Identification of differentially connected genes through network analysis on Breast Cancer

Fen Pei
University of Pittsburgh
fep7@pitt.edu

Seo-Jin Bang
Carnegie Mellon University
seojinb@andrew.cmu.edu

Abstract

Topological changes of a gene network across different conditions such as normal versus cancer can provide hints regarding the disrupted regulatory relationships or affected regulatory sub-networks specific to a phenotype of interest. This work aims to identify differentially connected genes (DCGs) that cause significant topological changes between two gene co-expression networks: breast invasive carcinoma versus normal samples.

1 Introduction

Breast cancer remains the most common malignancy in women worldwide and is the leading cause of cancer-related mortality. More than 1-2 million cases are diagnosed every year, affecting 10-12% of the female population and accounting for 500 000 deaths per year worldwide. Approximately 5-10% are thought to be inherited. The hereditary breast cancer syndrome includes genetic alterations in various susceptibility genes such as p53, PTEN, BRCA1, and BRCA2. Sporadic breast cancers result from a serial stepwise accumulation of acquired and uncorrected mutations in somatic genes, without any germline mutation playing a role. Oncogenes that have been reported to play an early role in sporadic breast cancer are MYC, CCND1 (Cyclin D1) and ERBB2 (HER2/neu). In sporadic breast cancer, mutational inactivation of BRCA1/2 is rare. However, non-mutational functional suppression could result from various mechanisms, such as hypermethylation of the BRCA1 promoter.

2 Method

In this section, we illustrate each procedure of the Differentially Connected Gene (DCG) analysis along with the Figure 2. In section 2.1, a weighted, undirected gene co-expression network for each group is constructed using gene expression profiles. In section 2.2, a Euclidean Commute Time Distance (ECTD) matrix for each group is computed which will be used to calculate the differential network measures for each gene within each network in section 2.3. In section 2.4, we propose several DCG statistics to evaluate connectivity difference of each gene between two networks using the differential network measures.

