NETWORK-BASED PATHWAY ENRICHMENT ANALYSIS FOR BIOLOGICAL INFERENCE OF HIGH-THROUGHPUT GENE EXPRESSION DATA

by

Pourya Naderi Yeganeh

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2019

Approved by:

Dr. M. Taghi Mostafavi

Dr. Mirsad Hadzikadic

Dr. Ann Loraine

Dr. Christine Richardson

Dr. Erik Saule

Dr. Mohamed Shehab

ABSTRACT

POURYA NADERI YEGANEH. Network-based Pathway Enrichment Analysis for Biological Inference of High-throughput Gene Expression Data. (Under the direction of DR. M. TAGHI MOSTAFAVI)

Pathway enrichment analysis models (PEM) are biological inference approaches that leverage annotated bio-molecular functions for interpreting the underlying processes of gene expression profiles from high-throughput genomic data. Common PEMs neglect the interactions among the gene/proteins and regard the known annotated functions as simple lists, even though the interactions are essential components of biological systems. Disregarding the interactions in the standard PEMs potentially results in inaccurate inference, especially when focusing on the biological pathways, which are important sub-classes of the biological knowledge. Network-based PEMs are emerging methods that account for the interactions in the biological networks to produce more informative functional interpretations. However, the methodologies that are used in the current network-based PEM do not necessarily capture the key features of the topological organization of pathways, including the upstream/downstream characteristics.

This research study devises a pathway enrichment analysis by using a novel graph model, Source/Sink centrality (SSC), to capture the network organizations in pathways effectively. The key idea of SSC is to measure the importance of a gene in both upstream and downstream of a pathway while accounting for the temporal/biochemical order of the interactions. We use SSC to derive a topological statistic for the importance of a given set of genes in the network of a pathway, and use this topological statistic to construct a network-based PEM, called Causal Disturbance Analysis (CADIA). The performance of CADIA is validated by showing that it uniquely produces relevant critical interpretations in multiple sets of experimental data, while other PEMs fail to do so. We also use synthetically generated data to

evaluate the specificity of CADIA in detecting pathway enrichments.

This research study also shows an exploratory evaluation of the SSC by hypothesizing that it can capture the topological organization of *a priori* known important genes. To this end, we investigate a battery of standard graph centrality models and their novel SSC extensions for describing the organization of cancer genes in the human pathways. From multiple perspectives, we show that the SSC extensions can distinguish between the topological positions of cancer and non-cancer genes. These results show that the SSC methodology contribute to the biological inference methods, as it can effectively capture the topological organization of a particular class of important genes in the biological pathways.

ACKNOWLEDGEMENTS

I have had a quite exciting journey at UNC Charlotte. As I reflect now on the past few years, I feel an overwhelming gratitude towards those who generously offered their support and mentorship. First and foremost, my advisor Dr. M. Taghi Mostafavi, who patiently and tirelessly fostered my growth, personally and professionally.

I would also like to thank my dissertation committee members, Drs. Mirsad Hadzikadic, Ann Loraine, Christine Richardson, Erik Saule, and Mohamed Shehab. I was fortunate to learn from and collaborate with Drs. Richardson and Saule as they generously offered their time and advice on my research. I would also like to thank Dr. Loraine for offering her support towards my research. I have also been privileged to collaborate with Ms. Zahra Bahrani-Mostafavi and Dr. David Tait since I started my Ph.D.

This work would have not been possible without the generous support of Department of Computer Science, Department of Bioinfomatics and Genomics, the Graduate School at UNC Charlotte, and Levine Cancer Institute at Carolinas Medical Center.

Last but not least, I would like to thank friends and family for providing an immeasurable support during these years. Particularly my parents, who selflessly invested their hopes and the time of their lives in me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CADIA Causal Disturbance Analysis

DE Differentially Expressed

DEG Differentially Expressed Genes

GO Gene Ontology

GSA Gene Set Analysis

GSEA Gene Set Enrichment Analysis

KEGG Kyoto Encyclopedia of Genes and Genomes

ORA Over Representation Analysis

PEM Pathway Enrichment Analysis Model

SPIA Signaling Pathway Impact Analysis

SSC Source/Sink Centrality

CHAPTER 1: INTRODUCTION

High-throughput gene expression technologies capture tens-of-thousands of sub-cellular signals and identify a detailed molecular snapshot of biological organisms at certain moments and conditions. Technologies such as Microarray and RNA-seq produce quantitative details of the mRNA contents associated with any gene in an experimental sample [1–3]. A typical high-throughput experiment measure the changes in each gene's expression profile, i.e., differential expressions, among a set of experimental samples from multiple conditions [4, 5]. However, the individual differential expressions are not independent events and, therefore, are not reflective of interconnected changes in sub-cellular functions [6]. For this reason, knowing the gene-level changes alone does not automatically yield an interpretation for the underlying biological functions and mechanisms.

Interpretation of the underlying cellular mechanisms is an integral task of a typical high-throughput gene expression study as it facilitates the extraction of high-level biological insight. The most common interpretation approach is to identify the association of the differential expressions with some known biological processes [1, 4, 7]. The known processes often come from *a priori* curated classes of genes and proteins – including cell functions, localization, disease drivers, and biological pathways [8, 9]. Currently, there are several major repositories that contain a detailed and comprehensive annotation of biological functions and classifications of genes/proteins, such as the Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [9, 10].

"Enrichment Analysis Models" is an umbrella term for the statistical biological inference methods that determine the prevalence (enrichment) of any *a priori* class

from the differential expressions [7]. There are two major methodologies of enrichment analysis; Over-representation analysis (ORA) and gene set analysis (GSA) [1, 11, 12]. ORA generally uses specific cut-off thresholds on gene-level change statistics to identify a set of differentially expressed (DE) genes. Then, ORA determines the enrichment of an *a priori* class if the frequency of DE genes of the class is higher than the global frequency of DE genes [13]. On the other hand, GSA avoids using cut-off thresholds and determines the enrichment of a class by evaluating the distribution of its members across a global sorted list of genes and their differential expression statistics [7, 11]. GSA detects the enrichment of a class if its members appear close together and incident at the extreme ends of the list [7]. Although ORA and GSA provide a systematic perspective for interpretation of gene expression data, they neglect the interactions among the genes/proteins [1]. This is major limitation of ORA and GSA, which potentially results in partial and inaccurate inference, especially when investigating the biological pathways.

The pathways are a special and important category of *a priori* classes; they are detailed networks of genes, proteins, and their interactions that carry out critical cellular functions [6, 14]. A well-established body of literature shows that the position of genes/proteins in the biological networks may determine their importance to biological organisms [15–17]. A gene might be centrally involved in several biological pathways and interact with many other genes. In such cases, the molecular-level changes of a central gene may cause dysfunction in its associated pathways. A well-studied case of this example is the dysfunction of TP53 gene that disrupt the DNA-repair and cell death programs in cancers [18]. The evidences on the topological organizations of biological networks and availability of the pathway annotations provides an opportunity to devise more informative and comprehensinve Enrichment Analysis models.

An emerging category of biological inference methods, also known as Network-based Pathway Enrichment Models (PEM), leverages the topology of the pathways for in-

terpreting gene expression data. Network-based PEMs take the underlying networks of pathways into account and are shown to detect unique and critical pathway enrichments that are observable through standard models [1, 12, 19–26]. Network-based PEMs often use graph theory or related modelings to emphasize on the importance of the differential expression patterns of the topologically central genes. While these models have provided improved inference perspectives, their abstractions of pathway organizations do not necessarily capture key topological features of pathways. Biological pathways, particularly signaling pathways, often have an upstream-to-downstream organization which indicates the interactions between the genes and proteins have a temporal and biochemical order.

This dissertation focuses on devising a comprehensive and informative pathway enrichment analysis pipeline by accounting for the organization of genes and their interactions. To capture the organization, this dissertation outlines the design of a novel graph theory concept and its usage as a statistical evidence for enrichment analysis. The presented work contains experimental, synthetic, and exploratory measures to verify the methodology.

The work of this dissertation is organized in three segments. The first segment addresses the shortcomings of the existing network-based PEMs by introducing a graph theory topological measure, Source/Sink Centrality (SSC), to capture the upstream-downstream organization of pathways [26]. SSC models a pathway as a directed graph and maps the graph nodes into real values by combining two distinct characteristics of the underlying network. The first is the importance of each network entity (gene) as a source of information, and the second is the importance as a receiver of information. We subsequently derive a topological statistical evidence by using SSC as an integral part of our network-based PEM, Causal Disturbance Analysis (CADIA). CADIA then combines the topological evidence and the standard ORA evidence to increase the sensitivity of its enrichment analysis pipeline.

The second segment of this dissertation focuses on the evaluation of CADIA. We test CADIA on multiple cancer gene expression datasets and contrast its performance against a battery of standard and the state-of-the-art PEMs. We also show CADIA's specificity by evaluating its false positive rate of enrichments when the method is tested on synthetically generated data.

The third segment focuses on exploratory approaches to show the use of Source/Sink modeling concept for capture topologically important genes. By extending multiple standard centrality measures using the Source/Sink concept, we show that the SSC framework can effectively distinguish the topological organization of known cancer genes from others in the human biological pathways [27].

Chapter 2 covers the related background/literature by overviewing differential expression analysis, the standard enrichment analysis models, and network-based PEM and their limitations. Chapter 3 constructs the analysis pipeline of CADIA by establishing the definition of Source/Sink centrality, deriving s topological evidence from SSC, combining SSC with ORA, and deriving a pathway enrichment statistic. Chapter 6 contains the experimental and synthetic validation of CADIA. Chapter 5 describes the exploratory approach for evaluating Source/Sink modeling and its utility for characterizing the topological organization of cancer genes in biological pathways. Chapter 6 summarizes this dissertations and outlines the future works necessary to be explored.

CHAPTER 2: Background and Literature Review

2.1 Differential Expression Analysis

A typical primary output of gene expression analyses, such as Microarray and mRNA sequencing, is a list of genes and their differential expression statistics. This statistic describes the changes of a gene's expression levels between two or more experimental conditions, e.g. cancer versus normal. A typical differential expression analysis may have two major procedures. The first is to calculate the statistic, and the second is to use a cut-off threshold for the statistical significance to determine a list of differentially expressed (DEG).

To illustrate the described procedure, consider a case where there are n_1 control and n_2 treatment samples. One simple differential expression analysis pipeline is to first hypothesize that the means of expressions in the control and treatment groups are equal for each gene. Subsequently, two-sample t-test can evaluate individual hypotheses under certain assumptions. For any gene, construct the following test-statistic:

$$T = \frac{\overline{X^c} - \overline{X^t}}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where T follows Student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom, $\overline{X^c}$ and $\overline{X^t}$ are the expression means of the control and the treatment groups, and s_p^2 is the pooled variance of the two groups. The output of this test is a p-value that corresponds to the probability of observing a more extreme difference between the group means. A gene is considered differentially expressed (DE) if its p-value is below some predefined threshold for which the null hypothesis of equal means is

rejected. More appropriate and complex differential expression analysis pipelines for various high-throughput platforms have been described in several manuscripts including in [4, 5, 28–31].

Differential expression analyses often test numerous hypothesis at the same time. To illustrate, consider a case where m distinct hypotheses are tested and have produced the p-values $P_1, P_2, \ldots, P_m$. Using a standard p-value threshold for testing each hypothesis (e.g $P_i < \alpha$) would result in falsely rejecting some of them beyond the desired threshold α (Type-I error). Multiple hypothesis testing methods find appropriate rejection thresholds that controls the Type-I errors. For example, Bonferroni correction states that if we set the rejection threshold at $\frac{\alpha}{m}$, then the probability of having at least one type-I error is less than α, which is a conservative method and is not practical for differential expression analysis. In practice, there are several more appropriate alternatives, such as the Benjamini Hochberg False Discovery Rate (FDR) that control the expected ratio of false rejections to the true rejections. Without loss of generality, assume the above list of p-values is sorted in an ascending order. For a desired α, FDR finds the largest k such that $P_k \leq \frac{\alpha.k}{m}$. Benjamini and Hochberg showed that for the first k hypothesis the expected value of false discovery rate is less than α. Different multiple hypothesis testing approaches and their use cases can be found in various manuscripts including in [32–35].

2.2 Biological Inference using annotated knowledge

The list of DEG may contain thousands of genes; a number that matches the systematic behavior of biological organisms. The DEG list is often a key data for extracting and interpreting the underlying cellular mechanisms in the experimental. A naïve approach to this end is to investigate the DEGs one by one and identify the biological functions associated with each. However, this approach is not informative as the changes in biological systems are often highly interconnected and coordinated [6]. Consequently, the task of finding a high-level biological interpretation for the DEGs

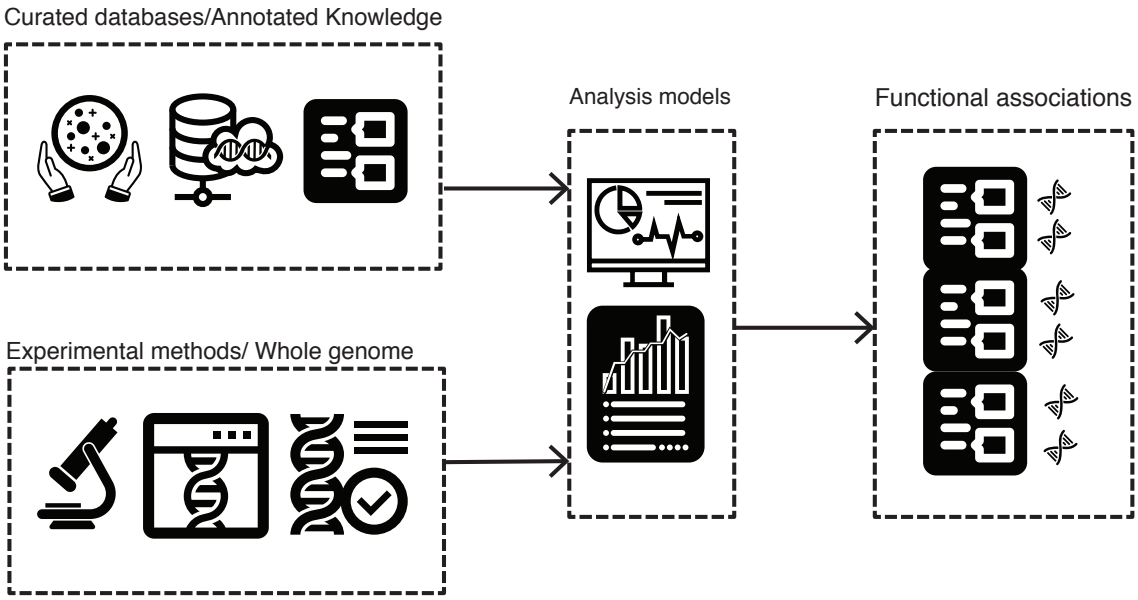

Figure 2.1: Enrichment Analysis uses two pieces of information. One is the list of differentially expressed genes from the experimental data and the other is the prior knowledge from curated databases such gene-chromosome, gene-pathways, and gene-drugs associations. Enrichment Analyses evaluate the relationship between the two and produce interpretation for the experimental data.

is not trivial and might be challenging.

Enrichment Analysis Models are systematic approach towards interpreting the gene expression data that highly organized *a priori* biological knowledge (Figure 2.1). The individual entities of *a priori* knowledge, also known as Gene Sets, are collections of genes that have been found to be associated with certain biological processes [10,36]. The most commonly used *a priori* biological knowledge repository is the Gene Ontology (GO), which contains a hierarchical representation of gene sets, their overlaps, and their associated genes [10]. While each gene set of GO is a simple list of genes, there are other datasets that may contain more detailed annotation of complex relationships for certain groups of genes set [9,37].

A special class of gene sets is biological pathways, which are of significant interest. From a bioinformatics perspective, the pathways are gene sets that contain additional information beyond class membership (Figure 2.2). While a generic gene set might

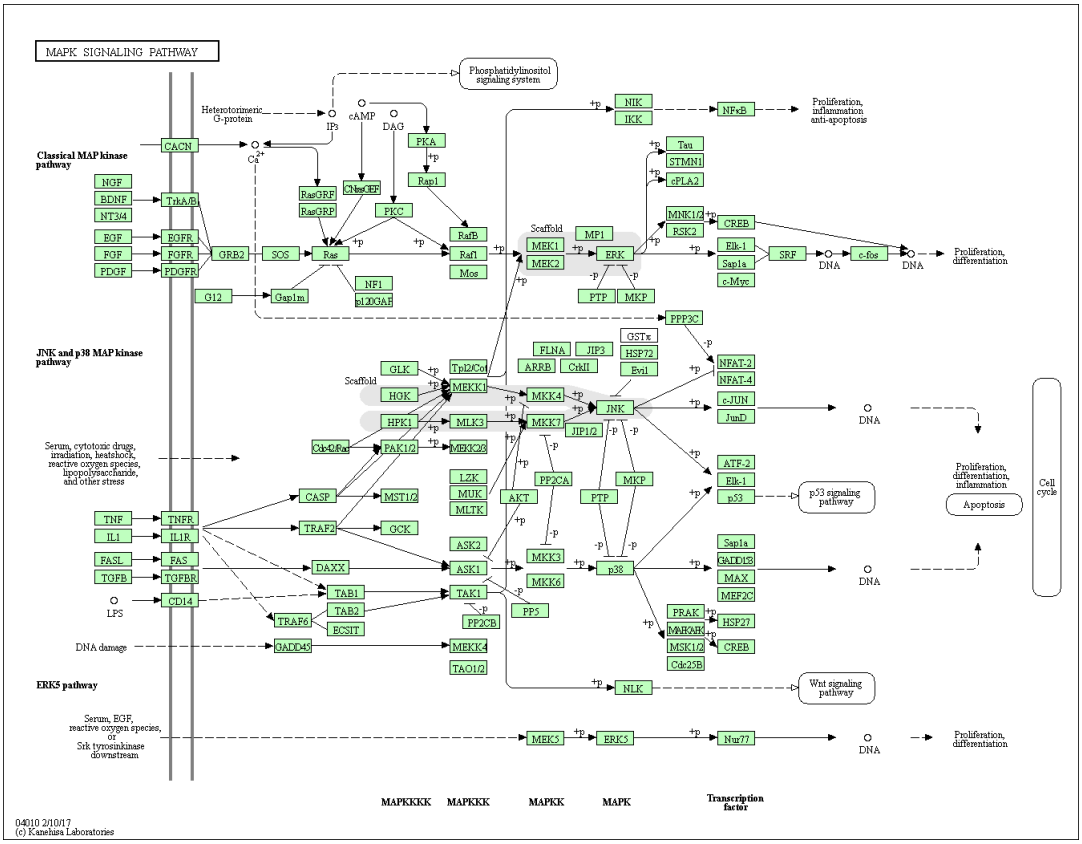

Figure 2.2: Biological pathways are a special class of gene sets that contain details of molecular interactions and their mechanisms. This example figure shows MAPK signaling which regulates several biological functions including cellular growth and proliferation. Pathways include a variety of entities, such as genes coded elements (in green), enzymes, or metabolites. The connections in the pathway diagrams may represent different types of interactions such as binding, phosphorylation, repression, and activation.

only be a simple list of gene identifiers, pathways contain details of interactions among the genes/protein such as direction or the type of interactions [9]. To date, a few hundreds/thousands pathways have been discovered/annotated in databases such as KEGG, BioCarta, and Reactome [9, 38].

From a biological perspective, the pathways are biochemical programs that regulate certain cellular functions [39]. Each pathway encompasses cascades of interactions between particular genes, proteins, and other biomolecular products (Figure 2.2). Though the changes of individual molecules may trigger variations in the cellular programs, many biological functions emerge from the systematic behaviour of entities

and interactions. This systematic behavior is a fundamental concept of the systems biology [40–42], and therefore, the concept of the pathways is essential for bridging the gap between the molecular activities and the biological functions. The use of pathways for the study of biological systems has a significant value for treatment, diagnosis, and prediction of diseases – notably that of cancer [18, 40, 43, 44]. Because of the importance of pathways in biological inference, Enrichment Analysis Models are sometimes referred to as Pathway Enrichment Analysis Models. This dissertations uses these terms interchangeably.

2.3 Pathway Enrichment Analysis Models

Pathway Enrichment Analysis Models (PEM) are among the most common techniques for biological inference [4, 7, 19, 21, 22, 25, 45–50]. As outlined in Figure 2.1, a typical PEM takes an input list of genes and their respective differential expressions, and assesses the prevalence (enrichment) of any pathway/gene-set from *a priori* datasets [4, 7–9, 37, 51]. The output of a typical PEM is a ranked list of the pathways and a significance score of enrichment for each pathway [7, 8, 19, 21, 22, 52]. For example, a PEM may report dysfunction of the pathways *cell cycle, cell death*, and *DNA repair* upon observing changes in activity of the gene *TP53*, which is associated with the aforementioned gene sets.

PEMs can roughly be classified into three generations. The first generation is over-representation analysis which evaluates the frequency of DEG in an annotated pathway [1, 11, 13]. The second generation is functional class scoring, also known as Gene Set Analysis, which assesses the distribution of a pathway's members across a sorted list of gene expressions profiles [7, 11, 53]. While the first and the second generation models provide useful inference, their methodology requires considering the pathways as simple sets and disregarding the interactions. A potential consequence of disregarding interactions is inaccurate inference due to not accounting for the knowledge of connection between genes' activities [19, 22].

The third category of PEMs, network-based models, account for the interactions among genes and proteins towards a comprehensive biological inference. Studies show that using the interactions can yield critical insight regarding the functionality of biological events [15–17,54–62]. The next few sections describe the technical details and advantages of each generation of PEM.

2.3.1 Over Representation Analysis

Over Representation Analysis is a simple method that determines whether the members of an *a priori* set of genes are over-represented among the DEG. The premise of ORA is that if there is no association between a pathway and DEG, we expect to observe the same frequency of DE genes in the universe (the set of all genes) and the pathway.

The first step of ORA is to identify a list of DEG by setting some cut-offs thresholds for differential expression statistics, e.g. p-value ≤ 0.05. ORA then builds a null hypothesis that the probability of a randomly selected gene to be DEG is independent of the probability of the gene belonging to a specific pathway.

Binomial testing is one simple approach to build ORA pipeline. Formally, let p denote the probability of a gene being DEG. Here, p is estimated from the total number of DEG divided by all of the genes in the high-throughput machinery. Suppose that k genes from a pathway with n members are found to be DEG. Under the assumption of null hypothesis, the probability of each gene in the pathway being belonging to DEG is equal to p. By this assumption, we can derive the probability of observing exactly k DEG in the pathway as:

$$P[X = k] := \binom{n}{k} p^k (1-p)^{n-k} \tag{2.1}$$

ORA evaluates the probability of observing k or more DEG from the pathway (P_{ora}). This p-value is a summation of the probability in formula 2.1 for all values of

size k and larger. Let X denote the number of observed DEG from the pathway:

$$P_{ora} := P[X \geq k] = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \tag{2.2}$$

This formula determines the enrichment of a pathway if we reject the null hypothesis that probability of being DEG is independent from belonging to the pathway. The binomial testing approach for ORA only works when meeting certain conditions on p and n because of the dependencies between the size of the pathway, size of the global gene list, and the size of DEG. Contingency table approaches, such as Fisher's Exact test or Chi-squared test, are more appropriate alternatives for ORA that produce more accurate results [13]. Suppose that K genes are found differentially expressed from the set of all genes with size N. To elaborate, suppose that k genes from a pathway with n members are found to be differentially expressed. The contingency table is constructed as shown in Table 2.1.

Table 2.1: Contingency table of observed differential expressions from high-throughput experiments and a given *a priori* pathway

	DE	Not DE	Total
In Pathway	k	n-k	n
Not in Pathway	K-k	N+k-n	N-n
Total	K	N-K	N

Fisher's exact test evaluates the probability of observing exactly k DEG's in a pathway. The p-values of Fisher's exact test come from hyper-geometric distribution and are obtained using the following formula:

$$\begin{aligned} P[X = k] &= \frac{\binom{n}{k}\binom{N-n}{K-k}}{\binom{N}{n}} \\ &\sim Hyper(k; K, n, N) \end{aligned} \tag{2.3}$$

Under the assumption of null hypothesis of this test, the probability of of observing

k DEG in a pathway determined by the global size of DEG, K, the size of the universe, N, and the size of the pathway, n. The formulation of hypergeometic probability distribution comes from the assumption of random sampling of DEG without replacement, i.e. k DEG are selected from n choices and the rest, $K-k$, are selected from $N-n$. Whereas, the formulation of binomial testing comes from the assumption of random sampling of DEG with replacement. The null hypothesis of Fisher's exact test is evaluated by finding the probability of observing k or more DEG in the pathway:

$$P_{ora} := P[X \geq k] = \sum_{i=0}^{n-k} \frac{\binom{n}{k+i}\binom{N-n}{K-k-i}}{\binom{N}{n}} \tag{2.4}$$

P_{ora} is calculated using the cumulative density function of hypergeometric distribution, which is the output of ORA and is often used for determining the enrichment. A pathway is enriched if we reject the null hypothesis with with some pre-defined cut-off threshold for P_{ora}.

Different approaches of ORA have specific advantages and disadvantages. The choice of method for ORA depends on the Gene Set size and the probability of DEG. There is also a computational trade-off depending on the size of the DEG and the size of the pathways which affects the choice of the model. Although contingency table tests provide more suitable problem formulation for ORA, they are shown to be statistically conservative [63, 64]. Given that ORA simultaneously test hundreds of different pathway enrichments, their cut-off threshold is determined by using multiple hypothesis testing methods such as False Discovery Rate or Bonferroni correction.

ORA is a very popular approach in gene expression studies. The background knowledge of pathways and gene sets are available from multiple biological databases. ORA allows for designing costume gene sets by simply compiling gene lists of interest.

However, ORA has some drawbacks [1]. ORA considers DEG as independent events and does not account for their dependencies. Also, ORA is based on subjective thresholds. In particular, DEG input is subject to change based on p-value and fold-change cut-offs. In such instances, the information with marginal thresholds values are lost, e.g. a gene would not be involved in ORA pipeline with a p-value of 0.5001 [1].

2.3.2 Gene Set Analysis

The second generation of PEMs, Gene Set Analysis (GSA), takes a different approach from ORA for biological inference and avoids using cut-offs on differential expression statistics. Instead, GSA uses a global sorted list of gene expressions to calculate the enrichment.

Gene Set Enrichment Analysis (GSEA) is the main GSA method [7]. The general idea of GSEA is leveraging all of the genes from high-throughput experiments. GSEA first sorts the list of all genes based on some criteria of significance and interest – For example, one can sort the global gene list based on fold-change, differential expression p-values, or correlation with certain phenotypic observations. In GSEA, the association of a pathway is with the experimental data is measured by the tendency of its member genes to appear on top (bottom of the list). The statistical evaluation of GSEA allows for conserving the correlation structure of the genes in the analysis [7].

GSEA models calculate an aggregate Enrichment Score (ES) for a functional set, F_i. Formally, suppose that $G =< g_1, g_2, \ldots, g_n >$ is a ranked list of genes and $R =< r_1, r_2, \ldots, r_n >$ is the ranking values, such as correlation with some phenotype, differential expression p-values, or fold change. The ES of a functional group is a running sum on the rank values. In particular, let j specify a threshold index for the ranked list. Define the two score, hits in the functional set, $H(i)$, and miss of the functional set, $M(i)$, as following:

$$H(i) = \sum_{g_j \in F, j \leq i} \frac{|r_j|}{N_R}, \ where \ N_R = \sum_{g_j \in F} |r_j| \tag{2.5}$$

$$M(i) = \sum_{g_j \notin F, j \leq i} \frac{1}{n - |F|} \tag{2.6}$$

Define ES as the maximum value of $|H(i) - M(i)|$. The null distribution of ES is generated by shuffling the sample labels and re-calculating r_j's and ES for F in each permutation. The statistical significance of ES is then calculated from this null distribution as the probability of observing a higher ES value. GSEA addresses some critical limitations of ORA. First, GSEA is free from cut-offs and, thus, is more objective. Second, GSEA can address genetic dependencies because the permutation-based null hypothesis testing preserves the correlation structure. GSEA however also has some disadvantages. Most notably, it is insensitive to the underlying topological structure of pathways. GSEA would produce the same results regardless of the the left most genes in the Figure 2.3 being highly connected or not.

2.4 Network-based Pathway Analysis

The third generation of PEMs, network-based models, use the biological interactions as an additional source of information [1]. As discussed, ORA and GSA are sensitive to the positionally (topologically) important gene/proteins and this would potentially deliver partial and inaccurate inference. In contrast, the premise of Network-based pathway analysis is that using the interactions can produce more accurate biological interpretation from the experimental data [1]. From a systems biology perspective, the interactions are integral parts of cellular functions. A strong body of literature suggests that the topological position of entities in biological networks can determine their importance to key functions in biological organisms [15, 16, 56]. For instance, Jeong et. al showed that the number of interactions associated with each

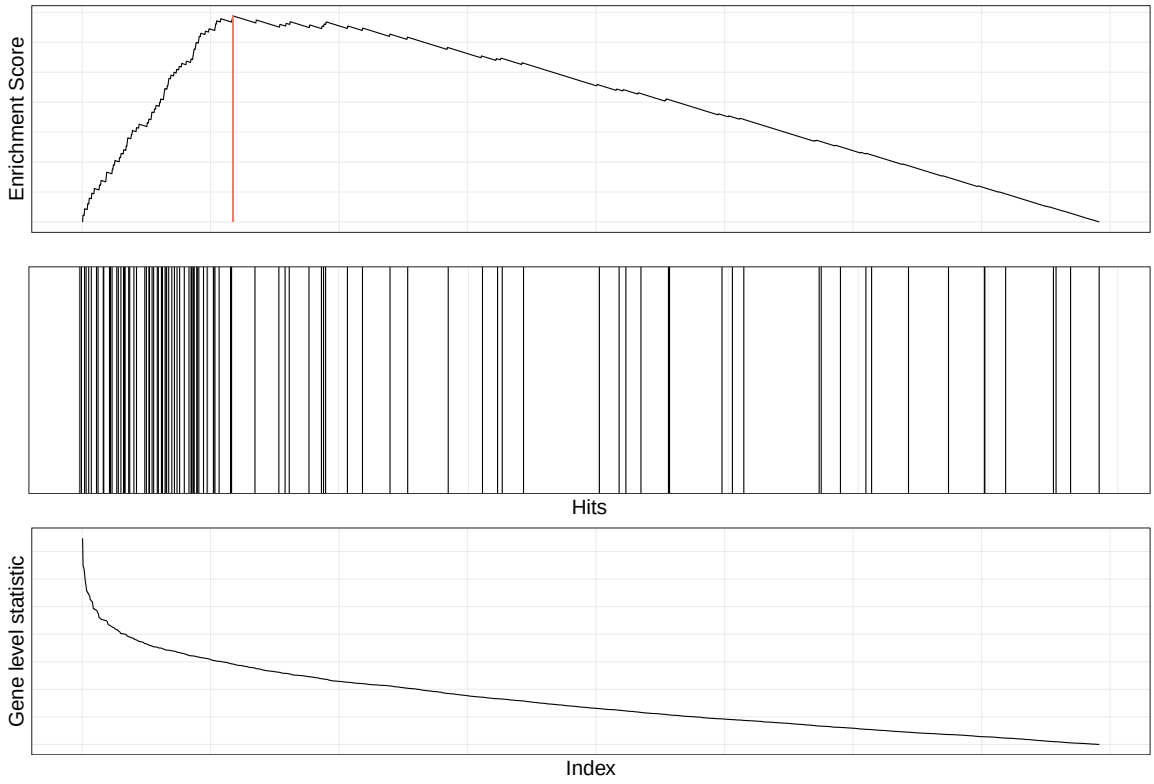

Figure 2.3: GSEA calculates an enrichment score based on the distribution a pathway elements in the list of all genes from High-throughput experiments. The X-axis of all three panels is the sorted gene list. The lower panel represents the sorted statistic, e.g. correlation with a phenotype. The leftmost is has the highest level of gene-level statistics and the rightmost is the lowest. The vertical lines in the middle panel represent pathway elements. The top figure is the ES. The red line is the maximum ES.

gene/protein (degree distribution) correlates with the probability that its removal would be lethal to the subject organisms [15]. The topological organization of biological networks provides an additional perspective for understanding and studying biological functions.

Several network-based PEM use graph theory concepts to account for the interactions between pathways [12]. Models such as "Signalling Pathway Impact Analysis" (SPIA) [19] and "EnrichNet" [22], leverage the topological properties of the network of genes and proteins for identifying the prevalence of *a priori* functional classes. The differences between various network-based models lie within their abstraction of topo-

logical importance. In the remaining parts of this section, we first overview the basic concepts of graph theory and then overviews some of the prominent network-based pathway analysis models to represent different perspectives.

