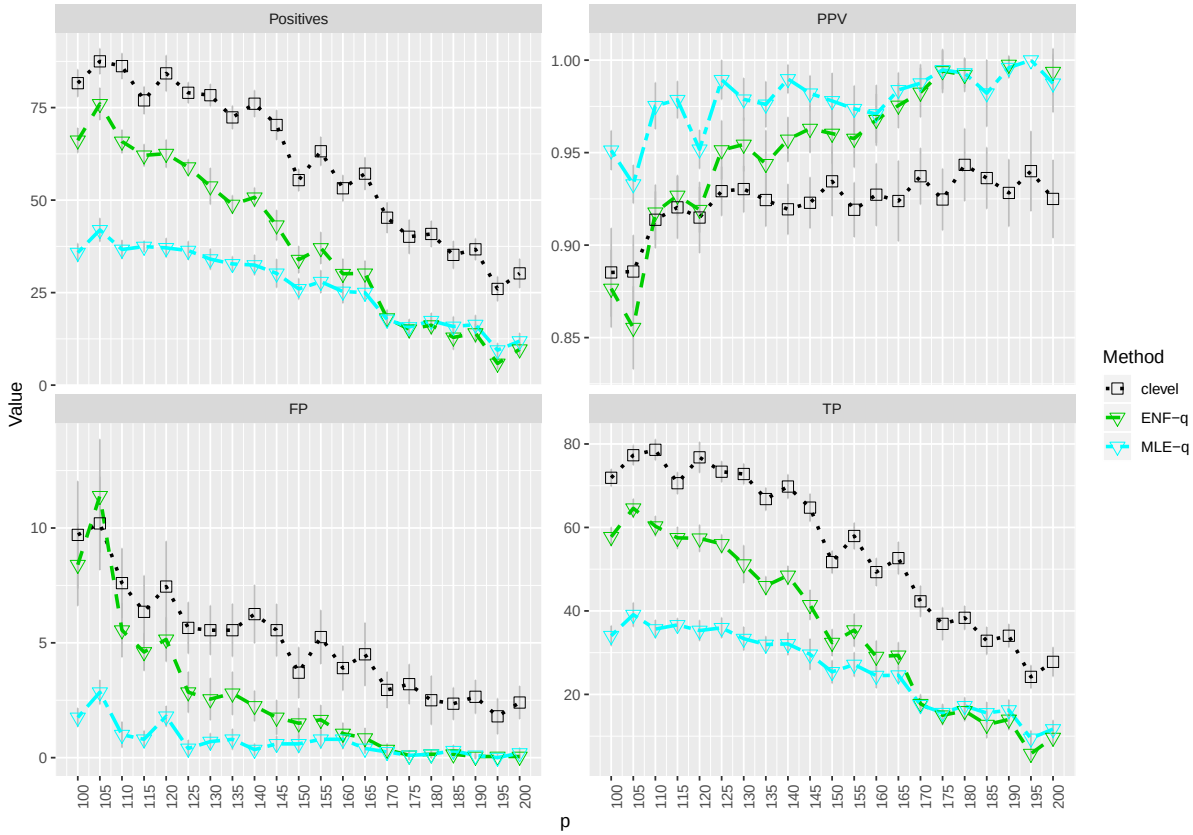

Figure 16: Inference at Alpha Equal to 0.05

Application on real dataset

