Centrality of cancer-related genes in human biological pathways: A graph analysis perspective

Pourya Naderi Yeganeh, Erik Saule, and M. Taghi Mostafavi*
Department of Computer science, The University of North Carolina at Charlotte,Charlotte, NC, USA
pnaderiy@uncc.edu, esaule@uncc.edu, *Corresponding: taghi@uncc.edu

***Abstract*—This study investigates standard and novel centrality models to identify the topological organization of cancer-related genes in biological pathways. We examined the linear relationship between the ratio of cancer-related genes and centrality rankings from different models. We also compared the cumulative distributions of centrality scores for cancer-related and non-cancer-related genes. Difference between the mean centrality scores of the two groups was tested in each pathway. The results show that when accounting for the directions of pathways and the importance of the interacting genes, the centrality of a gene correlates with the probability of cancer-relatedness. In particular, we show that the centrality measures we propose, namely Source-Sink PageRank and Source-Sink Katz, produce a distinction between the distribution of the two gene groups. Source-Sink PageRank shows the highest statistical power in differentiating between the means centrality values of two groups. The presented analysis provides a new perspective for understanding the topological organization of cancer-related genes.**

***Index Terms*—Biological Pathways, Graph Analysis, Network Analysis, Systems Biology**

I. Introduction

A main premise of systems biology is that the biological functions can arise as emergent properties of interaction webs of sub-cellular entities [1], [2]. Studies show that the topological position of the entities in networks can determine certain biological properties [1], [3], [4]. For example, Jeong et al. [1] have shown that the number of interactions of a node in the Protein-Protein Interaction (PPI) networks correlates with the probability of its removal being lethal to the organism [1], [4]. The topological properties of biological networks have widespread applications, including in pathway discovery [5], and enrichment analysis [6]–[10]. However, there is a gap of knowledge for the topological properties of key pathway regulators and the organization of genes in pathways.

This study investigates whether the network centrality models can differentiate between cancer-related genes and non-cancer-related genes. Cancers are diseases of pathways, and the dysfunction of cancer-related genes can result in dysfunction of their associated networks [11]. Here, we investigate these questions: 1- Does the number of interactions of a gene in associates with the probability of being cancer-related? 2- Does the probability of a gene being cancer-related associates with the topological importance of its interacting genes. 3- Does the direction of interactions gives information about the topological importance of cancer-related genes?

To answer these questions, we used three known standard centrality models – Degree, Katz, and PageRank [12]. In addition, we designed two novel centralities that address the shortcoming of existing models in biological pathways, namely, Source-Sink Katz and Source-Sink PageRank. These two novel models are capable of assigning node importance to both upstream and downstream ends of pathways, while accounting for directions of the interactions.

We took three statistical approaches to evaluate our hypotheses. 1- We investigated the linear relationship of gene rankings of each centrality with the probability of being cancer-related. 2- Compared the cumulative distribution of rankings for cancer-related genes versus non-cancer-related. 3- Compared the mean ranking of cancer-related versus non-cancer-related genes for each pathway. The results show some linear relationship between degree centrality and the probability of being cancer-related. Our analyses show that the spectral ranking of genes, particularly Source-Sink and undirected, exhibit stronger linear relationship with the probability of being cancer related. Pathway-by-pathway comparisons show unique patterns for distinguishing between cancer-related and non-cancer-related genes, with Source-Sink PageRank having the highest statistical power. We conclude that the cancer-related genes tend to have higher centrality when accounting for directionality and importance in both upstream and downstream of pathways.

II. Materials and Methods

A. Graph Modeling of Pathways

Let the graph, $G = (V, E)$, represent a pathway. $V(G) = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes. The set of edges is $E(G) = \{e_1, e_2, \ldots, e_m\}$, $e_k = (v_i, v_j)$, which are ordered pairs denoting the directions. A graph is undirected if the edges are unordered pairs. The *neighborhood* of a node, $N_G(v_i)$, is defined as $N_G(v_i) = \{v_j | (v_i, v_j) \in E(G)\}$. The degree of a node is the size of its neighborhood. For a directed graph, this notion of degree is out-degree, $Deg_{out}(v)$. Neighborhood and degree can be defined based on in-coming edges, i.e. in-degree, $Deg_{in}(v) = \big|\{u \mid (u, v) \in E\}\big|$. A graph, $G(V, E)$ can be represented as an adjacency matrix, A_G. Formally:

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

Transpose of a graph (G^T) is defined as a graph with reversed edges. Formally, $V(G^T) = V(G)$ and $E(G^T) = \{(u, v) | (v, u) \in E(G)\}$. For a transpose graph , $A_{G^T} = {A_G}^T$.

