Eigenvectors of Metabolic Networks and Gene Expression Profiles: Smoothness and Correlation

Yuhou Zhou
Department of Computer Science & Electrical Engineering
Jacobs University University
28201, Bremen
yu.zhou@jacobs-university.de

Abstract

I implement a workflow [1] to quantitatively assess smoothness of a colored graph and check the correlation between eigenvectors of a metabolic network and gene expression profiles. Rapaport et al. use gene network information as a priori to gene expression profoile analysis. This project expand previous work by quantifying the smoothness of a colored graph and checking correlation between eigenvectors of the metabolic network and gene expression profiles. The workflow of the experiment includes data preprocessing, plotting metabolic network, quantifying smoothness, compute correlation coefficients. Finally, a conclusion about these two points are drawn and some future work is discussed.

1 Introduction

Analysing gene expression profiles is a way to understand genetic phenomena. However, connecting genes and biological significance is a difficult task. By including a priori knowledge from a gene network to gene expression profiles can facilitate the process of understanding biological significance of genes, dissecting the genes responsible for a phenotype [1]. By looking at eigenvalues and eigenvectors of the Laplacian matrix of a gene network, we can attenuate high-frequency component of the gene network, and a distance metric can be defined based on it. In this report, Section 3 to section 5 give some background knowledge about spectral graphy theory, network biology, and BiGG database and GEO Omnibus database. Section 6 shows the pipeline of the experiments and their results. Section 7 restate the conclusions from the experiments, and section 8 opens the discussion for the work which could be further done.

2 Related Work

Gene network is any graph with genes as vertices, and edges between genes can represent various biological information. Metabolic gene networks use genes, which encode enzymes, as nodes, and the edges between two genes indicate that the product of a reaction catalysed by the enzyme of the first gene is the substrate of the reaction catalysed by the enzyme of the second gene. Beside the metabolic gene networks, there are gene regulatory network, protein-protein interaction network, gene co-expression networks, signaling networks, etc.

Spectral graph theory starts by associating matrices to graphs — notably, the adjacency matrix and the Laplacian matrix. The general theme is then, firstly, to compute or estimate the eigenvalues of such matrices, and secondly, to relate the eigenvalues to structural properties of graphs [2].

[1] Available on *https://github.com/yuhouzhou/eigengene*

Rapaport et al. proposed a method to integrate a priori, the knowledge of a gene network, in the analysis of gene expression data [1]. In the paper, the eigenvalues and eigenvectors of Laplacian from a metabolic network are firstly derived as the priori. They defined a metric which measures the distance between gene expression profiles. This metric uses the eigenvalues to attenuate high frequency components of the metabolic network. The paper further illustrates how such distance metric can be used for classification or regression purpose.

3 Spectral Decomposition

Spectral graph theory has a long history. Both algebraic method and geometric method utilize eigenvalues of the graph matrices. One of the main goals in graph theory is to deduce the principal properties and structure of a graph from its graph spectrum. The Laplacian of the graph G is the $n * n$ matrix, let d_v denote the degree of the vertex v [3]:

$$L(u,v) = \begin{cases} d_v & if\ u = v, \\ -1 & if\ u\ and\ v\ are\ adjacent, \\ 0 & otherwise. \end{cases}$$

L is the key concept of the spectral graph theory, It has some important properties: It is symmetric, positive semidefinite, and singular; Its eigenvalues λ_i are larger than 0, and the the smallest eigenvalue $\lambda_0 = 0$.

4 Network Biology

There are many different forms of interaction happening in biological organisms. These different kinds of interactions can be described as networks. It provides a mathematical representation of such connections. Adjacency matrix and Laplacian matrix are commonly used to study propoerties of networks. In order to understand networks better, some basic attributes of it should be known [4].

Veterx and Edge: A network is an abstract representation of a number of points connected by lines. Each point is usually called a vertex, and the lines are edges. In a terrestrial transportation network, some geographical locations can be vertices and the roads between them are edges. In protein-protein interaction networks, vertices are proteins and links represents mutual interactive relationship.

Directed or Undirected: Depending on the nature of the interactions, networks can be directed or undirected. In the above examples. A terrestrial transportation network can have roads which allow only one direction driving, so in many cases it is beneficiary to consider direction of edges in analysis. In a protein interaction network, the edges represent mutual relationship among proteins. Thus, such network is undirected.

Degree: The most elementary characteristic of a node is its degree, which tells us how many edges the vertex has to other vertices. In directed graph, for each node, the degree of one vertex is divided into incoming degree and outgoing degree. For the mentioned transportation network, for certain location, the roads which allow transportation coming in make incoming edges, and the number of them is the incoming degree.

4.1 Gene Network

In this project, experiments about the Laplacian property of a specific kind of biological network, gene network, was studied. Gene network is any graph with genes as vertices, and edges between genes can represent various biological information. Gene regulatory networks (DNA-protein interaction network) describes transcription factors, proteins binding to DNA, regulate the activities of genes. Metabolic gene networks use genes, which encode enzymes, as nodes, and the edges between two genes indicate that the product of a reaction catalysed by the enzyme of the first gene is the substrate of the reaction catalysed by the enzyme of the second gene. The main gene network used in this project is the metabolic network of yeast.