2.4.1 Graph Modeling of Pathways

Define a graph, $G = (V, E)$, as a pair of two sets, the nodes and the edges. The set of nodes, $V(G) = \{v_1, v_2, \ldots, v_n\}$, represents n distinct elements, and the set of edges, $E(G) = \{e_1, e_2, \ldots, e_m\}$, represents m distinct interactions between the nodes. Each edge, $e_k = (v_i, v_j)$, is an ordered pair that indicates a directed relationship from gene-encoded element v_i to v_j. This notion is known as the *directed graph*. A graph can be alternatively represented using a notation where the edges are unordered pairs, thus without directionality, namely the undirected graph. Figure 2.4 depicts a directed

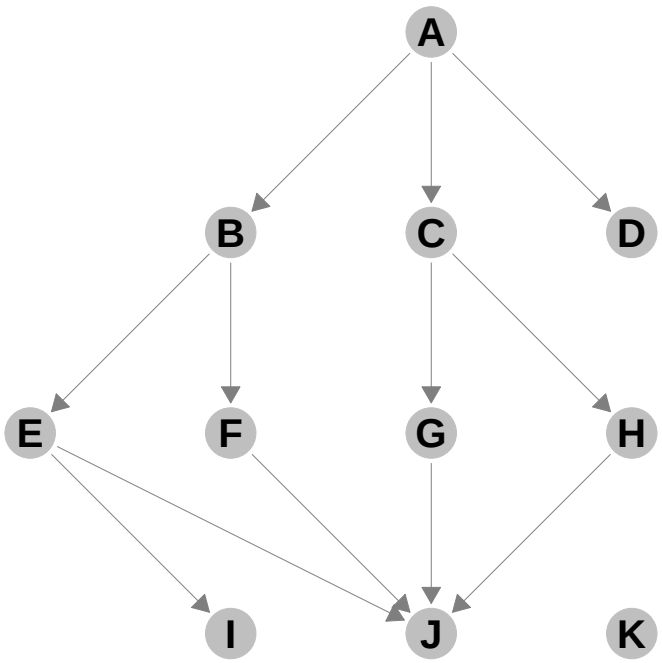

Figure 2.4: Common Enrichment Analyses, such as ORA and GSA, do not differentiate between the genes based on their position in the pathways. This figure illustrates the underlying graph of an example pathway. In this case, GSA and ORA regards the nodes A and K similarly. Network-based PEM address this issue by accounting for the underlying interactions in pathways. This figure illustrates a scenario of why it is essential to account for the underlying networks. Here, removal of the node A or J causes a more extreme disconnect in the graph compared to that of node K or D.

graph with the set of nodes $V = \{A, B, \ldots, K\}$ and the directed edges between them, e.g. (A, B) and (G, J). One can imagine an undirected graph using this example if the edge directions were removed.

A *walk*, w, is a sequence of the graph nodes $w = (v_i, \ldots, v_k, v_{k+1}, \ldots, v_j)$ in which any two consecutive vertices are connected by a link, $(v_k, v_{k+1}) \in E$. An ij-walk of a graph is a walk such that it starts at node i and ends at node j. The size of a walk, $|w|$, is the number of edges in a walk. For example in graph of Figure 2.4, the sequence (A, C, H, J) is an $AJ-$walk of size 3. A *cycle* is a walk were the start node and the end node are the same. An *acylic* graph is a graph that has no cycles.

For any graph, the *neighborhood* of a node v_i, $N(v_i)$, is the set of all adjacent nodes of v_i, $N(v_i) = \{v_j | (v_i, v_j) \in E(G)\}$. In a directed graph, the previous notion denotes the set of out-going edges. Degree of a node is defined as the size of its neighborhood, $Deg(v) = \big|N(v)\big|$. For a directed graph, the former notion of degree is called out-degree, $Deg_{out}(v)$. For example in Figure 2.13, $N(A) = \{B, C, D\}$ and thus. $DegA = 3$. Alternatively for a directed graph, neighborhood and degree can be defined based on in-coming edges, i.e. in-degree, $Deg_{in}(v) = \big|\{u \mid (u, v) \in E\}\big|$.

Any graph with n vertices has an equivalent representation of a $n \times n$ square matrix form, also known as the *adjacency matrix*, A_G. Formally:

$$[A_G]_{ij} = \begin{cases} 1, & (v_i, v_j) \in E \\ 0, & \text{otherwise} \end{cases} \tag{2.7}$$

As the above definition suggests, the adjacency matrix of an undirected graph is symmetric, $A_G = A_G^T$. This property does not necessarily hold for directed graphs. The adjacency matrix notation is useful for computational purposes, including for graph centrality measures. A graph *centrality* is a function, $C(v)$, from $V(G)$ to real numbers for describing a topological scoring (importance) of the nodes in a network, $C : V(G) \rightarrow \mathbf{R}$ [65]. Several graph centrality measures have been widely used for

topological description of networks across different disciplines [65]. Throughout this document, we will cover and explain some of the well-known centrality measures such as Katz, PageRank, and Degree; particularly, in Chapters 3 and 5.

2.4.2 Model Overviews

EnrichNet: is a network-based PEM by Glaab et al. that maps the DEG and the pathways genes to a global background network of protein-protein-interaction (PPI) to calculate the enrichments. In particular, EnrichNet initially calculates a random-walk distance between the DEG and the nodes of a target pathway [22]. A random walk initially starts at some particular nodes and at each step of time transitions randomly to a neighboring node. EnrichNet uses a specific definition of this concept, random walk-with-restart, which has an additional option to randomly jump to a predefined set of nodes at each transition step [66, 67]. EnrichNet than uses the vector of eventual probability distribution of the random walk being present at each node to calculate the distance. Formally:

$$P^{(t+1)} = (1-\alpha)P^{(0)} + (\alpha)P^{(t)}W \tag{2.8}$$

where $P^{(0)}$ is the initial state of the random walk and α is the transition probability. $P^{(t)}$ is the vector of the probabilities of being at each specific node. W is transition matrix which is the normalized adjacency matrix of overall PPI using the formula $W = D^{-1}A$, where $[D]_{ii} = max(Deg(v_i), 1)$. In EnrichNet, $P^{(0)}$ is determined by the set of DEG, $G = \{g_1, g_2, \ldots, g_m\}$. Enrichnet derives a distance based the probability of the random walk being present at each particular pathway node. It then divides the distance scores into n equal sized bins and defines the enrichment score as following:

$$\begin{aligned} Xd &= \sum_{i=1}^{n} \frac{P_{ic} - P_{ia}}{i \cdot n} \\ P_{ic} &= \frac{|\text{reference}_c \cap \text{target}_i|}{\sum_{j=1}^{n} |\text{reference}_c \cap \text{target}_j|} \end{aligned} \tag{2.9}$$

In the above equation, P_{ic} is the bin score and P_{ia} is the sum of distance scores across the whole PPI. Enrichnet calculates the statistical significance of Xd by finding specific thresholds using a regression fit of Xd values with P_{ora}.

SPIA: Tarca et al. also used the graph structure of the pathways to detect the enrichment of pathways. Their model (SPIA) finds the differential expressions and weights them according to the position in the pathway's network. In particular, SPIA calculates an accumulated perturbation in the pathway according to the positional weights and devises an enrichment pipeline according to this evidence. Formally, let $\Delta E =< \Delta e_1, \Delta e_2, \ldots, \Delta e_n >$ represent the vector of differential expression fold-changes for the genes in a pathway. SPIA defines a perturbation factor (PF) [19] for each gene, i, in a pathway as following:

$$PF(i) = \Delta e_i + \sum_{j:\ i \in N(j)} a_{ij} \cdot \frac{PF(j)}{|N(j)|} \tag{2.10}$$

In the above formulation, $N(j)$ is the neighborhood of gene j which itself is a neighbor of gene i. a_{ij} is the ij-th element of pathway adjacency matrix. For example, the PF of node J in Figure 2.4 would be its perturbation summed with the PF's of nodes E,F,G, and H weighted by their out degree. After calculating the PF's for all nodes, SPIA calculates an accumulated perturbation of a pathway node using the following formula:

$$Acc_i = PF(i) - \Delta e_i \tag{2.11}$$

Authors then show that the by using some algebraic manipulation the following formula to calculate the vector of Acc_i's:

$$Acc = B(I - B)^{-1}\Delta E \tag{2.12}$$

where B is the normalized connectivity matrix, $B_{n\times n} = AD^{-1}$. A is the adjacency

weighted matrix of the pathway graph and D is the diagonal degree matrix. SPIA then defines a value of accumulated perturbations, $\sum_i Acc_i$, and evaluates the probability of observing more extreme values, P_{pert}.

SPIA constructs a multi-evidence score by combining ORA p-values with P_{pert} [19]. This combined p-value, P_G, is then used as the measure of enrichment of selected annotated pathways from the KEGG database. The multi-evidence approach increases sensitivity and specificity of the model.

Cdist: This PEM by Naderi et al., Causal Disturbance (Cdist) models the DEG into the chains of biomolecular interactions. Cdist initially evaluates the number of walks in the pathway graph. Then, it measures the effect of DEG by removing the DE genes from the graph and recounting the number of remaining walks (Figure 2.4). To conserve the causal relationships and for computational purposes, each pathway was modeled as a directed acyclic graph. To formulate this idea, let A_P denote the adjacency matrix of the DAG for a pathway P with n nodes, the following equations can be used to calculate the number of all walks in the graph:

$$A_P^* = \sum_{j=0}^{n} A_P^j = (I - A_P)^{-1} \tag{2.13}$$

$$Psum(G_P) = \sum_{j,k} [A_P^*]_{jk} \tag{2.14}$$

where $Psum(G_p)$ is the number of total walks in the graph. To evaluate the structural effect of DEG on the pathway, Cdist calculates the fractions of walk that were deleted upon the removal of DE genes.

$$Cdist(P_i, S) = 1 - \frac{Psum(G_{p_i}[V_{p_i}/S])}{Psum(G_{p_i})} \tag{2.15}$$

P-values of Cdist, P_{cdist}, are calculated based on the null hypothesis of obtaining

more extreme Cdist values by using the same number of randomly selected pathway gene. Similar to SPIA, Cdist combines P_{cdist} and ORA p-values into test-statistic, as a representative of the enrichment score.

$$\chi_4^2 \sim -2[ln(P_{Cdist}) + ln(P_{ORA})] \tag{2.16}$$

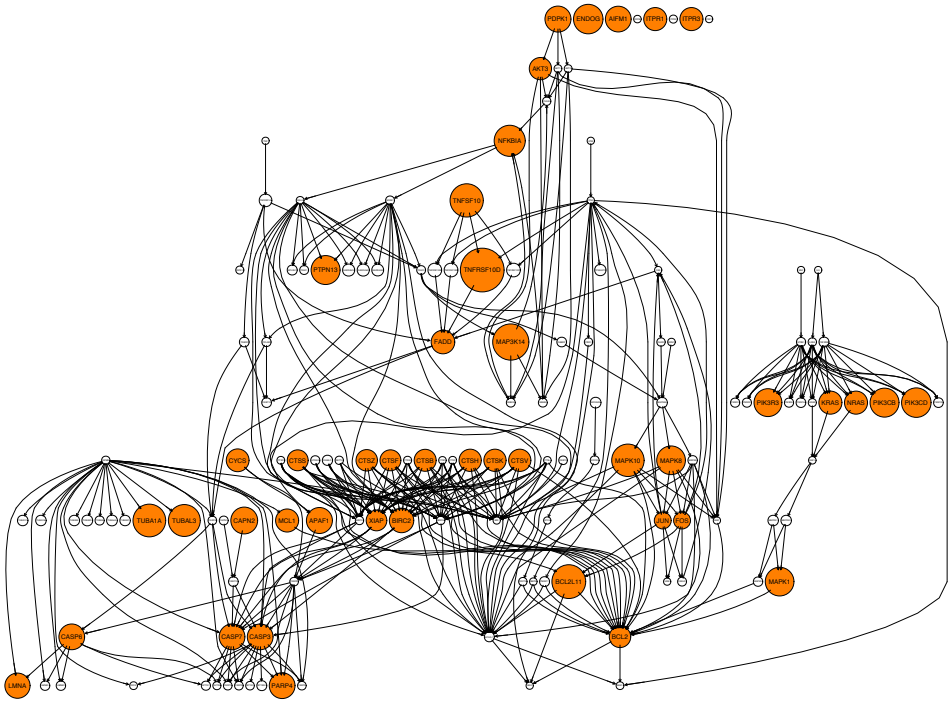

Figure 2.5: Cdist model successfully detects the enrichment of apoptosis pathway in colorectal cancer. Dysregulation of apoptosis in a well-known feature of cancer progression. In comparison, ORA is unable to identify the dysregulation of apoptosis. The results show instances where the structure of the pathways genes (highlighted) are informative in detecting unique pathway enrichments. Figure adopted from Naderi and Mostafavi [21]

The above formula is the combined score which is obtained by a Chi-squared test with 4 degress of freedom. Cdist uses KEGG database as a reference is capable of detecting disease pathways that are not observable using other related methods, such as ORA and SPIA (Figure 2.5) [19, 21].

NetGSA is a multivariate PEM [23] that also leverages the background knowledge of connections between genes and proteins. The main difference between NetGSA and the previously described models is that NetGSA directly models the gene expressions with respect to the underlying pathway network. Formally, let $Y_i = X_i + \epsilon_i$ denote the model of gene i's expressions, where X is signal and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ is noise. Each individual expression is then decomposed into two components of an unknown latent variable and a combined effect of other genes.

$$X_i = \gamma_i + \sum_{j|\ i \in N(j)} X_j \tag{2.17}$$

Where $\gamma_i \sim N(\mu_i, \sigma_\gamma^2)$ is the latent variable of unknown effects and μ_i is the mean expression of gene i without the effects of other genes. In the above formulation, the expression X_i relates to all the other genes, X_j, that have a direct connection to it. For example in Figure 2.4, the expression of node D has an effect from the expression of node A. The set of equations for all X_i's are convergent when certain criteria on the underlying pathway graph are met. Most notably, the pathway network being a Directed Acyclic Graph (DAG). The authors of NetGSA show that under certain conditions on the underlying graph the vector of expressions can be summarized as

$$Y = \Lambda\gamma + \epsilon \tag{2.18}$$

Where Λ is a matrix that its ji-th element is the number of paths between nodes i and j. Equivalently, $\Lambda = (1 - A^T)^{-1}$ where A is the adjacency matrix of the pathway. If the pathway graph is a DAG, then Λ uniquely exists. NetGSA then translates the expression equation into a Mixed Linear Model (MLM). The use of MLM enables NetGSA to handle multiple experimental conditions and the change of topology across conditions.

While network-based are powerful methods, they have a some limitations. First

is that these models work only when sufficient network information is available. In particular, these models are able to analyze the pathways when the map is available. In contrast, ORA and GSA are able to work on a gene set even without substantial information on the interaction.

Second is that their abstraction of pathway organizations do not necessarily capture the topological and functional properties of pathways. For instance, Enrichnet does not consider the direction of interactions. Also in SPIA, the differential expression of terminal pathway genes would have no effect on the *Acc* vector, while in many cases the downstream nodes deliver the definitive function of the pathways. For instance in ErbB signalling, a well-studied cancer pathway, there exist several critical oncogenes and tumor suppressors in both upstream and downstream ends [43, 68] — including ELK, JUN, ERBB2, etc. Likewise, most of the signaling pathways contain critical genes/proteins in their downstream that are key transcription factors of critical cellular functions. NetGSA also fails to capture the importance of nodes at both upstream and downstream of the pathways. Some of the methods require simplification of the underlying graph for pathway analysis. For example, NetGSA and Cdist require DAG modeling of the input data, which necessitates disregarding loops in the pathways. A follow-up modeling of NetGSA by Shojaie and colleagues address the issue of DAGs by re-defining the λ matrix, however, the issue of upstream-downstream organization still persists [24].

Many of the discussed network-based models follow some variations of known and well-studied centrality measures. For example, NetGSA and Cdist definitions are closely related Katz centrality [69]. Similarly, SPIA and EnrichNet are related to the PageRank centrality model [70]. Chapters 3 and 5 will have an extensive discussion on the graph centrality models and will show that how the network-based PEMs corresponds to known graph centrality model.

CHAPTER 3: Causal Disturbance Analysis (CADIA)

3.1 Introduction

The last chapter outlined the shortcomings of the graph methodologies in the existing network-based PEMs. In this chapter, we construct a novel graph-based model, Causal Disturbance Algorithm (CADIA), to effectively capture the topological organization of the pathway networks and produce informative enrichments. To this end, we define a new graph centrality model, Source/Sink Centrality (SSC), which is intended to hold the following characteristics:

- It is sensitive to the direction of interactions, and it conserves the structure and order of interactions. This is intended to address the issues of the PEM that do not consider the causal order of interactions such as Enrichnet [22].
- Differential expression of a gene has a stronger effect on its immediate targets compared to the distant ones. This concept allows for distinguishing between immediate and mediated relationships among the genes.
- Differential expression of a gene/protein relates to its relative position with respect to all of its downstream and upstream targets. This concept intends to address the gap of knowledge in the PEM where the differential expression of downstream nodes are regarded as topologically unimportant.

We derive a topological statistical evidence from the DE genes using SSC. The topological evidence evaluates how much the DE genes are important in the pathway structure. We then combine this topological evidence with a frequency-based evidence to construct an enrichment analysis pipeline with increased sensitivity towards critical differential expressions.

3.2 Source/Sink Centrality

Consistent with the definitions in Chapter 2, our model uses a directed graph representation for biological pathways. Formally, let $G = (V, E)$ represent the graph corresponding to a pathway. The set of vertices (nodes), $V(G) = \{v_1, v_2, \ldots, v_n\}$, represents n distinct gene-encoded elements. The set of edges (links) of a graph, $E(G) = \{e_1, e_2, \ldots, e_m\}$, represents m distinct directed interactions between the nodes, immediate or mediated by some none-gene-encode elements. Each edge, $e_k = (v_i, v_j)$, is an ordered pair that indicates a regulatory or causal relationship from gene-encoded element v_i to v_j.

To calculate a descriptive importance score for a gene/protein in a biological pathway, we introduce a novel graph centrality model that quantifies how the disturbance (perturbation, differential expressions, etc.) of a node can affect the activities of the downstream and upstream targets separately. In our model, the centrality of a node comes from two components. In particular, a node is central if it is either an important downstream receiver of signals (sink), or it is an important sender of signals (source). We define the weighted addition of these two concepts as the overall centrality of a node. Formally:

$$C_{ssc}(v) := C_{source}(v) + \beta C_{sink}(v) \tag{3.1}$$

where β is a positive real parameter to tune the relative contribution of the source and sink components. Next, we define the individual equations for calculating the Source and the Sink Components.

The source component, C_{source}, captures the importance of a node as a sender of signals. In this model, we assume that the change in one node sends signals to other nodes through edges of the graph. We also assume that a signal can travel through any existing route (chains of biochemical interactions) between a sender and a receiver.

The Source centrality captures the chains of biochemical interactions by using the concept of graph *walks*. For consistency, we consider a single node as a walk of length zero. We define the *walk-space* of a graph node, $\mathbf{W}_G(v)$, to be the set of all walks that start from the node, $\mathbf{W}_G(v) := \{w_i \mid w_i$: a vu-walk in G$\}$.

We assume that the signal decays proportionally to the distance that it travels. Consequently, the size of the signal that a hypothetical node X sends to another node Y depends on their distance. We capture the decay of the signal by using some parameter α where $0 < \alpha < 1$. We then derive the Source centrality of a node v by a weighted summation of the existing walks that start from v. Let k denote the length of an arbitrary walk from v. We define the additive contribution of each walk to the Source centrality as α^k. Formally:

$$C_{Source}(v) := \sum_{w_j:\ vu-walk\ of\ G} \alpha^{|w_j|} \tag{3.2}$$

The parameter α ensures three conditions. First, differential expression of a gene/protein has a greater effect on immediate targets compared to indirect ones. Second, the effect of a DE gene is the same on all immediate gene/proteins. Third, under certain conditions of α, the centrality is able to handle the graph loop. We elaborate on the conditions of α shortly after deriving a closed form solution of C_{Source}.

To calculate the Source centrality, we use the adjacency matrix of the graph, A_G. The total number of all ij walks of length k in a graph is the ij-th element of the k^{th} power of the adjacency matrix, $[A_G^k]_{ij}$. A proof of this can be constructed using induction and the properties of matrix multiplication. The total number of all walks of length k that start from a node v_i in the graph is obtained the following formula:

$$\begin{aligned} \sum_{\substack{w_j \in \mathbf{W}_G(v_i), \\ |w_j|=k}} 1 &= \sum_j [A_G^k]_{ij} \\ &= \delta^T(v_i) A_G^k \mathbb{1} \end{aligned} \tag{3.3}$$

The Formula 3.3 denotes the sum of all elements in the i-th row of the adjacency matrix. $\delta(v_i)$ is the Kronecker delta which is a vector of size n where i-th location is 1 and zero elsewhere. $\mathbb{1}$ is an $n \times 1$ column vector of size n with 1s for all elements. Formally:

$$\delta(v_i) = [\underbrace{0\ 0\ \dots\ 0}_{i-1}\ 1\ \underbrace{0\ 0\ \dots\ 0}_{n-i}]^T$$
$$\mathbb{1} = [\underbrace{1\ 1\ \dots\ 1}_{n}]^T$$

For computing C_{source}, we re-arrange the Equation 3.2 as following:

$$\begin{aligned} C_{Source}(v_i) &= \sum_{k=0}^{\infty} \sum_{\substack{w_j \in \mathbf{W}_G(v_i), \\ |w_j|=k}} \alpha^{|w_j|} \\ &= \sum_{k=0}^{\infty} \alpha^k \sum_{\substack{w_j \in \mathbf{W}_G(v_i), \\ |w_j|=k}} 1 \end{aligned} \tag{3.4}$$

The above formulations are summations over all existing walks of any length. In particular, the inner summation are on all the walks with a fixed length. For a closed form solution, we replace the inner sum with Formula 3.3:

$$\begin{aligned} C_{Source}(v_i) &= \sum_{k=0}^{\infty} \alpha^k \delta^T(v_i)[A_G^k]\mathbb{1} \\ &= \delta^T(v_i)\left[\sum_{k=0}^{\infty} \alpha^k A_G^k\right]\mathbb{1} \end{aligned} \tag{3.5}$$

A sufficient condition for the summation to be convergent is $\alpha \leq 1/\lambda_1$ where λ_1 is the largest positive eigenvalue of the adjacency matrix A_G. In particular, the condition $\alpha < 1/\lambda_1$ results in $\lim_{n\to\infty}[\alpha A]^k = 0$ and insures that the matrix $I - \alpha A_G$

is non-singular. A detailed proof on the convergence can be found in [71]. A choice of of $\alpha < 1/\lambda_1$ gives:

$$\begin{aligned}\sum_{k=0}^{\infty} \alpha^k A^k =& (I - \alpha A)^{-1}(I - \alpha A)[A^0 + \alpha A + \alpha^2 A^2 \dots] \\ =& (I - \alpha A)^{-1}\end{aligned} \tag{3.6}$$

It is possible for a finite graph to have an infinite number of walks. A graph has infinite number of walks if it has loops, i.e. non-zero walks that start and end at the same node. The above equation indicates that even if the graph has an infinite number of walks, the Source centrality converges by choosing proper values for α.

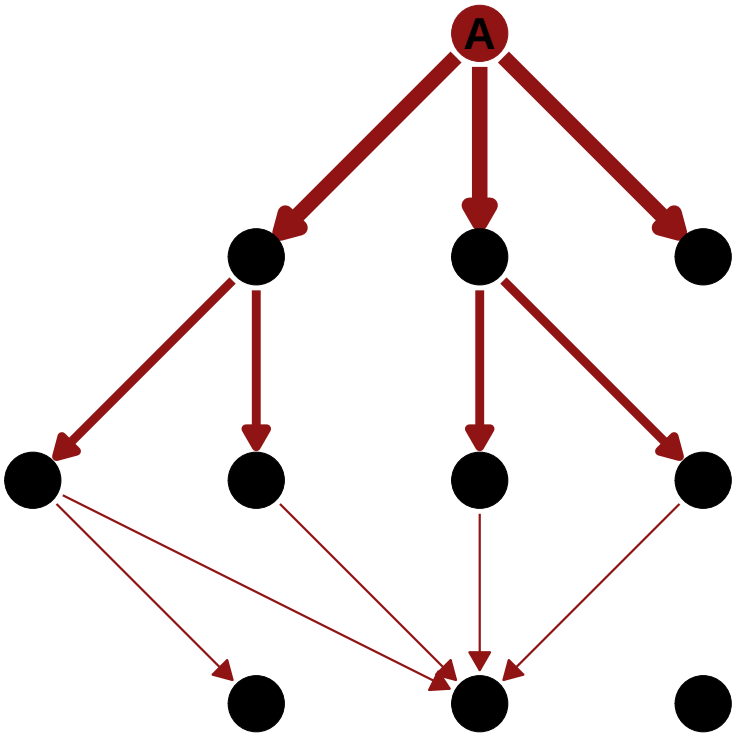

Figure 3.1: A graphical depiction of the Source centrality concept where the importance of the node is calculated A as a source of signals in the graph. The centrality of node A depends on the number of walks that start from it. Incoming red edges in a node denote that A is sending signals to that node. the thickness of the edges denote the strength of those signals, illustrating the relative weight of a walk.

Combining the Equations 3.5 and 3.6, we derive the closed form solution for the Source centrality as:

$$C_{source}(v_i) = \delta^T(v_i)(I - \alpha A_G)^{-1}\mathbb{1} \tag{3.7}$$

The sink component captures the importance of a node as a receiver of signals. In this model, we assume that a node can receive signals from other nodes through connections of the graph. We also assume that a signal can travel through any existing route (chains of biochemical interactions) between a sender and a receiver. We assume that the signal weakens if it has to travel larger distances.

Similar to the Source centrality, we use the concept of graph walks to capture the importance of a node as a sink. Having the assumptions stated above, we derive the Sink centrality of a node v by aggregating the existing weighted walks that end at v. We capture the relative contribution of each incoming walk relative to its length, which is by using some parameter α where $0 < \alpha < 1$. Formally:

$$C_{Sink}(v) := \sum_{w_j:\ uv-walk\ of\ G} \alpha^{|w_j|} \tag{3.8}$$

To calculate the Sink centrality of a node, we use the definition of transposed graphs. The transpose of a graph, G^T, is a graph with reversed edge directions. In this case, $V(G^T) = V(G)$ and $E(G^T) = \{(u, v)|(v, u) \in E(G)\}$. The adjacency matrix of a transposed graph is the transpose of the adjacency matrix, $A_{G^T} = A_G^T$. Any uv-walk of G is a vu-walk of G^T. Consider a walk $w = (v_1, v_2, \ldots, v_k)$ of G, then by definition, we have $(v_{i+1}, v_i) \in E(G^T)$. Therefore, $w' = (v_k, v_{k-1}, \ldots, v_1)$ is a walk in G^T. Using this property, we write Formula 3.8 as:

$$C_{Sink}(v) = \sum_{w_j \in \mathbf{W}_{G^T}(v)} \alpha^{|w_j|} \tag{3.9}$$

The use of the transposed graph allows to effectively calculate the Sink centrality in a procedure similar to the Source centrality. In particular, by rewriting the equations 3.4–3.8, we get:

$$C_{Sink}(v_i) = \delta^T(v_i)(I - \alpha A_G^T)^{-1}\mathbb{1} \tag{3.10}$$

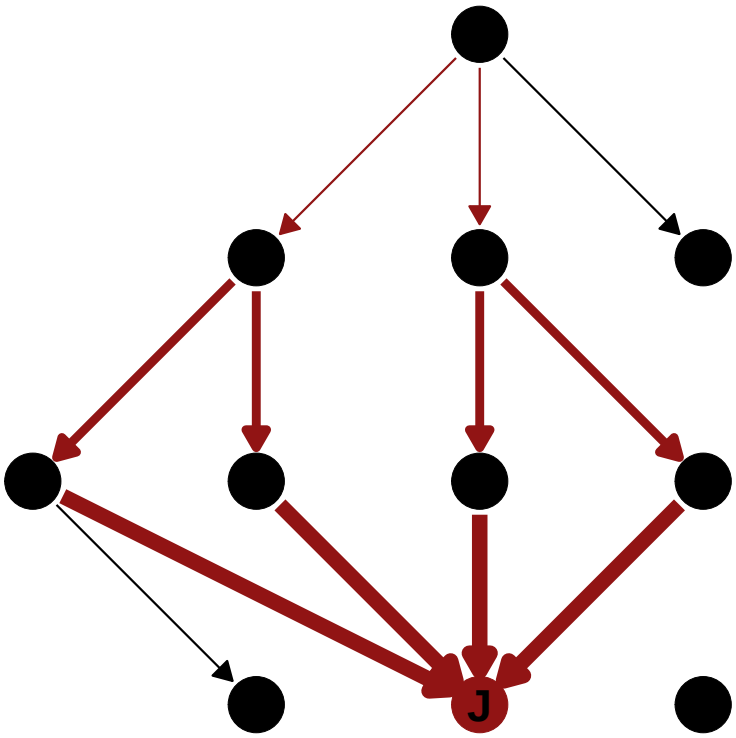

Figure 3.2: A graphical depiction of the Sink centrality concept where the importance of the node is calculated A as a sink of signals in the graph. The centrality of node J depends on the number of walks that end at it. Outgoing red edges in a node denote that it is sending signals to node J. the thickness of the edges denote the strength of those signals, illustrating the relative weight of a walk.

The convergence condition of the Sink component is the same as the Source component. In particular, the same α would also work for the transpose graph because the set of eigenvalues of a matrix and its transpose are equivalent. A proof of this can be constructed by showing that the characteristic polynomials of a matrix and its transposed are equal, $|(\lambda I - A^T)| = |(\lambda I - A)|$ [71].

The Source/Sink Centrality is derived by having the individual formulas for the Source and the Sink components. Plugging the above formulas into Formula 3.1 gives:

$$C_{ssc}(v_i) = \delta^T(v_i)\left[(I - \alpha A_G)^{-1} + \beta(I - \alpha A_G^T)^{-1}\right]\mathbb{1} \tag{3.11}$$

The above formulation captures the Source/Sink centrality (SSC) of a node by considering it as both a sender and a receiver. In pathway annotations, the upstream genes are mainly sender (source) of signals. Likewise, the downstream genes/proteins are mainly receiver of the signals.

When β takes small values ($\beta << 1$), the Source/Sink centrality shifts towards the capacity of the nodes as sources, where the SSC resembles the Source component (Eq. 3.2). When β grows larger, $1 << \beta$, the centrality shifts towards higher the sink capacity, where the SSC resembles the Sink Component (Eq. 3.8).

We can show that $\beta = 1$ ensures that $C_{ssc}(v)$ is the most distinct from the two individual Source and Sink components (Proof in Appendix A). Individual formulas for calculating source and sink centrality are closely related to the Katz-Bonacich centrality which is a popular centrality method in the study of social networks [69].

3.3 Constructing a PEM using Source/Sink Centrality

The next step of CADIA is to use the Source/Sink centrality for calculating an enrichment score for a set of DE genes. CADIA derives a topological evidence by

measuring the aggregated importance of the DE genes. CADIA also uses an additional evidence from ORA and combines it with Source/Sink Centrality to increase sensitivity, Similar to a methodology that was used in SPIA [19].

3.3.1 An Statistical Evidence from Source/Sink Centrality

CADIA derives a topological statistics from the Source/Sink Centrality. In contrast to the regular use of centrality models, where the individual centrality of the nodes is important, CADIA measures the centrality of the set of the DE genes. We define a notion of aggregated centrality for a subset of nodes to quantify this concept. Let $U = \{u_1, u_2, \ldots, u_m\}$ denote the DE genes of a pathway, $U \subset V(G)$. We measure the aggregate centrality, namely causal disturbance, of U by using the following:

$$Agg(U) := \prod_{u_i \in U} C_{ssc}(u_i) \tag{3.12}$$

The product in the above formula allows creating sensitivity towards the cases where the DE genes mainly have intermediate centrality values. The biological networks may contain hubs (nodes with extremely high centrality). A summation-based evaluation would dismiss the mainly-intermediate-centrality cases as non-significant in favor of the cases with few hubs and majority low-centrality. The product-based procedure (Formula 3.14) potentially disregards the instances where the set of DE genes contains only a few hubs, and the rest of the elements are unimportant nodes.

Accordingly, we derive a statistical significance of the aggregate score from a given set of DE genes. In particular, we evaluate the probability of observing a more extreme aggregate score by using a bootstrap sampling approach. Let $Agg(U_0)$ denote an observed causal disturbance of a pathway from m DE genes. The statistical significance of $Agg(U_0)$ is:

$$P_{ssc} = \mathbb{P}\Big\{ Agg(U) > Agg(U_0) \;\Big|\; |U| = |U_0| \Big\} \tag{3.13}$$

P_{ssc} denotes the probability of observing a higher aggregate Source/Sink score (causal disturbance) in a randomly selected subset U of $V(G)$ with size k. CADIA uses the probability density function (PDF) of $Agg(U)$ to extract the P_{ssc} by calculating the right-hand side area under the PDF curve. The PDF is based on sampling large enough different $Agg(U)$ values.

The Formula 3.12 can be re-written by taking logarithm of its right-hand side. Since the objective is to evaluate the extremeness of an observed aggregate score, the logarithm operation does not change P_{ssc} on the condition that $log(C_{ssc}(v_i)) \geq 0$. Formally:

$$\begin{aligned} Agg^*(U) :=& log\big(\prod_{u_i \in U} C_{ssc}(u_i) \big) \\ :=& \sum_{u_i \in U} log(C_{ssc}(u_i)) \end{aligned} \tag{3.14}$$

$$\begin{aligned} P_{ssc} =& \mathbb{P}\Big\{ Agg(U) > Agg(U_0) \;\Big|\; |U| = |U_0| \Big\} \\ =& \mathbb{P}\Big\{ Agg^*(U) > Agg^*(U_0) \;\Big|\; |U| = |U_0| \Big\} \end{aligned} \tag{3.15}$$

A property of P_{ssc} is that it remains invariant under a broad range of manipulations. For example, P_{ssc} is invariant to any positive scaling in the Formula 3.11. This allows for rearranging the definition of $C_{ssc}(\ )$ in a more symmetrical representation, the illustration and proof of this rearrangement is provided in the Appendix A.