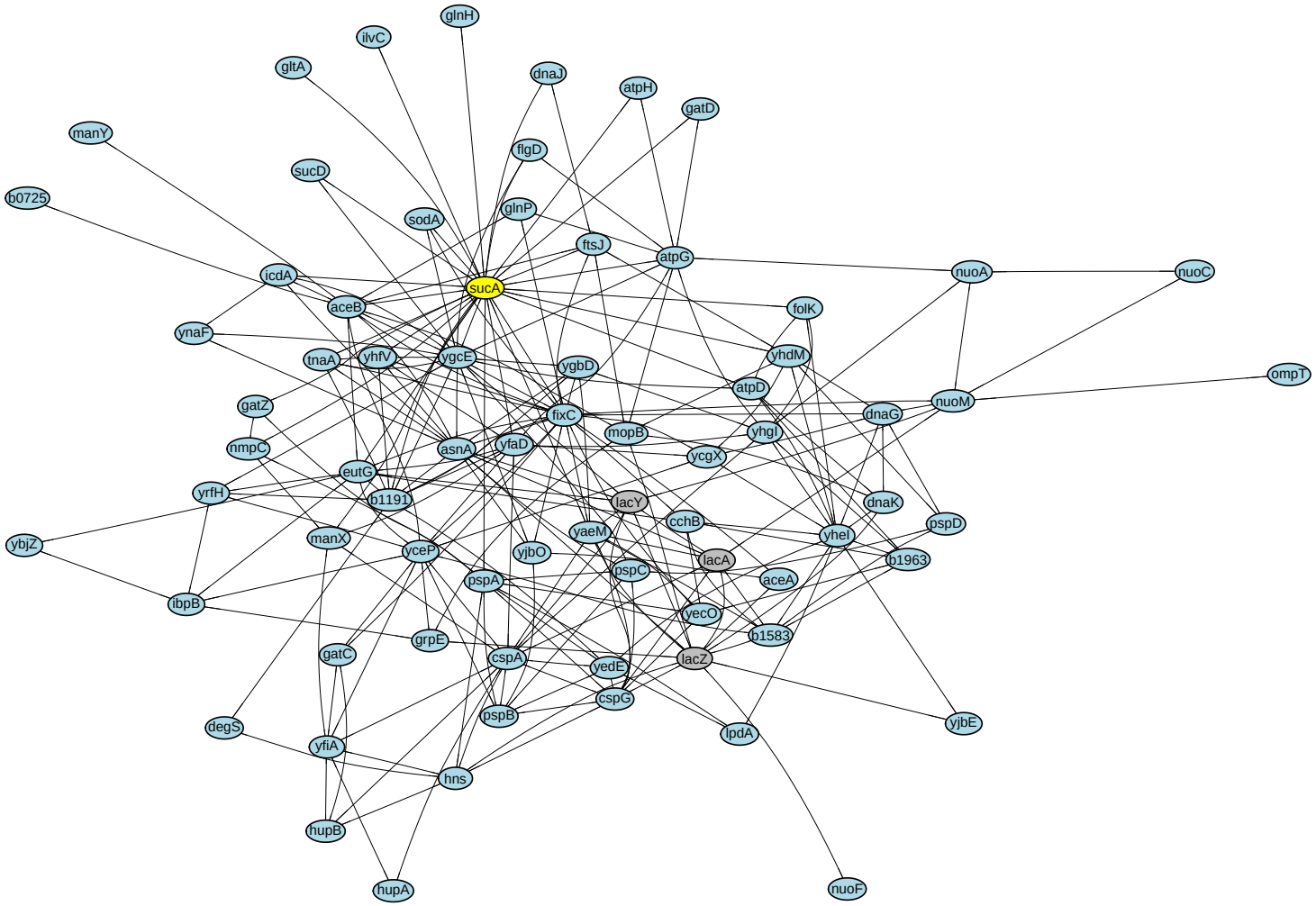

Figure 17: Network by Shrunk MLE (unadjusted)