B. Graph Centrality Models and Definitions

This study uses three standard centrality models to investigate its research hypothesis. In addition, this study investigates two novel centrality approaches, namely Source-Sink PageRank and Source-Sink Katz, which quantify the centrality of a node as a sender and receiver of biological information. The analysis was done using R packages *sna* and *Igraph* [13]. A description of each model is provided in the next few paragraphs.

Degree Centrality: In this model, centrality is the degree of the nodes. Studies show that degree centrality of nodes in PPI network of different organisms correlates with their essentiality, i.e. the likelihood of a protein's removal to be lethal for the model organism [1], [4]. In this study, degree centrality was calculated by combining in-degree and out-degree:

$$C_{deg}(v) = Deg_{in}(v) + Deg_{out}(v) \tag{2}$$

PageRank Centrality: is the probability distribution of a uniform random walk being present at each node. PageRank of a node is calculated based on the average centrality of its neighbors. Formally:

$$C_{pgr}(v) = \beta + \alpha \sum_{u \in \mathbf{N_G(v)}} \frac{C_{pgr}(u)}{|\mathbf{N_G(v)}|} \tag{3}$$

Where α and β are constant factors [12]. In this study, the term PageRank refers to directed PageRank. This study uses the factors $\alpha = 0.9$ and $\beta = 0.1$. PageRank is a spectral centrality, i.e. the importance of a node is relative to that of its neighbors. The spectral centralities have been used in pathway discovery and pathway enrichment analysis [5], [6].

Katz-Bonacich Centrality: is an spectral centrality model where the importance of a node is calculated relative to the sum of centrality of its neighbors [14]. Formally:

$$C_{katz}(v) = \beta + \alpha \sum_{u \in \mathbf{N_G(v)}} C_{katz}(u) \tag{4}$$

Where β and α are constant factors. If $\beta = 1$, then $a \leq 1/\lambda_1$ is a sufficient condition for convergence. λ_1 is the largest positive eigenvalue of the adjacency matrix. In this paper, Katz centrality refers to the directed graph. The parameters were set to $\alpha = 0.1$ and $\beta = 1$.

Source-Sink PageRank: this novel model is defined as measuring the centrality of graph nodes individually as sources and sinks of information. Directed centrality measures only give importance to upstream nodes and leave downstream ones with lowest importance (often zero). The literature suggests that some of the downstream nodes are critical components of pathways. We previously showed the utility of Source-Sink modeling in pathway enrichment analysis [Work not published/Under-Review]. This model is defined as the addition of two components, Source Centrality and Sink Centrality. Source component measures the importance of a node v relative to the nodes that v sends information to. In the case PageRank, the first component (Source), is:

$$C_{Src-pgr}(v) := C_{pgr}(v) \tag{5}$$

The second component, Sink Centrality, is:

$$C_{Sink-pgr}(v) = \beta' + \alpha' \sum_{u \in \mathbf{N_{G^T}(v)}} \frac{C_{Sink-pgr}(u)}{|\mathbf{N_{G^T}(v)}|} \tag{6}$$

The second component is closely related to PageRank except that it is defined on the transposed graph. Sink centrality models the centrality of a node as a receiver of information. After calculating Source and Sink Centrality values individually, the two components are summed as following:

$$C_{SS-pgr}(v) = C_{Src-pgr}(v) + \gamma C_{Sink-pgr}(v) \tag{7}$$

Where, γ is a parameter for the relative importance of Source versus Sink. This studies uses $\gamma = 1$, $\beta = \beta' = 0.1$, and $\alpha = \alpha' = 0.9$ for Source-Sink PageRank.