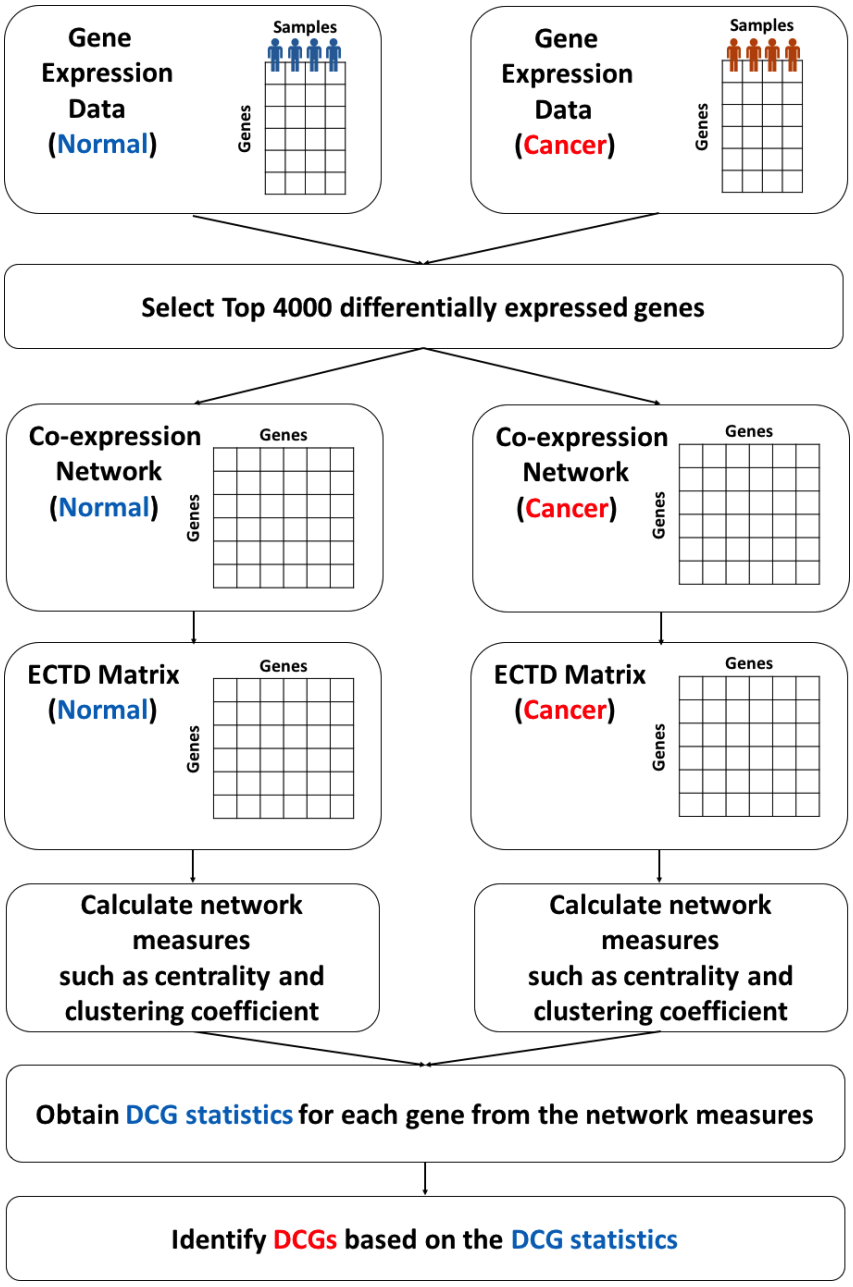

Figure 1: Flowchart of DCG analysis

2.1 Preliminaries

Given gene expression profiles for each group A and B, where n_A and n_B is the number of samples for group A and B respectively. Two sample t-test is performed on each genes to select differentially expressed genes between the two groups. The selected genes are used to construct a $p \times p$ gene co-expression matrix $\mathbf{N}_A$ for group A and $\mathbf{N}_B$ for group B such that:

$$\mathbf{N}_A = [|r_{A,ij}|]_{p\times p}$$
$$\mathbf{N}_B = [|r_{B,ij}|]_{p\times p}$$

where $r_{A,ij}$ and $r_{B,ij}$ are sample Pearson correlation coefficients between gene i and j for group A and B respectively; and p is the number of selected genes. Each co-expression matrix represents

each gene network $\mathcal{G}_A(\mathcal{V}, \mathcal{E}_A)$ and $\mathcal{G}_B(\mathcal{V}, \mathcal{E}_B)$ where $\mathcal{V}$ is the set of p nodes (genes); $\mathcal{E}_A$ and $\mathcal{E}_B$ are the set of edges in $\mathcal{G}_A$ and $\mathcal{G}_B$ respectively; an edge between a pair of genes are weighted with $|r_{A,ij}|$ and $|r_{B,ij}|$ respectively. Each weight represents the intensity of relationship between pairs of genes are non-negative and symmetric.

2.2 Euclidean Commute Time Distance

Euclidean Commute Time Distance (ECTD, [1]) is a distance metric between nodes of a weighted, undirected graph. ECTD matrix for each group is computed based on a random walk Markov chain model which transition probability is obtained from a Laplacian matrix of the gene co-expression network.

2.2.1 Laplacian Matrix of a Weighted Gene Network

An adjacency matrix $\mathbf{A}$ of each group is defined in a standard way as:

$$\mathbf{A_A} = [a_{A,ij}]_{p\times p} \begin{cases} |r_{A,ij}| & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

$$\mathbf{A_B} = [a_{B,ij}]_{p\times p} \begin{cases} |r_{B,ij}| & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

which is both positive and symmetric. The Laplacian matrix $\mathbf{L}$ of each group is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$ where $\mathbf{D} = diag(d_{ii})_{p\times p} = diag(\sum_{j=1}^{p} a_{ij})_{p\times p}$.

2.2.2 Random Walk Markov Chain Model

We define a random walk Markov chain model [2] along with each node in the gene co-expression network. Every gene is associated with a state of the Markov chain. The Laplacian matrix $\mathbf{L}$ is used to define a transition probability matrix $\mathbf{P}$ which is assigned to each edge. A random walker starts from node i at time t jumps from gene to gene with a single-step transition probability such that:

$$\mathbf{P}[s(t+1) = j | s(t) = i] = a_{ij}/d_{ii} = p_{ij}$$

where $s(t)$ is a random variable representing the state of the Markov chain at time t. Therefore, we define $\mathbf{P} = \mathbf{D^{-1}A}$.