3.3.2 Combining Source/Sink Centrality with ORA

To increase the sensitivity, CADIA uses an additional statistical evidence obtained from ORA, which is similar to the approach of SPIA [19]. Using the two evidences enables to investigate differential expressions from two simultaneous perspectives. 1– Is the frequency of the DE genes in a pathway unexpected? 2– Are the differential expressions topologically central to the pathway organization?

We then use the hypergeometric test to calculate the p-values of over-representation analysis (P_{ora}). As discussed in Section 2.3, the over-representation p-value of the pathway is defined as the probability of observing more DE genes in the pathway. Recall the p-value of over-representation, P_{ora}:

$$P_{ora} := \mathbb{P}\Big\{X > m\Big\} \sim Hyper(k, l, m, n) \tag{3.16}$$

where X is the random variable that denotes the number of DE genes in the pathway, m is the number of DE genes in the pathway, k is the total number of DE genes, l is the total size of the pathway, and n is the size of the universe.

P_{ora} and P_{ssc} are independent because given any m, the knowledge of P_{ora} does not add any information regarding P_{ssc}. A formal proof can be constructed by using the definition of Formula 3.13; P_{ssc} is independent from $\mathbb{P}\{X = |U_0|\}$ because the definition of P_{ssc} contains a condition of$\{|U| = |U_0|\}$. Similarly, P_{ssc} is independent from $\mathbb{P}\{X = |U_0| + i\}$ for all values of i. Also, the probabilities $\mathbb{P}\{X = |U_0| + i\}$ are mutually exclusive for all i's. Therefore, P_{ssc} and $\sum_i \mathbb{P}\{X = |U_0|+i\}$ are independent. Here, the summation of probability adds up to P_{ora}.

Given the independence of P_{ora} and P_{ssc}, it is possible to combine them into one test-statistic for producing higher statistical power. Fisher's method for meta-analysis uses Chi-square estimates to combine independent p-values [72]. In particular, let a random variable X indicate the product of P_{ora} and P_{ssc}. The chi-squared test

indicates the probability of observing smaller values for the product. This is similar to the methodology of SPIA for combining topological and ORA evidence [19]. Chi-squared test with four degrees of freedom [72] estimates this p-value as following:

$$\begin{aligned} P_{cadia} &= \mathbb{P}\Big\{X \leq P_{ora}.P_{ssc}\Big\} \\ &= -2[ln(P_{ssc}) + ln(P_{ora})] \sim \chi^2_4 \end{aligned} \tag{3.17}$$

where P_{cadia} denotes the combined probability of the topological evidence (P_{ssc}) and the ORA evidence (P_{ora}). P_{cadia} is the output of the enrichment analysis pipeline and we use it to determine the association of a known pathway with the experimental data. Since a typical input involves numerous pathways, the significance thresholds of P_{cadia} for enrichments is decided by appropriate multiple hypothesis testing criteria, such as $FDR < 0.05$.

From a computational perspective, the time complexity of CADIA is similar to that of SPIA. For a pathway of size n with m rounds of sampling, the time complexity is of $O(n^3 + m \cdot n)$. The n^3 component is for a one-time calculation of Source/Sink Centrality, which includes of a matrix inversion. The $m.n$ component depends on number of sampling rounds of n random DE gene (n is the upper-bound).

CHAPTER 4: Model Evaluation

This chapter focuses on evaluating CADIA for identification of informative pathway enrichments. First, we apply CADIA on real-world datasets and compare its results with the existing PEMs. Second, we evaluate CADIA by using it on synthetically generated list of differentially expressed genes. We hypothesize that if the data is randomly generated, then CADIA should not detect pathway enrichments beyond some margin of error. This is to investigate the false-positive rate of outputs of CADIA.

4.1 Methods of Experimental Data Evaluation

We use three real-world datasets for experimental evaluation of CADIA. For consistency, the three datasets were from mRNA expression microarray datasets, retrieved from the National Center for Biotechnology Information (NCBI) gene expression omnibus [73]. The datasets are cancer gene expression profiles, and the rationale for this choice is because of the abundance and depth of literature on signaling pathways in cancers. This allows to contrast results of CADIA against existing evidence and other methods [18]. We compared CADIA to other PEM including SPIA, ORA, GSA, and Enrichnet.

The first dataset is a microarray sample collection of ovarian tissues by Bowtell and colleague which contained 60 High-grade serous ovarian cancer and 30 Low malignant potential tumors [74]. This data was retrieved using the NCBI accession code GSE12172. The second dataset is a microarray sample collection from colorectal tissues by Mogushi and colleagues, retrieved using the NCBI accession code GSE21510. A subset of 25 normal colon tissues and 19 homogenized cancer tissues from this

dataset was selected for differential expression analysis [75]. The third dataset was from gastric cancer patients by Bing Ya and colleagues that contained 21 normal samples and 111 cancer samples, retrieved using the accession code GSE54129.

For each dataset, the log of RMA normalized mRNA expressions was used to calculate differential expressions. Limma package was used to calculate the significance and log-fold-change of each differential expression [76]. The p-values of each differential expression were subjected to multiple hypothesis testing using the Benjamini-Hochberg False Discovery Rate (FDR) [32]. Each dataset was subjected to a specific log-fold-change (FC) and FDR criteria for gene selection to create differential expression sets from multiple settings and different sizes. In particular, GSE12172 was subjected to the filtering criteria $|FC| < 1$ and $FDR < 0.05$. GSE21510 was subjected to the filtering criteria $|FC| < 1$ and $FDR < 0.005$. GSE54129 was subjected to the filtering criteria $|FC| < 3$ and $FDR < 0.05$.

4.1.1 Background Pathways

All the PEMs investigated this chapter used a prior set of biological pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [9]. SPIA and Enrichnet used internal list of pathway. For CADIA, ORA, and GSA, we used KEGGGraph package in R to parse pathway graphs [77]. All pathways were selected from KEGG classifications of *Environmental Information Processing*, *Cellular Processes*, *Organismal System*, *Human Diseases*, and *Drug Development*.

Some of the pathways potentially had incomplete information which may cause inconsistency in a PEM [25]. Therefore, we excluded incomplete pathway from the analysis to preserve consistency of graph analysis. The exclusion criteria were 1– pathways contained more than 50% abandoned nodes (without any edges), 2– their largest connected component was less than ten nodes, and 3– their edge count was less than 20. A total of 51 pathways exhibited these characteristics. Also, five pathways with the largest eigenvalue of more than 10 were excluded from analysis since they

imposed too small values of α for CADIA. A final set of 143 pathways passed the analysis criteria and were used for further analysis. A complete list of pathways in this study is provided in the Supplementary Table B.1.

4.1.2 Comparison of PEMs

We investigate the enrichment of KEGG pathway in the three datasets by using CADIA, ORA, GSA, Enrichnet, and SPIA. The methods calculate enrichment p-values for each pathway. We used the Benjamini-Hochberg False Discovery Rate (FDR) criteria to correct the p-values for multiple hypothesis testing correction when applicable. A pathway enrichment score was considered statistically significant if its respective FDR-corrected p-value was less than 0.05 ($FDR \leq 0.05$).

P-values of Source/Sink Centrality (P_{ssc}) in CADIA were calculated based on 10000 rounds of iteration for bootstrap sampling which can compute p-values as small as 10^{-4}. The parameters $\alpha = 0.1$ and $\beta = 1$ were used for calculating Source/Sink Centrality in CADIA. The choice of $\beta = 1$ ensures that Source/Sink centrality is maximally distinct from the Source component and the Sink component. As for the parameter α, we are interested in having the largest possible values, while preserving a reasonable coverage of pathways. Source/Sink centrality is closely related to Katz-Bonacich model, and prior studies have shown that the choice of α in Katz model can strongly affect the centrality rankings [69]. For these reasons, $\alpha = 0.1$ is the maximal choice to ensure the pathway coverage and the convergence of Source/Sink calculation (only 5 pathways had to be excluded because they required smaller choice of α)

Over-representation p-values (P_{ora}) were calculated using hypergeometric test. SPIA p-values were calculated using the SPIA R package [19]. GSA p-values were calculated using two available implementations GAGE and F-GSEA [53, 78]. Enrichnet pathway analysis was done by accessing its online portal [22]. All the data analysis in this chapter were performed in R and related Bioconductor packages when possible [79].

4.2 Methods of Synthetic Data Evaluation

We tested CADIA on random inputs of different sizes to verify that the results were not outcomes of false positives. We also tested ORA on the same randomly generated data to contrast P_{ora} and P_{ssc}. Ideally, a PEM should not detect significant enrichments for a randomly selected input. In practice, a test of random data may generate false positive. Therefore, it is desired to measure the rates using a controlled false positive criterion. The synthetic evaluation of CADIA to measure false positive rates is as following:

1. Set $n = 100$.
2. Select a random subset of n genes.
3. Calculate P_{ora}, P_{ssc} and P_{cadia}.
4. Evaluate the number of enriched pathways by each method ($FDR \leq 0.05$).
5. Repeat steps 2, 3, and 4 for 10 times and record the average number of false positives.
6. If $n \leq 5000$, do $n = n + 100$ and go to step 2.

CADIA and ORA parameters and the background data are described in subsection 4.1.2. The 10 repeats at each input size allows a more accurate estimate of the number of false positives. The 100–5000 range provides a variety of reasonable input sizes for measuring CADIA and ORA.

In addition, we applied Source/Sink centrality to the ErbB signaling pathway to showcase its ranking procedure. ErbB signaling is a suitable choice for an in-depth analysis because of 1– the existence of extensive literature on its mechanisms, 2– being a suitable example of upstream/downstream mechanisms [80]. 3– its relative small size for visualization purposes. We compared Source/Sink centrality to three

other well-known centrality models; Degree centrality, Betweenness centrality, and Katz centrality. A comprehensive description of these models and their applications can be found in the reference [65]. To compare these models, we used the ranking of the nodes produced by each centrality method. A higher rank value indicates higher centrality. In the case of having the same values, the minimum rank was assigned to all ties.

4.3 Results and Discussion

The experimental evaluation shows the ability of CADIA its in uniquely detecting critical enrichments. In particular, CADIA detected critical pathway enrichments for ovarian cancer, colorectal cancer, and gastric cancer that were not observable by SPIA, ORA, GSA, and Enrichnet – supported by evidence from the literature. Also, CADIA dismisses some pathway enrichments from SPIA and ORA, many of which do not have any particular association with the experimental data. Our synthetic data evaluation provide insight regarding the performance of CADIA and reliability of its results. The synthetic data evaluation shows that CADIA is not prone to make false positives above the expected level. Additional analysis provides insight regarding the performance of Source/Sink Centrality compared to standard centrality models.

4.3.1 Experimental Evaluation: Ovarian Cancer Dataset

Based on 1333 differentially expressed genes in the ovarian cancer dataset, CADIA uniquely identifies three pathways — *PI3K-AKT signaling, Focal Adhesion*, and *Ras signaling* pathways (Table 4.1). These are well-studied pathways in ovarian cancer [81, 82]. In particular, CADIA detects enrichment of PI3K-AKT signaling (FDR-corrected p-value $\leq 7.82 \times 10^{-3}$) by utilizing a P_{ssc} of $\leq 2.5 \times 10^{-3}$. PI3K-AKT is a cancer associated pathway that regulates many critical cellular mechanisms, including cellular proliferation, survival, and apoptosis [83–86]. PI3K-AKT is activated in ovarian cancer and it has been indicated for its utility for therapeutic approaches

[81, 83, 87],mabuchi2015pi3k,luo2003targeting. Similarly, CADIA detects enrichment of Focal Adhesion pathway (FDR-corrected p-value $\leq 2.41\times10^{-2}$) by utilizing a P_{ssc} of $\leq 9.80\times10^{-3}$. Focal Adhesion is well studied in cancers – particularly ovarian cancer –and is associated with cellular migration, proliferation, and differentiation [88]. In addition, CADIA detects enrichment of Ras signaling pathway (FDR-corrected p-value $\leq 2.20\times10^{-2}$) by utilizing a P_{ssc} of $\leq 2.00\times10^{-4}$. Ras signaling activates cellular proliferation and growth and is associated cancers [83, 89].

CADIA discards some pathways with insignificant Source/Sink topological evidence, some of which not having clear connections to ovarian cancer. For example, SPIA detects *cytokine-cytokine receptor interactions* pathway which is not detected by ORA or CADIA. Our literature search failed to identify any established results for the association of this pathway with ovarian cancer. On the other hand, CADIA is able to provide strong topological evidence for *Pathways in cancer* for which SPIA fails to provide topological evidence (Details in Supplementary Tables in Appendix B). Also, SPIA uniquely detects *mineral absorption* pathway for which the literature search failed to identify any established results for its association ovarian cancer. *Mineral absorption* was among the pathways that did not pass the quality criteria and did not qualify for CADIA because of its incomplete information.

Enrichnet fails to identify a number of pathways that were determined by ORA,SPIA, and CADIA. The significance threshold of Enrichnet's XD-Score for ovarian cancer data was 1.12 (Table 4.4). Also, Enrichnet fails to infer unique relevant pathway enrichments. For example, the literature search failed to find evidence for association of *Folate Biosynthesis* with ovarian cancer. For these reasons, we conclude that in this case of experimental evaluation Enrichnet does not perform better than any of the other methodologies.

On the ovarian cancer data, both GAGE and FGSEA identify a few of the pathways that were also discovered by SPIA, CADIA, and ORA (Tables 4.5 and 4.6). However,

Table 4.1: Statistically significant pathway enrichments identified by CADIA from the ovarian cancer sata (GSE12172)

Name[§]	ID	P_{ora}	P_{ssc}	**CADIA**[†]	FDR_{ora}[‡]
MicroR...	05206	3.66e-08	2.65e-01	2.70e-05	5.23e-06
Oocyte ...	04114	3.13e-04	3.00e-04	1.09e-04	1.49e-02
p53 sig...	04115	2.83e-07	4.80e-01	1.09e-04	2.02e-05
***PI3K-...**	04151	7.34e-03	2.50e-03	7.82e-03	8.31e-02
***Ras si...**	04014	3.65e-01	2.00e-04	2.20e-02	9.97e-01
***Focal...**	04510	1.01e-02	9.80e-03	2.41e-02	9.02e-02
Proges...	04914	4.82e-04	3.51e-01	3.19e-02	1.72e-02
Pathwa...	05200	2.30e-03	8.08e-02	3.19e-02	4.71e-02

[§] Names truncated for space limitation
[†] FDR corrected P_{cadia}
[‡] FDR corrected P_{ora}
[*] Unique to CADIA

GSEA requires less conservative significance thresholds. At the significance threshold of 0.05 GAGE finds only two pathways and FGSEA finds three pathways. There is no overlap between the result of the two methods. In contrast, the models investigated in the main document, discover larger number of pathways with many relevant cases.

4.3.2 Experimental Evaluation: Colorectal Cancer Dataset

Based on 2625 differentially expressed genes, CADIA uniquely detects six pathway enrichments in colorectal cancer data including *Apoptosis*, *Hippo Signaling* (Table 4.7). CADIA detects enrichment of Apoptosis pathway (FDR-corrected p-value $\leq 3.77 \times 10^{-2}$) by utilizing a P_{ssc} of $\leq 8.00 \times 10^{-3}$. Dysfunction of Apoptosis pathway – programmed cell death– is an important feature of cancers, and in particular colorectal cancer [18, 90]. Similarly, CADIA detects enrichment of *Hippo signaling* pathway (FDR-corrected p-value $\leq 3.77 \times 10^{-2}$) by utilizing a P_{ssc} of $\leq 5.64 \times 10^{-2}$. Hippo signaling is well-studied in human neoplasms and control cellular proliferation and apoptosis [91]. CADIA detects enrichment of *GnRH signaling* pathway (FDR-corrected p-value $\leq 3.77 \times 10^{-2}$) by utilizing a P_{ssc} of $\leq 1.81 \times 10^{-2}$. Literature evidence also show mechanisms in which GnRH signaling affects colorectal cancer [92]. Similarly,

Table 4.2: Statistically significant pathway enrichments identified by ORA from the ovarian cancer data (GSE12172)

Name	ID	CADIA†	**FDR$_{ora}$**‡
MicroRNAs in cancer	05206	2.70e-05	5.23e-06
p53 signaling pathway	04115	1.09e-04	2.02e-05
Oocyte meiosis	04114	1.09e-04	1.49e-02
Progesterone-mediated oocyt...	04914	3.19e-02	1.72e-02
***Proteoglycans in cancer**	05205	7.59e-02	3.29e-02
ECM-receptor interaction	04512	6.85e-02	4.63e-02
Pathways in cancer	05200	3.19e-02	4.71e-02

† FDR corrected P_{cadia}
‡ FDR corrected P_{ora}
* Unique to ORA

Table 4.3: Statistically significant pathway enrichments identified by SPIA from the ovarian cancer data (GSE12172)

Name	ID	**SPIA**†	CADIA†
***Cell cycle**	04110	6.38e-09	NA
p53 signaling pathway	04115	1.11e-04	1.09e-04
***Chemokine signaling pathway**	04062	4.85e-04	3.17e-01
***Mineral absorption**	04978	1.48e-02	NA
Oocyte meiosis	04114	1.73e-02	1.09e-04
***Cytokine-cytokine receptor...**	04060	1.73e-02	4.76e-01
Progesterone-mediated oocyt...	04914	3.77e-02	3.19e-02

† FDR corrected p-values
* Unique to SPIA
NA: Not Analyzed by CADIA

CADIA detects enrichment of *Phospholipase D signaling* pathway (fdr-corrected p-value $\leq 3.77 \times 10^{-2}$) by utilizing a P_{ssc} of $\leq 7.00 \times 10^{-3}$. Phospholipase D signaling is related to colorectal cancer through connections with Wnt signaling [93, 94].

The significance threshold of Enrichnet's XD-Score for colorectal cancer data were 1.78 (4.10). According to the authors and the portal, this score is equivalent of a $FDR < 0.05$. The over-representation p-values in the table were calculated by Enrichnet application. Enrichnet fails to identify a number of pathways that were determined by ORA,SPIA, and CADIA. Also, Enrichnet fails to infer unique relevant pathway enrichments. For these reasons, we conclude that in this case of experimental

Table 4.4: Statistically significant pathway enrichments identified by Enrichnet from the ovarian cancer data (GSE12172)

Name	ID	**XD.Score**†	FDR_{ora}*
Folate biosynthesis	00790	1.90	2.86e-01
DNA replication	03030	1.70	1.33e-02
Cell cycle	04110	1.53	6.83e-11
p53 signaling pathway	04115	1.53	2.42e-05
Bladder cancer	05219	1.25	4.82e-02
Prion diseases	05020	1.25	7.40e-02

† Significance score of Enrichnet (Threshold =1.12 for 95%)

* FDR corrected P_{ora} – as calculated in Enrichnet

Table 4.5: Statistically significant pathway enrichments identified by GSA–GAGE from the ovarian cancer data (GSE12172)

Name	ID	p.val†	FDR*
MicroRNAs in cancer	05206	8.86e-06	1.22e-03
p53 signaling pathway	04115	1.71e-05	1.22e-03
Oocyte meiosis	04114	1.19e-03	5.65e-02

* FDR corrected p-values – as calculated in GAGE

evaluation Enrichnet does not perform better than any of the other methodologies.

ORA uniquely detects enrichment of Thyroid hormone signaling pathway for which the literature search did not find any results in supports of its association with colorectral cancer. Similarly, SPIA detects pathways that not necessarily related to colorectal cancer such as Alzheimer's, Ameobiasis, Bile secretion, and Pancreatic cancer (Table 4.9). Some of these SPIA pathways were excluded from CADIA's analysis because of incomplete information (NA entries in Table 4.9). The other unique SPIA pathways were analyzed by CADIA but it did not find strong topological evidence (More detail in the Supplementary Tables B.2, B.3, and B.4). For example, SPIA detects a strong topological evidence for *Alzheimer's disease* pathway (pPERT $= 5 \times 10^{-6}$, Supplementary Table B.3) which is not necessarily related to colorectal cancer. "Alzheimer's" was among the pathways that were excluded from CADIA. The detection of Alzheimer's is an instance where the incomplete information causes irrel-

Table 4.6: Statistically significant pathway enrichments identified by GSA–FGSEA from the ovarian cancer data (GSE12172)

Name	ID	p.val[†]	FDR[*]
Oocyte meiosis	04114	5.96e-04	4.94e-02
Toll-like receptor signaling pathway	04620	3.19e-03	9.13e-02
Chemokine signaling pathway	04062	6.92e-04	4.94e-02
Progesterone-mediated oocyte maturation	04914	2.53e-03	9.13e-02
Influenza A	05164	2.66e-03	9.13e-02

[*] FDR corrected p-value – as calculated in FGSEA

evant outcomes for Network-based PEM by producing strong topological evidences.

In the colorectal cancer data, GAGE identifies two pathways at the significance threshold of 0.1 (4.11). The pathways were also discovered by SPIA, CADIA, and ORA. On the other hand, FGSEA finds over 60 pathways with the significance threshold of 0.05 (4.12). In this case, FGSEA identifies the enrichment of more than 30% of the pathways.

4.3.3 Experimental Evaluation: Gastric Cancer Dataset

Based on 133 differentially expressed genes, CADIA uniquely detects Wnt Signaling pathway in gastric cancer (Table 4.13). In the case of gastric cancer, CADIA detects Wnt signaling (FDR-corrected p-value $\leq 9.38 \times 10^{-3}$) by utilizing a P_{ssc} of $\leq 1.00 \times 10^{-4}$. Wnt signaling is among the most well-studied cancer pathways, and there is a plethora of evidence for its activation in cancers including gastric [95]. This case indicates that the DE genes of the Wnt pathway are substantially important in the structure and makes the case of why a structural pathway analysis can detect unique discoveries. Compared to CADIA, ORA detects enrichment of Renin Secretion and Vascular muscle contractions, for which the literature suggests no particular relevance to the disease. Similarly, SPIA detects Ameobiasis and Malaria. The literature search failed to identify any established results for the association of these pathway with gastric cancer. These pathways were among the list that did not pass the quality criteria and did not qualify for CADIA because of incomplete information.

Table 4.7: Statistically significant pathway enrichments identified by CADIA from the colorectal cancer data (GSE21510)

Name[§]	ID	P_{ora}	P_{ssc}	**CADIA**[†]	FDR_{ora}[‡]
Oocyt...	04114	6.02e-05	1.10e-03	8.30e-05	1.72e-03
p53 s...	04115	9.48e-08	4.34e-01	8.30e-05	1.36e-05
Pathw...	05200	1.17e-06	9.45e-01	7.77e-04	8.39e-05
Micro...	05206	5.09e-06	4.22e-01	1.08e-03	2.43e-04
PPAR ...	03320	1.09e-05	3.24e-01	1.37e-03	3.89e-04
HTLV-...	05166	1.31e-04	4.95e-01	1.64e-02	3.11e-03
Proge...	04914	8.65e-04	1.20e-01	1.88e-02	1.55e-02
***Olfa...**	04740	9.97e-01	1.00e-04	1.88e-02	9.97e-01
***Hipp...**	04390	6.58e-03	5.64e-02	3.77e-02	6.27e-02
***Phos...**	04072	6.01e-02	7.00e-03	3.77e-02	1.95e-01
***Apop...**	04210	4.56e-02	8.00e-03	3.77e-02	1.64e-01
Chemo...	04062	2.89e-03	1.12e-01	3.77e-02	3.47e-02
***GnRH...**	04912	2.03e-02	1.81e-02	3.77e-02	1.02e-01
***Vasc...**	04270	2.15e-02	2.12e-02	3.77e-02	1.02e-01
Small...	05222	7.35e-04	5.88e-01	3.77e-02	1.50e-02
***Calc...**	04020	2.51e-02	2.50e-02	4.70e-02	1.16e-01

[§] Names truncated for space limitations
[†] FDR corrected P_{cadia}
[‡] FDR corrected P_{ora}
[*] Unique to CADIA

The significance threshold of Enrichnet's XD-Score for colorectal cancer data were 0.72 (4.16). According to the authors and the portal, this score is equivalent of a $FDR < 0.05$. The over-representation p-values in the table were calculated by Enrichnet application. Enrichnet fails to identify a number of pathways that were determined by ORA,SPIA, and CADIA. Also, Enrichnet fails to infer unique relevant pathway enrichments. For these reasons, we conclude that in this case of experimental evaluation Enrichnet does not perform better than any of the other methodologies.

On the gastric cancer data, GAGE identifies 78 pathways and FGSEA identifies 91 pathways to be enriched (Tables 4.17 and 4.18). These values are more than half of the annotated pathways that were used.

The results show that GSA is either not discovering any significant pathways at the specified thresholds or it discovers numerous pathways (as much as 91 out of

Table 4.8: Statistically significant pathway enrichments identified by ORA from the colorectal cancer data (GSE21510)

Name[§]	ID	CADIA[†]	**FDR$_{ora}$**[‡]
p53 signaling pathway	04115	8.30e-05	1.36e-05
Pathways in cancer	05200	7.77e-04	8.39e-05
MicroRNAs in cancer	05206	1.08e-03	2.43e-04
PPAR signaling pathway	03320	1.37e-03	3.89e-04
Oocyte meiosis	04114	8.30e-05	1.72e-03
HTLV-I infection	05166	1.64e-02	3.11e-03
Small cell lung cancer	05222	3.77e-02	1.50e-02
Progesterone-mediated oocyte...	04914	1.88e-02	1.55e-02
Chemical carcinogenesis	05204	5.92e-02	2.95e-02
***TGF-beta signaling pathway**	04350	8.87e-02	3.02e-02
Chemokine signaling pathway	04062	3.77e-02	3.47e-02
***Proteoglycans in cancer**	05205	5.92e-02	3.47e-02
***Thyroid hormone signaling...**	04919	1.37e-01	4.73e-02

[§] Names truncated for space limitations
[†] FDR corrected P_{cadia}
[‡] FDR corrected P_{ora}
[*] Unique to ORA

143). Also, Enrichnet fails to discover multiple critical pathways that are discovered by CADIA, SPIA, and ORA.

4.3.4 Synthetic Data Evaluation

Figure 4.1 shows that the average false positive rates of the topological evidence (FDR-corrected P_{ssc}) is zero. When using ORA alone, the average false positive rate at some cases is not zero, but is below the FDR = 0.05 threshold. Similarly, the combined evidence (FDR-corrected P_{cadia}) produces small averages of false positive rates. Figure 4.1 shows that the controlled false positive rate of CADIA is consistent across a wide range of random DE genes input size (100–5000). These results indicate the specificity of CADIA, and ensure that the experimental data inferences are not results of false positive.

Additional synthetic data evaluation shows a uniform null distribution of P_{ssc} for the random DE genes sets. The uniform distribution of P_{ssc} shows that this topolog-

Table 4.9: Statistically significant pathway enrichments identified by SPIA from the colorectal cancer data (GSE21510)

Name	ID	**SPIA**†	CADIA†
***Cell cycle**	04110	4.86e-16	NA
p53 signaling pathway	04115	1.71e-05	8.30e-05
***RNA transport**	03013	1.27e-04	NA
PPAR signaling pathway	03320	3.54e-04	1.37e-03
***Mineral absorption**	04978	3.54e-04	NA
***Alzheimer's disease**	05010	9.23e-04	NA
HTLV-I infection	05166	1.46e-03	1.64e-02
***Amoebiasis**	05146	6.19e-03	NA
Oocyte meiosis	04114	7.68e-03	8.30e-05
***Bile secretion**	04976	9.12e-03	NA
Pathways in cancer	05200	9.12e-03	7.77e-04
***ECM-receptor interaction**	04512	1.21e-02	1.23e-01
Progesterone-mediated oocy...	04914	1.66e-02	1.88e-02
Small cell lung cancer	05222	1.91e-02	3.77e-02
Chemokine signaling pathway	04062	2.15e-02	3.77e-02
***Gap junction**	04540	2.76e-02	7.66e-02
***Transcriptional misregulat...**	05202	2.87e-02	NA
***Wnt signaling pathway**	04310	3.02e-02	8.87e-02
***Pancreatic secretion**	04972	4.99e-02	NA

† FDR corrected p-values
* Unique to SPIA
NA: Not Analyzed by CADIA

ical evidence is not biased towards making false-positives or false-negatives (Figure 4.3). In this figure, the large density at $P_{ssc} = 1$ is due to the large number of pathways that did not have enough DE genes for the topological analysis. P_{ssc} was calculated only if a pathway had two or more DE genes, and was reported as one otherwise.

Figure 4.2 displays the histogram of P_{ora} on the random DE genes. In this case, the density of the P_{ora} is higher for larger p-values, which is potentially because the hypergeometric test is conservative [13, 63, 64]. In this figure, the large density at $P_{ora} = 1$ is due to the large number of pathways that did not have any DE genes for the analysis (zero). Figure 4.4 displays the histogram of P_{cadia} on random DE genes. In this case, P_{cadia} values are close to a uniform distribution, which indicates that

Table 4.10: Statistically significant pathway enrichments identified by Enrichnet from the colorectal cancer data (GSE21510)

Name	ID	**XD.Score**†	FDR_{ora}*
DNA replication	03030	3.70	2.55e-08
Fatty acid metabolism	00071	2.59	3.90e-04
One carbon pool by folate	00670	2.34	1.33e-01
Cell cycle	04110	1.91	1.70e-11
Sulfur metabolism	00920	1.91	2.57e-01
Base excision repair	03410	1.83	1.17e-02

† Significance score of Enrichnet (Threshold =1.78 for 95%)
* FDR corrected P_{ora} – as calculated in Enrichnet

Table 4.11: Statistically significant pathway enrichments identified by GSA–GAGE from the colorectal cancer data ($FDR < 0.1$)(GSE21510)

Name	ID	p.val†	FDR*
Pathways in cancer	05200	2.49e-04	3.56e-02
p53 signaling pathway	04115	8.49e-04	6.07e-02

* FDR corrected p-values – as calculated in GAGE

CADIA is not biased towards making false rejection of the null hypothesis.

The synthetic data evaluation does not find any correlation (correlation estimate = 0.005 and p-value = 0.2) between P_{ssc} and P_{ora} (Figures 4.5). This figure displays the relationship of ORA and SSC p-values (Not FDR Corrected). Each point in the Figure 4.5 shows the two p-values for a pathway that had DE genes. The plot shows no linear relationship between the two p-values, as the test for correlation coefficient fails to reject the null hypothesis of no relationship.

Figure 4.6 displays the bootstrap (5×10^5 rounds) sampling for the aggregate scores (Formula 3.14) of the 31 DE genes (nodes) from focal adhesion (from Table 4.1). The pattern of normal distribution, in this case, is explainable by the central limit theorem. The aggregate score estimates a multiply of the mean of log-centrality values (Formula 3.14). With sufficiently large DE genes set, the normal distribution can replace the empirical estimation of P_{ssc}. Also, the random aggregate scores of $C(v_i)$s are independent and identically distributed (iid), having the necessary conditions for

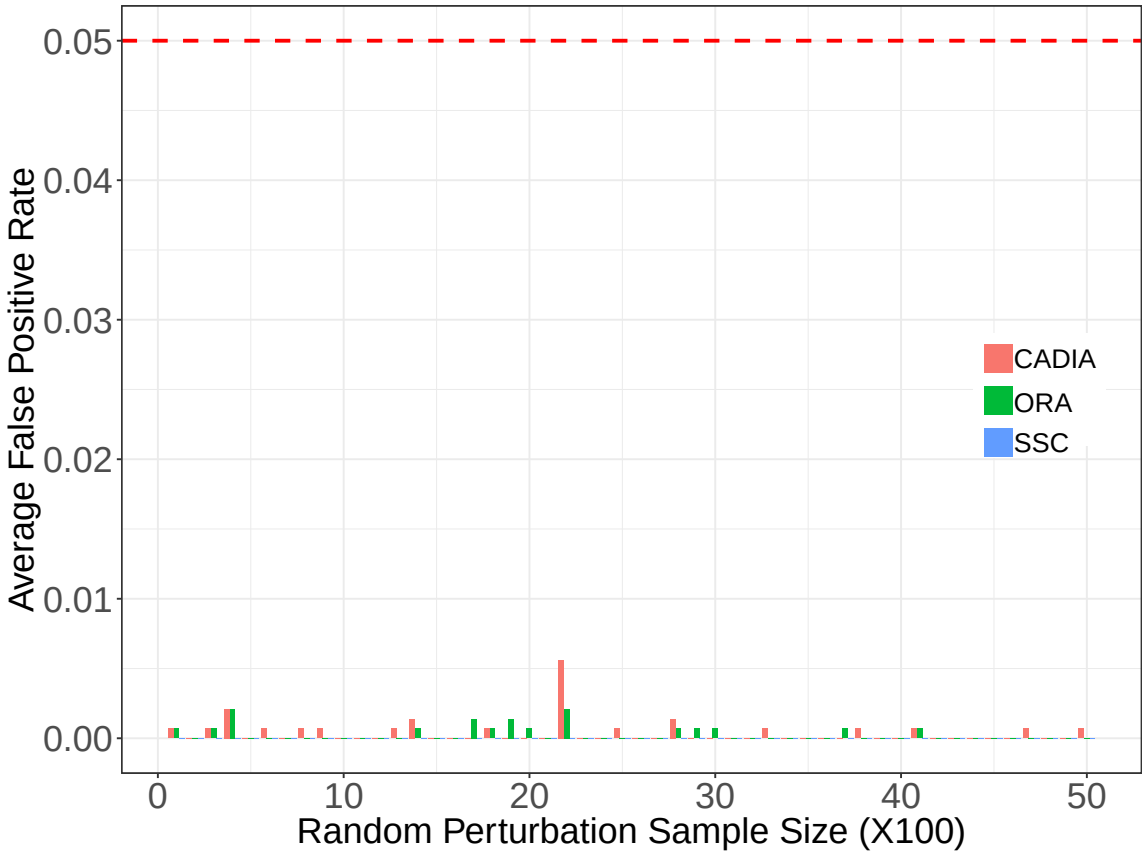

Figure 4.1: The number of false positives of ORA, SSC, and CADIA for different sizes of the randomly sampled DE genes (10 repeats). The Y-axis is the average number of false positives from a $FDR \leq 0.05$ threshold. The red dashed-line is the FDR control threshold. Figure adopted from Naderi and Mostafavi [26]

the central limit theorem. This normal distribution can lead to another route for showing that P_{ssc} and P_{ora} are independent. If P_{ssc} follows a normal distribution, based on mean $C(v_i)$s, then it is independent of the outcomes of the hypergeometric distribution in Formula 3.16.