5 Data Source

BiGG database contains a repository of high-quality, manually-curated genome-scale metabolic models (GEM), which can be used to predict metabolic pathway usage and growth phenotypes. BiGG centralized more than 75 GEM's [5]. The data set of a organism can be downloaded on its website in *xml, json,* and *mat* formats. A python library (BiGG[2]) is also provided for accessing the database programmatically. The syntax of the data set is intuitive and it can be extracted to structured form with little effort. Comparing to more extensive knowledge base, such as Kyoto Encyclopedia of Genes and Genomes (KEGG), the process of constructing metabolic network is more straightforward.

The Gene Expression Omnibus (GEO) database stores high-throughput gene expression profile data sets with detailed descriptions. [6] To recognize different records, a naming convention is applied. Series (GSExxx), Sample (GSMxxx), Platform (GPLxxx) and data set (GDSxxx). "A Platform record is composed of a summary description of the array or sequence and, for array-based Platforms, a data table defining the array template. A Sample record describes the conditions under which an individual Sample was handled, the manipulations it underwent, and the measurements derived from it. A Series record links together a group of related Samples and provides a focal point and description of the whole study. A fourth record type, referred to as data sets (GDSxxx), are assembled by the GEO curation staff from the three primary records."

6 Experiments

The data sets are downloaded from BiGG models. The main model is iND750 for Saccharomyces cerevisiae S288C, other models, used to compare with the main one, are iAM_pb448 for plasmodium berghei, iEC1344_C for Escherichia coli C, iMM904 for Saccharomyces cerevisiae S288C. Models are stored in JSON format.

6.1 Data Preprocessing for Metabolic Network

Every model contains 6 keys, namely "metabolites", "reactions", "genes", "id", "compartments", and "version". Every BiGG model has an ID, and that contains three parts: the two capital letters stands for the first letter of author's last names. the number attached indicated the number of genes in the model.

The "compartments" entry shows the abbreviation of the naming convention for "reactions" and "metabolites". For instance, 'c' stands for 'cytosol', 'g' stands for 'golgi apparatus', and 'e' stands for 'extracellular space'. Knowing this, the comparment where some rection takes place can be informed, such as "EX_met__L_e" is located at extracellular space.

The "metabolites" entry contains a list of metabolites involved in all the reactions. Important keys for preprocessing "id", "compartment" and "formula". The "compartment" entry can be used to study the network in a specific compartment, and by defining some rules, the "formula" helps filter out inorganic substances and leave the enzymes.

The "reaction" entry contains a list of information of reactions. Every reaction are represented as a dictionary. One of the important keys for preprocesing is "metabolites" (not the "metabolites" entry) which comprise the metabolites ID and their stoichiometry of the reaction. metabolite ID can be searched in "metabolites" entry. The positive sign of the stoichiometry means the metabolite is the product of the reaction, the negative sign means the substrate. The value of "gene_reaction_rule" are the ID's of genes which rule the reaction.

[2] Available on *https://github.com/SBRG/bigg_models*

The "gene" entry contains a list of genes of the model. The important keys for preprocessing is "id".

Algorithm 1: Extract edges between genes

```
Result: edges
initialize an empty table T;
for gene in "genes" entry do
    add gene name as index the table T;
    for reaction in "reactions" entry do
        if the reaction is ruled by the gene then
            record the reaction and its substrates and products to the gene;
            exclude inorganic substance;
        else
            pass
        end
    end
end
for gene a in T do
    for gene b in T do
        if substrate of a is product of b then
            gene a and gene b have an edge between them
        else
            pass
        end
    end
end
```

6.2 Plot Metabolic Network

After the prepocessing, there are 6001 edges among 684 nodes. The experiment use Python library NetworkX and Matplotlib to draw the graph. The Laplacian matrix can be easily extracted by the NetwrokX method *laplacian_matrix()*.

The $684 * 684$ Laplacian matrix of the iND750 network is

$$\begin{bmatrix} 12 & 0 & 0 & \dots & 0 & 0 & -1 \\ 0 & 30 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 30 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 3 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 12 & 0 \\ -1 & 0 & 0 & \dots & 0 & 0 & 26 \end{bmatrix}$$

The eigenvalues and eigenvectors can be computed by *numpy.linalg.eig()*. We, following Rapaport's notation, denote eigenvalues by $0 = \lambda_0 \leq \dots \lambda_k \cdots \leq \lambda_{n-1}$ and the corresponding eigenvectors by $e_1, \dots, e_n$, where n is the number of nodes in the network. The entries of an eigenvector are used to color the network. The color map is defined as: positive coefficients are marked in red, negative coefficients are in blue, and the intensity of the colour reflects the absolute values of the coefficients. The plot showing an example of Laplacian eigenvectors is presented on Figure 1. The workflow is repeated for iAM_Pb448, iEC1344, Recon3D models, showing on 9, 10, 11. They show a similar pattern of smoothness decreasing; only Recon3D model is rather homogeneous even with the largest eigenvalue. This was further explain in Section 6.3 A plot with a fixed step between eigenvalues to show the decreasing of smoothness is on Figure 12.