2.2.3 Expected Commute Time

Under the random walk Markov chain model, two basic quantities can be computed: the expected first-passage time and the expected commute time. [2] The expected first-passage time $m(k|i)$ is defined as the expected number of steps a random walker, starting in state i, will get to state k for the first time:

$$m(k|i) = \mathbf{E}[T_{ik}|s(0) = i]m(k|k) = 0$$

where T_{ik} it the minimum time until hitting state k such that $\min(t \geq 0 | s(t) = k \text{ and } s(0) = i)$. The expected commute time is defined as a related quantity of the expected first-passage time:

$$n(i, j) = m(j|i) + m(i|j)$$

Note that $n(i, j)$ is symmetric and non-negative measures by definition. Moreover, it has been shown that the expected commute time $n(i, j)$ is a distance metric. [1, 3, 4]

2.2.4 Computation of Euclidean Commute Time Distance

The Euclidean Commute Time Distance (ECTD) can easily be computed as a function of pseudo-inverse of the Laplacian matrix $\mathbf{L}$ and volume of the graph V_A and V_B. [1]

$$\begin{aligned} \mathbf{ECTD_A} &= [n_{A,ij}]_{p\times p} \\ &= [V_A(e_i - e_j)^T \mathbf{L}_A^-(e_i - e_j)]_{p\times p} \\ \mathbf{ECTD_B} &= [n_{B,ij}]_{p\times p} \\ &= [V_B(e_i - e_j)^T \mathbf{L}_B^-(e_i - e_j)]_{p\times p} \end{aligned}$$

where $V_A = \sum_{i,j} a_{A,ij}$ and $V_B = \sum_{i,j} a_{B,ij}$; e_i is a unit basis vector for each node i; $\mathbf{L}_A^-$ and $\mathbf{L}_B^-$ are a Moore-Penrose pseudo-inverse of the Laplacian matrix for each group. [5] Since $\mathbf{L}^-$ is a positive semi-definite matrix, n_{ij} is a distance metric in the Euclidean space of the gene network.

2.3 Differential Network Measures

Each node within a network can be characterized via several differential network measures. The measures attempts to quantify the 'importance' of a node within a network, however there is no unique interpretation of the 'importance'. For example, the 'importance' can be represented by potential for autonomy, control, risk, exposure, influence, belongingness, and so on. [6] Here we introduce three type of centrality measures and clustering coefficient based on ECTD matrix. Since the expected commute time represents dissimilarity between pairs of genes, the inverse of it used as a weight of the edges in each gene network. Those measures will be utilized to define DCG statistics in section 2.4.

2.3.1 Centrality Measures

- Degree Centrality
  In a binary graph, the degree centrality for node i is defined as the number of its neighbors. In general, degree centrality $c_d(i)$ is simply the sum of weight of edges including the gene i. [7, 8]

  $$c_d(i) = \sum_{j=1}^{p} \frac{1}{n(i,j)}$$

  By the definition, the degree centrality represents how each gene is involved in the weighted gene network along with its neighbors.

- Betweenness Centrality
  Betweenness centrality [9] is defined by the number of shortest paths going through a gene. In a binary network, it is defined as:

  $$c_b(i) = \sum_{u \neq v \neq i} \frac{\sigma_{uv}(i)}{\sigma_{uv}}$$

  where σ_{uv} is the total number of shortest paths from gene u to gene v and $\sigma_{uv}(i)$ is the number of those paths that pass through the gene i. We use a fast algorithm to calculate betweenness centrality proposed by Brandes (2001). [10] which allow to calculate the betweenness centrality on weighted networks.

- Closeness Centrality
  Closeness centrality [9] quantifies how many steps is required to access every other genes from a given gene. (i.e. the sum of distances to all other gene from a given gene) The closeness centrality has been generalized to weighted networks by Newman [11]. Therefore it is defined as:

  $$c_c(i) = \frac{1}{\sum_j dist(i,j)}$$

  where $dist(i,j)$ is the shortest distance between gene i and j. Note that the closeness centrality and degree centrality using ECTD matrix are very similar to each other since the expected commute time is a distance metric.