Figure 4.7, illustrates an example of the ability of Source/Sink Centrality in attributing importance to both upstream and downstream nodes. In the case of the ErbB pathway, the signal receptors associated genes EGFR and ERBB2 are critical sources that initialize activities (upstream), while the genes MYC, JUN, and ELK are critical endpoint receivers (downstream) [80]. Figure 4.7 shows the relative importance scores of Source/Sink centrality for each gene in the ErbB pathway. Genes

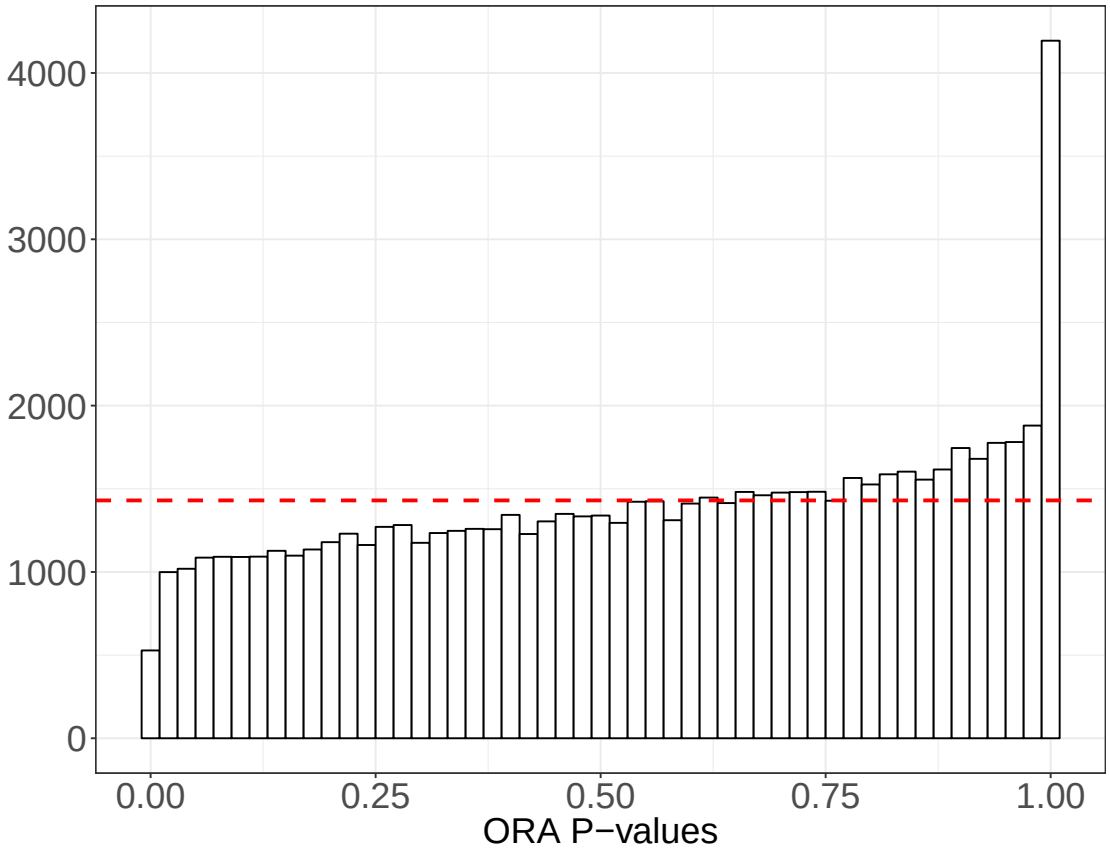

Figure 4.2: Synthetic data evaluation of ORA p-values. The plot shows the histogram of the p-values of pathway enrichments based on randomly selected DE genes. The red line shows the bar for a uniform distribution. Figure adopted from Naderi and Mostafavi, Supplementary Material [26].

at the upstream the pathway, including EGFR,ERBB1, and ERBB2, are recognized by Source/Sink as high centrality (Table 4.19). Also, Source/Sink centrality distinguished between the downstream nodes such as MYC, JUN, and ELK1. Other standard centrality measures assign low importance to terminal nodes of pathways. For example, ELK1, BAD, PTK2, MYC, and JUN would have the same centrality score regardless of their underlying biological functions and topological position in the graph (Figure 4.7 and Table 4.19).

A general centrality measure may fail to capture the downstream importance and assign low centrality values. This observation extends to the definition of topological importance in other network-based PEM. In SPIA for example, the downstream nodes

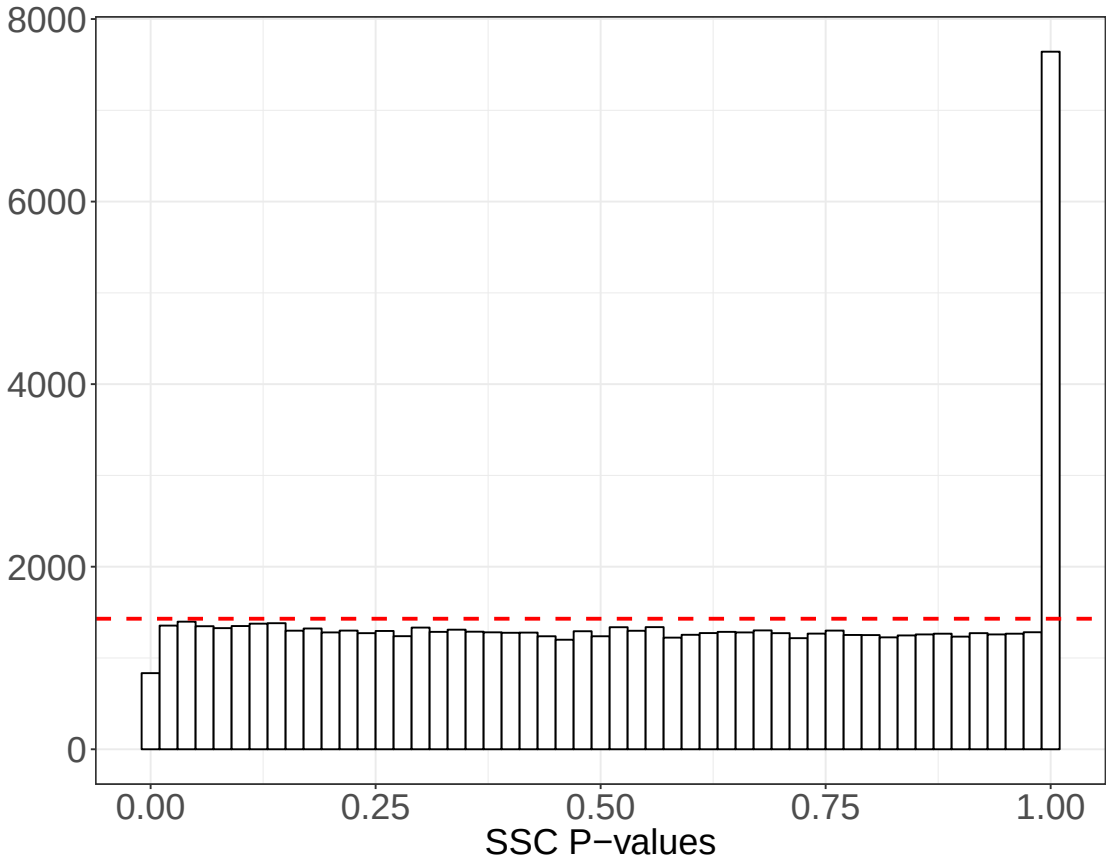

Figure 4.3: Synthetic data evaluation of SSC p-values. The plot shows the histogram of the p-values of pathway enrichments based on randomly selected DE genes. $P_{ssc} = 1$ for zero or one DE genes. The rest of the p-values follow a pattern close to uniform distribution. The red line shows the bar for a uniform distribution.Figure adopted from Naderi and Mostafavi, Supplementary Material [26].

will have the lowest importance because they have zero (or low) out-degree. A possible alternative solution is to sacrifice the network directions, like in that of Enrichnet [22]. The undirected graph approach will potentially deliver incomplete results because the topological features of the graph rely on the directions of the nodes. Evident by our experimental validation, addressing the issues with common centrality models in CADIA enables to detect unique pathway enrichments while delivering consistent results with a low false positive rate.

The three presented experimental test cases indicate that the use of Source/Sink centrality in CADIA enables detection of critical pathway enrichment from biolog-

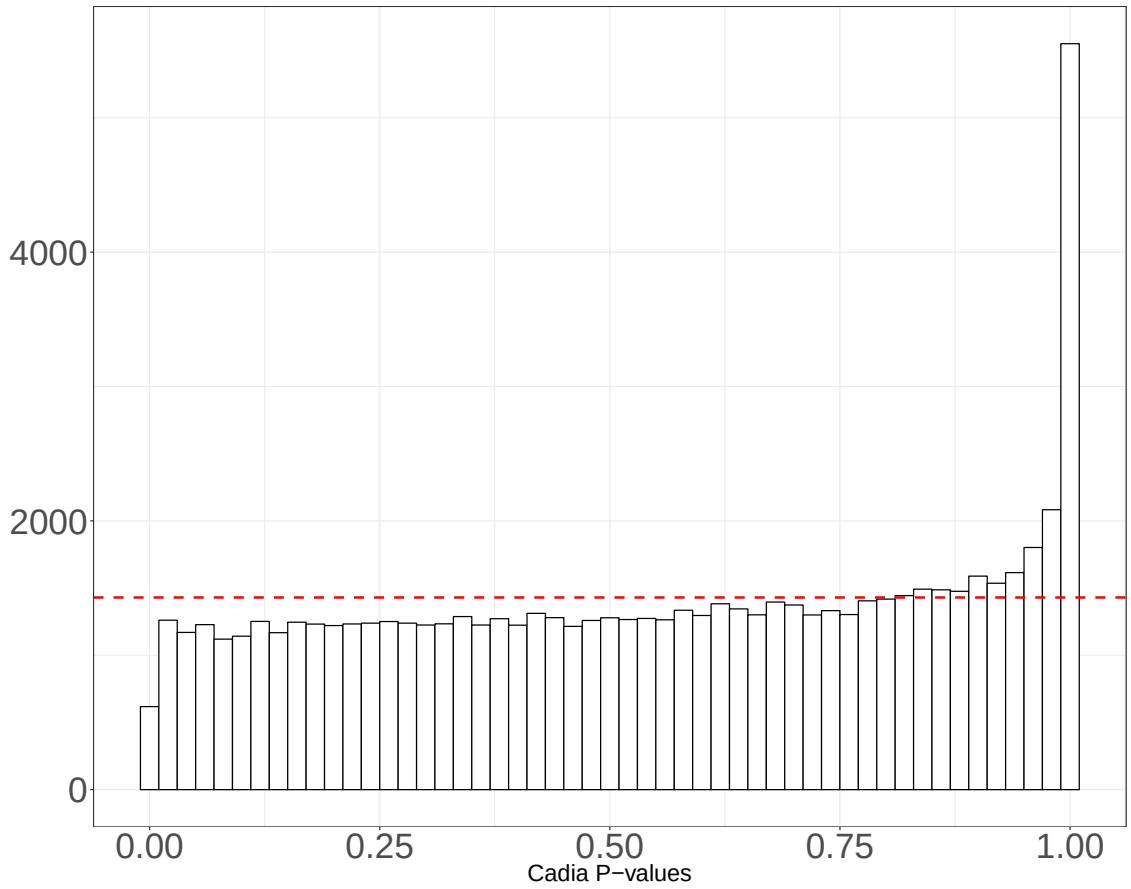

Figure 4.4: Synthetic data evaluation of CADIA p-values. The plot shows the histogram of the p-values of pathway enrichments based on randomly selected DE genes. $P_{cadia} = 1$ for zero or one DE gene. The rest of the p-values follow a pattern close to uniform distribution. The red line shows the bar for a uniform distribution.

ical data. Source/Sink centrality allows for attributing higher importance to the nodes that are missed by other network-based methods such as SPIA. Small p-values of Source/Sink centrality evidence indicates that CADIA is sensitive to differential expression of topologically central genes that are also important to a pathway's functionality. Although small p-values do not guarantee the dysfunction of any pathway, the support of literature for the experimental data shows the ability of CADIA in making an informative enrichments. The variety of differential expression set sizes in the experimental evaluation indicates the sensitivity of CADIA towards both small and large sets of DE genes. CADIA only requires a list of differentially expressed

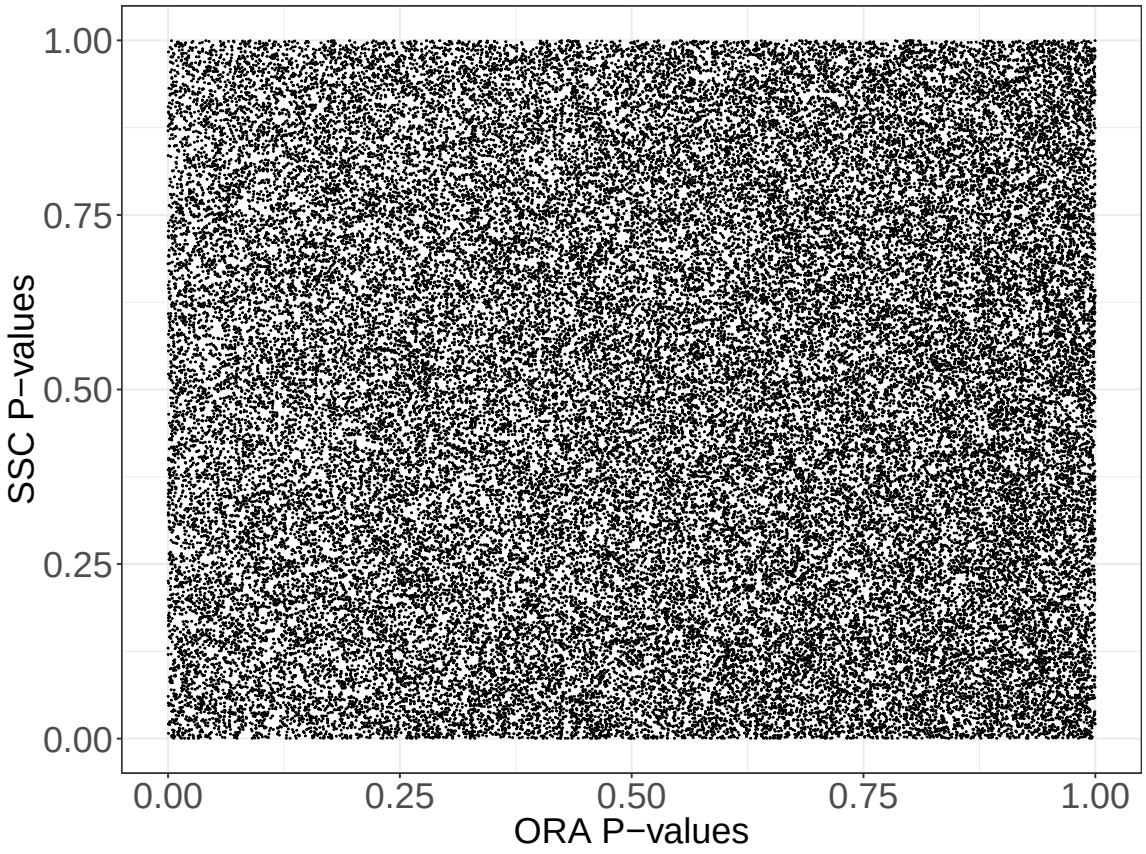

Figure 4.5: Synthetic data comparison of non-1 SSC p-values and ORA p-values. Each point denotes the enrichment p-values for a pathway with P_{ora} on X-axis and P_{ssc} on the Y-axis. The test for Pearson's correlation fails to find any linear relationship between the two values (correlation estimate = 0.005, p-value = 0.2). Figure adopted from Naderi and Mostafavi, Supplementary Material [26].

genes and a set of background pathways to produce the enrichment p-values (P_{cadia}). After the selection of differentially expressed genes, the method does not depend on a ranked list of genes nor their fold changes. Also, because of less limitations in pre-processing step, CADIA has a larger coverage in pathway analysis and is able to infer the enrichment of several critical pathways that are not included in SPIA analysis, such as "Ras Signaling" and "PI3K-Akt Signaling".

The exclusion of incomplete pathways in CADIA allows avoiding detecting inaccurate enrichments. With incomplete information, differential expression of any node with a non-minimal centrality score would produce small p-values for enrichment, and

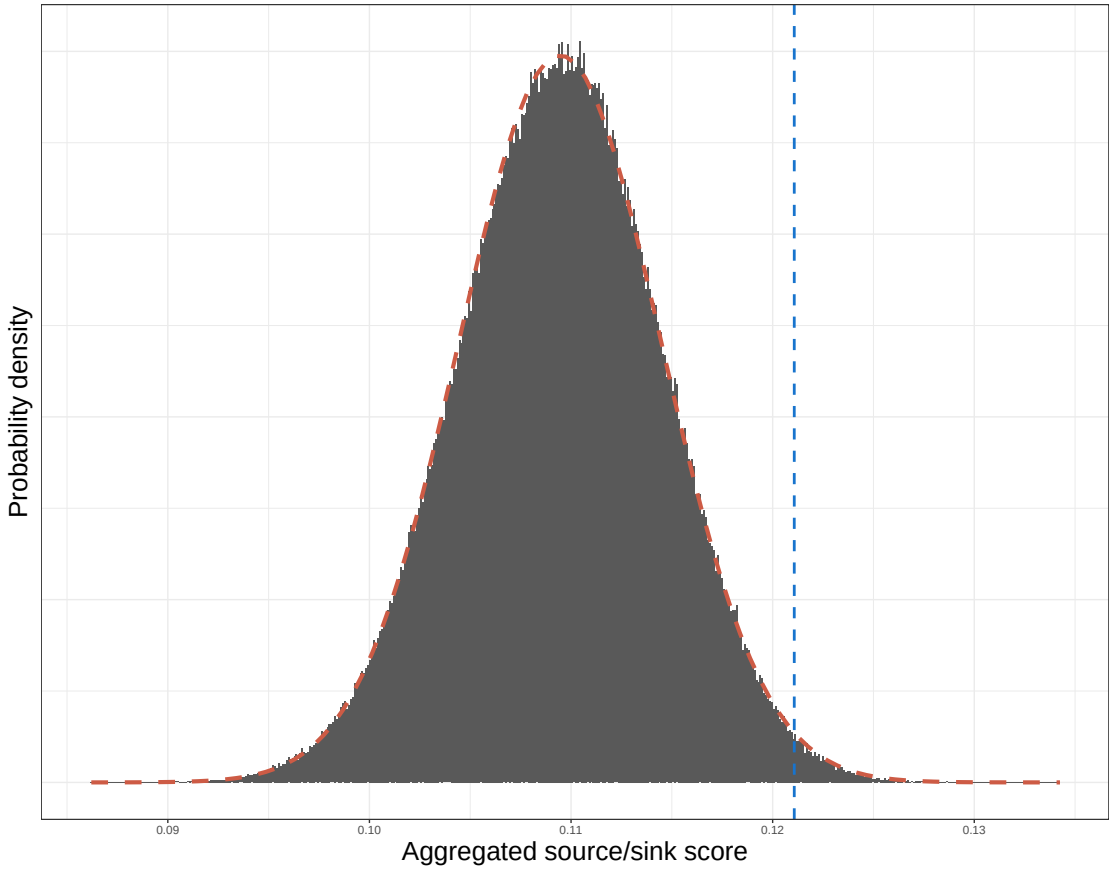

Figure 4.6: Null distribution of aggregate centrality score for calculating P_{ssc}. The figure is generated based on 31 DE genes in Focal adhesion pathway from ovarian cancer data (Table 4.1). The X-axis denotes the aggregate centrality score from Formula 3.14. The red dashed line indicates the normal distribution fit based on the observed mean and standard deviation of the null aggregate scores. The blue line the is experimental observation and its right-hand side area under the curve is P_{ssc}. Figure adopted from Naderi and Mostafavi [26].

subsequently, produce false positive. Network-based enrichment analyses are prone to producing incorrect inferences when the pathway information [25]. The results produced by SPIA show instances where network-based PEM are prone to make irrelevant inferences. Although we took a filtering approach to disregard incomplete pathways (See Supplementary Table in Appendix B), using predicted interactions could benefit CADIA's enrichment analysis in future developments. Readers interested in more details on the pathway processing and files may refer to Naderi and Mostafavi [21] and its supplementary codes, online at "https://github.com/pouryany/CADIA".

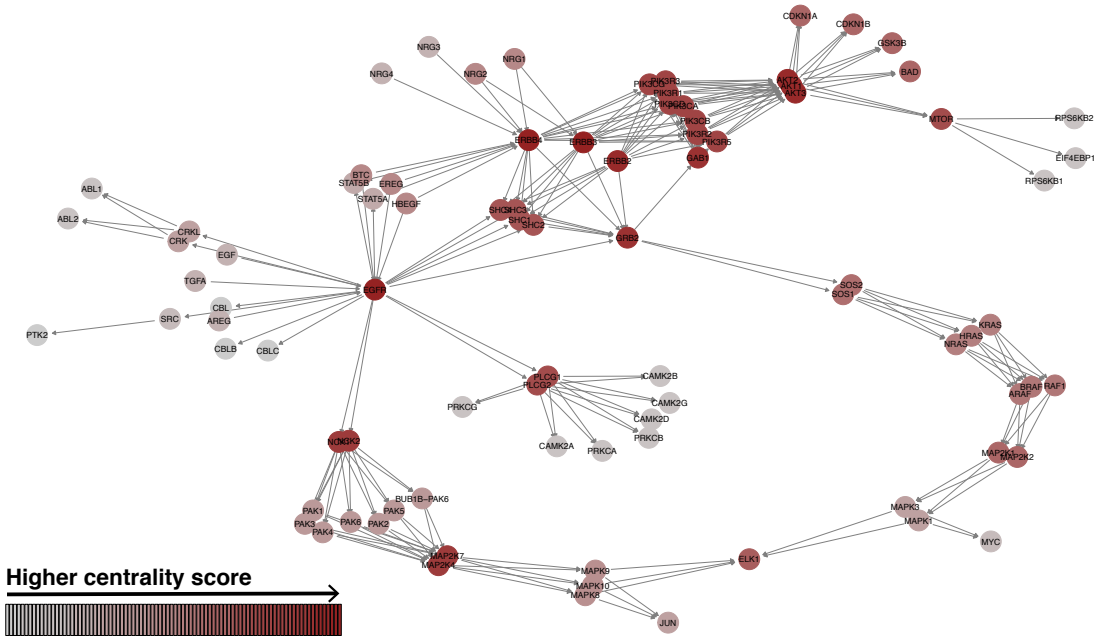

Figure 4.7: Application of Source/Sink centrality to ErbB signaling pathway. The color intensity indicates the ranking assigned by Source/Sink centrality. This figure shows the ability of Source/Sink centrality to the terminal nodes of the pathways such as ELK1, JUN, and BAD. A standard centrality score for directed graphs might assign zero importance to terminal nodes (See Table 4.19 for more details). Figure adopted from Naderi and Mostafavi [26].

CADIA leverages two independent ORA statistics (P_{ora}) and topological evidence (P_{ssc}). This approach is inspired by SPIA and produces increased sensitivity. We have shown that in multiple instances, the significance of the topological evidence P_{ssc} allows to compensate for the lack of strength in the over-representation evidence. Also, the lack of topological evidence allows to dismiss marginal over-representation evidences that might be irrelevant to the experimental data. While it is also possible to Source/Sink Centrality in GSA model through a methodology shown by Gu and colleagues [20], the choice of ORA allows to leverage two independent evidences simultaneously. Recent studies show that multi-evidence approaches for PEM can provide increased sensitivity and specificity [19, 96].

Table 4.12: Statistically significant pathway enrichments identified by GSA–FGSEA from the colorectal cancer data (GSE21510)

Name	ID	p.val	FDR*
Calcium signaling pathway	04020	2.03e-04	1.73e-03
Phospholipase D signaling pathway	04072	2.02e-04	1.73e-03
Sphingolipid signaling pathway	04071	2.03e-04	1.73e-03
cAMP signaling pathway	04024	2.04e-04	1.73e-03
cGMP-PKG signaling pathway	04022	2.04e-04	1.73e-03
Autophagy - animal	04140	2.03e-04	1.73e-03
Regulation of actin cytoskeleton	04810	2.03e-04	1.73e-03
Fc gamma R-mediated phagocytosis	04666	2.05e-04	1.73e-03
Insulin signaling pathway	04910	2.02e-04	1.73e-03
GnRH signaling pathway	04912	2.04e-04	1.73e-03
Adrenergic signaling in cardiomyocytes	04261	2.03e-04	1.73e-03
Gastric acid secretion	04971	2.01e-04	1.73e-03
Aldosterone-regulated sodium reabsorption	04960	2.00e-04	1.73e-03
Neurotrophin signaling pathway	04722	2.03e-04	1.73e-03
Choline metabolism in cancer	05231	2.04e-04	1.73e-03
Chemical carcinogenesis	05204	2.00e-04	1.73e-03
Insulin resistance	04931	2.05e-04	1.73e-03
Ras signaling pathway	04014	4.07e-04	2.94e-03
Rap1 signaling pathway	04015	4.08e-04	2.94e-03
MAPK signaling pathway	04010	4.11e-04	2.94e-03
Gap junction	04540	6.12e-04	3.98e-03
Oxytocin signaling pathway	04921	6.10e-04	3.98e-03
Alcoholism	05034	8.16e-04	4.86e-03
EGFR tyrosine kinase inhibitor resistance	01521	8.08e-04	4.86e-03
Vascular smooth muscle contraction	04270	1.22e-03	7.00e-03
Dopaminergic synapse	04728	1.63e-03	8.71e-03
Inflammatory mediator regulation of TRP channels	04750	1.64e-03	8.71e-03
Jak-STAT signaling pathway	04630	1.83e-03	9.35e-03
Hepatitis C	05160	2.03e-03	1.00e-02
Cholinergic synapse	04725	2.66e-03	1.27e-02
Apoptosis	04210	2.83e-03	1.31e-02
Toxoplasmosis	05145	3.07e-03	1.37e-02
Amphetamine addiction	05031	3.20e-03	1.39e-02
AMPK signaling pathway	04152	3.67e-03	1.54e-02
Bacterial invasion of epithelial cells	05100	3.84e-03	1.57e-02
Long-term potentiation	04720	4.42e-03	1.71e-02
Glioma	05214	4.42e-03	1.71e-02
Proteoglycans in cancer	05205	4.91e-03	1.85e-02
Circadian entrainment	04713	5.14e-03	1.88e-02

* FDR corrected P values

Table rows truncated for presentation

Table 4.13: Statistically significant pathway enrichments identified by CADIA from the gastric cancer data (GSE54129)

Name[§]	ID	P_{ora}	P_{ssc}	**CADIA**[†]	FDR_{ora}[‡]
ECM-r...	04512	1.44e-07	6.23e-01	2.29e-04	2.13e-05
Focal...	04510	2.44e-06	4.60e-01	8.15e-04	1.21e-04
Gastr...	04971	1.07e-06	9.94e-01	8.15e-04	7.91e-05
Chemi...	05204	1.23e-03	7.10e-03	4.08e-03	2.76e-02
***Wnt ...**	04310	2.76e-01	1.00e-04	9.38e-03	9.88e-01
PI3K-...	04151	2.13e-04	3.24e-01	1.81e-02	7.89e-03

[§] Names truncated for space limitations
[†] FDR corrected P_{cadia}
[‡] FDR corrected P_{ora}
[*] Unique to CADIA

Table 4.14: Statistically significant pathway enrichments identified by ORA from the gastric cancer data (GSE54129)

Name[§]	ID	CADIA[†]	$\mathbf{FDR_{ora}}$[‡]
ECM-receptor interaction	04512	2.29e-04	2.13e-05
Gastric acid secretion	04971	8.15e-04	7.91e-05
Focal adhesion	04510	8.15e-04	1.21e-04
PI3K-Akt signaling pathway	04151	1.81e-02	7.89e-03
***Renin secretion**	04924	1.35e-01	2.76e-02
Chemical carcinogenesis	05204	4.08e-03	2.76e-02
***AGE-RAGE signaling pat...**	04933	1.35e-01	2.76e-02
***Vascular smooth muscle...**	04270	1.26e-01	2.95e-02

[§] Names truncated for space limitations
[†] FDR corrected P_{cadia}
[‡] FDR corrected P_{ora}
[*] Unique to ORA

Table 4.15: Statistically significant pathway enrichments identified by SPIA from the gastric cancer data (GSE54129)

Name	ID	**SPIA**†	$CADIA^{\dagger}$
ECM-receptor interaction	04512	2.02e-09	2.22e-04
Gastric acid secretion	04971	5.39e-06	7.97e-04
Focal adhesion	04510	1.69e-05	7.97e-04
***TGF-beta signaling pathway**	04350	4.54e-03	4.06e-01
***Malaria**	05144	4.54e-03	NA
Cytokine-cytokine receptor....	04060	4.39e-02	5.09e-01
***Amoebiasis**	05146	4.43e-02	NA
Vascular smooth muscle con...	04270	4.43e-02	1.23e-01

† FDR corrected p-values
* Unique to SPIA

Table 4.16: Statistically significant pathway enrichments identified by EnrichNet from the gastric cancer data (GSE54129)

Name§	ID	**XD.Score**†	$FDR_{ora}{}^{*}$
Collecting duct acid secretion	04966	0.97	7.68e-02
Linoleic acid metabolism	00591	0.85	9.16e-02
Malaria	05144	0.82	7.39e-03
Gastric acid secretion	04971	0.77	9.23e-04
ECM-receptor interaction	04512	0.75	4.77e-04
Proximal tubule bicarbonate...	04964	0.74	3.79e-01

§ Names truncated for space limitations
† Significance score of Enrichnet (Threshold =0.72 for 95%)
* FDR corrected P_{ora} – as calculated in Enrichnet

Table 4.17: Statistically significant pathway enrichments identified by GSA–GAGE from the gastric cancer data (78, $FDR < 0.05$) (GSE54129)

Name	ID	p.val[†]	FDR[*]
Focal adhesion	04510	5.35e-08	4.63e-06
Pathways in cancer	05200	6.47e-08	4.63e-06
Proteoglycans in cancer	05205	2.26e-06	1.08e-04
AGE-RAGE signaling pathway in diabetic complications	04933	3.27e-06	1.17e-04
Chemical carcinogenesis	05204	5.66e-06	1.50e-04
PI3K-Akt signaling pathway	04151	6.29e-06	1.50e-04
Bacterial invasion of epithelial cells	05100	2.64e-05	5.39e-04
Osteoclast differentiation	04380	4.75e-05	7.80e-04
Rap1 signaling pathway	04015	4.98e-05	7.80e-04
Pertussis	05133	5.45e-05	7.80e-04
Adherens junction	04520	8.53e-05	1.03e-03
Chagas disease (American trypanosomiasis)	05142	8.64e-05	1.03e-03
ECM-receptor interaction	04512	1.07e-04	1.17e-03
TNF signaling pathway	04668	2.40e-04	2.35e-03
Hippo signaling pathway	04390	2.57e-04	2.35e-03
Regulation of actin cytoskeleton	04810	2.63e-04	2.35e-03
MicroRNAs in cancer	05206	2.80e-04	2.35e-03
Chemokine signaling pathway	04062	5.33e-04	3.89e-03
cGMP-PKG signaling pathway	04022	5.42e-04	3.89e-03
Legionellosis	05134	5.44e-04	3.89e-03
Gastric acid secretion	04971	6.04e-04	4.04e-03
Pancreatic cancer	05212	6.45e-04	4.04e-03
Salmonella infection	05132	6.50e-04	4.04e-03
Leishmaniasis	05140	9.88e-04	5.69e-03
MAPK signaling pathway	04010	9.94e-04	5.69e-03
Cytokine-cytokine receptor interaction	04060	1.06e-03	5.83e-03
HTLV-I infection	05166	1.19e-03	6.12e-03
Vascular smooth muscle contraction	04270	1.21e-03	6.12e-03
Shigellosis	05131	1.24e-03	6.12e-03
Chronic myeloid leukemia	05220	1.47e-03	7.00e-03
Thyroid hormone signaling pathway	04919	1.82e-03	8.02e-03
Staphylococcus aureus infection	05150	1.83e-03	8.02e-03
FoxO signaling pathway	04068	1.90e-03	8.02e-03
NF-kappa B signaling pathway	04064	1.91e-03	8.02e-03
Tuberculosis	05152	2.01e-03	8.20e-03
Phospholipase D signaling pathway	04072	2.17e-03	8.44e-03
Fc gamma R-mediated phagocytosis	04666	2.18e-03	8.44e-03
Platinum drug resistance	01524	2.71e-03	1.01e-02
Colorectal cancer	05210	2.76e-03	1.01e-02
Platelet activation	04611	2.86e-03	1.02e-02

[*] FDR corrected p-values – as calculated in GAGE. List truncated for space limitations.