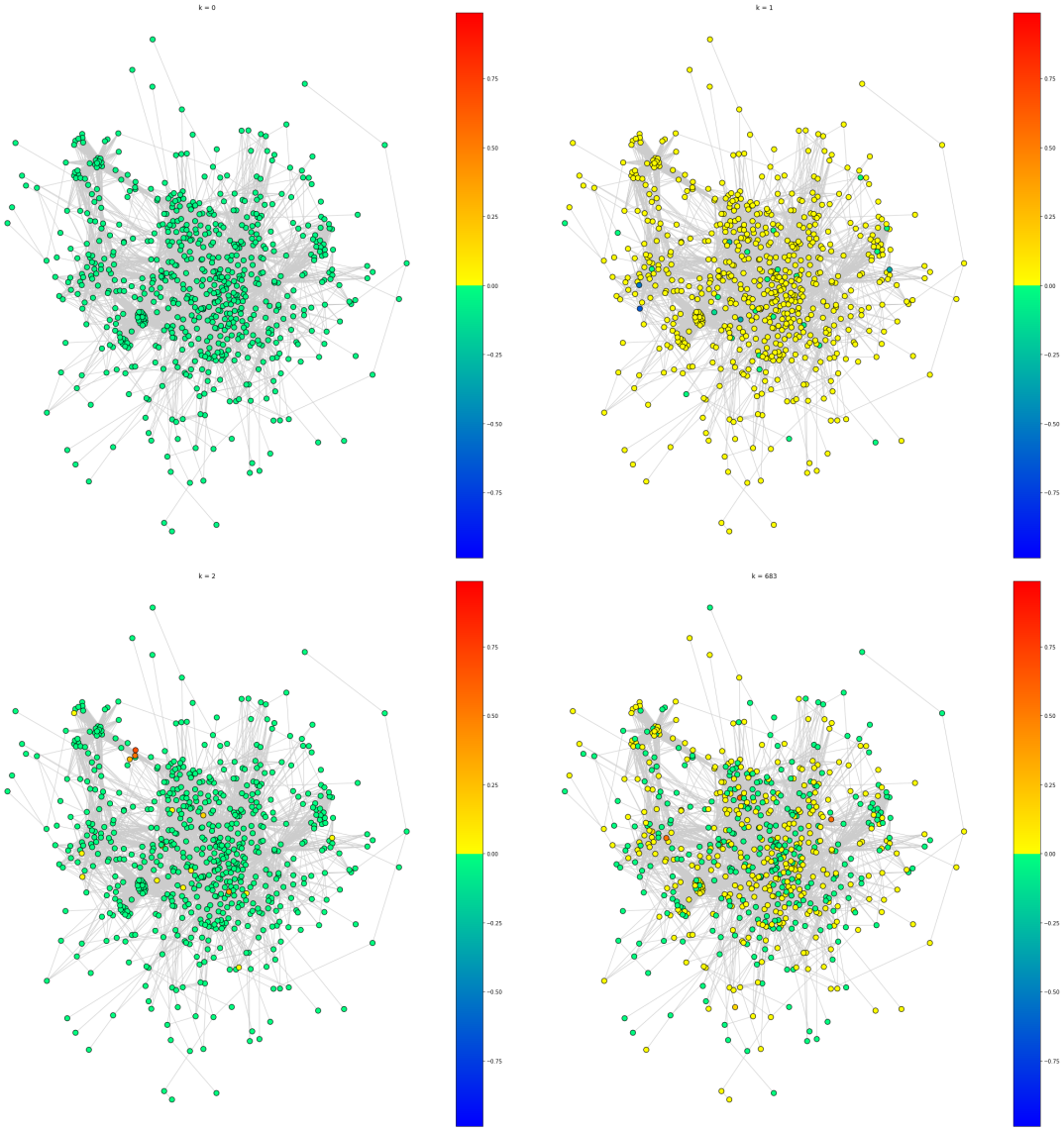

Figure 1: **Example of Laplacian eigenvectors.** On the upper-left side is the eigenvector associated with the smallest eigenvalue, on the upper-right side the one associated with the second smallest eigenvalues, on the lower-left side, the one associated with the third smallest eigenvalue while on the lower-right side is the one associated with the largest eigenvalue. The larger the eigenvalue, the less smooth the corresponding eigenvector.

6.3 Quantify Smoothness

Since the decreasing of smoothness is hard to distinguish between subplots by eyes when k is large, a quantitative way to assess the smoothness of the graph needs to be proposed. After remove the color bar and titles of the subplots, we use the file size of saved subplots to represent the homogeneity of the graph. The relation between indices k and the file sizes s are plotted as Figure 2. The x axis is standardized by $k = k/n$, and the file size is standardized by $s = \frac{s-min(S)}{max(S)}$, where S is the set of file sizes. After standardisation, x and y axes have the range from 0 to 1.

In order to check if the pattern of smoothness decreasing exists also in other metabolic network from different organisms. Above workflow was repeated for iAM_Pb448 and Recon3D models. Their

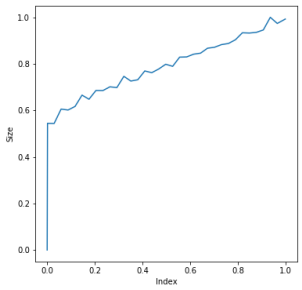

Figure 2: iND750 model (Saccharomyces cerevisiae S288C) smoothness change

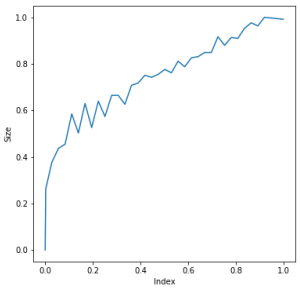

Figure 3: iAM_Pb448 model (Plasmodium berghei) smoothness change

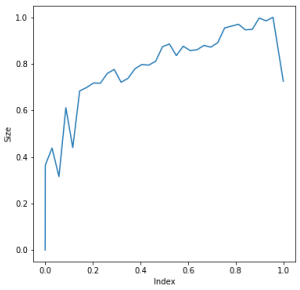

Figure 4: Recon3D model (Homo sapiens) smoothness change

results are showed on Figure 3 and Figure 4. Among different species, they are showing a similar pattern that the rate of decrease is large before first 10% of the indices, and then the rate become smaller. The previously observed homogeneous of Recon3D model with the largest eigenvalue is explained by the sudden drop of the y value on the end of x axis.