2.3.2 Clustering Coefficient

Clustering coefficient [12, 13] refers a fraction of neighbors of given gene that are interconnected each other, which is also called as transitivity. It quantifies how its neighbors are clustered together given a gene. Note that there are essentially two types of clustering coefficients: one for each gene and the other for a whole network. The clustering coefficient here refers the gene-level measurement. Also we use the generalized clustering coefficient suggested by Barrat [14] which can be calculated

on a weighted network:

$$CC(i) = \frac{1}{s_i(c_d(i)-1)} \sum_{j,h} \frac{w_{ij}+w_{ih}}{2} a_{ij}a_{ih}a_{jh}$$

where s_i is the strength of gene i; a_{ij} are elements of the adjacency matrix; $c_d(i)$ is the vertex degree; w_{ij} are the weights. Note that we use the inverse of expected commute time as a weight on each edge, so s_i, a_{ij}, and w_{ij} will be obtained from the ETCD.

2.4 Measures for Differentially Connected Gene

Based on the differential network measures introduced in section 2.3, we propose four measures to quantify connectivity difference of a gene between two different networks: degree differences, betweenness differences, closeness differences, and clustering coefficient differences. Note that the differential network measures are defined for each gene within a network. By comparing the measures from two different networks, we will identify differentially connected genes (DCGs).

2.4.1 Degree Differences

Degree differences is defined as an absolute difference between the degree centrality measures of group A and B, normalized by those sum:

$$DCG_{degree}(i) = \frac{|c_{A,d}(i) - c_{B,d}(i)|}{|c_{A,d}(i) + c_{B,d}(i)|}$$

where $c_{A,d}(i)$ and $c_{B,d}(i)$ are the degree centrality of gene i in group A and B respectively. Genes that have large difference in the degree centrality between two groups while overall degree centrality are small will be highly ranked based on the degree differences.

2.4.2 Betweenness Differences

Betweenness differences is defined as an absolute difference between the betweenness centrality measures of group A and B.

$$DCG_{betweenness}(i) = |c_{A,b}(i) - c_{B,d}(i)|$$

where $c_{A,b}(i)$ and $c_{B,b}(i)$ are the betweenness centrality of gene i in group A and B respectively. The betweenness difference ranks genes that has large difference in the betweenness centrality between two groups.

2.4.3 Closeness Differences

Closenness differences is defined as an absolute difference between the centralized closeness centrality of group A and B.

$$DCG_{closenness}(i) = |(c_{A,c}(i) - median(c_{A,c}(i))) - (c_{B,c}(i) - median(c_{B,c}(i)))|$$

where $c_{A,c}(i)$ and $c_{B,c}(i)$ are the closeness centrality of gene i in group A and B respectively; $median(c_{A,c}(i))$ and $median(c_{B,c}(i))$ are median of the closeness centrality of all genes in group A and B respectively. Genes which derivation of closenness centrality from its median is changed a lot between two groups will be highly ranked based on the closeness differences.

2.4.4 Clustering Coefficient Differences

$$DCG_{CC}(i) = \left| \log \frac{CC_A(i) \times c_{A,d}(i)}{CC_B(i) \times c_{B,d}(i)} \right|$$

where $CC_A(i)$ and $CC_B(i)$ are the clustering coefficient of gene i in group A and B respectively; $c_{A,d}(i)$ and $c_{B,d}(i)$ are the degree centrality of gene i in group A and B respectively. By the definition, clustering coefficient is inversely proportional to degree of genes. Therefore, effect of degree centrality of genes is controlled by multiplying the degree centrality to the clustering coefficient. The clustering coefficient differences is log ratio of those between two groups. As it increases, genes will be highly ranked.

3 Experimental Results

3.1 Data

The gene expression dataset we use is Breast invasive carcinoma dataset from The Cancer Genome Atlas (TCGA, https://tcga-data.nci.nih.gov/tcga), which contains expression profiles of 17,814 genes for 531 tumor and 62 normal samples. Missing values are imputed via K-nearest neighborhood (KNN, [15, 16]) method with $K = 10$. For each samples, it finds the 10 nearest neighbors using a Euclidean metric, then imputes the missing values by averaging values of its neighbors.