Application on real dataset

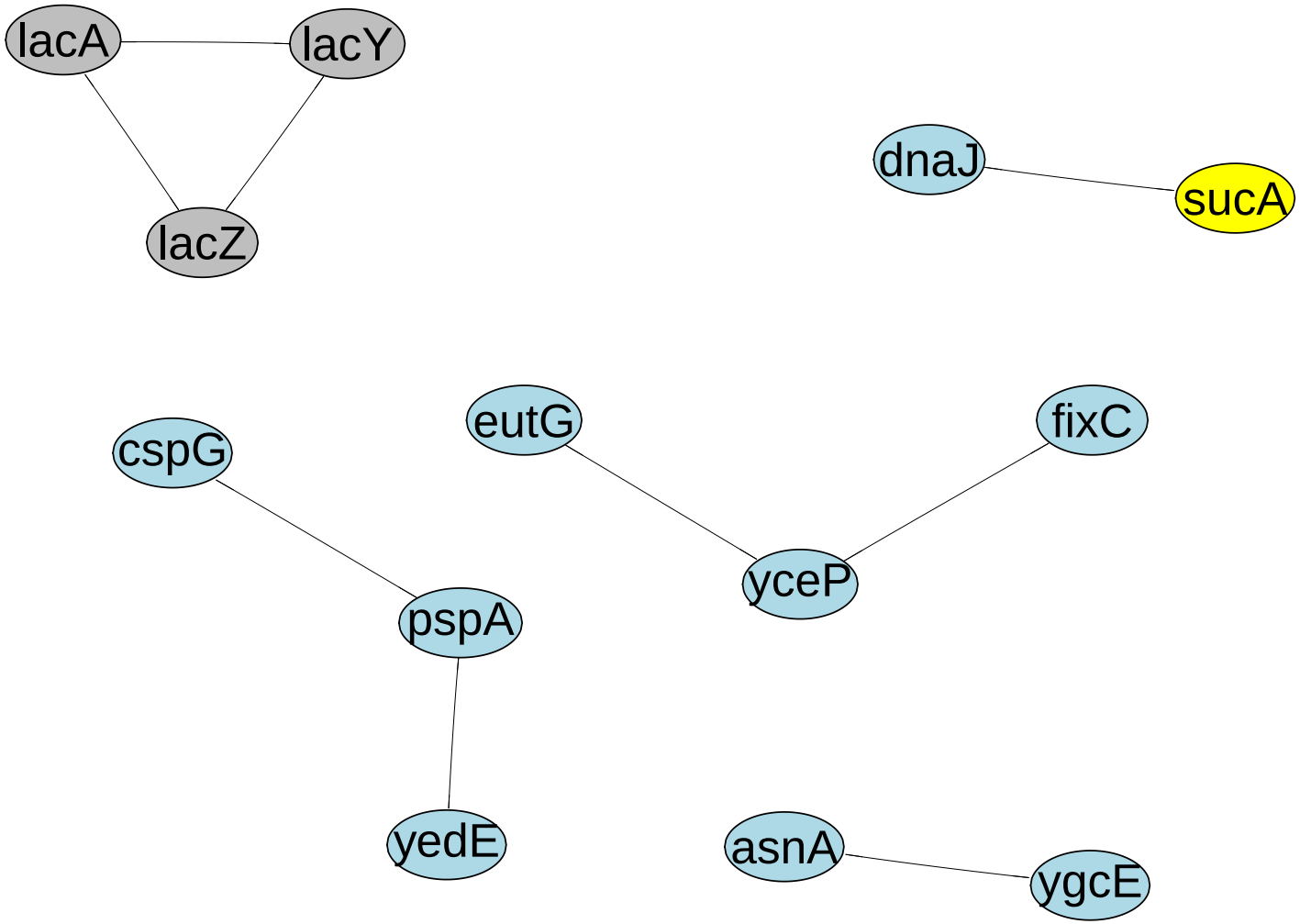

Figure 18: Network by Shrunk MLE (adjusted)

