Mining Differential Hubs in Homogenous Networks

Omar Odibat
Department of Computer Science
Wayne State University
Detroit, MI 48202 USA
odibat@wayne.edu

Chandan K Reddy
Department of Computer Science
Wayne State University
Detroit, MI 48202 USA
reddy@cs.wayne.edu

ABSTRACT

Networks have been extensively used to model various complex systems such as online social networks, co-authorship and citation networks and gene networks. Due to different kinds of variations such as temporal, spatial, topic and phenotypic variations, several variants of the same network may exist. For several practical problems, identifying the nodes that are changing between the networks provide vital information regarding the dynamics of the network states. Given two networks where the nodes are the same in both networks, but the edges are different, we consider the problem of identifying a set of hubs that best explain the differences between the two networks. To the best of our knowledge, this is the first work to address the problem of finding the differential hubs. To address this problem, we propose a novel ranking algorithm, *DiffRank*, which ranks the nodes of two networks based on their differential behavior between the two networks. We define new measures such as differential connectivity and differential centrality for each node. These measures are propagated through the network and are optimized to capture the local and global structural changes between two networks. We demonstrate the effectiveness of *DiffRank* on synthetic datasets and real-world applications including collaboration and biological networks. We show that *DiffRank* identifies meaningful and practically valuable information compared to some of the baseline methods that can be used for such a task.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - Data Mining; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Design

Keywords

Differential Network analysis, Ranking, Centrality, PageRank, Co-authorship Networks, Biological Networks.

MLG '11 San Diego, CA, USA

1. INTRODUCTION

Networks have been extensively used to model various complex systems such as online social networks, co-authorship and biological networks. These networks consist of data objects as the nodes and the interactions between the data objects as the edges. Studying such networks can provide valuable knowledge about the data objects and their interactions. The interactions between the data objects depend on the domain in which these data objects are studied. Considering two different domains, the same data objects can have two different sets of interactions (or edges) between them, and thus form two different networks. These are homogenous networks because they have single type of data objects in both networks. The interactions between the same data objects could change between the two networks due to different possible sources of variations such as:

- **Temporal variation:** networks evolve over time, which includes the addition and deletion of links and nodes [1]. Two non-overlapping snapshots (or mutually exclusive time intervals) could have unevenly evolving topological structure.
- **Topic variations:** an author can collaborate with other authors on one topic while collaborating with different authors on another topic [24].
- **Phenotypic variation:** normal and cancerous cells have the same set of genes, but some of these genes are differentially wired in the cancerous cells, which results in two different gene interaction networks [7].

Given two networks, ranking the nodes based on their differential behavior is a challenging problem. Most importantly, for several practical problems, identifying the differential hubs that are changing between the networks provides vital information regarding the dynamics of the network states. The differential hubs are the set of nodes that are responsible for the differences between two networks. In many scenarios, it is appealing to have a system that tracks and finds the differential nodes between two networks. Consider the following two applications:

Co-authorship networks: In scientific co-authorship networks, the nodes are authors of academic papers and the edges represent co-authorship (or collaboration) relationships between the authors. Two authors may have different relationships in two different research topics such as data mining and database [24]. The differential hubs in this case

include the authors who are highly active in one topic but not active in the other topic, or they may include the authors who are active in both topics but with different collaborators in each topic. Differential networking can also be used to analyze two co-authorship networks that are constructed from two mutually exclusive time intervals to identify the authors whose collaborations change over time.

Biological networks: Microarray studies are used to measure the expression level of thousands of genes under different conditions. These conditions could be different tissue types (normal vs cancerous), different subject types (e.g., male vs female), different group types (African-American and Caucasian American) [16], different stage of cancer (early stage vs developed stage) [21] or different time points [13]. Here, the nodes are the genes, and the edges represent the interactions between the genes. Since the genes that have strongly altered connectivity play an important role in the disease phenotype [7], finding the differential genes can be used in several applications such as identifying disease-causing genes and examining the effects of a certain treatment [7].

The main technical challenge of exploiting network structure to find the differential hubs is to find all the differences between two networks. A straightforward solution is to transfer this problem to solving the subgraph isomorphism problem. Unfortunately, this is not desirable as it is computationally infeasible, and it was shown that solving the subgraph isomorphism problem is NP-complete problem [23]. Instead, we propose *DiffRank*, as an efficient and approximate solution to find the differences between two networks.

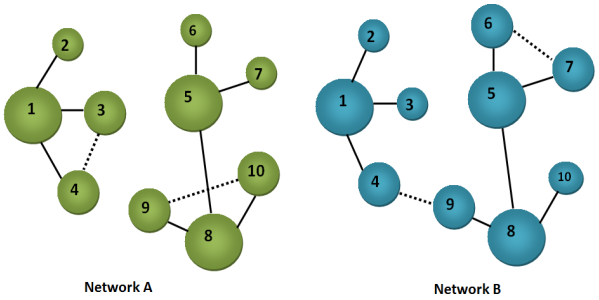


Figure 1: A simple illustration of differential network analysis. Network A and network B have the same nodes but different edges. The solid edges are common in both networks, while the dashed edges exist only in one network.