3.2 Experiment on Real Dataset

The DCG analysis is conducted to the TCGA Breast invasive carcinoma dataset. The 531 tumor samples are consist of a *Cancer* group, and 62 normal samples are consist of a *Normal* group. Two sample T-test is performed on each gene to find significant genes showing differential expression levels between two groups. Top 4000 DEGs are selected as significant genes, and used to construct co-expression matrix for each group. (Although it can be any numbers where the p-values are smaller than a significant level α, we choose the largest number possible to run in a computer.) ETCD matrix for each group is then constructed. The differential network measures are calculated via R-package called igraph. [17] Each DCG measures are calculated by comparing the network measures from two groups.

3.3 Result

The DCG measures are used to identify DCGs. Top k ranked genes having large values of the DCG measures are selected. The number of overlapped genes are examined in Figure 2. ($k = 20, 50$) Unlike other measures, the betweenness differences are obtained from the gene co-expression matrix but not from the ECTD matrix because most of the values from ECTD matrix is zero. Majority of genes are selected from all four measures, indicating that topological difference of those genes are very obvious in any measures. The completely overlapped genes among all DCG measures are listed

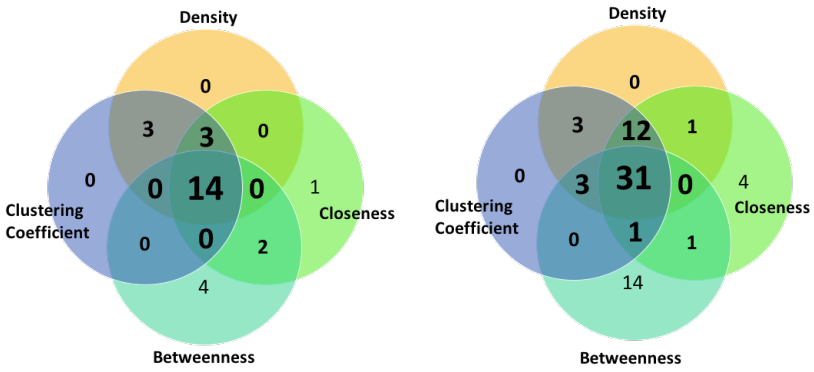

Figure 2: Overlapped genes between the DCG measures. Top 20 genes (left) are selected for each method. Total 27 genes are selected from at least one criteria. Total 50 genes (right) are selected for each method. Total 70 genes are selected from at least one criteria.

in Table 1. Note that the p-value from DEG analysis represents differential expression level between two groups while the DCG measures identify differential connectivity. Therefore, DCG measures can identify DCGs that have not been considered significant as much as other highly ranked genes in DEG analysis. Table 2 is a list of partially overlapped genes that are ranked high at least one DCG measure but not all DCG measures. Note that the partially overlapped genes are identified because at least one of differences in betweenness, closeness, or clustering coefficient, not because of the degree differences. Therefore, we consider those genes very interesting because it had been hard to be detected when you only consider the module centrality to evaluate its connectivity status. That is, the partially overlapped genes are not easy to be identified by previous approaches on gene co-expression network that has only focused on hub genes connected with large number of neighbors,

Table 1: List of completely overlapped genes between the DCG measures. Top 20 genes are selected for each method. P-values obtained from DEG analysis are adjusted using a method proposed by Benjamini and Hochberg (2005). [18] Genes are ranked based on the p-values.

Gene	P-value (BH)	Rank in DEG Analysis
$APBB1$	$2.23E-20$	2069
$ATAD1$	$1.40E-14$	3881
$C2orf37$	$1.68E-27$	647
$CAPRIN1$	$1.02E-13$	3790
$CKAP2$	$2.48E-39$	205
$KIAA0859$	$4.29E-27$	1609
$LOC388284$	$8.14E-17$	2405
$LRFN3$	$3.09E-13$	3751
$NUP155$	$2.10E-27$	1151
$PCGF5$	$7.41E-23$	1611
RBJ	$5.86E-21$	1685
$RFWD3$	$1.16E-27$	1040
$STAU1$	$3.77E-16$	2889
$TBRG4$	$6.74E-17$	2780

while the completely overlapped genes might be identified. Moreover, there are few more interesting genes identified based on the betweenness difference measure. Figure 3 represent the betweenness difference measures obtained from the ECTD matrix and the gene co-expression matrix. Majority of genes has zero values for the betweenness difference measure because the betweenness centralities from both the cancer and the normal ECTD matrix are zero. Unlike the majority of genes, three genes ($NSMCE4A$, $ATE1$, $APBB$) show large changes in the betweenness difference measures. Especially, $ATE1$ and $NSMCE4A$ show large difference even though there are almost no difference in the betweenness difference measures obtained from the gene co-expression matrix. Those two genes are also highly ranked from other measures: density differences, closeness differences, and clustering coefficient.

