# Network Based Gene Set Analysis

Ali Shojaie and George Michailidis, Department of Statistics, University of Michigan

## 1. Abstract

Development of high throughput technologies including DNA microarrays has facilitated the study of cells and living organisms. The challenge is no longer to identify the genes or proteins that are differentially expressed, but rather to find sub-systems that interact with each other in response to given environmental conditions. Study of these interacting sub-systems has provided an invaluable source of additional information that can be used to better understand the complex mechanisms of life. In this paper, we propose a latent variable model that directly incorporates the external information about the underlying network. We then use the theory of mixed linear models to present a general inference framework for the problem of testing the significance of subnetworks.

## 2. Problem Statement

We would like to know whether a ***Genetic Pathway*** is ***involved*** in responding to changes in environmental conditions or in specific cell functions. For simplicity consider two classes: Control and Treatment. We would like to know whether the ***behavior*** of genes in the pathway is different under treatment compared to control.

## 3. Gene Set Enrichment Analysis

*Subramanian et. al.* (2005) proposed a new method to test the significance of *a priori* defined sets of genes. In short, this method finds an association measure relating each gene to the class labels (Traditional Single Gene Analysis), ranks all the genes in the list based on this measure and then computes a Kolomogorov-Smirinov type summary measure for each gene set. The significance of gene set is assessed by a permutation based test which permutes the class labels. The main challenge in analyzing gene sets is to determine the ***null*** and ***alternative*** hypotheses. *Tian et. al.* (2005) and *Efron and Tibshirani* (2007) argue that there are two different type of hypotheses that should be evaluated when studying the significance of gene sets:

- **Q1** Are the genes in the gene set randomly selected among the set of all genes?
  This hypothesis is tested by permuting the genes and considering random selection of genes (***Row Permutation***).
- **Q2** Do genes in a gene set have the same pattern of association with the phenotype as the rest of the genes?
  This hypothesis is tested by permuting the class labels to see whether the pattern of association would remain the same if samples belong to random classes (***Column Permutation***).

***Note***: Methods of Gene Set Analysis, do not directly use the ***interactions*** between genes; their goal is to preserve the correlation structure among genes.

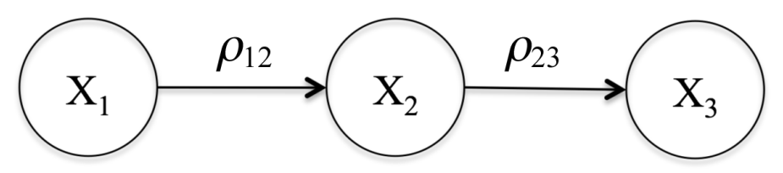
## 4. Network Based Analysis of Gene Sets

Consider the mRNA gene expression data $\mathcal{D}$ organized in a $p \times n$ matrix, with the first $n_1$ columns corresponding to control samples and the remaining $(n - n_1)$ columns to treatment samples. Let $Y_i$ be the $i$th sample in the expression data ($i$th column of $\mathcal{D}$). Represent the gene network as a directed graph $G = (V, E)$ and let $A$ be the weighted adjacency matrix of the graph with the nonzero element $A_{ij} \neq 0$ representing directed edges with the weight (correlation) between the two vertices $i$ and $j$. Let $Y_i = X_i + \varepsilon_i$, where $X_i$ represents the signal and $\varepsilon_i \sim N_p(0, \sigma_\varepsilon^2 I_p)$ the noise. Define the latent variable $\gamma_i \sim N_p(\mu, \sigma_\gamma^2 I_p)$ for every i, then we can write:

$$Y_i = \Lambda\gamma_i + \varepsilon_i, \;\; \Rightarrow \;\; Y_i \sim N_p(\Lambda\mu, \sigma_\gamma^2 \Lambda\Lambda' + \sigma_\varepsilon^2 I_p)$$