Application on real dataset

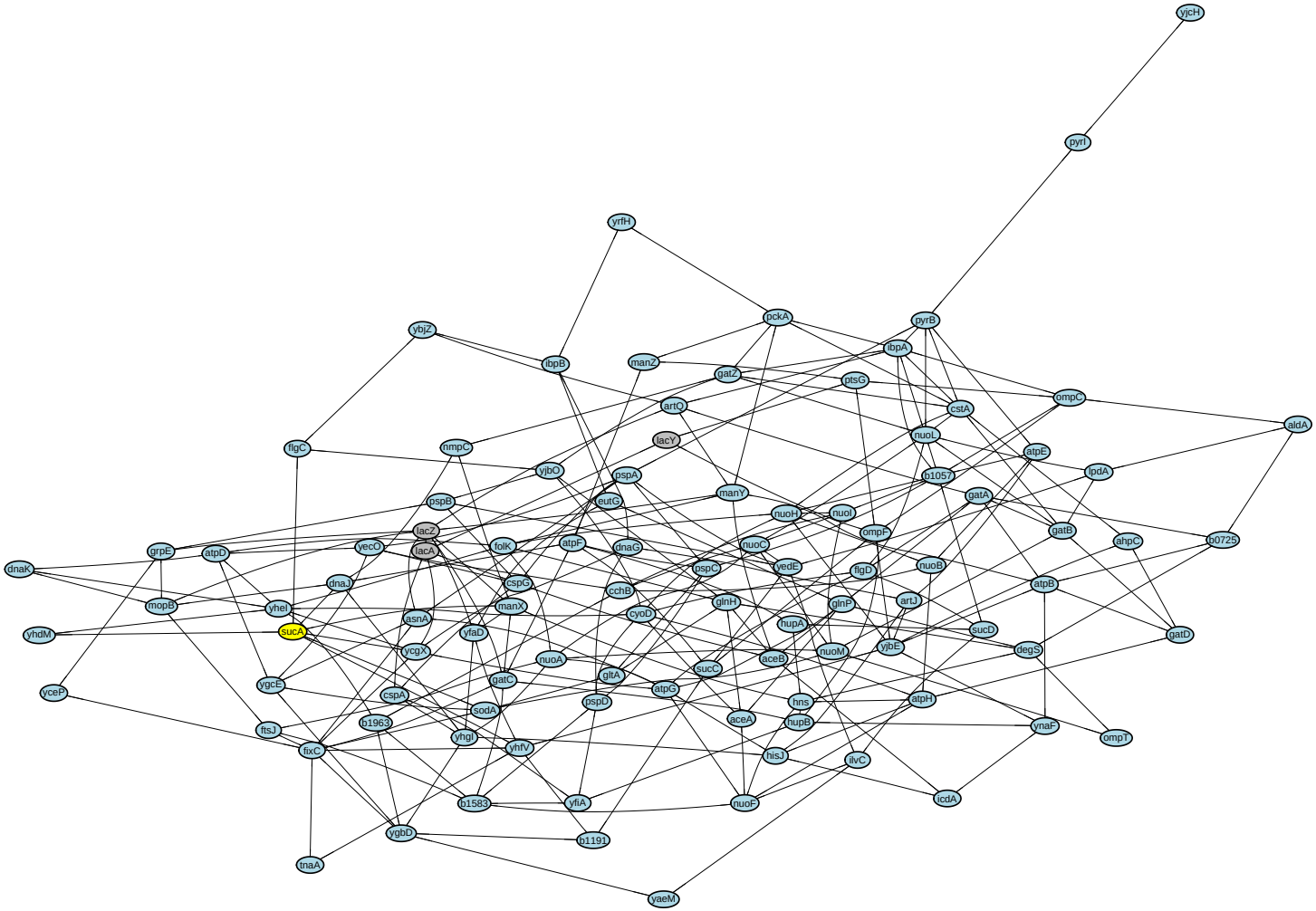

Figure 19: Network by 0-PCG

Section 5

Discussion

Summary

- In this study, we apply a novel network inference method c-level partial correlation graph on the gene expression data
- Compared to other partial correlation based methods, c-level PCG is able to test edges more powerfully
  - It can detect as many significant nodes as less conservative methods and control the FDR like more conservative methods
- Besides, unlike other existing methods, c-level PCG can be use to construct other hypothesis test, such as $H_0 : \hat{\rho}_{j_1,j_2} <= 0.25$