Detailed biological evaluation of all of those selected genes will be examined in section 3.4.

3.4 Biological Evaluation

To evaluate the results of proposed approach. We identify enriched biological GO terms and pathways of the DCGs.

4 Conclusions

How to dsafsdfsdfasdfdsjflkds cite example: [19]

The distribution of measures are very similar between two groups. if it is not we forceto make it. Also we expectedthat olu

References

Table 2: List of partially overlapped genes between the DCG measures. Top 20 genes are selected for each method. P-values obtained from DEG analysis are adjusted using a method proposed by Benjamini and Hochberg (2005). [18] Genes are ranked based on the p-values.

Gene	P-value (BH)	Rank in DEG Analysis
$ATE1$	$8.22E-16$	3139
$GSTP1$	$3.84E-38$	1049
$NSMCE4A$	$2.45E-13$	4030
$ATP13A1$	$2.10E-18$	1895
$FAM10A5$	$6.71E-16$	2953
$GGPS1$	$1.58E-18$	2571
$RXFP4$	$3.39E-14$	3592
$WDR8$	$5.54E-17$	2787
$BRMS1$	$7.01E-20$	1777
$C20orf121$	$1.03E-18$	2425
$FKSG24$	$6.88E-22$	1559
$GON4L$	$2.13E-14$	3543
$WSB1$	$4.73E-24$	1235

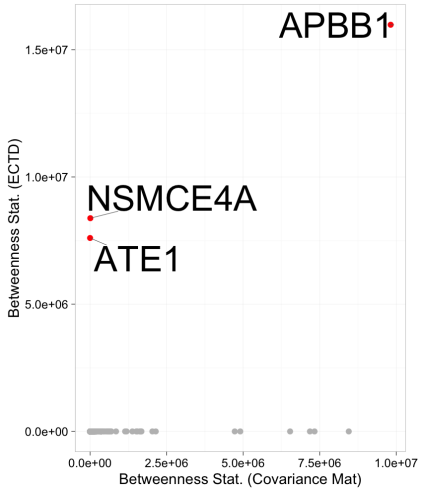

Figure 3: The betweenness difference measures obtained from the ECTD matrix (y-axis) versus from the gene co-expression matrix (x-axis).

[1] Marco Saerens, Francois Fouss, Luh Yen, and Pierre Dupont. The principal components analysis of a graph, and its relationships to spectral clustering. In *Machine Learning: ECML 2004*, pages 371–383. Springer, 2004.

[2] James R Norris. *Markov chains*. Number 2008. Cambridge university press, 1998.

[3] F Göbel and AA Jagers. Random walks on graphs. *Stochastic processes and their applications*, 2(4):311–336, 1974.

[4] Douglas J Klein and Milan Randić. Resistance distance. *Journal of Mathematical Chemistry*, 12(1):81–95, 1993.

[5] Arnold Dresden. The fourteenth western meeting of the american mathematical society. *Bull. Amer. Math. Soc.*, 26(9):385–396, 06 1920.

[6] Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, 2010.

[7] Mark EJ Newman. Analysis of weighted networks. *Physical Review E*, 70(5):056131, 2004.

[8] Tore Opsahl, Vittoria Colizza, Pietro Panzarasa, and Jose J Ramasco. Prominence and control: the weighted rich-club effect. *Physical review letters*, 101(16):168702, 2008.

[9] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[10] Ulrik Brandes. A faster algorithm for betweenness centrality*. *Journal of mathematical sociology*, 25(2):163–177, 2001.

[11] Mark EJ Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132, 2001.

[12] R Duncan Luce and Albert D Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, 1949.

[13] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.

[14] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004.

[15] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[16] Trevor Hastie, Robert Tibshirani, Gavin Sherlock, Michael Eisen, Patrick Brown, and David Botstein. Imputing missing data for gene expression arrays, 1999.

[17] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[18] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[19] Omar Odibat and Chandan K Reddy. Ranking differential hubs in gene co-expression networks. *Journal of bioinformatics and computational biology*, 10(01):1240002, 2012.