***Example (two level binary tree)***:

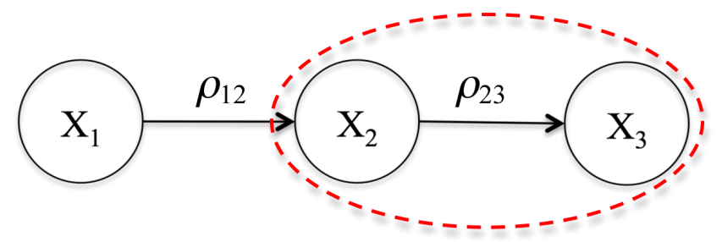
$$X_1 = \gamma_1$$
$$X_2 = \rho_{12}X_1 + \gamma_2 = \rho_{12}\gamma_1 + \gamma_2$$
$$X_3 = \rho_{23}X_2 + \gamma_3 = \rho_{23}\rho_{12}\gamma_1 + \rho_{23}\gamma_2 + \gamma_3$$



For References, Questions and comments please contact `shojaie@umich.edu`

## 5. New Inference Procedure

This model can be represented as a *Mixed Linear Model* (MLM) in the form $\mathbf{Y} = \mathbf{\Gamma}\beta + \mathbf{\Pi}\gamma + \varepsilon$. For any given contrast vector, $l$, we can test $H_0 : l\beta = 0 \;\; vs. \;\; H_1 : l\beta \neq 0$ using the test statistic $T = \frac{l\hat{\beta}}{\sqrt{l\hat{C}l'}}$. Under $H_0$, T has approximately $t$ distribution with degrees of freedom, $\nu$, which should be estimated from data. To test the significance of each pathway, we need to include interactions among genes in the set, while excluding confounding effect of other genes in the network:



**Proposition 1.** ***Optimal Choice of Contrast Vector****:*
*Consider a* $1 \times p$ *indicator vector* $\mathrm{b}$. *Then* $(\mathrm{b}\Lambda \cdot \mathrm{b})\gamma$ *is not affected by any node not in* $\mathrm{b}$ *while includes the effects of nodes in* $\mathrm{b}$ *on each other.*

## 6. Analysis of Yeast GAL Data

- *Ideker et. al. (2001)* data on yeast *Saccharomyces cerevisiae* Galactose Utilization Pathway: data available on 343 genes and 419 interactions (correlations among genes also available).
- 2 sets of expression levels available for each of 9 different perturbations of GAL genes and the wild-type yeast: presence of galactose (gal+) and absence of galactose (gal –).
- GSEA only finds the *Galactose Utilization Pathway* to be significant, Net-GSA find several other pathways.