6.4 Gene Expression Profile

Gene expression profiles are downloaded from GEO database. In this experiment, one microarray data set "Expression data of Saccharomyces cerevisiae treated with clioquinol" [7] and one high-throughput sequencing data set "Gene expression analysis using RNA Sequencing of the Saccharomyces cerevisiae BY4741 H4-T30 and S47 mutant strains upon heat and osmotic stress" [8] are used. The data cleaning was performed to find the common gene in both metabolic network and gene expression profile. The visualization of the gene expression profiles are on Figure 5 and Figure 6.

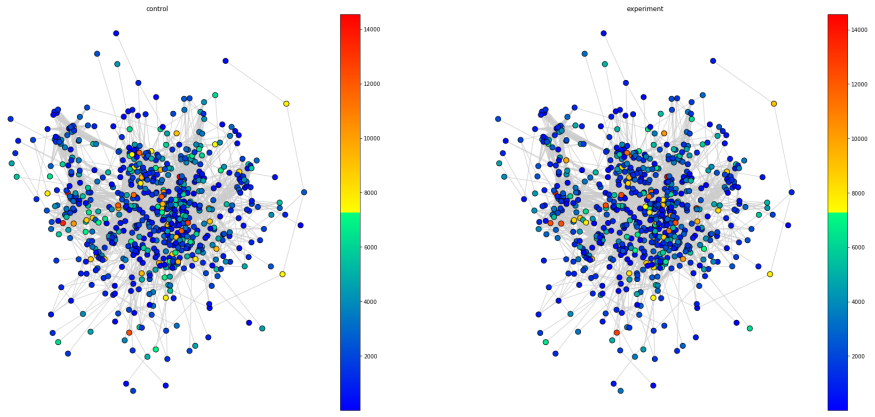

Figure 5: **iND750 microarray data set.** Expression data of Saccharomyces cerevisiae treated with clioquinol.

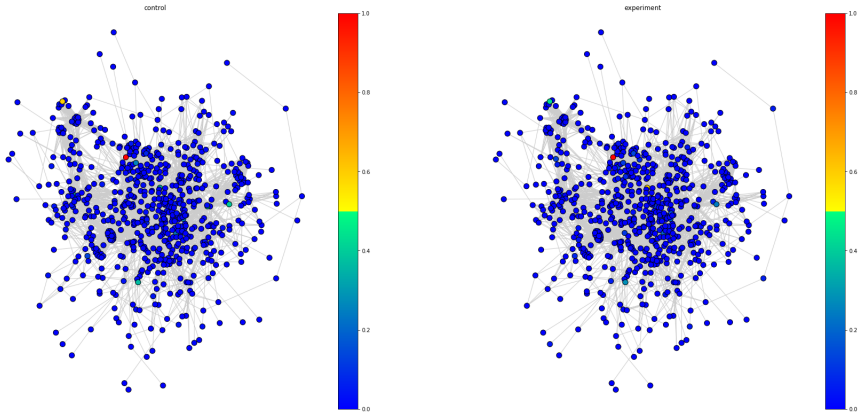

Figure 6: **iND750 high-throughput sequencing data set.** Gene expression analysis using RNA Sequencing of the Saccharomyces cerevisiae BY4741 H4-T30 and S47 mutant strains upon heat and osmotic stress

6.5 Correlation between Eigenvectors and Gene Expression Profiles

The correlation between eigenvectors and gene expression profiles are based on Pearson correlation coefficient and Spearman's correlation coefficient. Pearson correlation coefficient is defined as $r = \frac{cov_{xy}}{s_x s_y} = \frac{\sum(x_i - \overline{x})((y_i - \overline{y})}{(N-1)s_x s_y}$, where N is the number of samples, s is the standard deviation. It is the standardized covariance whose range is between -1 and 1. Spearman's correlation coefficient first takes the rank of the variables and applies the rank to Pearson function.

The final coefficient is the maximum values of all coefficients (both Pearson and Spearman's) between each eigenvector and the gene expression profile vector. The result is shown on the Table 1.

data set	Correlation coefficient	P value
Miscroarray	0.1897602896124568	2.7242330200126107e-05
RNA sequencing	0.2717323999012909	6.592013500421056e-13

Table 1: Maximum Correlation Coefficient

In order to validate the result, the entries of gene expression profile vector are shuffled, and then the shuffled vector is compared to the eigenvectors. The process is conducted for 100 times, and thus we yield a distribution of the result. It is shown on Figure 7 and Figure 8.