Table 4.18: Statistically significant pathway enrichments identified by GSA (FGSEA) from the gastric cancer data (91 pathways, $FDR < 0.05$) (GSE54129)

Name	ID	p.val[†]	FDR[*]
Ras signaling pathway	04014	2.70e-04	9.84e-04
Rap1 signaling pathway	04015	2.66e-04	9.84e-04
MAPK signaling pathway	04010	2.74e-04	9.84e-04
NF-kappa B signaling pathway	04064	2.39e-04	9.84e-04
TNF signaling pathway	04668	2.44e-04	9.84e-04
cAMP signaling pathway	04024	2.63e-04	9.84e-04
cGMP-PKG signaling pathway	04022	2.59e-04	9.84e-04
Cytokine-cytokine receptor interaction	04060	2.75e-04	9.84e-04
ECM-receptor interaction	04512	2.36e-04	9.84e-04
Focal adhesion	04510	2.65e-04	9.84e-04
Regulation of actin cytoskeleton	04810	2.70e-04	9.84e-04
Complement and coagulation cascades	04610	2.37e-04	9.84e-04
Platelet activation	04611	2.49e-04	9.84e-04
Toll-like receptor signaling pathway	04620	2.41e-04	9.84e-04
T cell receptor signaling pathway	04660	2.44e-04	9.84e-04
Th1 and Th2 cell differentiation	04658	2.38e-04	9.84e-04
Th17 cell differentiation	04659	2.43e-04	9.84e-04
Chemokine signaling pathway	04062	2.64e-04	9.84e-04
Oxytocin signaling pathway	04921	2.54e-04	9.84e-04
Vascular smooth muscle contraction	04270	2.48e-04	9.84e-04
Axon guidance	04360	2.63e-04	9.84e-04
Osteoclast differentiation	04380	2.54e-04	9.84e-04
Central carbon metabolism in cancer	05230	2.35e-04	9.84e-04
MicroRNAs in cancer	05206	2.57e-04	9.84e-04
Proteoglycans in cancer	05205	2.65e-04	9.84e-04
Breast cancer	05224	2.57e-04	9.84e-04
Inflammatory bowel disease (IBD)	05321	2.35e-04	9.84e-04
Dilated cardiomyopathy	05414	2.37e-04	9.84e-04
AGE-RAGE signaling pathway in diabetic complications	04933	2.41e-04	9.84e-04
Shigellosis	05131	2.35e-04	9.84e-04
Pertussis	05133	2.34e-04	9.84e-04
Staphylococcus aureus infection	05150	2.29e-04	9.84e-04
Tuberculosis	05152	2.62e-04	9.84e-04
HTLV-I infection	05166	2.75e-04	9.84e-04
Measles	05162	2.51e-04	9.84e-04
Hepatitis B	05161	2.56e-04	9.84e-04
Toxoplasmosis	05145	2.46e-04	9.84e-04
Leishmaniasis	05140	2.35e-04	9.84e-04
Chagas disease (American trypanosomiasis)	05142	2.42e-04	9.84e-04
Endocrine resistance	01522	2.41e-04	9.84e-04

[*] FDR corrected p-values – as calculated in FGSEA. List truncated for space limitations.

Table 4.19: Centrality scores of different algorithms for ErbB Signalling pathway

Gene	SSC	Bet	Deg	Katz	Gene	SSC	Bet	Deg	Katz
EGFR	88	87	88	88	EREG	43	1	36	77
ERBB4	87	86	87	87	HBEGF	43	1	36	77
ERBB3	86	66	85	85	NRG1	41	1	36	75
ERBB2	85	1	85	85	NRG2	41	1	36	75
AKT1	82	79	77	72	MAPK10	38	51	36	31
AKT2	82	79	77	72	MAPK8	38	51	36	31
AKT3	82	79	77	72	MAPK9	38	51	36	31
GRB2	81	88	60	71	BUB1B-PAK6	31	44	36	43
GAB1	80	85	84	84	PAK1	31	44	36	43
MAP2K4	78	67	60	56	PAK2	31	44	36	43
MAP2K7	78	67	60	56	PAK3	31	44	36	43
NCK1	76	72	80	82	PAK4	31	44	36	43
NCK2	76	72	80	82	PAK5	31	44	36	43
PIK3CA	68	58	60	63	PAK6	31	44	36	43
PIK3CB	68	58	60	63	MAPK1	29	56	36	31
PIK3CD	68	58	60	63	MAPK3	29	56	36	31
PIK3CG	68	58	60	63	JUN	28	1	1	1
PIK3R1	68	58	60	63	CRK	26	41	36	31
PIK3R2	68	58	60	63	CRKL	26	41	36	31
PIK3R3	68	58	60	63	STAT5A	24	1	1	1
PIK3R5	68	58	60	63	STAT5B	24	1	1	1
MTOR	67	82	60	52	AREG	21	1	26	53
PLCG1	65	54	80	80	EGF	21	1	26	53
PLCG2	65	54	80	80	TGFA	21	1	26	53
SHC1	61	1	26	27	NRG3	19	1	26	50
SHC2	61	1	26	27	NRG4	19	1	26	50
SHC3	61	1	26	27	SRC	18	41	26	26
SHC4	61	1	26	27	MYC	17	1	1	1
ELK1	60	1	1	1	ABL1	8	1	1	1
BAD	56	1	1	1	ABL2	8	1	1	1
CDKN1A	56	1	1	1	CAMK2A	8	1	1	1
CDKN1B	56	1	1	1	CAMK2B	8	1	1	1
GSK3B	56	1	1	1	CAMK2D	8	1	1	1
MAP2K1	54	74	36	38	CAMK2G	8	1	1	1
MAP2K2	54	74	36	38	PRKCA	8	1	1	1
SOS1	52	83	60	61	PRKCB	8	1	1	1
SOS2	52	83	60	61	PRKCG	8	1	1	1
ARAF	49	69	36	40	EIF4EBP1	5	1	1	1
BRAF	49	69	36	40	RPS6KB1	5	1	1	1
RAF1	49	69	36	40	RPS6KB2	5	1	1	1
HRAS	46	76	60	58	CBL	2	1	1	1
KRAS	46	76	60	58	CBLB	2	1	1	1
NRAS	46	76	60	58	CBLC	2	1	1	1
BTC	43	1	36	77	PTK2	1	1	1	1

CHAPTER 5: Topological Organizations in Pathways

In the last two chapters we outlined a network-based analysis pipeline based on a novel graph centrality method, SSC. We showed that the use of topological evidence from SSC allows to make unique and informative inference. This section attempts to provide insight on why the concept of Source/Sink modeling is informative. To this end, we extended the notion of Source/Sink centrality modeling by mixing with existing standard centrality models and use them for an in-depth investigation of the topological organization of the human pathways.

In particular, we assemble a battery of standard and novel graph centrality methods and investigate whether these topological models can differentiate between the organization of cancer-related genes and non-cancer-related genes in the pathways. The rationale for choosing the cancer-related genes is the intuition that cancers are regarded as diseases of pathways, i.e. cancers are primarily driven by perturbation/alteration of pathways [18, 90]. Subsequently, the dysfunction one or more cancer-related genes can result in dysfunction of their associated pathways [18]. Therefore, understanding the topological position of cancer associated genes may reveal insight regarding the topological organization of key pathway drivers/regulators.

We address our research questions by using four known standard centrality models Degree, Katz, PageRank, and Laplacian, as well as, their possible extensions with the Source/Sink technique. The rationale for choosing these specific centrality methods is to investigate the relationship of genes with their property of being cancer-related from certain perspectives. In particular, four models enable to investigate the importance of a gene with respect to the number of interactions, the importance of interacting genes, and the direction of interactions. The Source/Sink technique, as outlined in

the chapter 3, enables assigning node importance in both upstream and downstream ends of pathways, while accounting for directions of the interactions.

We design three statistical approaches to evaluate each of the centrality models. A challenge in our analysis is the existence of many biological pathways. We address this challenge by evaluating global and individual patterns of centralities across pathways. In particular, 1– We investigate the linear relationship of gene rankings, according to each centrality model, with the probability of being cancer related. 2– Compare the cumulative distribution of the rankings of cancer-related genes versus that of non-cancer-related genes. 3– Compare the mean ranking of cancer-related versus non-cancer-related genes for each pathway by two-sample testing.

The analysis results show that the ranking of pathway genes, based on the number of interactions, correlates with the probability of being cancer-related [27]. We show that, the Source/Sink modeling increases the linear relationship of gene centralities with their probability of being cancer across all models. Pathway-by-pathway comparisons shows that each model has unique pattern for distinguishing between cancer-related and non-cancer-related genes, Source/Sink PageRank shows the highest statistical power as its number of hypothesis rejections are the highest.

Our analysis shows that the cancer-related genes in tend to have higher centrality; particularly, when accounting for directionality and importance in both upstream and downstream of pathways using Source/Sink modeling. These results can potentially be incorporated with existing graph-based pathway analysis models in order to increase biological relatedness.

5.1 Graph Modeling of Pathways

In this chapter we investigate four standard known centrality models to reflect the research hypothesis– Degree, PageRank, Laplacian, and Katz. Each of these models have been applied to the problem of pathway enrichment analysis in some format [19, 21, 23, 24]. We investigate different variations of each model, Undirected,

Source, and Sink when applicable. In addition to these models, we derive the concepts Source/Sink PageRank, Source/Sink Katz, and Source/Sink Laplacian to account for importance in both upstream and downstream of the pathways. A brief description of each model along with justification for choosing them are provided in following.

Consistent with our definitions in Chapters 2 and 3, let $G = (V, E)$ represent a pathway where $V(G) = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes. The set of edges $E(G) = \{e_1, e_2, \ldots, e_m\}$ is a collection of pairs of nodes, $e_i \in V \times V$.

Degree Centrality: In this model, the centrality of a node is the number of its neighbors. Degree centrality has been well-studied in the context of biological network, most particularly for protein-protein interaction networks (PPI). Studies show that degree centrality of node in PPI of different organisms correlates with its essentiality, meaning the likelihood of a protein's removal, e.g. knockdown, to be lethal for the model organism [15, 62, 97]. Since the input pathways of this study are directed networks, degree centrality was calculated by combines in-degree and out-degree of a node. This value is the same as the degree centrality in the underlying undirected graph. In this case:

$$C_{deg}(v) = Deg_{in}(v) + Deg_{out}(v) \tag{5.1}$$

PageRank Centrality: PageRank is a member of spectral centrality measures where the importance of a nodes is a function of the centrality of its neighbors. PageRank describes the probability distribution of a uniform random walk with restart being present at each node of a graph after a large number of steps [65, 67, 70]. Formally, the vector of PageRank C_{pgr} is defined as:

$$\begin{aligned} C_{pgr} &= \lim_{t\to\infty} P^{(t)} \\ &\text{subject to:} \\ P^{(t+1)} &= (1-\alpha)P^{(0)} + (\alpha)P^{(t)}D^{-1}A \end{aligned} \tag{5.2}$$

Where $P^{(0)}$ is the probability vector of initial states of the random walk and α is the transition probability. $P^{(t)}$ is the probability vector the random walk being at each specific node. D is the diagonal degree matrix such that $[D]_{ii} = max(Deg_{out}(v_i), 1)$. A is the adjacency matrix of the graph. In graph theory terms, the PageRank of a node v is based on the PageRank of the nodes with links to v, divided by their out degrees. Formally:

$$C_{pgr}(v_i) = \beta_i + \alpha \sum_{u|v_i \in \mathbf{N_G(u)}} \frac{C_{pgr}(u)}{|\mathbf{N_G(u)}|} \tag{5.3}$$

In the above Formula, β_i's are constant values indicating the probability of restarting at node v_i. Formula 5.3 can be extended to the cases where β_i's are arbitrary parameters [65]. In this case, the output of the algorithm is not necessarily a probability distribution. The formula 5.3 can be expressed in a vectorized format as following:

$$C_{pgr} = \beta + A^T D^{-1} C_{pgr} \tag{5.4}$$

where C_{pgr} is the vector of centralities and β is the vector of initial values. A closed form solution of Formula 5.3 can be achieved by rearranging and solving for C_{pgr} [65]. Formally:

$$C_{pgr} = (I - \alpha A^T D^{-1})^{-1} \beta \tag{5.5}$$

PageRank can be used for both directed and undirected graphs. The notion of PageRank in Formula 5.2 is closely related to the distance measure of Enrichnet [22]. Similarly, Formula 5.5 is closely related to the definition of the Perturbation Factor in SPIA [19].

PageRank Sink: we define the PageRank Sink component as the standard PageRank of a directed graph. The original concept of PageRank, as described by Brin and Page, measures the importance of a website based on the importance of the websites that have a link to it [70]. Likewise, in the Sink component of the PageRank, the downstream nodes have the higher importance. This is because a random walk will not be present at any node without incoming edges, unless by a restart event (Formula 5.2). The PageRank Sink centrality captures the importance of a node as a receiver of information. Formally we define the Sink PageRank centrality as:

$$C_{Sink-pgr}(v) := C_{pgr}(v) \tag{5.6}$$

PageRank Source: As mentioned, the standard PageRank for directed graphs would produce the minimal importance score for the nodes with no incoming edges, and the upstream nodes of pathways fall into this category. We derive a PageRank Source model that captures the importance of a source node by modifying the underlying graph. Formally, the PageRank Source is:

$$C_{Src-pgr}(v_i) = \beta'_i + \alpha' \sum_{u|v_i \in \mathbf{N_{G^T}(u)}} \frac{C_{Sink-pgr}(u)}{|\mathbf{N_{G^T}(u)}|} \tag{5.7}$$

The above formula essentially calculates the standard PageRank on the transposed of a given graph. A closed form solution of Formula 5.7 can be achieved by rearranging and solving for the vector of centralities of all nodes [65]. Define the diagonal in-degree matrix, D', of G such that $[D']_{ii} = max(1, Deg_{in}(v_i))$. Formally:

$$C_{Src-pgr} = (I - \alpha' A D'^{-1})^{-1} \beta' \tag{5.8}$$

Source/Sink PageRank: The fundamental concept of Source/Sink modeling is measure the centrality of nodes as both sources and sinks of information. Using directed centrality measures only gives importance to either upstream nodes or downstream ones. By adopting the Source/Sink concept to the PageRank, we define a new model based on addition of two individual centrality components, the Source and the Sink. After calculating Source and Sink Centrality values individually, the two components are summed as following:

$$C_{SS-pgr}(v) = C_{Src-pgr}(v) + \gamma C_{Sink-pgr}(v) \tag{5.9}$$

In the above Formula, γ represents a parameter for indicating the relative importance of Source versus Sink.

Katz-Bonacich Centrality: is another member of the spectral family of centrality models. In this type of model, the centrality of a node is calculated relative to the sum of centrality of its neighbors. Formally:

$$C_{katz}(v_i) = \beta_i + \alpha \sum_{u \in \mathbf{N_G(v_i)}} C_{katz}(u) \tag{5.10}$$

In the above formula, β is a constant factor and α is dampening factor. It can be shown that the convergence of the Formula 5.10 depends on the largest eigenvalue of the adjacency matrix. It can be shown that $\alpha < 1/\lambda_1$ is a sufficient condition for convergence, with λ_1 being the largest positive eigenvalue of the adjacency matrix. Rearranging Formula 5.10 gives a closed form solution of Katz centrality. Formally:

$$C_{katz} = (I - \alpha A)^{-1} \beta \tag{5.11}$$

Throughout this document, Katz centrality refers to the directed graph. Katz-Bonacich centrality is closely related to the formulations of Cdist and NetGSA for pathway enrichment analysis [21, 23].

Source/Sink Katz: This model is also defined in a similar fashion to Source/Sink PageRank. The **Source Katz** component is defined as the Katz centrality component of the directed graph.

$$C_{Src-ktz}(v) := C_{katz}(v) \tag{5.12}$$

The **Sink Katz** component is defined as the Katz centrality of the transposed graph.

$$\begin{aligned} C_{Sink-ktz}(v_i) &:= \beta'_i + \alpha' \sum_{u \in \mathbf{N_G^T(v_i)}} C_{Sink-ktz}(u) \\ C_{Sink-ktz} &= (I - \alpha' A^T)^{-1} \beta' \end{aligned} \tag{5.13}$$

Katz Source/Sink Centrality is then defined as:

$$C_{SS-ktz}(v) = C_{Src-ktz}(v) + \gamma C_{Sink-ktz}(v) \tag{5.14}$$

It can be shown that Source and Sink components have the same convergence criteria. When $\beta = \beta'$ and $\alpha = \alpha'$, the Source/Sink Katz centrality is equal to the Source/Sink model defined in Chapter 3.

Laplacian Centrality: Laplacian graph influence measures are a family of models that capture the amount of effect a node has on the other nodes. These measures are the core of the heat diffusion kernels of graphs [56, 67, 98, 99]. Graph Laplacians are generally defined for undirected graphs [56, 67]. There are modifications for directed graphs either on strongly connected graphs or directed acyclic graphs [99, 100]. In this study, we use a specific version for directed graphs by Shojaie and Michailidis for pathway enrichment analysis [24]. For an adjacency matrix of a directed graph, A, define the weight normalized matrix L using a positive real value d as following:

$$L_{ij}(d) = \frac{A_{ij}}{d + \sum_{j=1}^{n} |A_{ij}|} \tag{5.15}$$

$$L = \lim_{d \to 0} L(d) \tag{5.16}$$

Define the influence matrix, L^*, as the geometric series of L. In the case of undirected graphs, this notion is related to the concept of normalized Laplacian and heat diffusion kernels [67].

$$L^* = \sum_{i=0}^{\infty} L^i \tag{5.17}$$

On the condition of convergence, the above summation can be written as:

$$L^* = \lim_{d \to 0} (I - L(d))^{-1} \tag{5.18}$$

According to Shojaie and Michailidis, choice of d as small as 0.01 would produce consistent and stable results. However, to eliminate the need for the parameter d, we rewrite an equivalent formulation for the matrix L as :

$$L := D^{-1}A \tag{5.19}$$

where D is the diagonal degree matrix with the same definition as in D of PageRank. As noted in [65], for undirected graphs, the solution to the matrix L in a matrix geometric series uniquely exist. That is, the matrix L^* from Formula 5.17 is only guaranteed to uniquely exist when we use the symmetric matrix of the undirected graph.

However, the case might be different for directed graphs. Therefore, including a shrinking factor, $\alpha < 1$, that ensures the convergence in a geometric summation. We then re-define:

$$L := \alpha D^{-1} A \tag{5.20}$$

Using the above Formula, we define the Laplacian centrality of a node as the aggregated influence of a node i on all other nodes. This is obtained from Formula 5.17:

$$\begin{aligned} C_{lap} &= L^* \mathbb{1} \\ &= (I - \alpha D^{-1} A)^{-1} \mathbb{1} \end{aligned} \tag{5.21}$$

Like other models, we define the centrality in 4 formats. The **Undirected Laplacian** is obtained by using the adjacency matrix of the undirected graph in Formula 5.21. It is possible to show that the undirected laplacian produces a constant value for all centrality values in connected components of a graph. However, we have analyzed the pathways' data by including this model since pathways may contain isolated nodes and multiple connected components. The **Laplacian Source** component is defined as the Laplacian centrality of the directed graph:

$$C_{lap-Src} := C_{lap} \tag{5.22}$$

The **Laplacian Sink** component is the Laplacian centrality of the transposed graph:

$$C_{lap-Sink} := (I - \alpha' D'^{-1} A^T)^{-1} \mathbb{1} \tag{5.23}$$

The **Source/Sink Laplacian** is then defined as the sum of the two components:

$$C_{lap-ssc} := C_{lap-Src} + \gamma C_{lap-Sink} \tag{5.24}$$

5.2 Model Evaluations

This section details the process of applying each centrality model to a background set of biological pathways and a background list of cancer-related genes. For each model, we will contrast the Undirected, Source, Sink, and the Source/Sink variations. And, we investigate which one of the variations are more informative regarding the topological organization of the cancer genes.

In particular, different variations of each centrality model are examined through three aspects. 1– The linear relationship between the centrality scores of a particular variation and the probability of genes being cancer related. 2– The distribution of centrality scores of cancer-related genes and non-cancer-related genes (normal). 3– The mean difference between the centrality scores cancer-related genes versus normal genes for each pathway. Since the subjects of study are multiple pathways, rather than a single global graph, normalization and ranking procedures were used to create a unified framework.

5.2.1 Background Pathways

Human pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG) were retrieved (n = 330, As of August 2018). Pathways with more than 1000 nodes and more than 4000 interactions were excluded from any further analysis. Pathways with equal or less than 20 nodes or 20 edges were neglected from analysis (n = 85). Also, pathways with largest eigenvalues more than 10 (n = 15) were excluded from analysis in order to maintain consistent centrality calculations. In addition, pathways with a single unique value for any of the centrality measures (e.g. all degrees being 10) were excluded from the analysis (n= 11). The pathways were retrieved and analyzed using R-packages "KEGGGraph" and "Pathview" [77, 101]. In addition, pathways with 5 or less cancer associated genes were excluded from analysis for consistency of p-value calculations (n = 64). The final set of pathways contained 155 entries.

Cancer-related genes were retrieved from relevant gene familiy classifications of Broad Institute's MSigDB (n = 417, 06-06-18) [7]. The MSigDB's gene family classifications of the cancer related genes were *Oncogenes*, *Tumor Suppressors*, and *Translocated cancer genes*. Cancer Gene Census from Sanger Institute was used as an additional reference list for cancer-related genes (n= 719, 06-06-18) [102]. The union of these two sets were used as the reference cancer gene list (n = 733). Overall, a total of 19001 nodes were analyzed after pathway preprocessing, having 4474 distinct genes. There were 3798 cancer related nodes, associated with 397 unique cancer genes in the dataset.

This studies uses $\gamma = 1$, $\beta = \beta' = \mathbb{1}$, and $\alpha = \alpha' = 0.85$ for different variations of PageRank. For Katz centrality variations, the parameter setting was $\alpha = \alpha' = 0.1$, $\beta = \beta' = \mathbb{1}$ and $\gamma = 1$. We did not analyze for Undirected Katz because of limitation of the largest eigenvalues. For Laplacian centrality variations, the parameters were $\gamma = 1$, $\beta = \beta' = \mathbb{1}$, and $\alpha = \alpha' = 0.85$.

5.2.2 Regression Analysis

For each pathway, the nodes were ranked using all of the centrality measures. The centrality ranks of each pathway were placed in 100 quantiles. The 100th quantile indicates most central genes in a pathway and 1st quantile indicates the lowest importance. Let $C_{a,j}(v_i)$ denote the centrality of a node v_i in pathway j using model a. The quantile ranking of a node i, $Q_j(v_i)$, is then defined as:

$$Q_j(v_i) = \left\lceil 100 \times \frac{C_{a,j}(v_i)}{|V_j|} \right\rceil \tag{5.25}$$

In the above formula, V_j is the total number of nodes in pathway j. The quantile ranking allows to compare the centrality rankings among all pathways because different pathways have different number of nodes. For example, a gene (namely X) with a rank score of 20 in a pathway with 20 nodes, then it is the most central; A gene

(namely Y) with a rank score of 20 in a pathway with 100 nodes are among the least important nodes. When quantile transformation is applied, X would have a quantile score of 100 and Y would have 20, allowing to compare how central they are between different pathways.

To investigate the relationship between cancer-relatedness of a gene and its centrality, the proportion of cancer-related genes were calculated on each quantile across all pathways. Let Q_{ij} denote the set of genes belonging to i-th quantile in pathway j — $Q_{ij} = \{v \mid v \in V_j, Q_j(v) = i\}$. Let R denote the set of all cancer-related genes. The ratio of cancer related genes for i-th quantile, F_i^c, is defined as:

$$F_i^c = \frac{\sum_j \left|\{v \mid v \in R \cap Q_{ij}\}\right|}{\sum_j \left|\{v \mid v \in Q_{ij}\}\right|} \tag{5.26}$$

Although some genes were occurring in multiple pathways, each occurrence was treated as an unique gene because the purpose was to evaluate the centrality with respect to pathways. F_i^c was then tested against the level of quantile for assessing linear relationships. In the below formula, i indicates the index value of a quantile group, e.g. 1 for the 1st quantile and 10 for the 10th quantile. Let a_1 and a_0 be the the coefficients of the linear regression. Formally:

$$F_i^c = a_1 \cdot i + a_0 \tag{5.27}$$

For each centrality measure the above linear regression was fitted and the adjusted R-squared (coefficient of determination) were evaluated. In addition, Pearson correlation of $Q_(v_i)$ values between each centrality measure were calculated to outline the differences between the models.

5.2.3 Comparison of Cumulative Densities

To compare the distribution of centrality values from a global perspective, the centrality scores were normalized within each pathway using the following formula:

$$N_{a,j}(v_i) = \frac{C_a^j(v_i) - \mu_{a,j}}{\sigma_{a,j}} \tag{5.28}$$

where $\mu_{a,j}$ and $\sigma_{a,j}$ are the mean and standard deviation of centrality scores of pathway j using method a. Accordingly, $N_{a,j}(v_i)$ is the normalized centrality score of node v_i in pathway j, using the centrality method a. The normalized score for all pathways were placed in 100 quantiles. The distribution of quantile scores for the types of genes "Cancer" and "Non-cancer" were compared by Kolmogorov-Smirnov (KS) test on cumulative distribution function (CDF) of cancer-related and normal genes. The p-values were calculated based on the alternative hypothesis of the CDF of the cancer-related lying below that of the normal. In this test, the CDF of all genes combined would follow a straight line.

5.2.4 Pathway-wise Two-Sample Testing

For each pathway, the difference of the mean raw centrality values between cancer-related genes and none cancer genes were evaluated using Welch's t-test. Formally:

$$\begin{aligned} t &= \frac{\hat{\mu}_{a,c} - \hat{\mu}_{a,n}}{\sqrt{\frac{s_{a,c}^2}{N_c} + \frac{s_{a,n}^2}{N_n}}} \\ H_0 : &\quad \mu_{a,c} = \mu_{a,n} \\ H_A : &\quad \mu_{a,c} > \mu_{a,n} \end{aligned} \tag{5.29}$$

where $\hat{\mu}_{a,c}$ and $\hat{\mu}_{a,n}$ are the estimated means of centrality values for cancer and normal genes by model a. Similarly, $s_{a,c}^2$ and $s_{a,n}^2$ are the variance estimates of the centrality scores of cancer and normal genes, using model a. N_c and N_n denote the sample size of cancer genes and normal genes. H_0 is the null hypothesis of cancer

and normal genes having the same mean. H_A is the alternative hypothesis where the cancer genes have a higher mean.

Since the underlying distribution of the centrality values is unknown, we also used Wilcox non-parametric test to evaluate the null hypothesis of cancer and normal gens having the same mean. Wilcox test ranks individual observations and evaluates the difference between the sum of the ranking in two classes of the hypothesis.

For each centrality model, the p-values from Formula 5.29 and Wilcox test were calculated across all pathways. Because of the large number of pathways, multiple hypothesis testing criteria was used to determine significant p-values. In particular, Benjamini-Hochberg False Discovery Rate was applied to all calculated p-values for each centrality method to control type-I error at %5 ($FDR < 0.05$) [32]. The same procedure was applied to both parametric and non-parametric approaches. The sets of significant pathways for each centrality model were contrasted against each other.

5.3 Results

5.3.1 Regression Analysis

Degree centrality: As evident Figure 5.1, for the higher values of quantile scores the fraction of cancer genes tend to be higher, and low degree quantile scores exhibit lower fractions of cancer genes. The analysis supports this observation by showing a linear relationship between the scores and the ratio of cancer genes with an adjusted R-squared of 0.25 (Figure 5.1). The regression analysis shows an statistically significant positive coefficient of 1.45×10^{-3} for the quantile scores (p-value $= 1.07 \times 10^{-7}$, Table 5.1).

Katz centrality: Figure 5.2 shows that individual Source and Sink components of Katz centrality fail to capture a linear relationship between the quantile centrality scores and the fraction of cancer genes. The standard Katz centrality for directed graph (the Source component) does not find evidence (p-value $= 0.244$) for linear relationship between the quantile scores and the ratio of cancer genes (Figure 5.2).

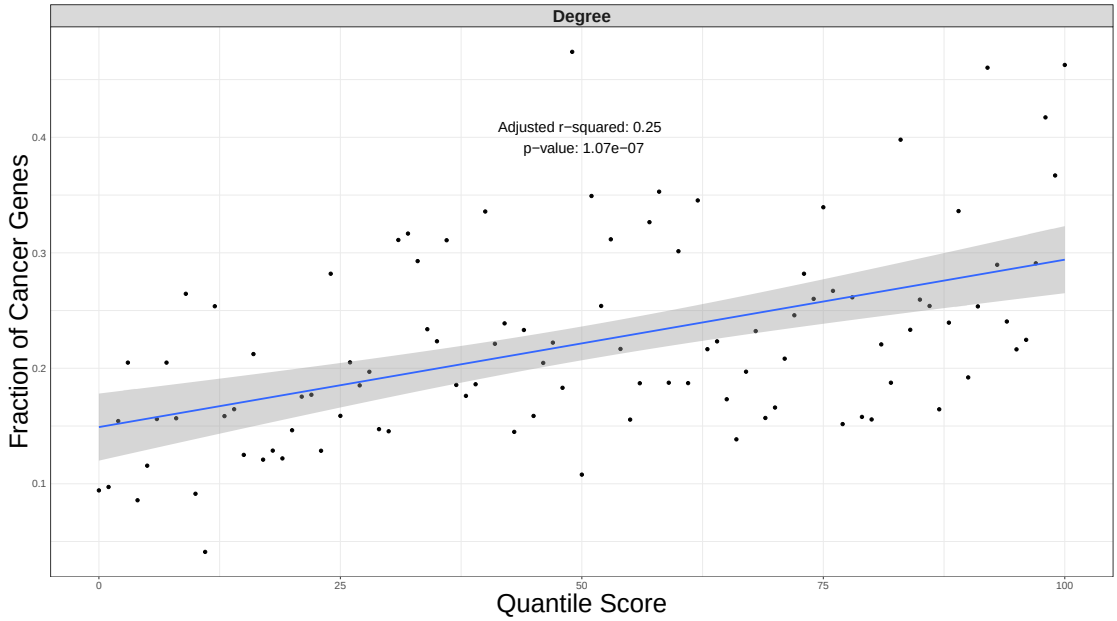

Figure 5.1: Linear regression of quantile-scores versus the ratio of cancer-related genes for Degree centrality model. In this case, higher centrality quantile score denotes a higher fraction of cancer-genes. X-axis represents the quantile-scores generated by Formula 5.25. Y-axis represents the fraction of all genes that are cancer related (Formula 5.26). The blue line represent the regression line from Formula 5.27. The gray band represent 95 confidence interval of the linear regression.

In this case, the regression model accounts for an insignificant fraction of the linear model variance (adj-$R^2 = 0.015$). Similarly, the Katz Sink Component fails to detect a linear relationship between the quantile score and the ratio of cancer genes (adj-$R^2 = 0.0014$, p-value $= 0.719$). In contrast, the combined value of the two components, Source/Sink Katz, shows that the linear relationship explains a statistically significant portion of the variance, more compared to Degree centrality (adj-$R^2 = 0.36$). In this case, the regression analysis shows an statistically significant positive coefficient of 1.42×10^{-3} for the quantile scores (p-value $= 2.29 \times 10^{-11}$, Table 5.1).

The correlation analysis of all variations of Katz centrality and Degree centrality shows insight regarding their underlying mechanisms (Figure 5.3). When comparing Katz-Source component to the other options, it is observable that the nodes with no Source importance have a varying range of centrality across different models. The dense bands at Katz-Source values of zero in Figure 5.3 exhibit the instances

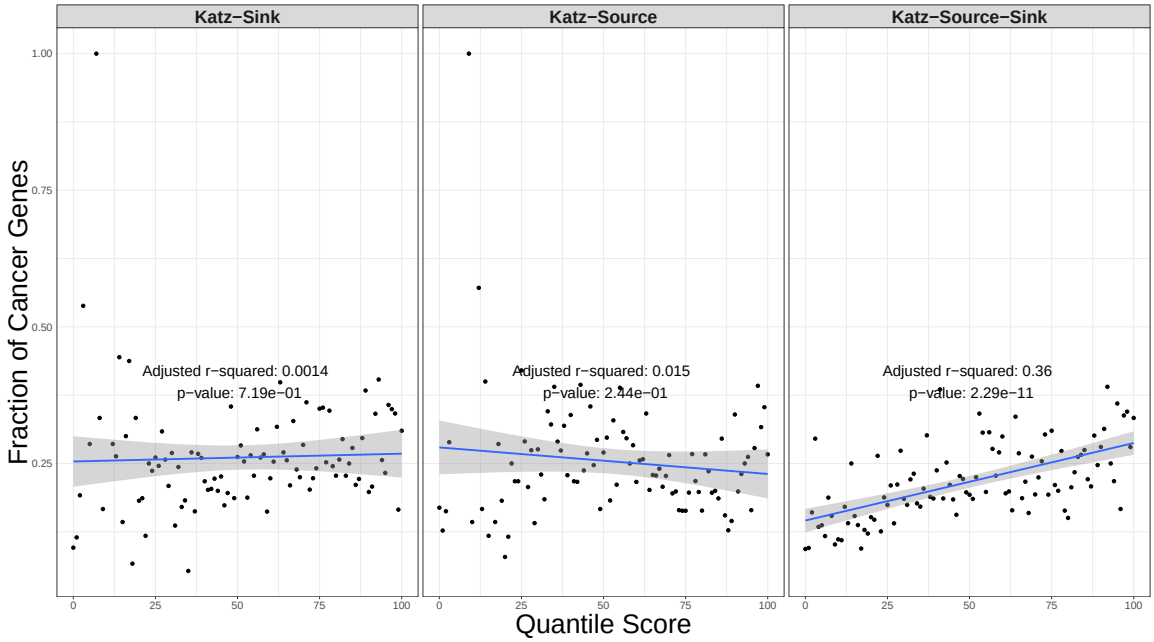

Figure 5.2: Linear regression of quantile-scores versus the ratio of cancer-related genes for Katz centrality model. X-axis represents the quantile-scores generated by Formula 5.25 Y-axis represents the fraction of all genes that are cancer related (Formula 5.26). The blue line represent the regression line from Formula 5.27. The gray band represent 95 confidence interval of the linear regression.

where the Katz-Source centrality is unable to assign importance to the nodes that are distinguished by other models.