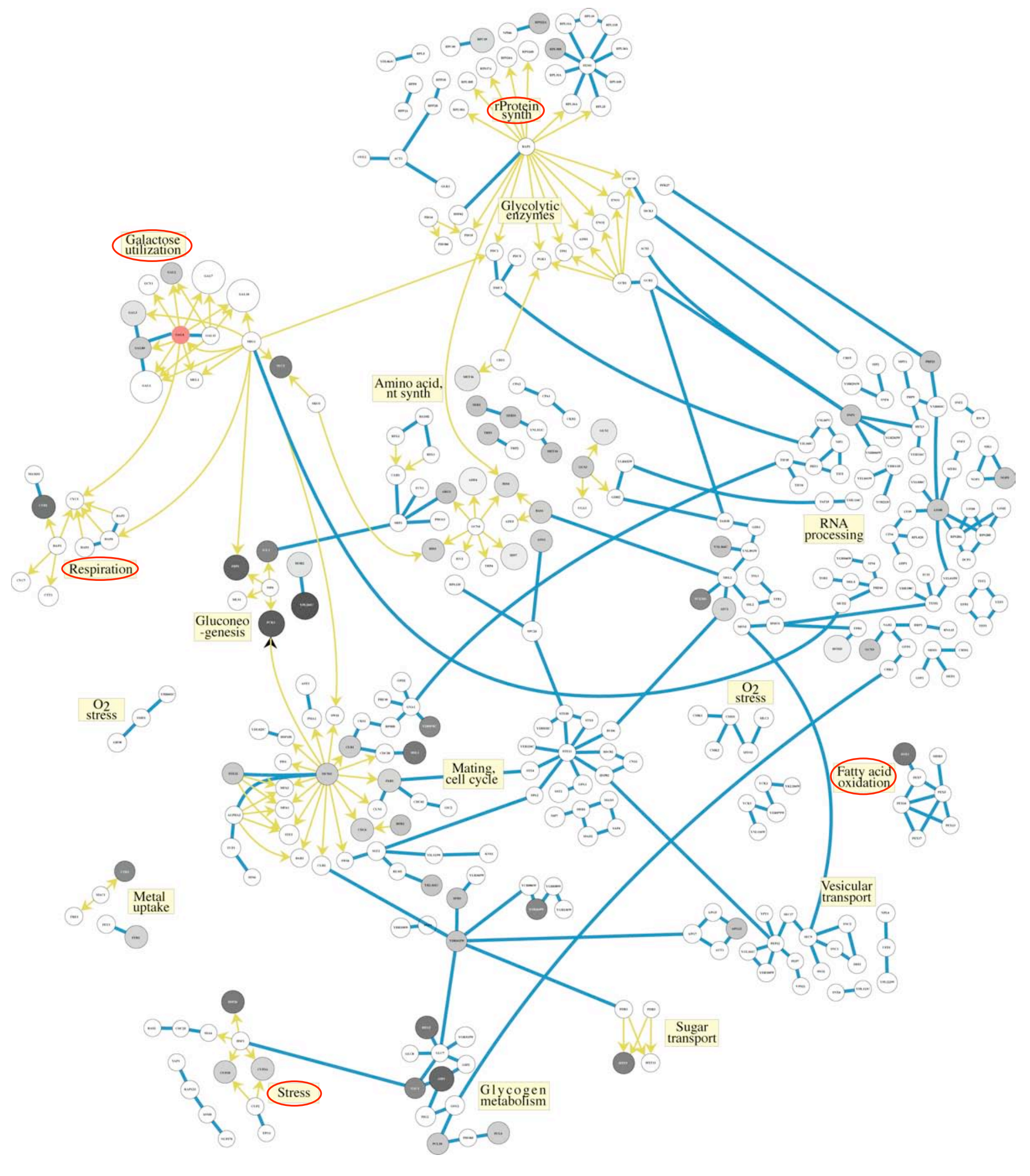


Figure 1: Yeast gene network with significant pathways indicated with RED ovals .

## 7. Analysis of Complex Experiments

- A number of extensions are needed to generalize the model for analysis of complex experiments, including time-course gene expression data
- To analyze general gene networks, use the (a version of) Laplacian matrix of the network instead of the original adjacency matrix
- Use a Block-Relaxation algorithm for estimation of parameters of MLM, extend to distributed estimation over subnetworks
- Use contrast matrices and F-test to simultaneously test multiple effects of a subnetwork
- Temporal correlations are handled using the covariance matrix of the error