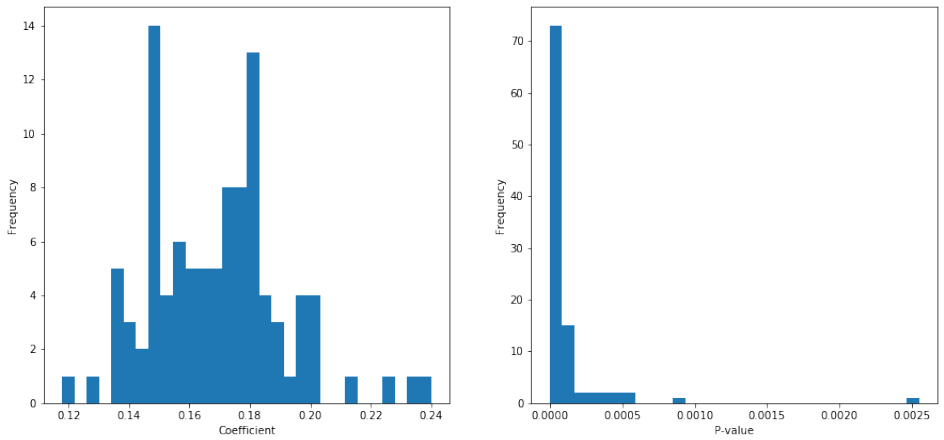

Figure 7: **Distibution of coefficients and p-values of iND750 microarray data set**

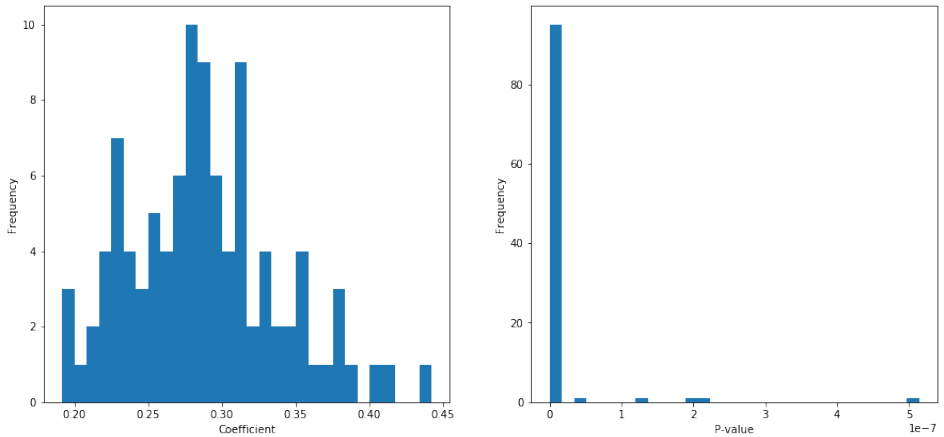

Figure 8: **Distibution of coefficients and p-values of iND750 data set high-throughput sequencing data set.**

The two modes of the distribution of microarray coefficient are 0.15 and 0.18, the mode of the distribution of RNA-sequencing coefficient is around 0.275. The modes are close to the coefficients, which are got from comparing gene expression profiles and eigenvectors, so we cannot object that the correlated coefficient are insignificant, and the maximum value is yield by coincidence. The conclusion is that the eigenvectors are not significantly correlated with gene expression profiles.

7 Conclusions

From the experiment results, I conclude that the smoothness is overall decreasing, when colored graphs use eigenvectors corresponding to larger eigenvalues. The decreasing rate is high before around ten percent. After the plump of the rate, it declines slowly. There is little difference between species, in terms of such smoothness decreasing pattern.

The correlation between the eigenvectors and gene expression profiles is weak, because the correlation coefficient is falling into the interval around the mode of correlation coefficient distribution of randomly shuffled gene expression vectors.

8 Discussion

In this project, only undirected graphs are considered. A Directed graph (digraph) also can be used to describe metabolic network, as in a chemical reaction, substrates and products are normally not reversible. Akman et al. introduced how Spectral Functional-Digraph Theory can be applied to Gene Regulatory Network. [9] The method studies adjacency matrix of the directed graph instead of Laplacian matrix properties. A directed graph can be presented not only by wiring diagram but also phase space, which is more tractable. Considering adjacency matrix and phase space for further analysis is promising.

Acknowledgements

This work is mentored by Professor Marc-Thorsten Hütt. I thank him for giving me a guiding hand, and his opening to my mistakes and inexperience. I could not implement this project successfully and have the inspiration to explore more in the field, without his patient mentoring and careful explanation.