Similarly, the Katz-Sink component assigns zero values to many nodes which have higher importance by the other models (the dense band at Katz-Sink = 0, Figure 5.3). The Source Component and the Sink component values of Katz do not show any strong correlation, which indicates that they might potentially produce radically different characterization of the graphs. The complete comparison of correlations among all models can be found in the Supplementary Figure B.1.

Laplacian centrality: Figure 5.4 shows that individual Source and Sink components of the Laplacian centrality fail to capture any linear relationship between the quantile centrality scores and the fraction of cancer genes. The standard Laplacian centrality for directed graph (the Source component) does not find evidence (p-value = 0.142) for linear relationship of the quantile scores and the ratio of cancer genes

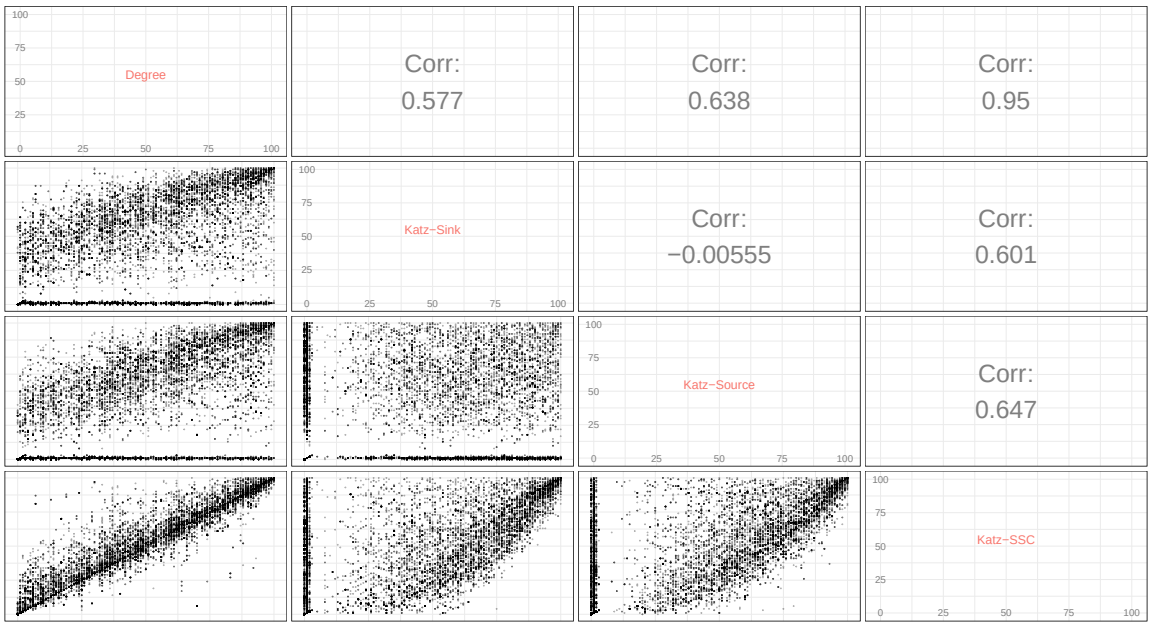

Figure 5.3: Correlation of quantile scores of all genes in human pathways between variations of Katz centrality model and Degree centrality.

(Figure 5.4). In this case, the linear model accounts for an insignificant fraction of the variance (adj-$R^2 = 0.023$). Similarly, the Laplacian Sink Component fails to detect a linear relationship between the quantile score and the ratio of cancer genes (adj-$R^2 = 0.0006$, p-value $= 0.817$). The Undirected Laplacian model also does not exhibit any relationship between the quantile scores and the fraction of cancer genes (adj-$R^2 = 0.00018$, p-value $= 0.895$). In contrast, the combined value of the two components, Source/Sink Katz, shows that the linear relationship explains a statistically significant portion of the variance, more compared to Degree centrality (adj-$R^2 = 0.46$). In this case, the regression analysis shows an statistically significant positive coefficient of 1.86×10^{-3} for the quantile scores (p-value $= 9.9 \times 10^{-15}$, Table 5.1).

The correlation analysis of all variations of Laplacian centrality (Figure 5.5) shows similar patterns to those of Katz centrality. It is observable that the nodes with no Source importance have a varying range of centrality across different models. The dense bands at Source values of zero in Figure 5.5 exhibit the instances which the Source variation is unable to assign importance to the nodes that are distinguished

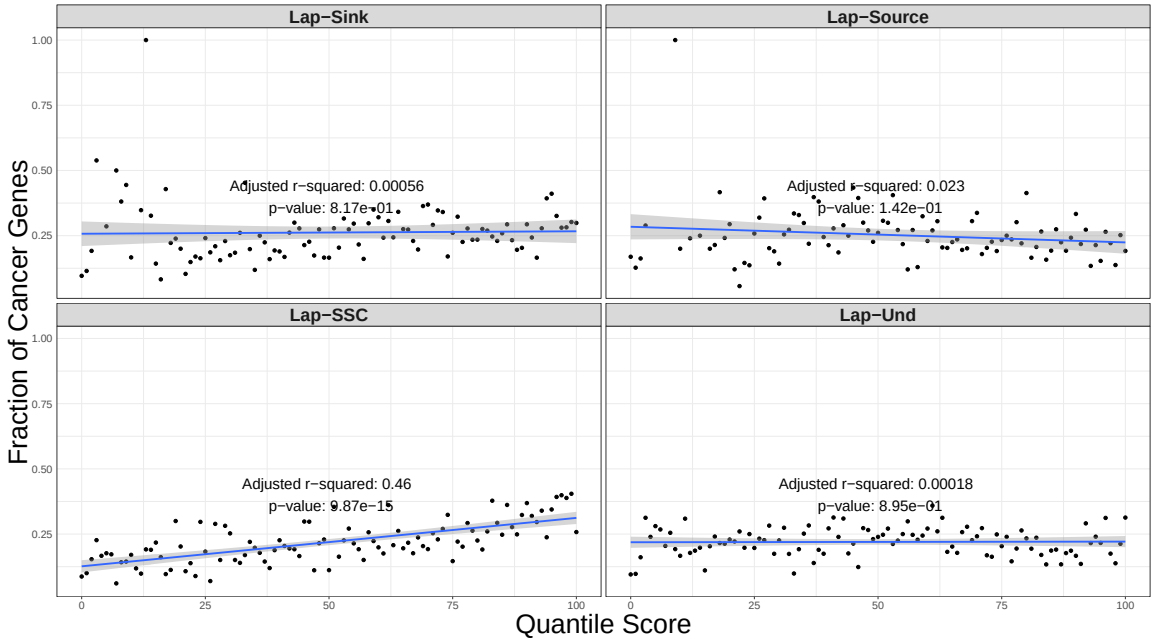

Figure 5.4: Linear regression of quantile-scores versus the ratio of cancer-related genes for Laplacian centrality model. X-axis represents the quantile-scores generated by Formula 5.25 Y-axis represents the fraction of all genes that are cancer related (Formula 5.26). The blue line represent the regression line from Formula 5.27. The gray band represent 95 confidence interval of the linear regression.

by other models.

Similarly, the Sink component assigns zero values to many nodes which have higher importance by the other models (the dense band at Lap-Sink = 0). The Source Component and the Sink component values of Laplacian centrality do not show any strong correlation, which indicates that they produce different characterization of graphs. Interestingly, the correlation coefficient between Laplacian Source/Sink and Undirected Laplacian is relatively low (Figures 5.5 and Supplementary Figure B.1). This also indicates the radically different characterizations from each model.

PageRank centrality: Figure 5.6 shows that individual Source and Sink components of the PageRank capture the linear relationship between the quantile centrality scores and the fraction of cancer genes. In particular, the standard PageRank centrality for directed graph (the Sink component) finds an evidence (p-value $= 1.57 \times 10^{-4}$) for linear relationship between the quantile scores and the ratio of cancer genes (Fig-

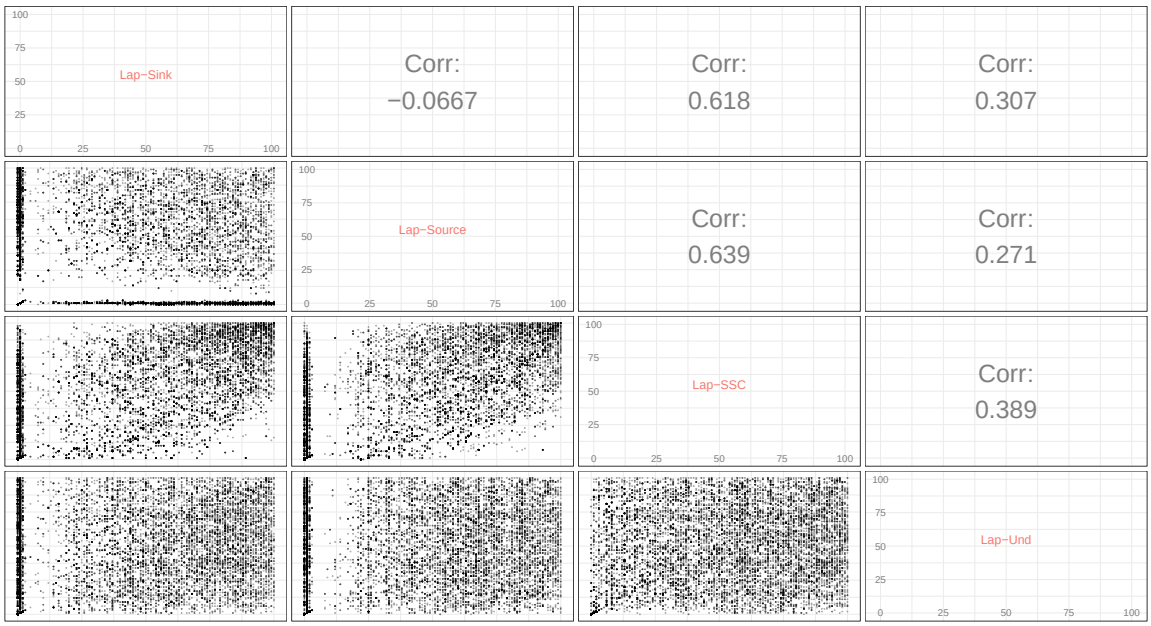

Figure 5.5: Correlation of quantile scores of all genes in human pathways between variations of Laplacian centrality model.

ure 5.6). In this case, the regression model finds an adjusted-R^2 of 0.14 for the linear relationship. The linear regression coefficient for the centrality scores is 1.64×10^{-3}. Similarly, the Source Component detects a linear relationship between the quantile score and the ratio of cancer genes (adj-$R^2 = 0.41$, p-value $= 2.89 \times 10^{-12}$). In this case, the linear regression coefficient for the centrality scores is 2.07×10^{-3}. The Undirected PageRank exhibit a stronger linear relationship between the quantile scores and the fraction of cancer genes (adj-$R^2 = 0.66$, p-value $= 1.02 \times 10^{-24}$) with a regreession coefficient for the centrality scores is 2.34×10^{-3}. The combined value of the two components, Source/Sink PageRank, shows the strongest linear relationship and explains a statistically significant portion of the variance, more compared to Degree centrality (adj-$R^2 = 0.74$). In this case, the regression analysis shows a coefficient of 2.71×10^{-3} for the quantile scores (p-value $= 4.2 \times 10^{-31}$, Table 5.1).

The correlation analysis of all variations of PageRank centrality (Figure 5.7) show similar patterns to those of Katz and Laplacian. It is observable that the nodes with no Source importance have a varying range of centrality across different models.

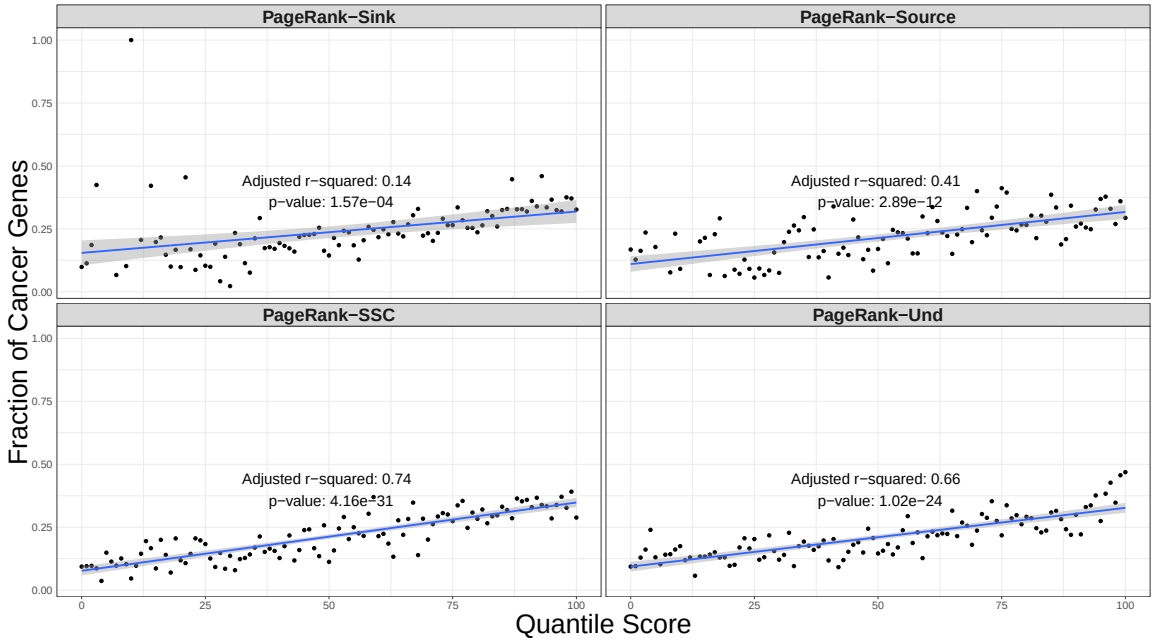

Figure 5.6: Linear regression of quantile-scores versus the ratio of cancer-related genes for PageRank centrality model. X-axis represents the quantile-scores generated by Formula 5.25 Y-axis represents the fraction of all genes that are cancer related (Formula 5.26). The blue line represent the regression line from Formula 5.27. The gray band represent 95 confidence interval of the linear regression.

The dense bands at Source values of zero in Figure 5.7 exhibit the instances which the Source variation is unable to assign importance to nodes. Similarly, the Sink component assigns zero values to many nodes which have higher importance by the other models (the dense band at Pgr-Sink = 0). The Source Component and the Sink component values of PageRank centrality do not show any strong correlation, which indicates that they produce different characterization of graphs. Unlike Laplacian, the Source/Sink and the Undirected variations of PageRank exhibit some considerable correlation. Interestingly, nodes with low Source/Sink PageRank would not have a high Undirected PageRank, but the reverse of this relationship does not exist (Figure 5.7).

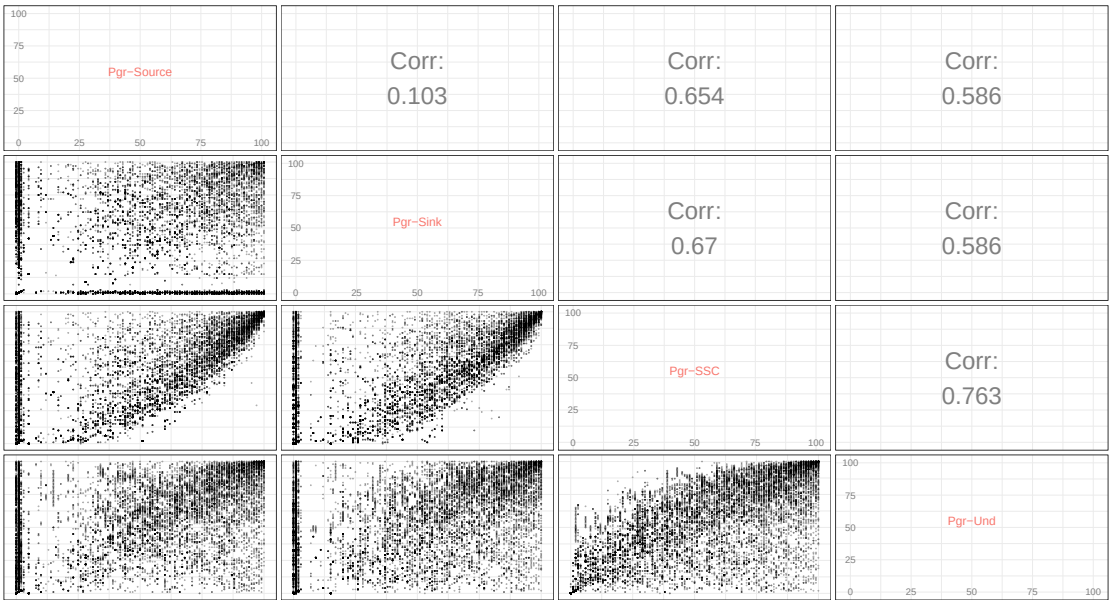

Figure 5.7: Correlation of quantile scores of all genes in human pathways between variations of PageRank centrality model and Degree centrality.

5.3.2 Comparison of CDF's

The analysis of cumulative density function (CDF) of quantile scores outlines the differences between scoring of cancer-related and normal genes (Figure 5.8). For all of the centrality models and their variations, the (CDF) of cancer genes lies below that of normal genes. This observation is supported by strong evidence from Kolmogorov-Smirnov test as displayed in Figure 5.8.

The null hypothesis of normal genes and cancer genes having the same distribution is rejected in favor of the alternative hypothesis that the CDF of cancer genes lies below that of non-cancer genes– p-values provided in Figure 5.8. The difference between CDFs shows that cancer-genes tend to have higher centrality overall. Also, the Kolmogorov-Smirnov test provides an evidence that Source/Sink PageRank creates a better separation between the two class compared to the other models (Figure 5.9). In particular, the test suggest that the distribution of quantile-scores produced by Source/Sink PageRank for cancer-related lies below that of any other model (p-values in Table 5.2).

Table 5.1: Linear regression fit of quantile scores and the ratio of cancer genes

Centrality	term	estimate	std.error	statistic	p.value
Degree	(Intercept)	1.49e-01	1.46e-02	1.02e+01	4.34e-17
Degree	Coefficient	1.45e-03	2.53e-04	5.73e+00	1.07e-07
Katz-Sink	(Intercept)	2.54e-01	2.33e-02	1.09e+01	2.26e-18
Katz-Sink	Coefficient	1.43e-04	3.95e-04	3.61e-01	7.19e-01
Katz-Source	(Intercept)	2.79e-01	2.48e-02	1.12e+01	5.69e-19
Katz-Source	Coefficient	-4.86e-04	4.14e-04	-1.17e+00	2.44e-01
Katz-Source/Sink	(Intercept)	1.46e-01	1.09e-02	1.34e+01	6.50e-24
Katz-Source/Sink	Coefficient	1.42e-03	1.88e-04	7.54e+00	2.29e-11
Lap-Sink	(Intercept)	2.57e-01	2.38e-02	1.08e+01	2.69e-18
Lap-Sink	Coefficient	9.42e-05	4.05e-04	2.32e-01	8.17e-01
Lap-Source	(Intercept)	2.84e-01	2.44e-02	1.16e+01	1.02e-19
Lap-Source	Coefficient	-5.99e-04	4.05e-04	-1.48e+00	1.42e-01
Lap-SSC	(Intercept)	1.26e-01	1.18e-02	1.07e+01	3.18e-18
Lap-SSC	Coefficient	1.86e-03	2.04e-04	9.10e+00	9.87e-15
Lap-Und	(Intercept)	2.19e-01	1.08e-02	2.03e+01	5.35e-37
Lap-Und	Coefficient	2.47e-05	1.86e-04	1.32e-01	8.95e-01
PageRank-Sink	(Intercept)	1.54e-01	2.49e-02	6.21e+00	1.45e-08
PageRank-Sink	Coefficient	1.64e-03	4.17e-04	3.94e+00	1.57e-04
PageRank-Source	(Intercept)	1.10e-01	1.54e-02	7.14e+00	2.11e-10
PageRank-Source	Coefficient	2.07e-03	2.57e-04	8.05e+00	2.89e-12
PageRank-SSC	(Intercept)	7.69e-02	9.24e-03	8.33e+00	4.80e-13
PageRank-SSC	Coefficient	2.71e-03	1.60e-04	1.70e+01	4.16e-31
PageRank-Und	(Intercept)	9.35e-02	9.82e-03	9.52e+00	1.22e-15
PageRank-Und	Coefficient	2.34e-03	1.70e-04	1.38e+01	1.02e-24

5.3.3 Pathway-wise Two-Sample Testing

Pathway by pathway analysis outlines the utility of each centrality method for distinguishing between Cancer-related (cancer) genes and non-cancer-related (non-cancer) genes. In Tables 5.3 and 5.4, the diagonal elements indicate the number of pathways with higher mean centrality of cancer-related genes (rejected hypothesis) for each model. The off-diagonal entries indicate the number of rejected hypothesis by both models that correspond to the row and the column.

Under the normal distribution assumption, each method rejects some null hypotheses of cancer genes having the same mean with the non-cancer genes (Alternative: greater mean for cancer, $FDR < 0.05$). In particular, Source/Sink Katz and Degree

Table 5.2: Kolmogorov Smirnov test of CDF of cancer genes for contrasting PageRank Source/Sink with other centrality measures

Centrality models	H_a : Pgr.SSC above ks.p-values	H_a : Pgr.SSC below ks.p-values
katz.source	1	9.5e-48
katz.sink	1	5e-15
katz.ssc	1	6.1e-31
degree	0.99	2.1e-31
pgr.source	1	4.8e-12
pgr.sink	0.6	9.7e-05
pgr.und	0.57	0.00063
lap.source	1	1.7e-48
lap.sink	1	4.1e-15
lap.ssc	1	3.4e-12
lap.und	0.76	8.2e-54

identify five pathways each, having four of them common between both and no overlap with Sink PageRank or Source/Sink PageRank. On the other hand, Source Katz centrality only identifies two pathways, both identified by degree and Source/Sink Katz. Sink Katz identifies six pathways. Similarly, Source/Sink PageRank and Undirected PageRank identify five and eight pathways, with only one pathway in common. Tha laplacian family shows the highest statistical power. In particular, the Sink, the Source, the Source/Sink, and Undirected Laplacian centralities identify 14, 8, 17, and 17 pathways.

Using Wilcox rank sum test increases the number of rejected hypotheses. For all methods, except Laplacian family, the number of rejected null hypotheses ($FDR < 0.05$) increases (Table 5.4). Source/Sink PageRank and Undirected PageRank show the highest statistical power by detecting 32 and 30 pathways. The overlap between Undirected and Source/Sink PageRank is limited to 17 pathways, showing that two methods produce a considerable number of different pathways.

5.4 Discussion

This chapter investigated the explanatory power of different centrality models with respect to cancer genes. The analysis showed the differences between topological position of cancer and non-cancer genes. In particular, our findings assert three topological properties of cancer-related genes in human biological pathways.

The number of connections (Degree centrality) of a gene in a biological pathway is related to its probability being cancer-related. Regression analysis supports this hypothesis by finding a statistically significant linear relationship between quantile-transformed degree centrality and the ratio of cancer genes. This result indicates that cancer-related genes tend to have higher degree in the organization of biological pathways.

Regression analysis also shows that spectral importance determines the ratio of cancer genes, particularly, when formulated in Source/Sink modeling. Using individual source or sink components of directed Katz and directed Laplacian produces no evidence for linear relation of centrality with the ratio of cancer genes. When the importance is measured only in Source or Sink directions, many of the cancer genes are given low importance– as demonstrated in the correlation plots 5.3, 5.5, and 5.7. However, when Katz and Laplacian centralities are measured and in Source/Sink formulation, the linear relationship becomes statistically significant (Table 5.1). This improvement is particularly because of assigning centrality values to nodes that are terminal but topologically important as receivers of information.

Similarly, Source/Sink PageRank produces a stronger R-squared (0.74) compared to Undirected, Source, and Sink PageRank (Figure 5.6). The higher adjusted R^2 and regression coefficient of Source/Sink compared to Undirected PageRank can be because of SS-PageRank being sensitive to the organization of the directions in the network. It turns out that for every one of the centrality models the adjusted R^2 and the slope of the linear regression coefficient increase when using the Source/Sink

framework. This is highly consistent with our hypothesis of the organization of genes in the pathways. This is highly consistent with our research hypothesis that the Source/Sink concept can capture the topological organization of important genes in pathways.

Kolmogorov-Smirnov test shows that each centrality produces some differentiation between the centrality of cancer genes and non-cancer genes. Although each model has its own specific underlying distribution of centrality, PageRank Source/Sink shows the highest distinction between the group of genes (Table 5.2 and Figure 5.9). This indicates that, overall, Source/Sink PageRank assigns a higher non-parametric quantile to cancer-related genes compared to the other models. However, one has to note that the differences between the centrality models might be subtle and each of their scoring might be superior to the others in certain ranges over the distribution. Standard applications in computational biology, particularly PEMs, either use undirected measures or directed graph measures where disregards terminal nodes – e.g. [19, 22, 56]. The presented results show that using directions while giving importance to terminal nodes in pathways may give higher explanatory power. This results might be of particular interest to the research in evolutionary organization of biological networks and pathways.

Pathway-by-pathway analysis shows that the Source/Sink PageRank has the highest statistical power to distinguish cancer genes from non-cancer in pathways (Tables 5.3 and 5.4). Although all of the methods have some overlap with each other, their differences indicate the uniqueness of each centrality model's evidence for distinguishing cancer genes from non-cancer. Higher statistical power of non-parametric test is because the underlying distribution of the centrality scores is often non-normal. Also, in the non-parametric approach, the Source/Sink variations show stronger statistical power compared to the other alternatives for Katz and PageRank. This suggest that further analysis of the underlying distributions of centrality score may reveal useful

insight for leveraging centrality models and finding more descriptive transformations. We have provided kernel density plots of normalized centrality scores of some of the models in the Supplementary Figures B.2, B.3, B.4, and B.5 for interested readers.

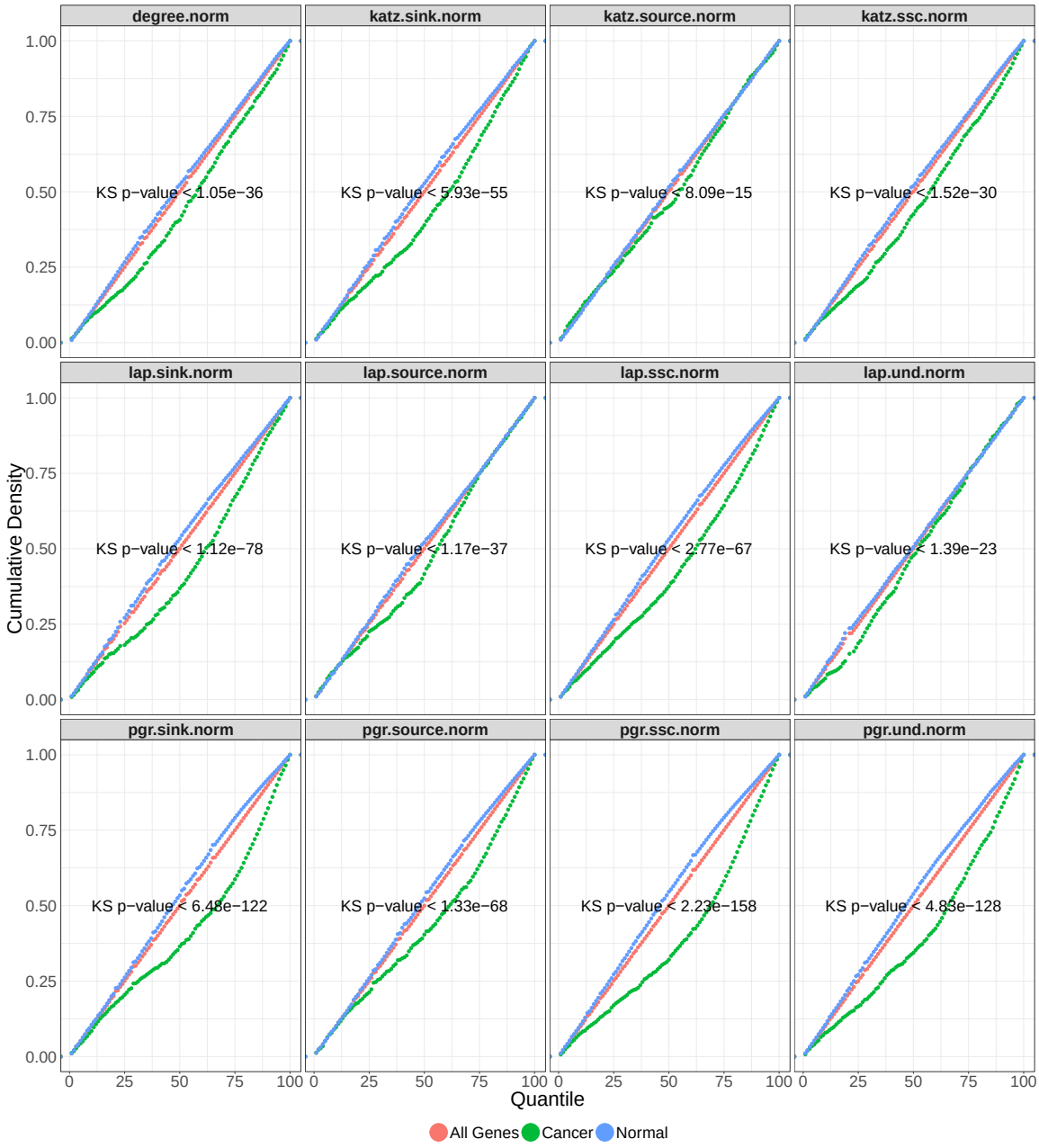

Figure 5.8: Comparison of cumulative density between cancer-related genes and normal genes. The data points represent the quantile-scores calculated based on normalized centrality (Formula 5.28) across all pathways.

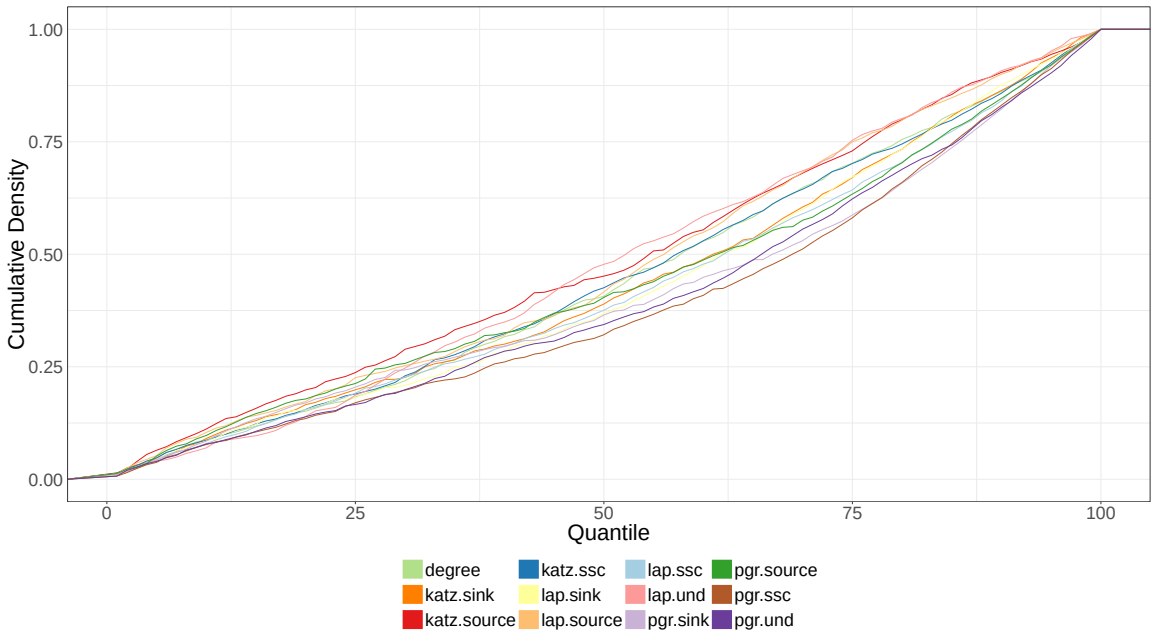

Figure 5.9: Kolmogorov-Smirnov test for CDF of cancer genes with alternative hypothesis of the CDF of the Source/Sink PageRank lying below that of other centralities.