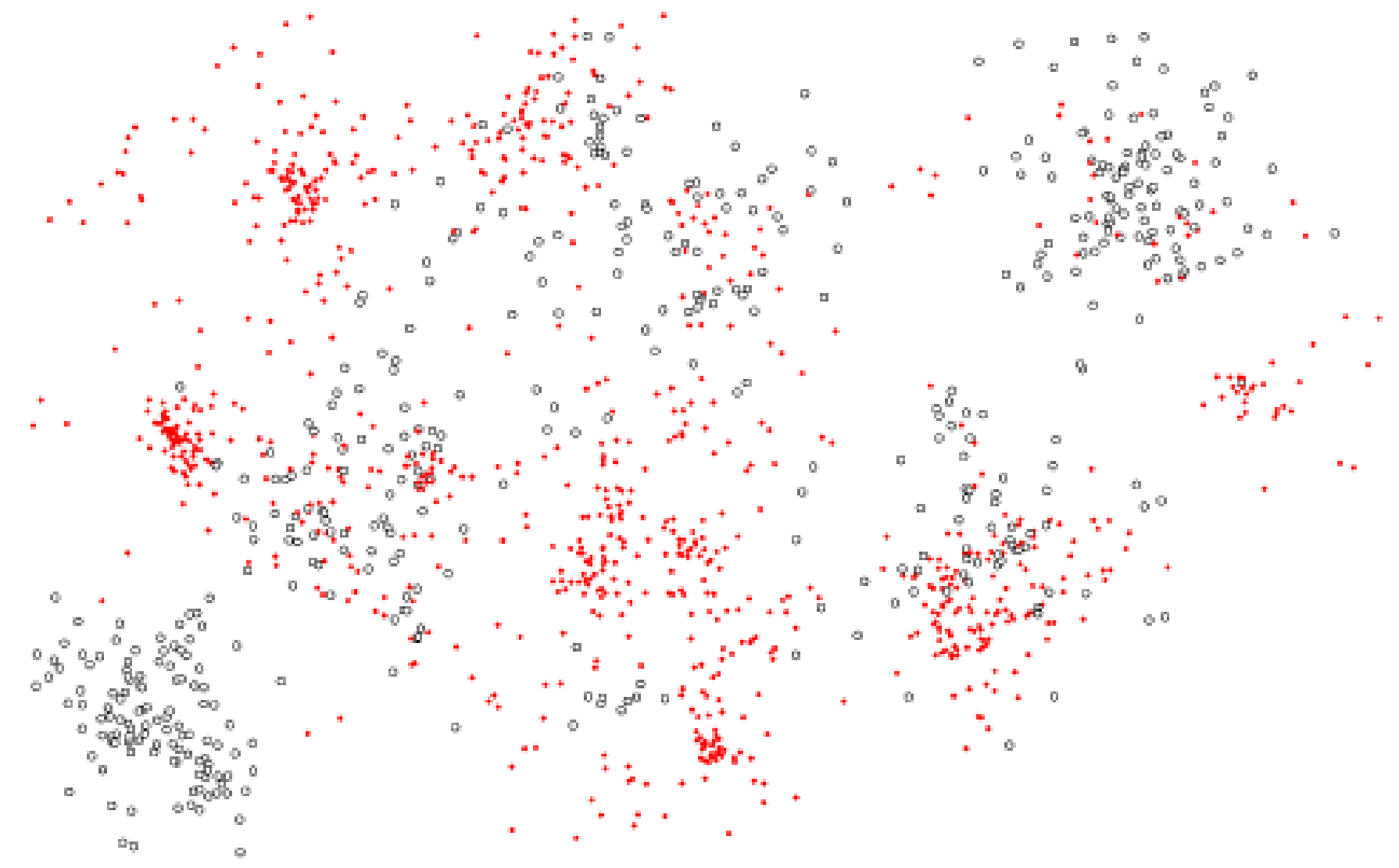
## 7. Yeast ESR (*Gasch et al. (2000)*)

| Experiment | Obs. Time |
|---|---|
| Mild Heat Shock | 5, 15, 30 min |
| *29C to 33C, no sorbitol* | |
| Mild Heat Shock | 5, 15, 30 min |
| *29C +1M sorbitol, 33C +1M sorbitol* | |
| Mild Heat Shock | 5, 15, 30 min |
| *29C +1M sorbitol, 33C no sorbitol* | |