9 Supplementary

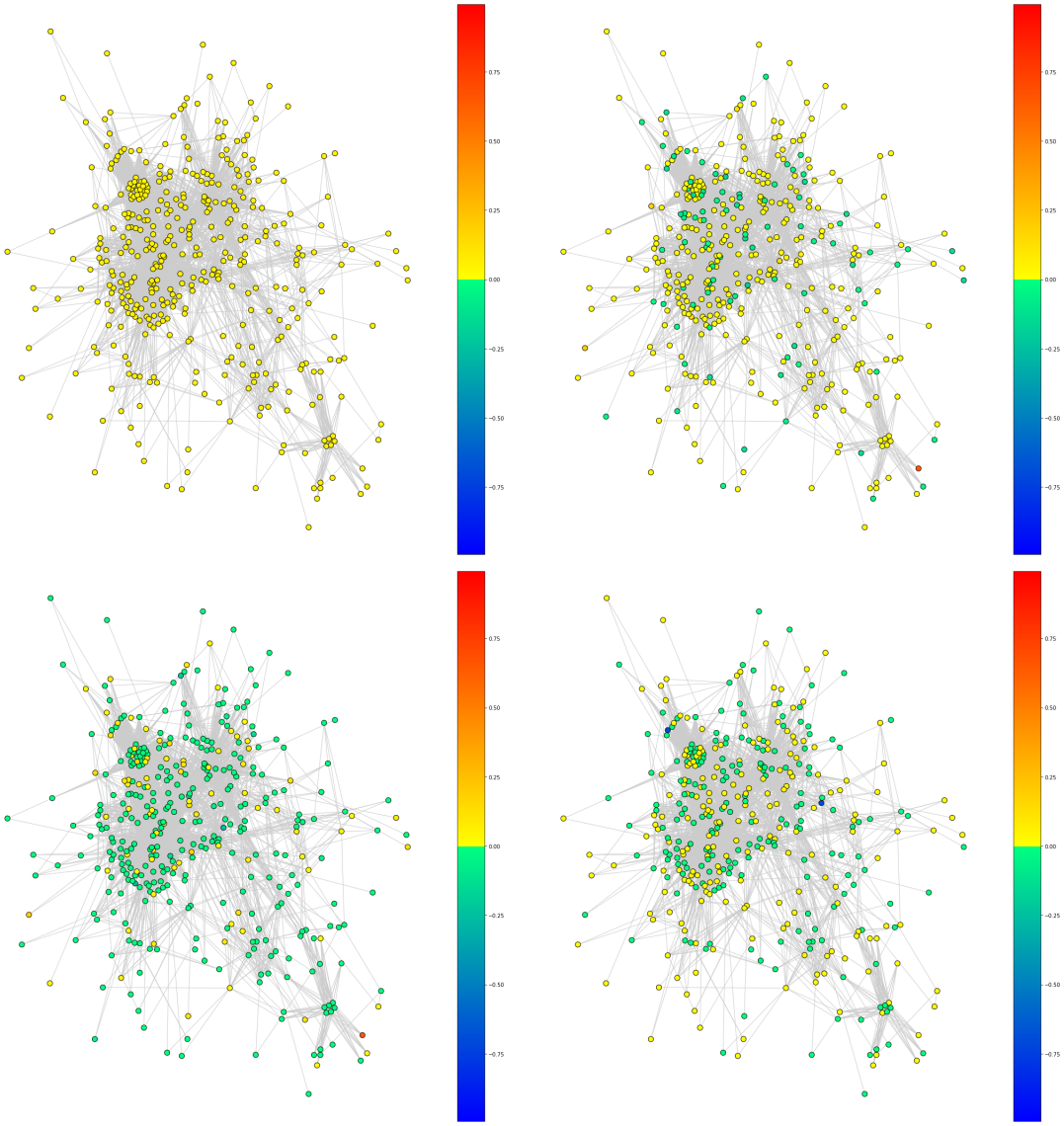

Figure 9: iAM_Pb448 metabolic network

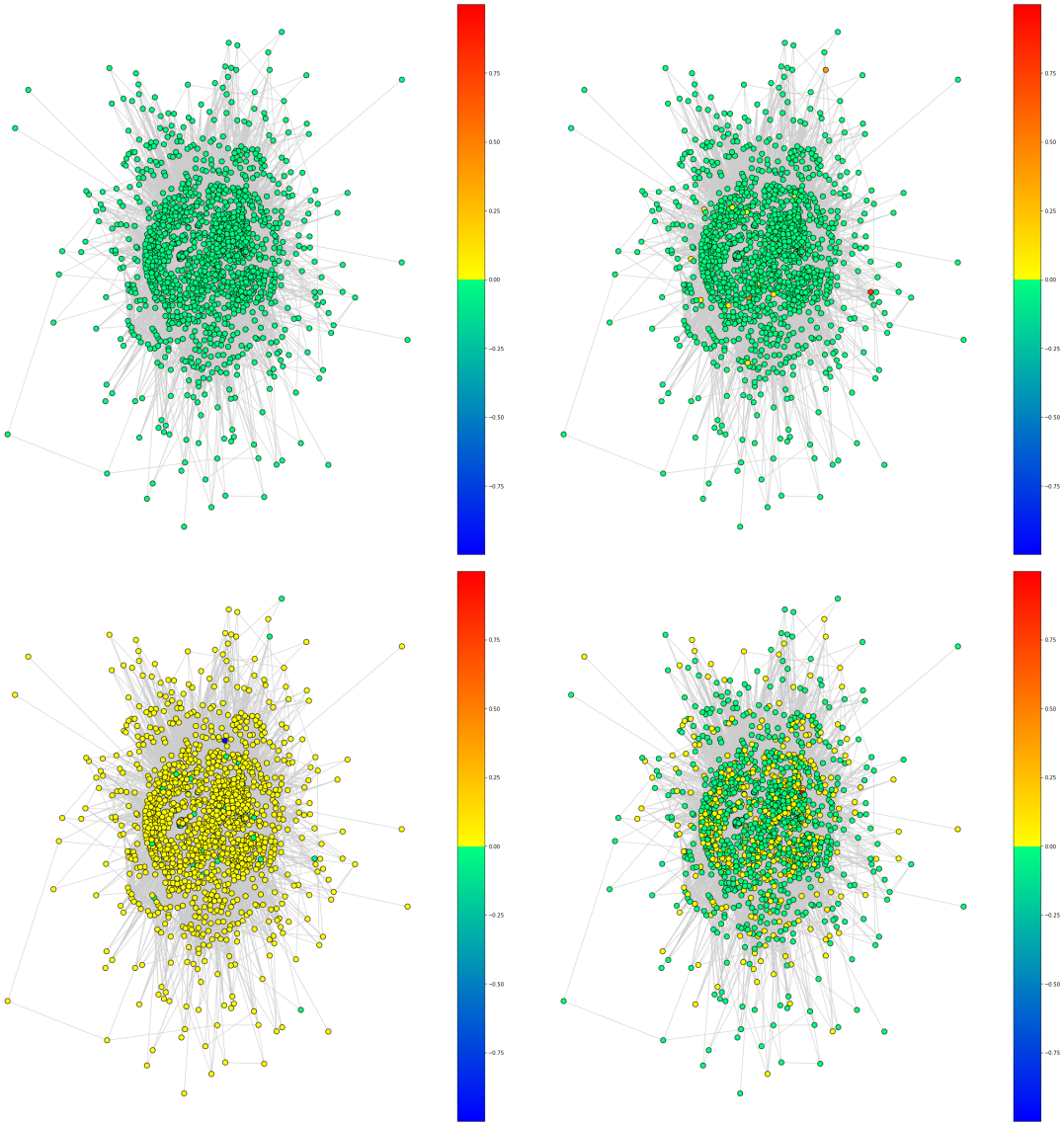

Figure 10: iEC1344 metabolic network

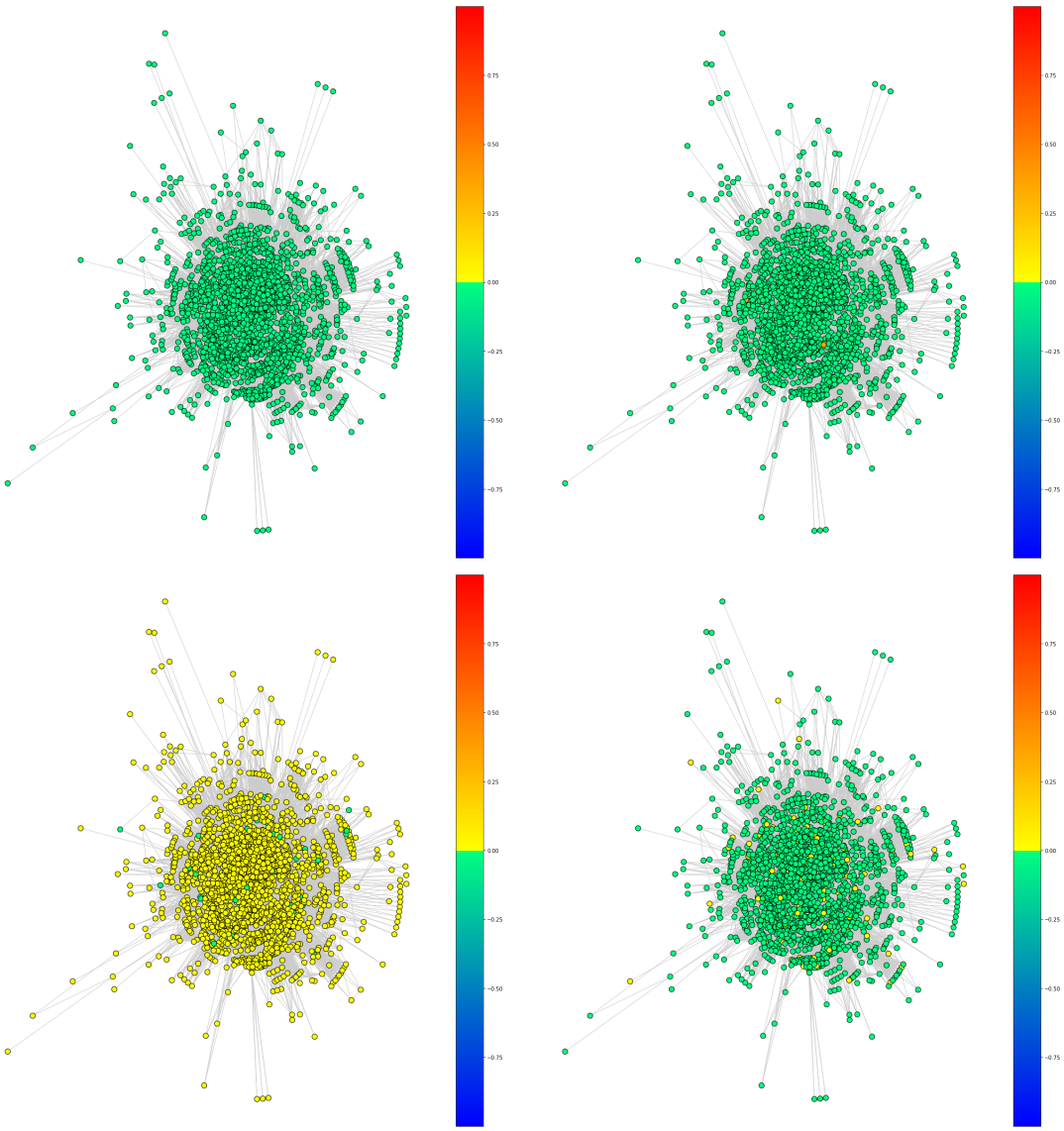

Figure 11: Recon3D metabolic network

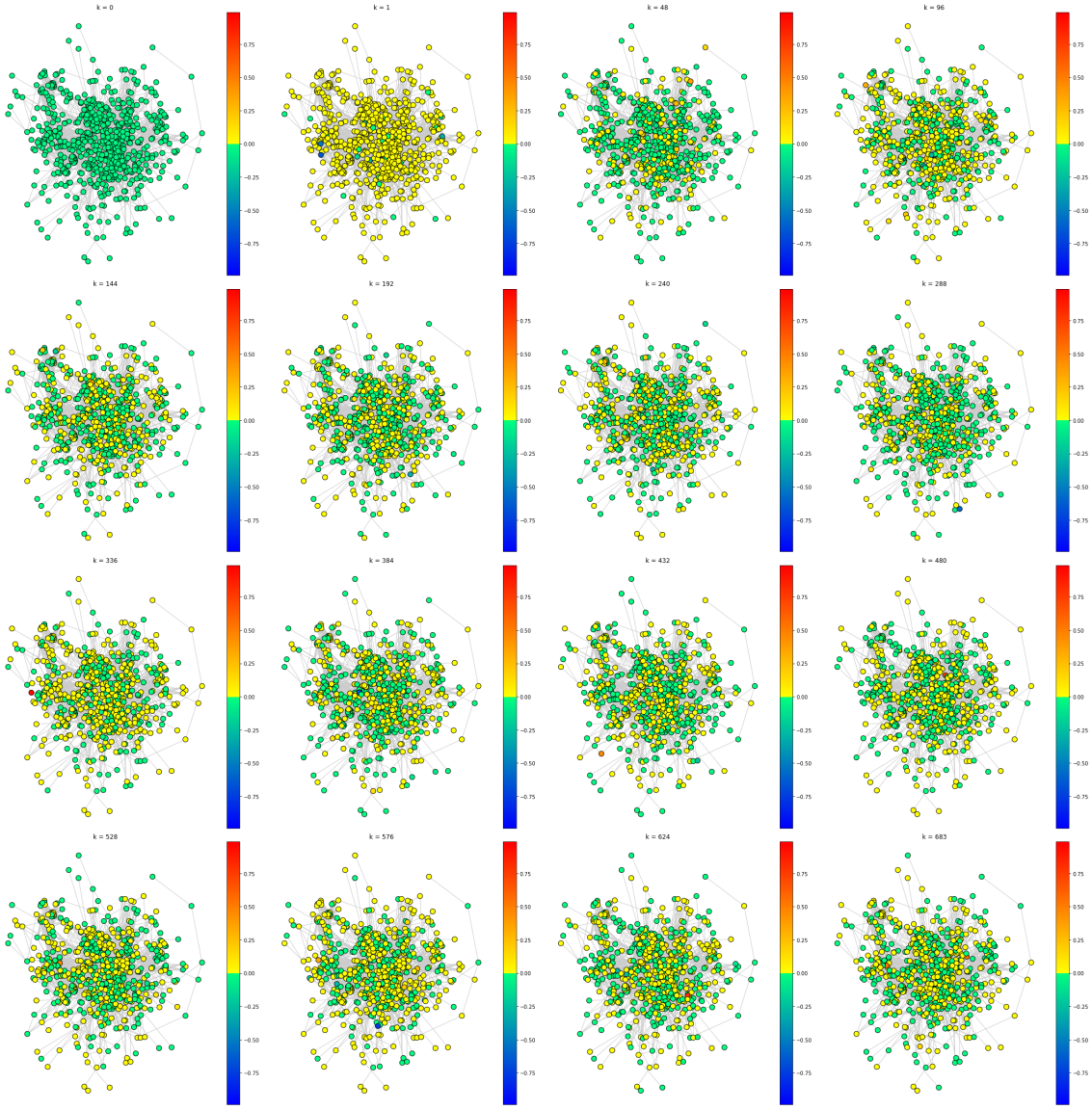

Figure 12: **Decreasing of the Smoothness.** Set a fixed step between graphs, the trend of smoothness decreasing is more obvious. The decreasing rate also seems to decline as index get larger.

References

[1] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J.-P. Vert, "Classification of microarray data using gene networks," *BMC Bioinformatics*, vol. 8, no. 1, Dec. 2007. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-35