Source-Sink Katz: Similar to Source-Sink PageRank, the Source component is the Katz centrality of the directed graph:

$$C_{Src-ktz}(v) := C_{katz}(v) \tag{8}$$

The sink component is the Katz centrality of G^T.

$$C_{Sink-ktz}(v) := \beta' + \alpha' \sum_{u \in \mathbf{N_{G^T}(v)}} C_{Sink-ktz}(u) \tag{9}$$

Katz Source-Sink Centrality is then defined as:

$$C_{SS-ktz}(v) = C_{Src-ktz}(v) + \gamma C_{Sink-ktz}(v) \tag{10}$$

It can be shown that Source and Sink components have the same convergence criteria [Not published/Under-Review]. This studies uses $\gamma = 1$, $\beta = \beta' = 0.1$, and $\alpha = \alpha' = 1$.

C. Statistical Evaluations

The difference between centrality scores of cancer-related genes and non-cancer related in each model were investigated through three approaches. Since the subjects of study are multiple pathways, normalization and ranking procedures were used to create a unified framework.

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

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

Where $|V_j|$ is the number of nodes in pathway j. The quantile ranking allows comparing pathways with different number of nodes. In addition, 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:

$$ {F_i}^c = \frac{\sum_j \left|\{v \mid v \in R \cap Q_{ij}\}\right|}{\sum_j \left|\{v \mid v \in Q_{ij}\}\right|} \tag{11} $$

Here, each gene occurrence in a pathway was treated as an unique gene. ${F_i}^c$ was then tested against the quantile score (i) for assessing the linear relationships. Formally:

$$ {F_i}^c = a_1.i + a_0 \tag{12} $$

Where a_1 and a_0 are the coefficients of the linear regression. For each centrality model, the adjusted r-squared (coefficient of determination) was evaluated. In addition, Pearson correlation of $Q(v_i)$ values between each centrality measure were calculated.

2) Comparison of Cumulative Densities: To compare the distribution of centrality values from a global perspective, the centrality scores, $C(v_i)$, were normalized within each pathway as following:

$$ NS_{a,j}(v_i) = \frac{C_{a,j}(v_i) - \mu_{a,j}}{\sigma_{a,j}} \tag{13} $$

Where $\mu_{a,j}$ and $\sigma_{a,j}$ are the mean and standard deviation of centrality scores of model a for pathway j. $NS_{a,j}(v_i)$ is the normalized centrality score of model a for node v_i in pathway j. The normalized score across all pathways were placed in 100 quantiles. The distribution of quantile scores for the "Cancer-related" and "Non-cancer" groups were compared by contrasting their cumulative distribution function (CDF) using a Kolmogorov-Smirnov (KS) test. The analysis was limited to the top performing models from regression analysis. The alternative hypothesis was the CDF of the cancer-related lying below that of the non-cancer.

3) Pathway-wise testing: Welch's t-test was used to contrast between the estimated means of centrality values for cancer genes and normal genes, individually for each centrality model and for each pathway. Here, the null hypothesis is two groups having the same mean. The alternative hypothesis is cancer genes having a higher mean. Also, non-parametric Wilcox test was used as well with the same null and alternative hypotheses.

For each centrality model, the p-values from Welch and Wilcox tests were calculated for each of the 155 pathways. Benjamini-Hochberg False Discovery Rate was applied to all calculated p-values for each centrality method to control type-I errors ($FDR < 0.05$) [15].

D. Biological Data Processing

Human pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) were retrieved (n = 330, as of August 2018). Pathways with $\leq$20 nodes or 20 edges were neglected from analysis (n = 85). Also, pathways with largest eigenvalues $\geq$10 (n = 15) were excluded to maintain consistent centrality calculations. Pathways with a single unique value for any of the centrality measures were excluded from the analysis (n= 10). The pathways were parsed using R-packages "KEGGGraph" and "Pathview" [16], [17]. 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 156 entries.

Cancer-related genes were retrieved from MSigDB (n = 417 as of 06-06-18) [18]. The cancer-related gene included were *Oncogenes*, *Tumor Suppressors*, and *Translocated cancer genes*. Cancer Gene Census from Sanger Institute was used as an additional reference (n= 719 as of 06-06-18) [19]. The union of these two sets were used as the reference cancer gene list (n = 733). 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.