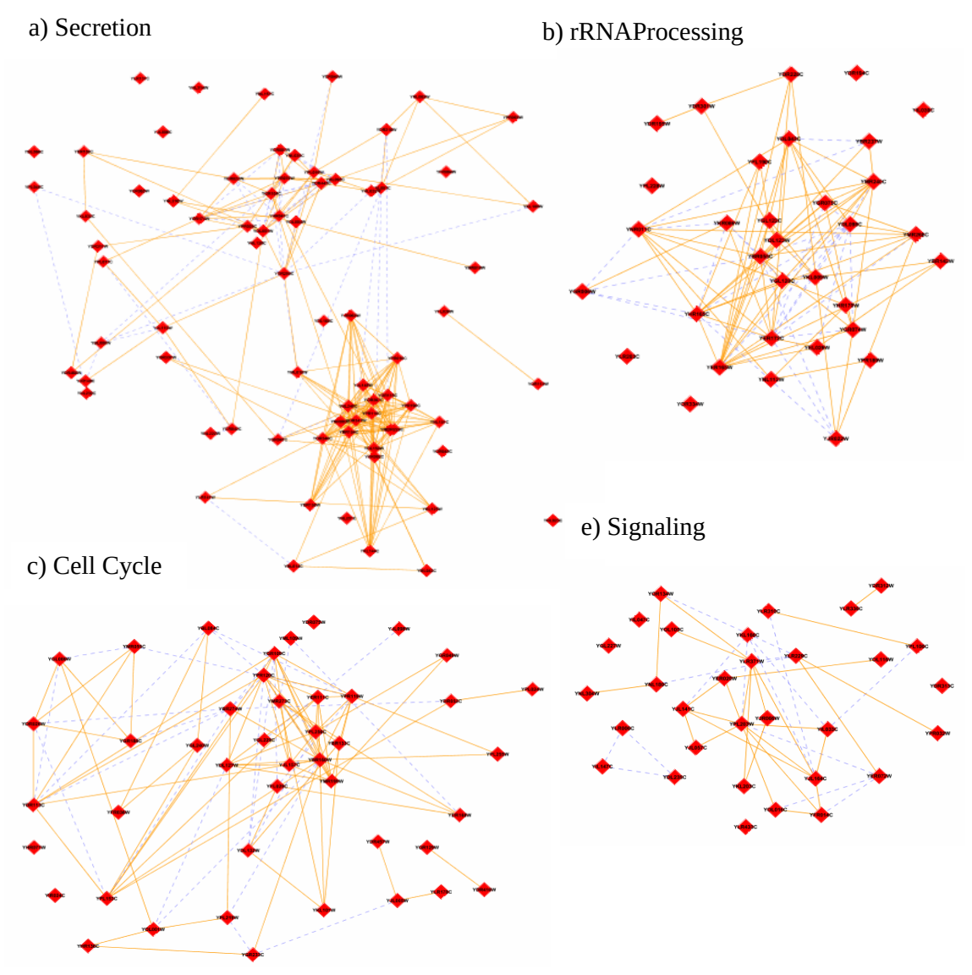
### Network Information:

- Use YeastNet (*Lee et al.* (2007)) for gene-gene interactions (102,000 interactions among 5,900 yeast genes)
- Use independent experiments of *Gasch et al.* to estimate weights
- Use GO to define gene sets

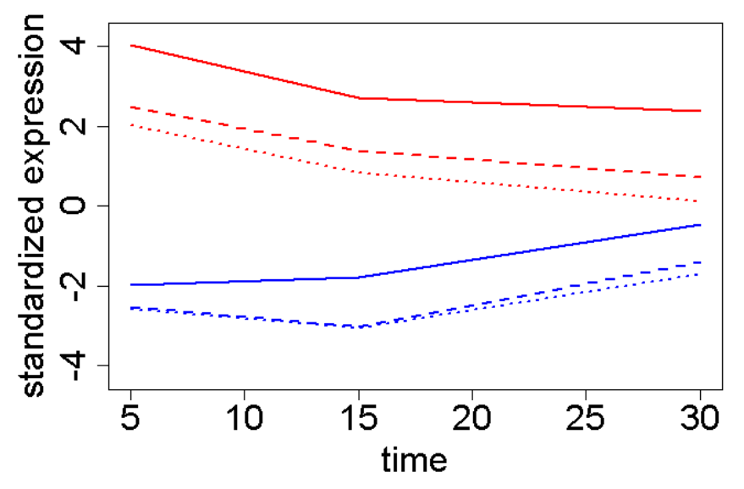
### Network with Significance Information



### Subnetworks



Average expression profile of significant pathways (induced and suppressed). Solid, dashed and dotted lines indicate the first, second and third experimental conditions, respectively.