Limitations

- In the future, there are still much study that we can work on
  - The selection of tunning parameter is not satistfying
  - More simulation studies can be conducted to investigate how the network structures affect inference performance
  - Variables (genes) in gene expression data is not typical normally distributed variable, data can be simulated from other distribution
  - Measure of dependency can be generalized, since partial correlation is a very good measure of conditional dependency, but it can not capture non-linear association
  - Due to the experimental setting of genomics study, hierarchical model can be considered

Acknowledgement

This project is supervised by Dr. Liu and Dr. Qiu. I appreciate their time and support.

Also, I want to thank Dr. Wang and Wenting for their nice comments.

Many thanks for their help!

Thanks

Questions?

Reference I

Albert, Réka, and Albert-László Barabási. 2002. "Statistical Mechanics of Complex Networks." *Reviews of Modern Physics* 74 (1): 47.

Ansorge, Wilhelm J. 2009. "Next-Generation Dna Sequencing Techniques." *New Biotechnology* 25 (4): 195–203.

Barabasi, Albert-Laszlo, and Zoltan N Oltvai. 2004. "Network Biology: Understanding the Cell's Functional Organization." *Nature Reviews Genetics* 5 (2): 101.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300.

Reference II

Bernal, Victor, Rainer Bischoff, Victor Guryev, Marco Grzegorczyk, and Peter Horvatovich. 2019. "Exact Hypothesis Testing for Shrinkage Based Gaussian Graphical Models." *Bioinformatics*.

Boccaletti, Stefano. 2010. *Handbook on Biological Networks*. Vol. 10. World Scientific.

Efron, Bradley. 2005. "Local False Discovery Rates." Division of Biostatistics, Stanford University.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2008. "Sparse Inverse Covariance Estimation with the Graphical Lasso." *Biostatistics* 9 (3): 432–41.

Gu, Zuguang, Chenyu Zhang, and Jin Wang. 2012. "Gene Regulation Is Governed by a Core Network in Hepatocellular Carcinoma." *BMC Systems Biology* 6 (1): 32.

Reference III

Heller, Michael J. 2002. "DNA Microarray Technology: Devices, Systems, and Applications." *Annual Review of Biomedical Engineering* 4 (1): 129–53.

Peng, Jie, Pei Wang, Nengfeng Zhou, and Ji Zhu. 2009. "Partial Correlation Estimation by Joint Sparse Regression Models." *Journal of the American Statistical Association* 104 (486): 735–46.

Qiu, Yumou, and Xiao-Hua Zhou. 2018. "Estimating c-Level Partial Correlation Graphs with Application to Brain Imaging." *Biostatistics*.

Schaefer, Juliane, Rainer Opgen-Rhein, and Korbinian Strimmer. 2009. "GeneNet: Modeling and Inferring Gene Networks." *URL Http://CRAN. R-Project. Org/Package= GeneNet. R Package Version* 1 (4).

Schäfer, Juliane, and Korbinian Strimmer. 2004. "An Empirical Bayes Approach to Inferring Large-Scale Gene Association Networks." *Bioinformatics* 21 (6): 754–64.

Reference IV

———. 2005. "A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics." *Statistical Applications in Genetics and Molecular Biology* 4 (1).

Schmidt-Heck, W, R Guthke, S Toepfer, H Reischer, K Duerrschmid, and K Bayer. 2004. "Reverse Engineering of the Stress Response During Expression of a Recombinant Protein." In *Proceedings of the Eunite Symposium*, 10–12.

Wang, YX Rachel, and Haiyan Huang. 2014. "Review on Statistical Methods for Gene Network Reconstruction Using Expression Data." *Journal of Theoretical Biology* 362: 53–61.

Yu, Donghyeon, MinSoo Kim, Guanghua Xiao, and Tae Hyun Hwang. 2013. "Review of Biological Network Data and Its Applications." *Genomics & Informatics* 11 (4): 200.