III. Results and Discussion

A. Linear relationship between topological rankings and the ratio of cancer-related genes

Regression analysis provides insight regarding the linear relationship of the quantile-scores of centrality models with the ratio of cancer genes (Figure 1). The number of connections (Degree) of a gene in a biological pathway is related to its probability being cancer-related. Regression analysis finds a statistically significant (Adjusted r-squared = 0.26, p-value = 5.56×10^{-8}) evidence of linear relationship between ranks of 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.

The results show that Source-Sink modeling has a stronger evidence of linear relationship between the centrality scores and the ratio of cancer-genes. For Katz centrality (as defined in Formula 4), there is not enough evidence for linear relationship between the quantile-score and ratio of cancer genes (Adjusted r-squared = 0.009). When the importance is measured only in upstream-to-downstream direction, many of the cancer genes are given low importance (Figure 1). However, when Katz centrality is measured and summed from both upstream and downstream directions (Source-Sink Katz), the linear relationship explains a statistically significant portion of the variance (Adjusted r-squared 0.36, p-value = 4.32×10^{-11}) – more than that of Degree centrality. 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 linear relationship (Adjusted r-squared = 0.74) compared to PageRank and Undirected PageRank, Adjusted r-squared of 0.21 and 0.65. These results show that spectral importance determines the ratio of cancer genes, particularly, when considered in both upstream and downstream directions. The strong descriptive power of Source-Sink is potentially because of

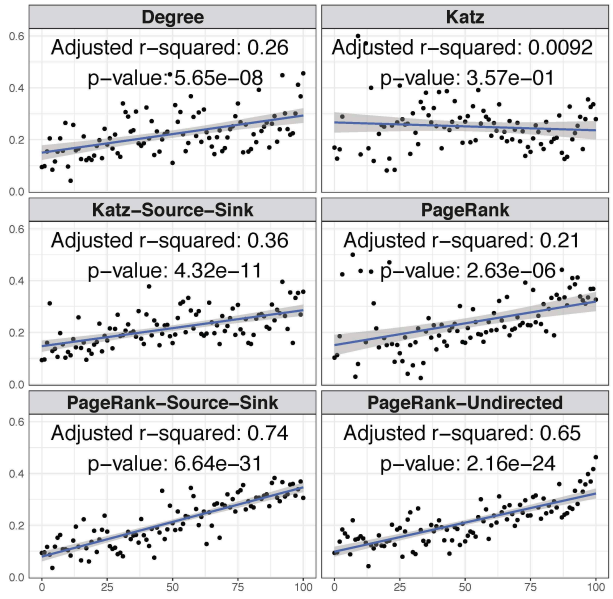

Fig. 1. Linear regression of quantile-scores versus the ratio of cancer-related genes. X-axis represents the scores generated by Formula 11. Y-axis represents the ratio of cancer-related genes (Formula 11). The blue line is the regression fit (Formula 12). The gray band is the 95% confidence interval.

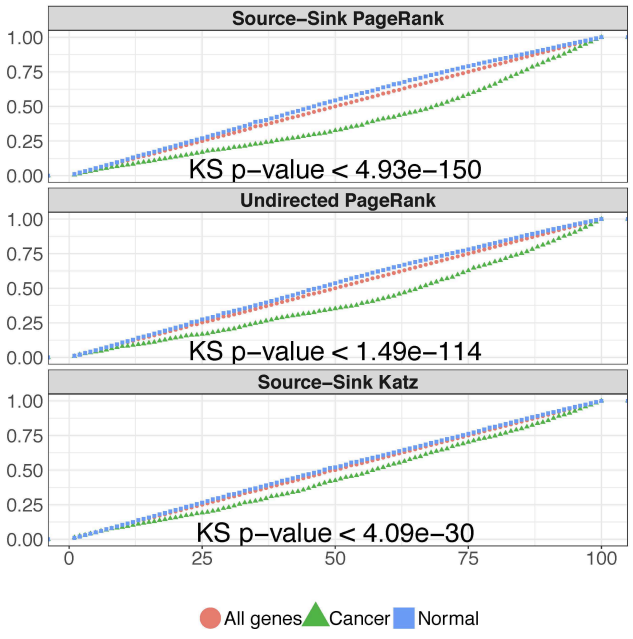

Fig. 2. Comparison of cumulative density between cancer-related genes and normal genes. The data points represent the quantile-scores calculated based on normalized centrality (Formula 13) across all pathways.