Table 5.3: Pathways identified with higher mean centrality for cancer genes by t-test

	Deg	Katz			Lap				Pgr			
	Deg	Sink	Src	SSC	Sink	Src	SSC	Und	Src	Snk	SSC	Und
Degree	5	0	2	4	0	2	1	0	1	0	0	2
Katz-Snk		6	0	1	5	0	1	0	0	2	2	1
Katz-Src			2	2	0	2	1	0	1	0	0	0
Katz-SSC				5	1	2	1	0	1	0	0	1
Lap-Snk					14	0	4	1	0	3	3	1
Lap-Src						8	3	1	1	0	0	0
Lap-SSC							17	4	1	2	2	1
Lap-Und								17	0	1	1	2
Pgr-Src									2	1	1	1
Pgr-Snk										5	3	1
Pgr-SSC											5	2
Pgr-Und												8

Table 5.4: Pathways identified with higher mean centrality for cancer genes by Wilcox test

	Deg	Katz			Lap				Pgr			
	Deg	Sink	Src	SSC	Sink	Src	SSC	Und	Src	Snk	SSC	Und
Degree	10	2	4	10	2	2	2	0	3	2	4	10
Katz-Sink		13	1	3	10	1	3	1	2	12	11	9
Katz-Source			8	5	1	5	2	0	5	3	3	7
Katz-SSC				14	3	3	2	0	4	4	6	12
Lap-Sink					19	2	6	3	2	13	12	8
Lap-Source						8	4	0	7	4	4	5
Lap-SSC							16	2	6	7	7	4
Lap-Und								5	1	2	4	2
Pgr-Source									15	7	11	10
Pgr-Sink										24	20	14
Pgr-SSC											32	17
Pgr-Und												30

CHAPTER 6: Summary and Future Directions

In this dissertation, we explored devising a new perspective for pathway enrichment analysis by focusing on the shortcomings of the existing models. In particular, we argued that existing models fail to capture the topological importance of the genes, particularly with respect to the upstream-downstream organization of pathways. We introduced a novel graph methodology, Source/Sink centrality, to effectively capture the organization of a pathway and used it to derive topological statistic from differential expressions. We then used this topological statistic in combination with the classical over-representation analysis to create a sensitive network-based PEM, CADIA. CADIA takes an unordered list of DEG and produces a list of p-value that can determine the enrichment of pathways.

Through experimental data validation on three cancer datasets, we showed that CADIA is able to uniquely detect critical pathway enrichments while other standard and the state-of-the-art PEM fail to do so. By synthetic data evaluation, we showed the high specificity of CADIA, which indicates its unique pathway enrichments are not results of false discovery. As the pathways data collections grow and become more complete, CADIA will be able to produce more precise and effective inferences. The presented methodology can contribute to the applications of drug target discovery and biomarker discovery, as it concerns pathway analysis with respect to the underlying topology.

We sought exploratory approaches to show the utility of the Source/Sink concept in differentiating the topological organization of a particular class of important genes in biological pathways. The existing graph-theory approaches for the analysis of biological networks only consider one of the Source and the Sink components in their

methodologies. From multiple aspects, we showed that the Source/Sink extensions of different standard centrality models have increased statistical power to attribute higher importance to cancer genes in the human biological pathways. Although these results are not definitive for the choice of appropriate network modeling for PEM, we have evidently shown that Source/Sink concept is superior in various aspects and can be applied to increase sensitivity to the underlying biological patterns. The choice of underlying topological model for a pathway analysis model depends of several assumptions as well as the source of input data. The results presented in this dissertation are also useful for the researches in identifying the evolutionary patterns in the topology of the biological networks.

Although significant efforts have been made in nearly two decades of research on biological inference models, there are still some open areas that can lead to further improvement. In our opinion, further improvements in the following areas will lead to more comprehensive and informative pathway enrichment analysis solutions.

- Moving towards differential network biology: There is a strong need for a standalone pathway-level statistics that describes the changes of certain biological function across different experimental conditions [103]. These hypothetical statistics should enable to move beyond individual gene-level changes towards pathway-level changes as primary targets of therapeutics and diagnostics. Pathway analysis methods such as ORA, GSA, SPIA, Cdist, EnrichNet, and CADIA all rely on a global information for analyzing pathway-level changes. One can imagine a point in the future where there exist specialized diagnostic panels that only measure the differential expression of few genes and produce critical medical insights. In such a case, it is not unreasonable to assume some pathway driven insight may play a key role in identifying diagnostic panels. A promising start point of having a pathway level statistics will be a framework similar to NetGSA, upon the condition of re-evaluating the mechanisms of capturing the

topology of the graphs. In such a scenario, we believe that the framework of Source/Sink Centrality may prove beneficial and informative.

- The work presented here can be extended to different organisms and different classes of genes. It will be valuable to see whether the presented results could be replicated for different sets of biological pathways from different databases. Needless to say that any attempt at these questions requires appropriate level of availability and consistency in the software and annotations.

REFERENCES

[1] P. Khatri, M. Sirota, and A. J. Butte, "Ten years of pathway analysis: current approaches and outstanding challenges," *PLoS computational biology*, vol. 8, no. 2, p. e1002375, 2012.

[2] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, p. 57, 2009.

[3] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, T. P. Speed, *et al.*, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[4] D. K. Slonim and I. Yanai, "Getting started in gene expression microarray analysis," *PLoS Comput Biol*, vol. 5, no. 10, p. e1000543, 2009.

[5] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences*, vol. 98, no. 9, pp. 5116–5121, 2001.

[6] T. Werner, "Bioinformatics applications for pathway analysis of microarray data," *Current opinion in biotechnology*, vol. 19, no. 1, pp. 50–54, 2008.

[7] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[8] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[9] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "Kegg: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Research*, vol. 45, no. D1, pp. D353–D361, 2017.

[10] G. O. Consortium *et al.*, "The gene ontology (go) database and informatics resource," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D258–D261, 2004.

[11] R. Mathur, D. Rotroff, J. Ma, A. Shojaie, and A. Motsinger-Reif, "Gene set analysis methods: a systematic comparison," *BioData mining*, vol. 11, no. 1, p. 8, 2018.

[12] C. Mitrea, Z. Taghavi, B. Bokanizad, S. Hanoudi, R. Tagett, M. Donato, C. Voichita, and S. Draghici, "Methods and approaches in the topology-based analysis of biological pathways," *Frontiers in physiology*, vol. 4, p. 278, 2013.

[13] I. Rivals, L. Personnaz, L. Taing, and M.-C. Potier, "Enrichment or depletion of a go category within a class of genes: which test?," *Bioinformatics*, vol. 23, no. 4, pp. 401–407, 2006.

[14] R. K. Curtis, M. Orešič, and A. Vidal-Puig, "Pathways to the analysis of microarray data," *TRENDS in Biotechnology*, vol. 23, no. 8, pp. 429–435, 2005.

[15] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[16] U. Alon, "Network motifs: theory and experimental approaches," *Nature Reviews Genetics*, vol. 8, no. 6, pp. 450–461, 2007.

[17] V. Janjić, R. Sharan, and N. Pržulj, "Modelling the yeast interactome," *Scientific reports*, vol. 4, 2014.

[18] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *cell*, vol. 144, no. 5, pp. 646–674, 2011.

[19] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J.-s. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero, "A novel signaling pathway impact analysis," *Bioinformatics*, vol. 25, no. 1, pp. 75–82, 2009.

[20] Z. Gu, J. Liu, K. Cao, J. Zhang, and J. Wang, "Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes," *BMC systems biology*, vol. 6, no. 1, p. 56, 2012.

[21] P. Naderi Yeganeh and M. T. Mostafavi, "Use of structural properties of underlying graphs in pathway enrichment analysis of genomic data," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 279–284, ACM, 2017.

[22] E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, and A. Valencia, "Enrichnet: network-based gene set enrichment analysis," *Bioinformatics*, vol. 28, no. 18, p. i451, 2012.

[23] A. Shojaie and G. Michailidis, "Analysis of gene sets based on the underlying regulatory network," *Journal of Computational Biology*, vol. 16, no. 3, pp. 407–426, 2009.

[24] A. Shojaie and G. Michailidis, "Network enrichment analysis in complex experiments," *Statistical applications in genetics and molecular biology*, vol. 9, no. 1, 2010.

[25] J. Ma, A. Shojaie, and G. Michailidis, "Network-based pathway enrichment analysis with incomplete network information," *Bioinformatics*, vol. 32, no. 20, pp. 3165–3174, 2016.

[26] P. Naderi Yeganeh and M. T. Mostafavi, "Causal disturbance analysis: A novel graph centrality based method for pathway enrichment analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, 2019.

[27] P. N. Yeganeh, E. Saule, and M. T. Mostafavi, "Centrality of cancer-related genes in human biological pathways: A graph analysis perspective," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 214–218, IEEE, 2018.

[28] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[29] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical applications in genetics and molecular biology*, vol. 3, no. 1, pp. 1–25, 2004.

[30] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome biology*, vol. 11, no. 10, p. R106, 2010.

[31] G. A. Churchill, "Using anova to analyze microarray data.," *Biotechniques*, vol. 37, no. 2, pp. 173–5, 2004.

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

[33] J. D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 3, pp. 479–498, 2002.

[34] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of statistics*, pp. 1165–1188, 2001.

[35] J. J. Goeman and A. Solari, "Multiple hypothesis testing in genomics," *Statistics in medicine*, vol. 33, no. 11, pp. 1946–1978, 2014.

[36] C. Wu, C. Orozco, J. Boyer, M. Leglise, J. Goodale, S. Batalov, C. L. Hodge, J. Haase, J. Janes, J. W. Huss, *et al.*, "Biogps: an extensible and customizable portal for querying and organizing gene annotation resources," *Genome Biol*, vol. 10, no. 11, p. R130, 2009.

[37] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, *et al.*, "The string database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D561–D568, 2011.

[38] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews, *et al.*, "Reactome: a knowledgebase of biological pathways," *Nucleic acids research*, vol. 33, no. suppl_1, pp. D428–D432, 2005.

[39] D. L. Nelson, A. L. Lehninger, and M. M. Cox, *Lehninger principles of biochemistry*. Macmillan, 2008.

[40] H. Kitano, "Systems biology: a brief overview," *Science*, vol. 295, no. 5560, pp. 1662–1664, 2002.

[41] S. Zhao and R. Iyengar, "Systems pharmacology: network analysis to identify multiscale mechanisms of drug action," *Annual review of pharmacology and toxicology*, vol. 52, p. 505, 2012.

[42] S. I. Berger and R. Iyengar, "Network analyses in systems pharmacology," *Bioinformatics*, vol. 25, no. 19, pp. 2466–2472, 2009.

[43] Y. Yarden and G. Pines, "The erbb network: at last, cancer therapy meets systems biology," *Nature reviews Cancer*, vol. 12, no. 8, pp. 553–563, 2012.

[44] M. A. Swartz, N. Iida, E. W. Roberts, S. Sangaletti, M. H. Wong, F. E. Yull, L. M. Coussens, and Y. A. DeClerck, "Tumor microenvironment complexity: emerging roles in cancer therapy," *Cancer research*, vol. 72, no. 10, pp. 2473–2480, 2012.

[45] R. A. Irizarry, C. Wang, Y. Zhou, and T. P. Speed, "Gene set enrichment analysis made simple," *Statistical methods in medical research*, vol. 18, no. 6, pp. 565–575, 2009.

[46] C. Backes, A. Keller, J. Kuentzer, B. Kneissl, N. Comtesse, Y. A. Elnakady, R. Müller, E. Meese, and H.-P. Lenhof, "Genetrail–advanced gene set enrichment analysis," *Nucleic acids research*, vol. 35, no. suppl 2, pp. W186–W192, 2007.

[47] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic acids research*, vol. 37, no. 1, pp. 1–13, 2009.

[48] Y. Liu, M. Koyutürk, J. S. Barnholtz-Sloan, and M. R. Chance, "Gene interaction enrichment and network analysis to identify dysregulated pathways and their interactions in complex diseases," *BMC systems biology*, vol. 6, no. 1, p. 65, 2012.

[49] J. Han, X. Shi, Y. Zhang, Y. Xu, Y. Jiang, C. Zhang, L. Feng, H. Yang, D. Shang, Z. Sun, *et al.*, "Esea: discovering the dysregulated pathways based on edge set enrichment analysis," *Scientific reports*, vol. 5, p. 13044, 2015.

[50] P. Tamayo, G. Steinhardt, A. Liberzon, and J. P. Mesirov, "The limitations of simple gene set enrichment analysis assuming gene independence," *Statistical methods in medical research*, vol. 25, no. 1, pp. 472–487, 2016.

[51] A. Chatr-Aryamontri, B.-J. Breitkreutz, S. Heinicke, L. Boucher, A. Winter, C. Stark, J. Nixon, L. Ramage, N. Kolas, L. O'Donnell, *et al.*, "The biogrid interaction database: 2013 update," *Nucleic acids research*, vol. 41, no. D1, pp. D816–D823, 2013.

[52] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics*, vol. 18, no. suppl 1, pp. S233–S240, 2002.

[53] W. Luo, M. S. Friedman, K. Shedden, K. D. Hankenson, and P. J. Woolf, "Gage: generally applicable gene set enrichment for pathway analysis," *BMC bioinformatics*, vol. 10, no. 1, p. 161, 2009.

[54] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

[55] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of escherichia coli," *Nature genetics*, vol. 31, no. 1, pp. 64–68, 2002.

[56] F. Vandin, E. Upfal, and B. J. Raphael, "Algorithms for detecting significantly mutated pathways in cancer," *Journal of Computational Biology*, vol. 18, no. 3, pp. 507–522, 2011.

[57] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[58] J.-D. J. Han, N. Bertin, H. Tong, D. S. Goldberg, *et al.*, "Evidence for dynamically organized modularity in the yeast protein-protein interaction network," *Nature*, vol. 430, no. 6995, p. 88, 2004.

[59] M. R. Said, T. J. Begley, A. V. Oppenheim, D. A. Lauffenburger, and L. D. Samson, "Global network analysis of phenotypic effects: protein networks and toxicity modulation in saccharomyces cerevisiae," *Proceedings of the National Academy of Sciences*, vol. 101, no. 52, pp. 18006–18011, 2004.

[60] R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, and T. Ideker, "Conserved patterns of protein interaction in multiple species," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 6, pp. 1974–1979, 2005.

[61] H. Yu, D. Greenbaum, H. X. Lu, X. Zhu, and M. Gerstein, "Genomic analysis of essentiality within protein networks," *TRENDS in genetics*, vol. 20, no. 6, pp. 227–231, 2004.

[62] X. He and J. Zhang, "Why do hubs tend to be essential in protein networks," *PLoS Genet*, vol. 2, no. 6, p. e88, 2006.

[63] R. B. D'agostino, W. Chase, and A. Belanger, "The appropriateness of some common procedures for testing the equality of two independent binomial populations," *The American Statistician*, vol. 42, no. 3, pp. 198–202, 1988.

[64] J. H. McDonald, *Handbook of biological statistics*, vol. 2. sparky house publishing Baltimore, MD, 2009.

[65] M. Newman, *Networks: an introduction*. Oxford University Press, 2010.

[66] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *The American Journal of Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.

[67] F. Chung, "The heat kernel as the pagerank of a graph," *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, pp. 19735–19740, 2007.

[68] M. Volm, W. Rittgen, and P. Drings, "Prognostic value of erbb-1, vegf, cyclin a, fos, jun and myc in patients with squamous cell lung carcinomas," *British journal of cancer*, vol. 77, no. 4, p. 663, 1998.

[69] P. Bonacich and P. Lloyd, "Eigenvector-like measures of centrality for asymmetric relations," *Social networks*, vol. 23, no. 3, pp. 191–201, 2001.

[70] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[71] C. D. Meyer, *Matrix analysis and applied linear algebra*, vol. 71. Siam, 2000.

[72] A. Birnbaum, "Combining independent tests of significance," *Journal of the American Statistical Association*, vol. 49, no. 267, pp. 559–574, 1954.

[73] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall, *et al.*, "Ncbi geo: archive for high-throughput functional genomic data," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D885–D890, 2009.

[74] M. S. Anglesio, J. M. Arnold, J. George, A. V. Tinker, R. Tothill, N. Waddell, L. Simms, B. Locandro, S. Fereday, N. Traficante, *et al.*, "Mutation of erbb2 provides a novel alternative mechanism for the ubiquitous activation of ras-mapk in ovarian serous low malignant potential tumors," *Molecular cancer research*, vol. 6, no. 11, pp. 1678–1690, 2008.

[75] S. Tsukamoto, T. Ishikawa, S. Iida, M. Ishiguro, K. Mogushi, H. Mizushima, H. Uetake, H. Tanaka, and K. Sugihara, "Clinical significance of osteoprotegerin expression in human colorectal cancer," *Clinical cancer research*, vol. 17, no. 8, pp. 2444–2450, 2011.

[76] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for rna-sequencing and microarray studies," *Nucleic acids research*, vol. 43, no. 7, pp. e47–e47, 2015.

[77] J. D. Zhang and S. Wiemann, "Kegggraph: a graph approach to kegg pathway in r and bioconductor," *Bioinformatics*, vol. 25, no. 11, pp. 1470–1471, 2009.

[78] A. Sergushichev, "An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation," *BioRxiv*, p. 060012, 2016.

[79] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, *et al.*, "Bioconductor: open software development for computational biology and bioinformatics," *Genome biology*, vol. 5, no. 10, p. R80, 2004.

[80] Y. Yarden and M. X. Sliwkowski, "Untangling the erbb signalling network," *Nature reviews Molecular cell biology*, vol. 2, no. 2, pp. 127–137, 2001.

[81] P. N. Yeganeh, C. Richardson, Z. Bahrani-Mostafavi, D. L. Tait, and M. T. Mostafavi, "Dysregulation of akt3 along with a small panel of mrnas stratifies high-grade serous ovarian cancer from both normal epithelia and benign tumor tissues," *Genes & cancer*, vol. 8, no. 11-12, pp. 784–798, 2017.

[82] J. Downward, "Targeting ras signalling pathways in cancer therapy," *Nature Reviews Cancer*, vol. 3, no. 1, p. 11, 2003.

[83] R. C. Bast Jr, B. Hennessy, and G. B. Mills, "The biology of ovarian cancer: new opportunities for translation," *Nature Reviews Cancer*, vol. 9, no. 6, p. 415, 2009.

[84] J. Luo, B. D. Manning, and L. C. Cantley, "Targeting the pi3k-akt pathway in human cancer," *Cancer cell*, vol. 4, no. 4, pp. 257–262, 2003.

[85] K. D. Courtney, R. B. Corcoran, and J. A. Engelman, "The pi3k pathway as drug target in human cancer," *Journal of clinical oncology*, vol. 28, no. 6, p. 1075, 2010.

[86] A. De Luca, M. R. Maiello, A. D'Alessio, M. Pergameno, and N. Normanno, "The ras/raf/mek/erk and the pi3k/akt signalling pathways: role in cancer pathogenesis and implications for therapeutic approaches," *Expert opinion on therapeutic targets*, vol. 16, no. sup2, pp. S17–S27, 2012.

[87] P. N. Yeganeh and M. T. Mostafavi, "Use of machine learning for diagnosis of cancer in ovarian tissues with a selected mrna panel," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2429–2434, IEEE, 2018.

[88] A. K. Sood, J. E. Coffin, G. B. Schneider, M. S. Fletcher, B. R. DeYoung, L. M. Gruman, D. M. Gershenson, M. D. Schaller, and M. J. Hendrix, "Biological significance of focal adhesion kinase in ovarian cancer: role in migration and invasion," *The American journal of pathology*, vol. 165, no. 4, pp. 1087–1095, 2004.

[89] A. A. Adjei, "Blocking oncogenic ras signaling for cancer therapy," *Journal of the National Cancer Institute*, vol. 93, no. 14, pp. 1062–1074, 2001.

[90] B. Vogelstein and K. W. Kinzler, "Cancer genes and the pathways they control," *Nature medicine*, vol. 10, no. 8, p. 789, 2004.

[91] D. Pan, "The hippo signaling pathway in development and cancer," *Developmental cell*, vol. 19, no. 4, pp. 491–505, 2010.

[92] E. Carlsson, A. Ranki, L. Sipilä, L. Karenko, W. Abdel-Rahman, K. Ovaska, L. Siggberg, U. Aapola, R. Ässämäki, V. Häyry, *et al.*, "Potential role of a navigator gene nav3 in colorectal cancer," *British journal of cancer*, vol. 106, no. 3, p. 517, 2012.

[93] D. W. Kang, K.-Y. Choi, *et al.*, "Phospholipase d meets wnt signaling: a new target for cancer therapy," *Cancer research*, vol. 71, no. 2, pp. 293–297, 2011.

[94] D. W. Kang, B. H. Lee, Y.-A. Suh, Y.-S. Choi, S. J. Jang, Y. M. Kim, K.-Y. Choi, *et al.*, "Phospholipase d1 inhibition linked to upregulation of icat blocks colorectal cancer growth hyperactivated by wnt/β-catenin and pi3k/akt signaling," *Clinical Cancer Research*, vol. 23, no. 23, pp. 7340–7350, 2017.

[95] J. Mao, S. Fan, W. Ma, P. Fan, B. Wang, J. Zhang, H. Wang, B. Tang, Q. Zhang, X. Yu, *et al.*, "Roles of wnt/β-catenin signaling in the gastric cancer stem cells proliferation and salinomycin treatment," *Cell death & disease*, vol. 5, no. 1, p. e1039, 2015.

[96] M. Alhamdoosh, M. Ng, N. J. Wilson, J. M. Sheridan, H. Huynh, M. J. Wilson, and M. E. Ritchie, "Combining multiple tools outperforms individual methods in gene set enrichment analyses," *Bioinformatics*, vol. 33, no. 3, pp. 414–424, 2017.

[97] E. Zotenko, J. Mestre, D. P. O'leary, and T. M. Przytycka, "Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality," *PLoS Comput Biol*, vol. 4, no. 8, p. e1000140, 2008.

[98] R. I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete input spaces," in *ICML*, vol. 2, pp. 315–322, 2002.

[99] F. Chung, "Laplacians and the cheeger inequality for directed graphs," *Annals of Combinatorics*, vol. 9, no. 1, pp. 1–19, 2005.

[100] F. Bauer, "Normalized graph laplacians for directed graphs," *Linear Algebra and its Applications*, vol. 436, no. 11, pp. 4193–4222, 2012.

[101] W. Luo and C. Brouwer, "Pathview: an r/bioconductor package for pathway-based data integration and visualization," *Bioinformatics*, vol. 29, no. 14, pp. 1830–1831, 2013.

[102] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, "A census of human cancer genes," *Nature Reviews Cancer*, vol. 4, no. 3, p. 177, 2004.

[103] T. Ideker and N. J. Krogan, "Differential network biology," *Molecular systems biology*, vol. 8, no. 1, p. 565, 2012.

APPENDIX A: Additional notes and proofs

A goal of Source/Sink centrality is to be as distinct as possible from the Source component and the Sink component. Intuitively, the parameter choice of $\beta = 1$ should create the most distinction.

It is possible to derive the optimal value for β. Recall the Source and Sink components from as:

$$C_{src} = (I - \alpha A)^{-1}.\mathbb{1}_{n\times 1}$$

and

$$C_{sink} = (I - \alpha A^T)^{-1}.\mathbb{1}_{n\times 1}$$

Without loss of generalization, define the Source/Sink centrality as

$$C_{ssc} = \frac{1}{1+\beta}C_{src} + \frac{\beta}{1+\beta}C_{sink} \tag{A.1}$$

The above formulation does not change any calculations in CADIA. In particular, the division by non-zero constant $1+\beta$ does not affect the calculation of the aggregate score for pathway enrichment.

Proof. As $Agg(U)$ is defined in the main document, the P_{ssc} of any subset of nodes, U, is invariant under positive scaling of $Agg(U)$. Define $r > 0$, then the following holds true:

$$\begin{aligned} P_{ssc} =& \mathbb{P}\Big\{ Agg(U) > Agg(U_0) \;\Big|\; |U| = |U_0| \Big\} \\ =& \mathbb{P}\Big\{ r \cdot Agg(U) > r \cdot Agg(U_0) \;\Big|\; |U| = |U_0| \Big\} \end{aligned} \tag{A.2}$$

The above equations hold because for any positive scalar r:

$$Agg(U) > Agg(U_0) \Leftrightarrow r \cdot Agg(U) > r \cdot Agg(U_0) \tag{A.3}$$

Let C'_{ssc} denote any positive scaling of Source/Sink centrality ($C'_{ssc} = s \cdot C_{ssc},\ s > 0$). Define a new aggregate score as following:

$$Agg_1(U) = \prod_{u_i \in U} C'_{ssc}(u_i) \tag{A.4}$$

Then, the following holds true:

$$\begin{aligned} \prod_{u_i \in U} C'_{ssc}(u_i) &= \prod_{u_i \in U} s \cdot C_{ssc}(u_i) \\ &= s^{|U|} \cdot \prod_{u_i \in U} C_{ssc}(u_i) \\ &= s^{|U|} Agg(U) \end{aligned} \tag{A.5}$$

The above Formula shows that a new aggregate score is a scaling of the original aggregate score. Thus, by the property of invariance, P_{ssc} is equal for both $Agg()$ and $Agg_1()$.

□

Define the distance from Source and Sink as $|| C_{ssc} - C_{src} ||_2$ and $|| C_{ssc} - C_{sink} ||_2$. Here, $|| \ . \ ||_2$ is the L-2 norm of the matrix. Define the distinction of Source/Sink centrality as the product of the distances from Source component and Sink component. Then the problem of finding the optimal β parameter for maximizing the distinction is:

$$\max \quad || C_{ssc} - C_{src} || \ . \ || C_{ssc} - C_{sink} ||$$

$$\text{s.t.} \qquad \beta > 0$$

By plugging the values from Eq.1, the distance of the Source/Sink centrality vector from the Source centrality vector and Sink centrality vector can be written as:

$$\begin{aligned} || \, C_{ssc} - C_{src} \, || = & \quad || \, \frac{\beta}{1+\beta}(-C_{src} + C_{sink}) \, || \\ || \, C_{ssc} - C_{sink} \, || = & \quad || \, \frac{1}{1+\beta}(C_{src} - C_{sink}) \, || \end{aligned} \tag{A.6}$$

The numerical coefficients from the L-2 norm can be extracted. We can multiply the internal by the scalar -1. Then, we have:

$$\begin{aligned} || \, C_{ssc} - C_{src} \, || = & \quad \frac{\beta}{1+\beta} \, || \, (C_{src} - C_{sink}) \, || \\ || \, C_{ssc} - C_{sink} \, || = & \quad \frac{1}{1+\beta} \, || \, (C_{src} - C_{sink}) \, || \end{aligned} \tag{A.7}$$

By substituting the above equations into the optimization problem we have

$$\begin{aligned} & \max \, \frac{\beta}{1+\beta} \frac{1}{1+\beta} \, || \, (C_{src} - C_{sink}) \, ||^2 \\ & s.t. : \beta > 0 \end{aligned} \tag{A.8}$$

$|| \, (C_{src} - C_{sink}) \, ||$ is constant and only depends on the underlying graph. We then optimize by solving for derivative of β.

$$\begin{aligned} \frac{\partial}{\partial \beta} \frac{\beta}{(1+\beta)^2} \, || \, (C_{src} - C_{sink}) \, || &= 0 \\ \frac{-2\beta}{(1+\beta)^3} + \frac{1}{(1+\beta)^2} &= 0 \\ \beta &= 1 \end{aligned} \tag{A.9}$$

One can show that the presented results hold for any matrix norm and are not limited only to L-2 norms. The original notation of Source/Sink centrality, as described in the main document, accepts any β that is a non-negative real number. A

corollary of the representation in Formula A.1 in this supplementary material is that Source/Sink Centrality can be reformatted into a more relatively symmetric representation for computing the aggregate score ($Agg()$). Formally, by defining a variable $z = \frac{1}{1+\beta}$, one can show:

$$C_{ssc} = z \cdot C_{src} + (1 - z) \cdot C_{sink} \tag{A.10}$$

The above notation allows for redefining the tuning parameter into a variable that is in the domain $[0, 1]$. In this case, the optimal distinction between the Source and the Sink components happens at $z = 0.5$.

APPENDIX B: Pathways list and information

	Name	nodes	dges	eigen
1	Glycolysis / Gluconeogenesis	68	277	3.90
2	Citrate cycle (TCA cycle)	30	98	2.45
3	Pentose phosphate pathway	30	154	5.67
4	Pentose and glucuronate interconversions	34	70	1.26
5	Fructose and mannose metabolism	33	137	1.73
6	Galactose metabolism	31	69	0.00
7	Ascorbate and aldarate metabolism	27	38	0.00
8	Fatty acid biosynthesis	13	18	1.62
9	Fatty acid elongation	30	51	3.68
10	Fatty acid degradation	44	172	5.15
11	Synthesis and degradation of ketone bodies	10	32	0.00
12	Steroid biosynthesis	19	31	2.25
13	Primary bile acid biosynthesis	17	29	1.76
14	Ubiquinone and other terpenoid-quinone biosynthesis	11	11	1.41
15	Steroid hormone biosynthesis	59	878	16.60
16	Oxidative phosphorylation	133	132	0.00
17	Arginine biosynthesis	21	45	0.00
18	Purine metabolism	174	5424	42.75
19	Caffeine metabolism	5	10	1.41
20	Pyrimidine metabolism	101	1608	19.91
21	Alanine, aspartate and glutamate metabolism	35	111	1.00
22	Glycine, serine and threonine metabolism	40	135	3.51
23	Cysteine and methionine metabolism	45	127	2.47

Continued on next page

	Name	nodes	dges	eigen
24	Valine, leucine and isoleucine degradation	48	199	3.28
25	Valine, leucine and isoleucine biosynthesis	4	0	0.00
26	Lysine degradation	59	193	5.57
27	Arginine and proline metabolism	50	142	3.78
28	Histidine metabolism	23	33	0.00
29	Tyrosine metabolism	36	185	8.18
30	Phenylalanine metabolism	17	38	0.00
31	Tryptophan metabolism	40	136	4.95
32	Phenylalanine, tyrosine and tryptophan biosynthesis	5	6	1.41
33	beta-Alanine metabolism	31	148	5.65
34	Taurine and hypotaurine metabolism	11	27	0.00
35	Phosphonate and phosphinate metabolism	6	6	0.00
36	Selenocompound metabolism	17	35	2.69
37	D-Glutamine and D-glutamate metabolism	5	6	0.00
38	D-Arginine and D-ornithine metabolism	1	0	0.00
39	Glutathione metabolism	56	606	13.10
40	Starch and sucrose metabolism	36	190	4.79
41	N-Glycan biosynthesis	49	108	0.00
42	Other glycan degradation	18	0	0.00
43	Mucin type O-glycan biosynthesis	31	134	4.49
44	Other types of O-glycan biosynthesis	22	0	0.00
45	Mannose type O-glycan biosynthesis	23	39	1.00
46	Amino sugar and nucleotide sugar metabolism	48	146	5.57
47	Neomycin, kanamycin and gentamicin biosynthesis	5	0	0.00

Continued on next page

	Name	nodes	dges	eigen
48	Glycosaminoglycan degradation	19	29	0.00
49	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	20	31	4.08
50	Glycosaminoglycan biosynthesis - keratan sulfate	14	0	0.00
51	Glycosaminoglycan biosynthesis - heparan sulfate / heparin	24	0	0.00
52	Glycerolipid metabolism	61	826	17.11
53	Inositol phosphate metabolism	73	756	5.08
54	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	25	38	1.00
55	Glycerophospholipid metabolism	97	1157	15.74
56	Ether lipid metabolism	47	522	7.16
57	Arachidonic acid metabolism	63	720	12.84
58	Linoleic acid metabolism	29	233	8.39
59	alpha-Linolenic acid metabolism	25	24	1.00
60	Sphingolipid metabolism	47	485	9.08
61	Glycosphingolipid biosynthesis - lacto and neolacto series	27	193	7.14
62	Glycosphingolipid biosynthesis - globo and isoglobo series	15	29	1.41
63	Glycosphingolipid biosynthesis - ganglio series	15	44	3.22
64	Pyruvate metabolism	39	181	2.00
65	Glyoxylate and dicarboxylate metabolism	30	50	2.33
66	Propanoate metabolism	32	72	0.00
67	Butanoate metabolism	28	72	0.00

Continued on next page

	Name	nodes	dges	eigen
68	One carbon pool by folate	20	150	6.14
69	Thiamine metabolism	16	27	0.00
70	Riboflavin metabolism	8	16	0.00
71	Vitamin B6 metabolism	6	10	2.10
72	Nicotinate and nicotinamide metabolism	30	206	3.54
73	Pantothenate and CoA biosynthesis	19	50	3.62
74	Biotin metabolism	3	1	0.00
75	Lipoic acid metabolism	3	4	1.00
76	Folate biosynthesis	26	83	2.85
77	Retinol metabolism	67	882	12.26
78	Porphyrin and chlorophyll metabolism	42	90	1.73
79	Terpenoid backbone biosynthesis	22	57	3.56
80	Nitrogen metabolism	17	5	0.00
81	Sulfur metabolism	9	14	1.00
82	Aminoacyl-tRNA biosynthesis	66	14	0.00
83	Metabolism of xenobiotics by cytochrome P450	76	1245	26.46
84	Drug metabolism - cytochrome P450	72	550	10.43
85	Drug metabolism - other enzymes	79	209	4.58
86	Biosynthesis of unsaturated fatty acids	23	0	0.00
87	Metabolic pathways	1293	0	0.00
88	Carbon metabolism	116	0	0.00
89	2-Oxocarboxylic acid metabolism	18	0	0.00
90	Fatty acid metabolism	48	0	0.00
91	Biosynthesis of amino acids	74	0	0.00
92	EGFR tyrosine kinase inhibitor resistance	79	229	0.00