[2] B. Nica, *A Brief Introduction to Spectral Graph Theory*. Zuerich, Switzerland: European Mathematical Society Publishing House, May 2018. [Online]. Available: http://www.ems-ph.org/doi/10.4171/188

[3] F. R. K. Chung, *Spectral graph theory*, ser. Regional conference series in mathematics. Providence, R.I: Published for the Conference Board of the mathematical sciences by the American Mathematical Society, 1997, no. no. 92.

[4] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, Feb. 2004. [Online].

Available: http://www.nature.com/articles/nrg1272

[5] Z. A. King, J. Lu, A. Dräger, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. O. Palsson, and N. E. Lewis, "BiGG Models: A platform for integrating, standardizing and sharing genome-scale models," *Nucleic Acids Research*, vol. 44, no. D1, pp. D515–D522, Jan. 2016. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1049

[6] E. Clough and T. Barrett, "The Gene Expression Omnibus database," *Methods in molecular biology (Clifton, N.J.)*, vol. 1418, pp. 93–110, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4944384/

[7] "GEO Accession viewer." [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130549

[8] "GEO Accession viewer." [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17257

[9] D. Akman and F. Akman, "Spectral Functional-Digraph Theory, Stability, and Entropy for Gene Regulatory Networks," *Frontiers in Applied Mathematics and Statistics*, vol. 4, Jul. 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fams.2018.00028/full