being sensitive to the organization of the original network in terms of information flow and directions. Standard applications either use undirected or a directed modeling which disregards terminal nodes – e.g. [5], [6]. The presented results show that using directions while giving importance to terminal nodes in pathways may give higher explanatory power.

TABLE I
PATHWAYS WITH HIGHER CENTRALITY OF CANCER GENES BY T-TEST

	SS-Katz	Degree	Katz	Pgr	SS-Pgr	Und. Pgr
SS-Katz	5	4	2	0	0	1
Degree		5	2	0	0	2
Katz			2	0	0	0
Pgr				4	3	1
SS-Pgr					6	1
Und. Pgr						7

TABLE II
PATHWAYS WITH HIGHER CENTRALITY OF CANCER GENES BY WILCOX TEST

	SS-Katz	Degree	Katz	Pgr	SS-Pgr	Und. Pgr
SS-Katz	14	10	5	4	5	12
Degree		10	4	2	4	10
Katz			8	3	3	7
Pgr				24	19	12
SS-Pgr					31	15
Und. Pgr						27

B. Difference between the centrality ranking distribution of cancer-related and non-cancer-related genes

Analysis of cumulative density of quantile scores outlines the differences between scoring of cancer-related and normal genes for Source-Sink PageRank, Undirected PageRank, and Source-Sink Katz (Figure 2). The CDF of cancer genes lies below that of normal genes for all three models. This observation is supported by Kolmogorov-Smirnov test, as displayed in Figure 2. The null hypothesis of two groups having the same distribution is rejected and the CDF of cancer genes lying below that of normal genes – p-values of 4.93×10^{-150}, 1.49×10^{-114}, and 4.09×10^{-30}. The difference between the CDF shows that cancer-genes tend to have higher centrality in all three models. For example in Source-Sink PageRank, only $\sim$30% of cancer genes have a quantile score below 50, compared to $\sim$55% normal genes for the same cut-off. These results support the linear regression results that a higher centrality value indicates higher probability of being cancer-related. CDF analysis and regression analysis show that Source-Sink PageRank best captures the topological organization of cancer genes in biological pathways.

C. Pathways with higher mean of centrality values for cancer-related genes

Pathway by pathway analysis outlines the statistical power of each centrality method for distinguishing between Cancer-related (Cancer) genes and non-cancer-related (Normal) genes. In Tables I and II, the diagonal elements indicate the number of pathways with higher mean centrality of cancer-related genes (rejected hypothesis) for each model. The non-diagonal entries indicate the number of rejected hypothesis by both respective models in row and column index.

Using Welch's test (Table I), Source-Sink Katz and Degree identify 5 pathways each. These methods have no overlap with PageRank or Source-Sink PageRank. On the other hand, Katz centrality only identifies two pathways, both identified by degree and Source-Sink Katz. Using Wilcox test increases the

TABLE III
CORRELATION BETWEEN CENTRALITY VALUES

	Deg	SS-Katz	Pgr	SS-Pgr	Und. Pgr
Deg	1	0.95	0.5	0.67	0.91
SS-Katz		1	0.53	0.74	0.86
Pgr			1	0.66	0.57
SS-Pgr				1	0.74
Und. Pgr					1

statistical power ($FDR < 0.05$). Source-Sink Katz, Degree, and Katz detect 14, 10, and 8 pathways with higher mean values of ranks for cancer genes. In this case, the pathways that are detected by Degree centrality are also detected by Source-Sink Katz. This shows that Source-Sink Katz has a higher power compared to Degree, and is more informative.