Continued on next page

	Name	nodes	dges	eigen
93	Endocrine resistance	98	290	2.47
94	Antifolate resistance	31	12	0.00
95	Platinum drug resistance	73	79	1.17
96	ABC transporters	44	0	0.00
97	Ribosome biogenesis in eukaryotes	105	2	0.00
98	Ribosome	153	0	0.00
99	RNA transport	171	295	0.00
100	mRNA surveillance pathway	91	156	0.00
101	RNA degradation	79	73	0.00
102	RNA polymerase	31	0	0.00
103	Basal transcription factors	45	0	0.00
104	DNA replication	36	0	0.00
105	Spliceosome	134	0	0.00
106	Proteasome	45	0	0.00
107	Protein export	23	0	0.00
108	PPAR signaling pathway	74	271	0.00
109	Base excision repair	33	0	0.00
110	Nucleotide excision repair	47	0	0.00
111	Mismatch repair	23	0	0.00
112	Homologous recombination	41	20	0.00
113	Non-homologous end-joining	13	0	0.00
114	Fanconi anemia pathway	54	83	1.00
115	MAPK signaling pathway	295	1961	1.59
116	ErbB signaling pathway	85	203	0.00
117	Ras signaling pathway	232	1570	0.00

Continued on next page

	Name	nodes	dges	eigen
118	Rap1 signaling pathway	206	1206	2.14
119	Calcium signaling pathway	183	528	0.00
120	cGMP-PKG signaling pathway	163	374	0.00
121	cAMP signaling pathway	198	664	0.00
122	Cytokine-cytokine receptor interaction	294	373	0.00
123	Chemokine signaling pathway	185	1640	0.00
124	NF-kappa B signaling pathway	95	172	2.03
125	HIF-1 signaling pathway	100	280	0.00
126	FoxO signaling pathway	132	433	0.00
127	Phosphatidylinositol signaling system	99	1890	28.58
128	Sphingolipid signaling pathway	118	253	0.00
129	Phospholipase D signaling pathway	146	409	0.00
130	Neuroactive ligand-receptor interaction	277	47	0.00
131	Cell cycle	124	618	11.00
132	Oocyte meiosis	124	424	1.53
133	p53 signaling pathway	72	86	0.00
134	Ubiquitin mediated proteolysis	137	0	0.00
135	Sulfur relay system	8	8	0.00
136	SNARE interactions in vesicular transport	34	42	0.00
137	Autophagy - other	32	59	0.00
138	Mitophagy - animal	65	100	1.00
139	Autophagy - animal	128	335	1.26
140	Protein processing in endoplasmic reticulum	165	78	1.00
141	Lysosome	123	0	0.00
142	Endocytosis	244	233	0.00

Continued on next page

	Name	nodes	dges	eigen
143	Phagosome	152	247	0.00
144	Peroxisome	83	7	0.00
145	mTOR signaling pathway	151	565	0.00
146	PI3K-Akt signaling pathway	354	3089	0.00
147	AMPK signaling pathway	120	318	1.00
148	Apoptosis	136	346	1.67
149	Longevity regulating pathway	89	245	0.00
150	Longevity regulating pathway - multiple species	62	184	0.00
151	Apoptosis - multiple species	33	0	0.00
152	Ferroptosis	40	10	1.00
153	Necroptosis	162	331	0.00
154	Cellular senescence	160	427	0.00
155	Cardiac muscle contraction	78	18	1.32
156	Adrenergic signaling in cardiomyocytes	144	847	0.00
157	Vascular smooth muscle contraction	121	273	0.00
158	Wnt signaling pathway	146	846	0.00
159	Notch signaling pathway	48	142	0.00
160	Hedgehog signaling pathway	47	174	3.61
161	TGF-beta signaling pathway	84	173	0.00
162	Axon guidance	175	523	5.67
163	VEGF signaling pathway	59	158	0.00
164	Apelin signaling pathway	137	847	0.00
165	Osteoclast differentiation	128	288	2.00
166	Hippo signaling pathway	154	589	2.51
167	Hippo signaling pathway - multiple species	29	55	0.00

Continued on next page

	Name	nodes	dges	eigen
168	Focal adhesion	199	1816	3.74
169	ECM-receptor interaction	82	521	0.00
170	Cell adhesion molecules (CAMs)	144	584	21.00
171	Adherens junction	72	170	4.03
172	Tight junction	170	3	0.00
173	Gap junction	88	227	1.00
174	Signaling pathways regulating pluripotency of stem cells	139	434	1.26
175	Complement and coagulation cascades	79	83	1.00
176	Platelet activation	123	280	0.00
177	Antigen processing and presentation	77	373	13.00
178	Renin-angiotensin system	23	1	0.00
179	Toll-like receptor signaling pathway	104	217	0.00
180	NOD-like receptor signaling pathway	168	286	1.00
181	RIG-I-like receptor signaling pathway	70	147	0.00
182	Cytosolic DNA-sensing pathway	63	74	0.00
183	C-type lectin receptor signaling pathway	104	278	1.00
184	JAK-STAT signaling pathway	162	3208	9.12
185	Hematopoietic cell lineage	97	0	0.00
186	Natural killer cell mediated cytotoxicity	131	350	1.00
187	IL-17 signaling pathway	93	14	0.00
188	Th1 and Th2 cell differentiation	92	256	2.89
189	Th17 cell differentiation	107	200	2.04
190	T cell receptor signaling pathway	101	254	1.22
191	B cell receptor signaling pathway	71	137	1.00

Continued on next page

	Name	nodes	dges	eigen
192	Fc epsilon RI signaling pathway	68	151	0.00
193	Fc gamma R-mediated phagocytosis	91	198	0.00
194	TNF signaling pathway	110	139	0.00
195	Leukocyte transendothelial migration	112	792	23.00
196	Intestinal immune network for IgA production	49	18	0.00
197	Circadian rhythm	31	95	3.16
198	Circadian entrainment	96	780	0.00
199	Thermogenesis	229	228	1.53
200	Long-term potentiation	67	339	4.64
201	Synaptic vesicle cycle	63	37	2.00
202	Neurotrophin signaling pathway	119	357	2.81
203	Retrograde endocannabinoid signaling	148	672	0.00
204	Glutamatergic synapse	114	468	3.76
205	Cholinergic synapse	112	506	0.00
206	Serotonergic synapse	115	414	0.00
207	GABAergic synapse	88	506	0.00
208	Dopaminergic synapse	131	599	0.00
209	Long-term depression	60	226	0.00
210	Olfactory transduction	419	4238	0.00
211	Taste transduction	83	32	0.00
212	Phototransduction	28	45	0.00
213	Inflammatory mediator regulation of TRP channels	99	132	0.00
214	Regulation of actin cytoskeleton	213	987	3.64
215	Insulin signaling pathway	137	412	0.00
216	Insulin secretion	85	99	0.00

Continued on next page

	Name	nodes	dges	eigen
217	GnRH signaling pathway	93	237	0.00
218	Ovarian steroidogenesis	49	50	0.00
219	Progesterone-mediated oocyte maturation	99	177	0.00
220	Estrogen signaling pathway	137	383	1.81
221	Melanogenesis	101	389	0.00
222	Prolactin signaling pathway	70	191	1.84
223	Thyroid hormone synthesis	74	112	0.00
224	Thyroid hormone signaling pathway	116	339	0.00
225	Adipocytokine signaling pathway	69	200	1.22
226	Oxytocin signaling pathway	152	419	0.00
227	Glucagon signaling pathway	103	422	8.00
228	Regulation of lipolysis in adipocytes	54	108	0.00
229	Renin secretion	65	69	0.00
230	Aldosterone synthesis and secretion	96	275	0.00
231	Relaxin signaling pathway	130	556	2.67
232	Cortisol synthesis and secretion	64	129	0.00
233	Parathyroid hormone synthesis, secretion and action	106	242	1.65
234	Type II diabetes mellitus	46	124	3.30
235	Insulin resistance	107	224	2.13
236	Non-alcoholic fatty liver disease (NAFLD)	149	136	2.28
237	AGE-RAGE signaling pathway in diabetic complications	99	299	2.96
238	Cushing syndrome	154	509	0.00
239	Type I diabetes mellitus	43	3	0.00

Continued on next page

	Name	nodes	dges	eigen
240	Maturity onset diabetes of the young	26	32	0.00
241	Aldosterone-regulated sodium reabsorption	37	37	0.00
242	Endocrine and other factor-regulated calcium reabsorption	47	44	1.00
243	Vasopressin-regulated water reabsorption	44	46	0.00
244	Proximal tubule bicarbonate reclamation	23	6	0.00
245	Collecting duct acid secretion	27	0	0.00
246	Salivary secretion	90	34	0.00
247	Gastric acid secretion	75	111	0.00
248	Pancreatic secretion	96	18	0.00
249	Carbohydrate digestion and absorption	44	7	0.00
250	Protein digestion and absorption	90	0	0.00
251	Fat digestion and absorption	41	16	0.00
252	Bile secretion	71	26	0.00
253	Vitamin digestion and absorption	24	0	0.00
254	Mineral absorption	51	6	0.00
255	Cholesterol metabolism	50	36	1.73
256	Alzheimer disease	171	67	0.00
257	Parkinson disease	142	25	0.00
258	Amyotrophic lateral sclerosis (ALS)	51	49	0.00
259	Huntington disease	193	26	1.00
260	Prion diseases	35	19	0.00
261	Cocaine addiction	49	91	0.00
262	Amphetamine addiction	68	248	0.00
263	Morphine addiction	91	549	0.00

Continued on next page

	Name	nodes	dges	eigen
264	Nicotine addiction	40	0	0.00
265	Alcoholism	180	757	0.00
266	Bacterial invasion of epithelial cells	74	101	0.00
267	Vibrio cholerae infection	50	27	0.00
268	Epithelial cell signaling in Helicobacter pylori infection	68	47	0.00
269	Pathogenic Escherichia coli infection	55	56	0.00
270	Shigellosis	65	104	1.95
271	Salmonella infection	86	173	0.00
272	Pertussis	76	111	0.00
273	Legionellosis	55	37	0.00
274	Leishmaniasis	74	136	0.00
275	Chagas disease (American trypanosomiasis)	102	225	2.95
276	African trypanosomiasis	35	28	0.00
277	Malaria	49	7	0.00
278	Toxoplasmosis	113	179	0.00
279	Amoebiasis	96	66	0.00
280	Staphylococcus aureus infection	56	46	0.00
281	Tuberculosis	179	463	0.00
282	Hepatitis C	131	183	0.00
283	Hepatitis B	144	259	1.00
284	Measles	132	248	1.00
285	Human cytomegalovirus infection	225	744	0.00
286	Influenza A	171	270	0.00
287	Human papillomavirus infection	339	1628	1.82

Continued on next page

	Name	nodes	dges	eigen
288	Human T-cell leukemia virus 1 infection	255	622	1.41
289	Kaposi sarcoma-associated herpesvirus infection	186	401	2.20
290	Herpes simplex infection	185	255	0.00
291	Epstein-Barr virus infection	201	266	5.00
292	Human immunodeficiency virus 1 infection	212	611	5.00
293	Pathways in cancer	526	1948	2.04
294	Transcriptional misregulation in cancer	186	12	0.00
295	Viral carcinogenesis	201	4	0.00
296	Chemical carcinogenesis	82	491	4.05
297	Proteoglycans in cancer	201	596	1.26
298	MicroRNAs in cancer	299	518	0.00
299	Colorectal cancer	86	149	1.26
300	Renal cell carcinoma	69	98	0.00
301	Pancreatic cancer	75	128	0.00
302	Endometrial cancer	58	89	0.00
303	Glioma	71	178	0.00
304	Prostate cancer	97	270	0.00
305	Thyroid cancer	37	57	0.00
306	Basal cell carcinoma	63	316	0.00
307	Melanoma	72	252	0.00
308	Bladder cancer	41	46	0.00
309	Chronic myeloid leukemia	76	158	0.00
310	Acute myeloid leukemia	66	156	0.00
311	Small cell lung cancer	93	231	0.00
312	Non-small cell lung cancer	66	137	1.41

Continued on next page

	Name	nodes	dges	eigen
313	Breast cancer	147	488	0.00
314	Hepatocellular carcinoma	168	559	7.42
315	Gastric cancer	149	431	0.00
316	Central carbon metabolism in cancer	65	130	0.00
317	Choline metabolism in cancer	99	194	0.00
318	Asthma	31	4	0.00
319	Autoimmune thyroid disease	53	5	0.00
320	Inflammatory bowel disease (IBD)	65	81	2.04
321	Systemic lupus erythematosus	133	30	0.00
322	Rheumatoid arthritis	90	13	0.00
323	Allograft rejection	38	24	0.00
324	Graft-versus-host disease	41	51	0.00
325	Primary immunodeficiency	37	0	0.00
326	Hypertrophic cardiomyopathy (HCM)	83	38	0.00
327	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	72	11	1.00
328	Dilated cardiomyopathy (DCM)	90	142	0.00
329	Viral myocarditis	59	23	0.00
330	Fluid shear stress and atherosclerosis	139	390	0.00

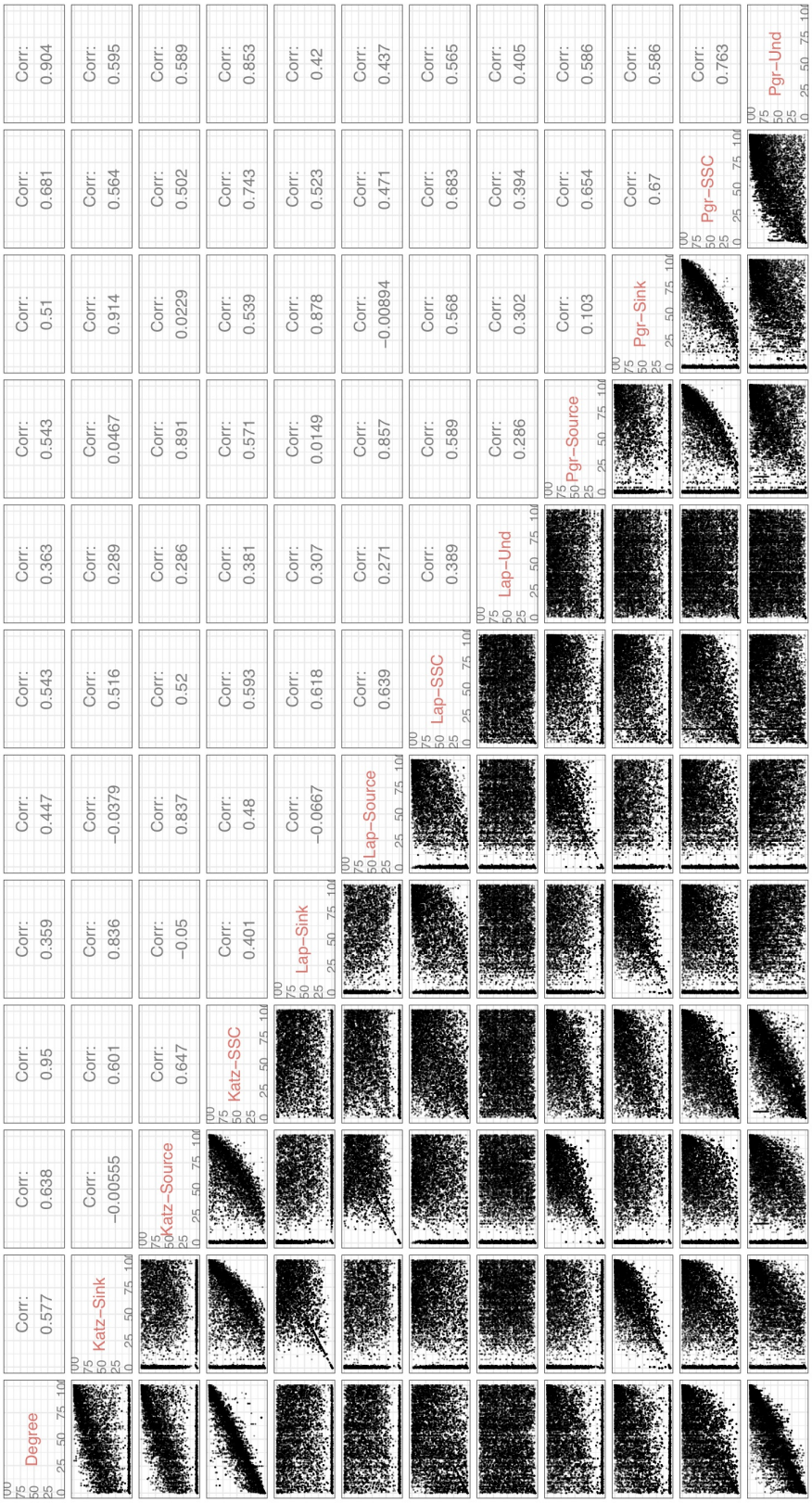

Figure B.1: Correlation of quantile scores between centrality models.

Table B.2: Detailed comparison of SPIA and CADIA for Ovarian cancer $FDR < 0.05$

Name.SPIA	ID	pNDE	pPERT	pGFdr	Name.CADIA	P_ORA	P_SSC	cadia	ORAFDR
Cell cycle	04110	1.83e-12	9.43e-01	6.38e-09	NA	NA	NA	NA	NA
p53 signaling ...	04115	2.22e-07	4.44e-01	1.11e-04	p53 signaling ...	2.83e-07	4.80e-01	1.09e-04	2.02e-05
Chemokine sign...	04062	1.46e-01	5.00e-06	4.85e-04	Chemokine sign...	1.95e-01	4.33e-02	3.17e-01	6.98e-01
Mineral absorp...	04978	4.02e-05	1.00e+00	1.48e-02	NA	NA	NA	NA	NA
Oocyte meiosis	04114	7.20e-04	9.00e-02	1.73e-02	Oocyte meiosis	3.13e-04	3.00e-04	1.09e-04	1.49e-02
Cytokine-cytok...	04060	2.49e-02	3.00e-03	1.73e-02	Cytokine-cytok...	3.08e-02	6.92e-01	4.76e-01	2.32e-01
Progesterone-m...	04914	3.71e-04	5.69e-01	3.77e-02	Progesterone-m...	4.82e-04	3.51e-01	3.19e-02	1.72e-02
Pathways in ca...	05200	2.31e-03	2.85e-01	8.04e-02	Pathways in ca...	2.30e-03	8.08e-02	3.19e-02	4.71e-02
Focal adhesion	04510	1.07e-02	9.17e-01	4.50e-01	Focal adhesion	1.01e-02	9.80e-03	2.41e-02	9.02e-02
NA	05206	NA	NA	NA	MicroRNAs in c...	3.66e-08	2.65e-01	2.70e-05	5.23e-06
NA	04151	NA	NA	NA	PI3K-Akt signa...	7.34e-03	2.50e-03	7.82e-03	8.31e-02
NA	04014	NA	NA	NA	Ras signaling ...	3.65e-01	2.00e-04	2.20e-02	9.97e-01

Table B.3: Detailed comparison of SPIA and CADIA for colorectal cancer $FDR < 0.05$

Name.SPIA	ID	pNDE	pPERT	pGFdr	Name.CADIA	P_ORA	P_SSC	cadia	ORAFDR
Cell cycle	04110	1.47e-18	5.40e-02	4.86e-16	NA	NA	NA	NA	NA
p53 signaling ...	04115	6.75e-08	1.95e-01	1.71e-05	p53 signaling ...	9.48e-08	4.34e-01	8.30e-05	1.36e-05
RNA transport	03013	2.98e-07	5.66e-01	1.27e-04	NA	NA	NA	NA	NA
PPAR signaling...	03320	6.43e-06	1.26e-01	3.54e-04	PPAR signaling...	1.09e-05	3.24e-01	1.37e-03	3.89e-04
Mineral absorp...	04978	8.70e-07	1.00e+00	3.54e-04	NA	NA	NA	NA	NA
Alzheimer's di...	05010	5.93e-01	5.00e-06	9.23e-04	NA	NA	NA	NA	NA
HTLV-I infection	05166	2.85e-05	2.01e-01	1.46e-03	HTLV-I infection	1.31e-04	4.95e-01	1.64e-02	3.11e-03
Amoebiasis	05146	7.98e-05	4.02e-01	6.19e-03	NA	NA	NA	NA	NA
Oocyte meiosis	04114	2.86e-04	1.62e-01	7.68e-03	Oocyte meiosis	6.02e-05	1.10e-03	8.30e-05	1.72e-03
Bile secretion	04976	3.67e-04	1.71e-01	9.12e-03	NA	NA	NA	NA	NA
Pathways in ca...	05200	9.89e-05	7.05e-01	9.12e-03	Pathways in ca...	1.17e-06	9.45e-01	7.77e-04	8.39e-05
ECM-receptor i...	04512	2.85e-03	3.70e-02	1.21e-02	ECM-receptor i...	2.13e-02	1.57e-01	1.23e-01	1.02e-01
Progesterone-m...	04914	3.80e-03	4.30e-02	1.66e-02	Progesterone-m...	8.65e-04	1.20e-01	1.88e-02	1.55e-02
Small cell lun...	05222	1.35e-03	1.53e-01	1.91e-02	Small cell lun...	7.35e-04	5.88e-01	3.77e-02	1.50e-02
Chemokine sign...	04062	4.18e-03	6.10e-02	2.15e-02	Chemokine sign...	2.89e-03	1.12e-01	3.77e-02	3.47e-02
Gap junction	04540	2.02e-02	1.80e-02	2.76e-02	Gap junction	1.81e-02	7.39e-02	7.66e-02	9.96e-02
Transcriptiona...	05202	4.07e-04	1.00e+00	2.87e-02	NA	NA	NA	NA	NA
Wnt signaling ...	04310	9.34e-04	4.92e-01	3.02e-02	Wnt signaling ...	5.17e-03	3.82e-01	8.87e-02	5.28e-02
Pancreatic sec...	04972	1.61e-03	5.38e-01	4.99e-02	NA	NA	NA	NA	NA
GnRH signaling...	04912	7.30e-02	2.00e-02	7.47e-02	GnRH signaling...	2.03e-02	1.81e-02	3.77e-02	1.02e-01
Apoptosis	04210	7.91e-02	2.70e-02	8.63e-02	Apoptosis	4.56e-02	8.00e-03	3.77e-02	1.64e-01
Vascular smoot...	04270	2.76e-02	2.80e-01	1.98e-01	Vascular smoot...	2.15e-02	2.12e-02	3.77e-02	1.02e-01
Calcium signal...	04020	2.89e-02	4.09e-01	2.49e-01	Calcium signal...	2.51e-02	2.50e-02	4.70e-02	1.16e-01
Olfactory tran...	04740	9.94e-01	6.11e-01	9.94e-01	Olfactory tran...	9.97e-01	1.00e-04	1.88e-02	9.97e-01
NA	05206	NA	NA	NA	MicroRNAs in c...	5.09e-06	4.22e-01	1.08e-03	2.43e-04
NA	04390	NA	NA	NA	Hippo signalin...	6.58e-03	5.64e-02	3.77e-02	6.27e-02
NA	04072	NA	NA	NA	Phospholipase ...	6.01e-02	7.00e-03	3.77e-02	1.95e-01

Table B.4: Detailed comparison of SPIA and CADIA for Gastric cancer $FDR < 0.05$

Name.SPIA	ID	pNDE	pPERT	pGFdr	Name.CADIA	P_ORA	P_SSC	cadia	ORAFDR
ECM-receptor i...	04512	1.93e-07	5.00e-06	2.02e-09	ECM-receptor i...	1.44e-07	6.25e-01	2.22e-04	2.06e-05
Gastric acid s...	04971	1.07e-06	7.00e-03	5.39e-06	Gastric acid s...	1.07e-06	9.92e-01	7.97e-04	7.65e-05
Focal adhesion	04510	2.56e-06	1.50e-02	1.69e-05	Focal adhesion	2.44e-06	4.66e-01	7.97e-04	1.16e-04
TGF-beta signa...	04350	3.67e-04	6.60e-02	4.54e-03	TGF-beta signa...	2.26e-02	2.20e-01	4.06e-01	2.93e-01
Malaria	05144	2.70e-05	1.00e+00	4.54e-03	NA	NA	NA	NA	NA
Cytokine-cytok...	04060	4.10e-02	1.00e-02	4.39e-02	Cytokine-cytok...	4.40e-02	2.14e-01	5.09e-01	4.19e-01
Amoebiasis	05146	1.25e-03	4.11e-01	4.43e-02	NA	NA	NA	NA	NA
Vascular smoot...	04270	1.36e-03	4.22e-01	4.43e-02	Vascular smoot...	1.60e-03	4.59e-01	1.23e-01	2.85e-02
Wnt signaling ...	04310	3.04e-01	6.00e-03	8.83e-02	Wnt signaling ...	2.76e-01	1.00e-04	9.07e-03	9.55e-01
NA	05204	NA	NA	NA	Chemical carci...	1.23e-03	5.90e-03	3.33e-03	2.67e-02
NA	04151	NA	NA	NA	PI3K-Akt signa...	2.13e-04	3.16e-01	1.70e-02	7.62e-03

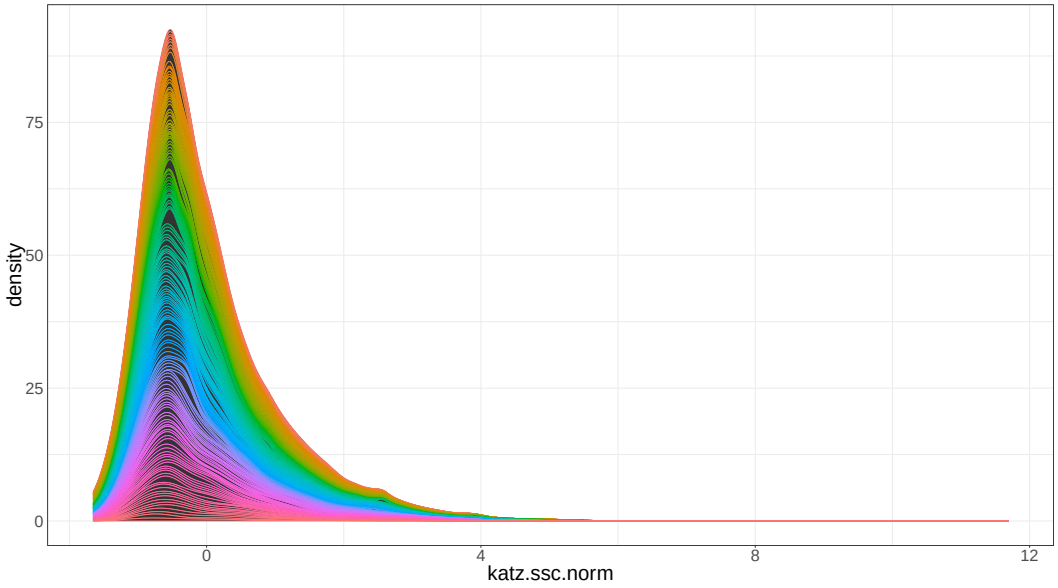

Figure B.2: The density plot of Katz Source/Sink centrality normalized values. Each color represents a different pathway.

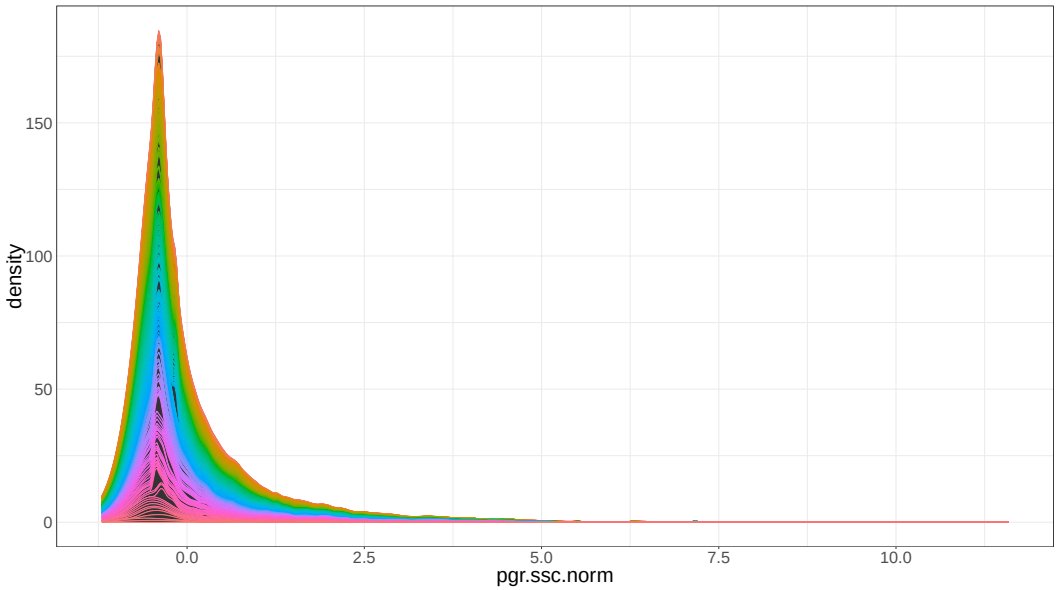

Figure B.3: The density plot of PageRank Source/Sink centrality normalized values. Each color represents a different pathway.

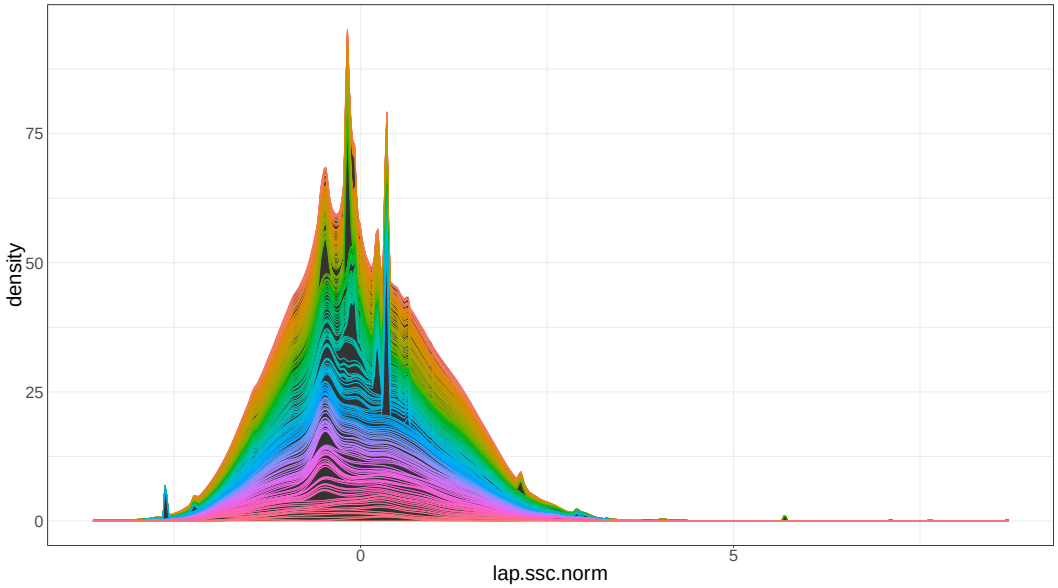

Figure B.4: The density plot of PageRank Source/Sink centrality normalized values. Each color represents a different pathway.

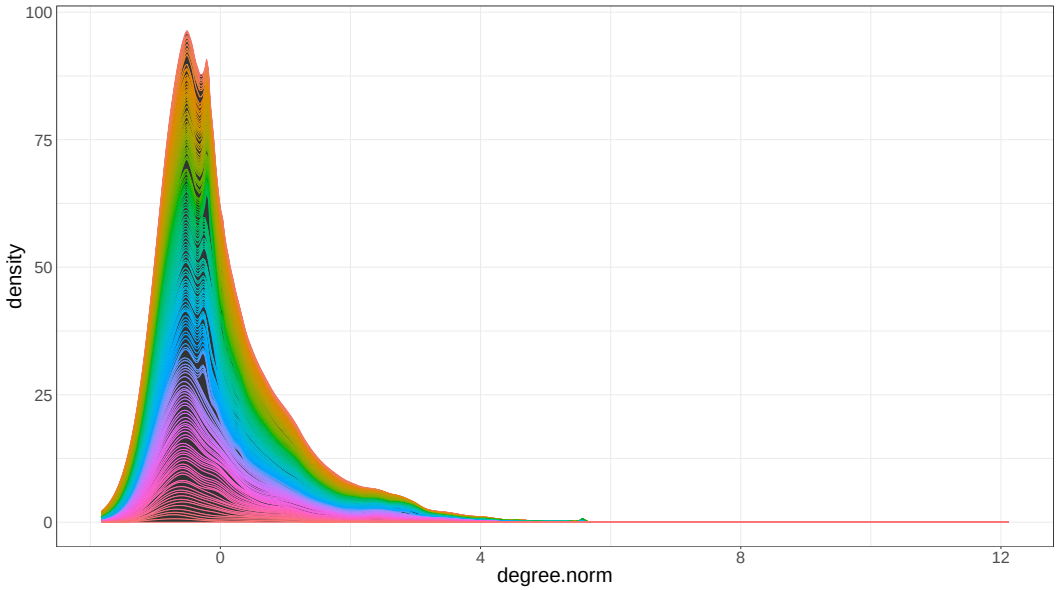

Figure B.5: The density plot of Degree centrality normalized values. Each color represents a different pathway.