## 8. Simulation Studies

### Simulation 1: NetGSA vs GSEA

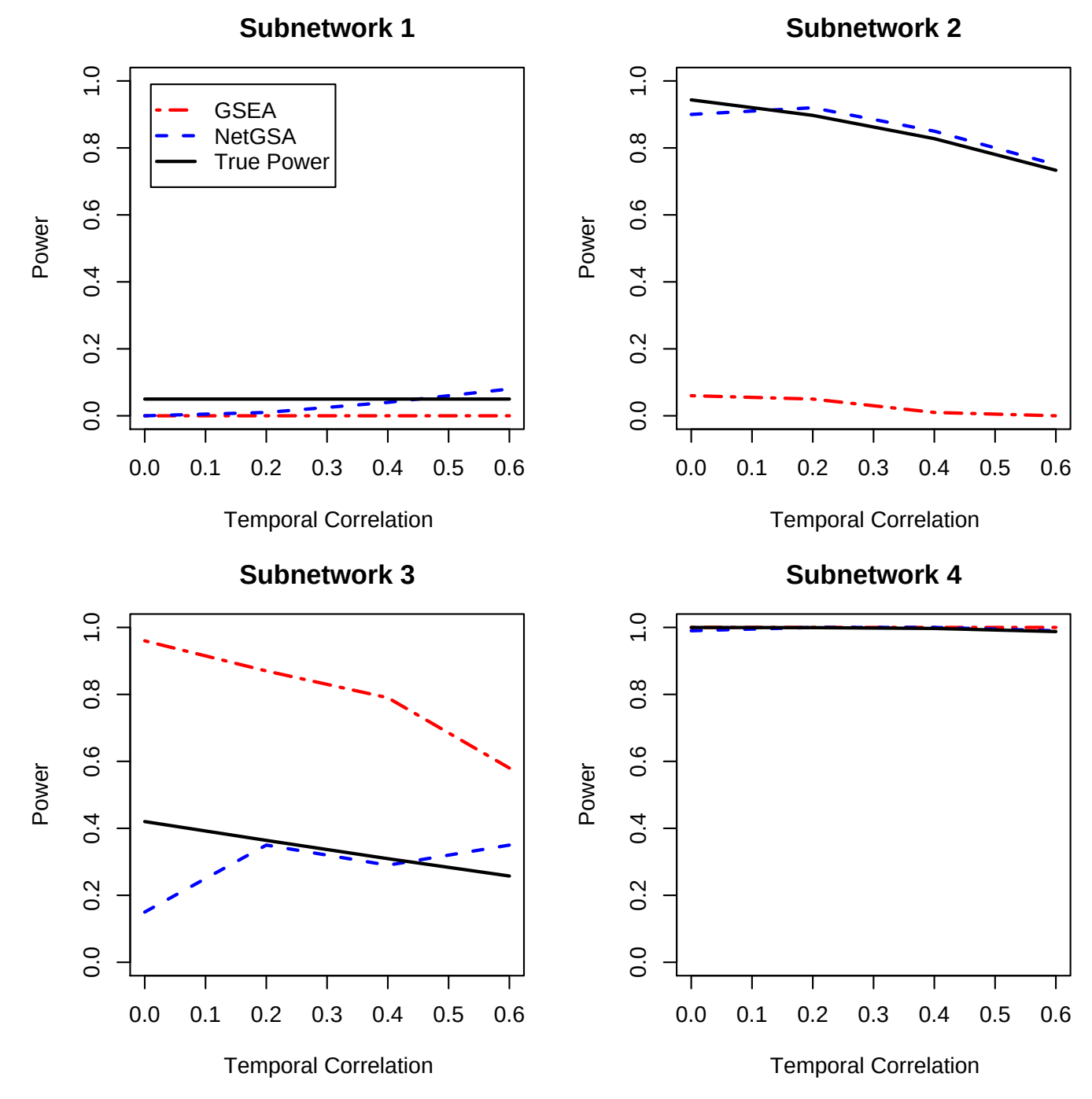


Figure 2: Estimated and true powers with Temporal Correlation

### Simulation 2: Noise in Network

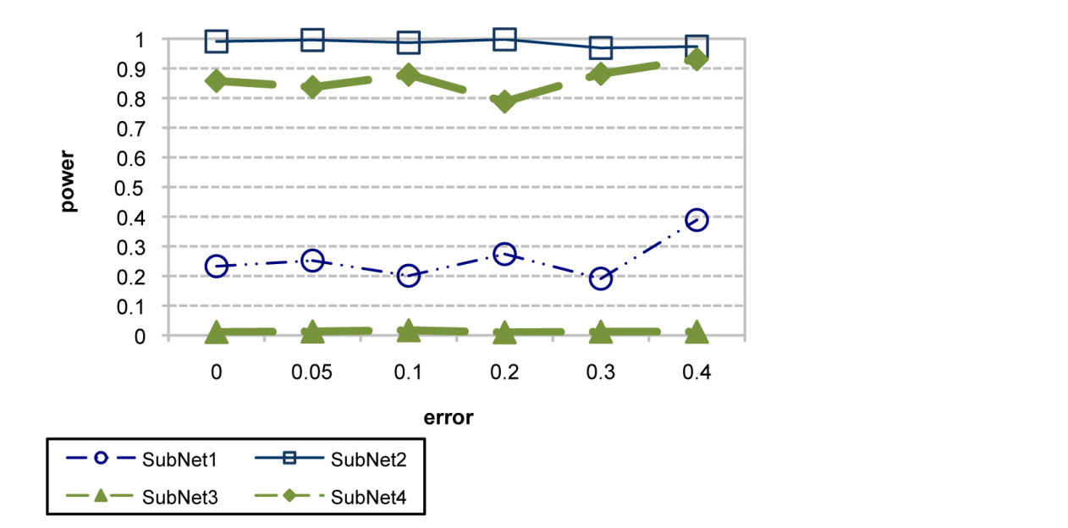


Figure 3: Estimated powers with random noise

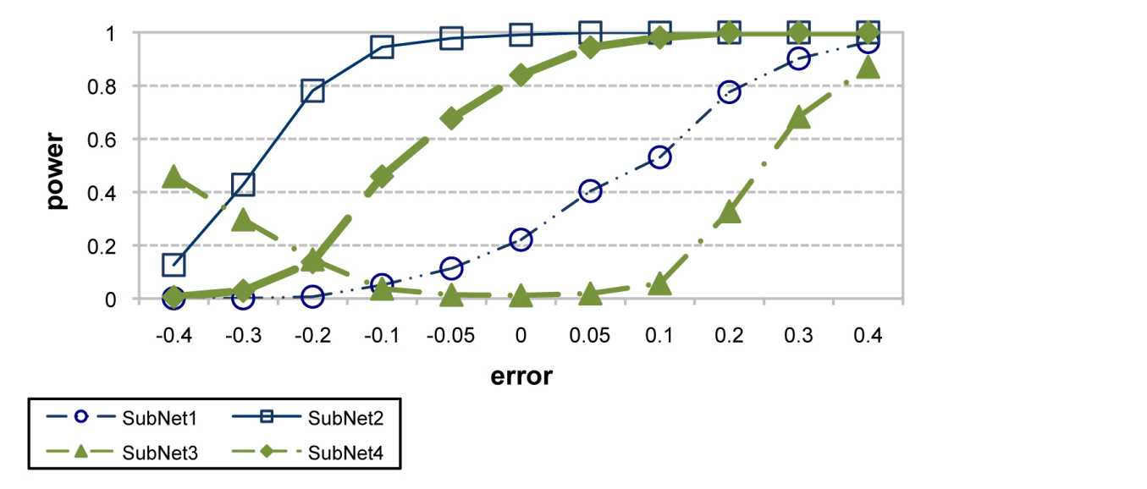


Figure 4: Estimated powers with systematic noise

## 9. Conclusion

- Need to formulate null and alternative hypotheses that test both change in the expression levels and in the network.
- The proposed method when is used with the optimal choice of contrast vector has better power properties than the methods of gene set analysis.
- Provides a general framework for studying a variety of phenotypes and can be extended to analysis of time series mRNA data and change in the network over time.
- Require external information on the weighted adjacency matrix of the network. Binary network information can be used to efficiently learn the weights
- When no external information available, nonparametric methods like the GSEA might result in better inference properties.