Using Welch's test, Source-Sink PageRank and Undirected PageRank identify 6 and 7 pathways, with only 1 pathway in common. By using Wilcox test, PageRank, Source-Sink PageRank, and Undirected PageRank detect 24, 31, and 27 pathways each. Source-Sink PageRank and PageRank have an overlap of 19 pathways, indicating that most of detections from PageRank are also detected by Source-Sink PageRank. The overlap between Undirected and Source-Sink PageRank is limited to 15 pathways, showing that two methods produce a considerable number of different pathways. These differences indicate the uniqueness of each centrality model for distinguishing cancer genes from non-cancer. Non-parametric tests have higher statistical power because the distribution of the centrality scores is non-normal. Further analysis of the distributions may reveal useful insight for finding more descriptive transformations and tests.

The correlation analysis between the ranking produced by each centrality model provides insight regarding the relationship between their procedures (Table III). Degree centrality, Source-Sink Katz, and Undirected PageRank are highly correlated. The correlation of Source-Sink PageRank with other models is not as high in comparison, this indicates the difference between the ranking procedures of Source-Sink PageRank and other models. For example, the correlation coefficient of Source-Sink PageRank and undirected PageRank is 0.74 (Table III). In addition, PageRank produces lower correlations with the other models. This is because terminal nodes in PageRank are all assigned with lowest possible centrality values and a considerable number of nodes have PageRank centrality of zero.

IV. CONCLUSIONS

This study compared the explanatory power of different centrality models with respect to cancer-related genes. The analysis showed the differences between topological position of cancer-related and non-cancer-related (normal) genes. Our findings assert three topological properties of cancer-related genes in human biological pathways.

We have concluded that cancer genes tend to have higher degree and their organization follows a spectral pattern. We have also concluded that the direction of interactions creates a better description for topological importance of cancer genes in biological pathways, particularly when the topological importance is measured as both receiver and sender of information. The results have also shown that the presented novel methodology, the Source-Sink PageRank, is highly descriptive of the topological organization of cancer genes in biological networks. The presented evidence suggest that network-based pathway analysis methods should consider the topological importance of genes in a directed format in both downstream and upstream ends of pathways.

AVAILABILITY

The R codes of this study is shared for reproducibility at: https://github.com/pouryany/OncoCentrality

ACKNOWLEDGEMENT

This research has been supported by the CCI at UNC-Charlotte and in part by a National Science Foundation grant, CCF- 1652442.

REFERENCES

[1] H. Jeong *et al.*, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.
[2] U. Alon, "Network motifs: theory and experimental approaches," *Nature Reviews Genetics*, vol. 8, no. 6, p. 450, 2007.
[3] V. Janjić, R. Sharan, and N. Pržulj, "Modelling the yeast interactome," *Scientific reports*, vol. 4, 2014.
[4] X. He and J. Zhang, "Why do hubs tend to be essential in protein networks," *PLoS Genet*, vol. 2, no. 6, p. e88, 2006.
[5] 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.
[6] A. L. Tarca *et al.*, "A novel signaling pathway impact analysis," *Bioinformatics*, vol. 25, no. 1, pp. 75–82, 2009.
[7] P. Naderi Yeganeh and M. T. Mostafavi, "Use of structural properties of underlying graphs in pathway enrichment analysis of genomic data," in *Proc. of ACM-BCB*, 2017, pp. 279–284.
[8] 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.
[9] Z. Gu *et al.*, "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.
[10] C. Mitrea *et al.*, "Methods and approaches in the topology-based analysis of biological pathways," *Frontiers in physiology*, vol. 4, p. 278, 2013.
[11] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *cell*, vol. 144, no. 5, pp. 646–674, 2011.
[12] M. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010.
[13] C. T. Butts *et al.*, "Social network analysis with sna," *Journal of Statistical Software*, vol. 24, no. 6, pp. 1–51, 2008.
[14] P. Bonacich and P. Lloyd, "Eigenvector-like measures of centrality for asymmetric relations," *Social networks*, vol. 23, no. 3, pp. 191–201, 2001.
[15] 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.
[16] 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.
[17] 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.
[18] A. Subramanian *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *PNAS*, vol. 102, no. 43, pp. 15 545–15 550, 2005.
[19] P. A. Futreal *et al.*, "A census of human cancer genes," *Nature Reviews Cancer*, vol. 4, no. 3, p. 177, 2004.