Toy Example: Figure 1 shows a simple illustration of the concept of differential network analysis using two unweighted and undirected networks. The top three hubs from network A, based on their degrees, are 1, 5 and 8, which are the same top three hubs in network B. However, for the differential analysis purpose, we would like to see the nodes 4 and 9 ranked top in the list because they are responsible for the major differences between the networks. Node 4 has the same degree in both networks, so looking only at the degrees of nodes in each network individually does not help in identifying the differential hubs. Moreover, some edges are more important than others. For instance, the edge between node 4 and node 9 is important because it connects two subnetworks, but the edge between node 6 and node 7 is less important. In this paper, we propose a new algorithm to find the differential hubs that are responsible for the changes in the connectivity and the topological structure between two networks.

Our goal is to identify the differential hubs by analyzing two interaction networks. We combine differential network analysis with ranking in one framework and propose a novel ranking algorithm, *DiffRank*, which ranks the nodes of two networks based on their differential behavior in the two networks. To achieve this goal, we define novel measures such as differential connectivity and differential centrality for each node. These measures are propagated through the network and are optimized to capture the changes in the local and global structures between two networks.

Contributions: The main contributions of this paper can be summarized as:

1. We propose *DiffRank* algorithm to rank the hubs of two networks based on their differential behavior in the two networks and to identify the differential hubs.
2. We propose two novel differential measures:
   (a) A local structure measure, *differential connectivity*, to capture the local differences between two networks based on their weighted edges.
   (b) A global structure measure, *differential betweenness centrality*, to capture the global differences between two networks based on the shortest paths
3. We develop a simulator for generating synthetic differential scale-free networks based on two models to evaluate the proposed algorithm.
4. We apply the proposed algorithm on different real-world datasets including the *DBLP* dataset and a lung cancer dataset.

The proposed algorithm has two salient features. First, it can effectively capture the differences in both local and global structures between two networks. Second, it iteratively propagate the novel differential scores through the network until convergence to obtain accurate rankings for all the nodes. We show that *DiffRank* is motivated by and well reflects the existing observations about the differences between two networks. Empirical experiments on three different applications show that our approach is effective and outperforms various baselines.

The remainder of this paper is organized as follows. Section 2 presents problem formalization. Section 3 details the proposed algorithm and analyzes its properties. We present performance evaluation and simulation in Section 4. The results on real-world datasets are reported in Section 5. Section 6 reviews related work. Finally, we offer conclusions and research directions in Section 7.

2. PRELIMINARIES

2.1 Problem Formulation

We will now introduce the notations to be used in the rest of the paper; then, we formally present the problem statement. Given two networks represented by graphs $G^A(V, E^A)$ and $G^B(V, E^B)$, where V is the set of N nodes and E^k is the set of edges in G^k, $k \in \{A, B\}$. An edge between two nodes u and v, with a weight $w^k(u, v)$ in G^k, determines

the strength of the interaction between the two nodes. The weight of each edge must be a non-negative value, 0 if the nodes are not connected to each other, or 1 in unweighted graphs. In this work, we focus our discussion to undirected networks with no self-links.

Problem Formulation: *Given two networks, G^A and G^B, the goal is to find the differential hubs that best explain the differences between the two networks. The final output of DiffRank is a vector*

$$\Pi =< \pi_1, \pi_2, ..., \pi_N >$$

where π_v denotes the rank of the differential node v.

2.2 Basic Intuition

A reasonable and accurate model for differential networks should not only capture the changes in the local structure, but also the changes in the global structure. Before formally introducing the algorithm, we first explain several key observations that motivate our approach.

Connectivity: The connectivity, or the degree, of a node is the number of other nodes that it is connected to. Nodes with the highest number of edges, known as the hubs, play an essential role in the analysis of networks. Pair-wise comparisons of the degree of each node in the two networks, as proposed in [12], may not lead to accurately identifying the differential hubs. For example, node 4 in Figure 1 has the same degree in both networks but the edges are different.

Centrality: Centrality is important in understanding many networks such as social networks [5], co-authorship networks [8] and biological networks [14]. Moreover, central nodes can have high influence on their neighbors [27]. Betweenness Centrality (BC) can be used to measure the centrality for each node, which is proportional to the sum of the shortest paths passing through it [11]. If P_{st} is the number of shortest paths from node s to node t, where $s \neq t$, and $P_{st}(v)$ is the number of shortest paths from s to t that pass through a node v, where $s \neq v$ and $t \neq v$, then the BC of the node v, $BC(v)$, can be computed as $\sum_{s \neq t} \frac{P_{st}(v)}{P_{st}}$ [10].

Identifying the shortest paths between two nodes is critical in several applications, such as social and biological networks [14], and the influence maximization problem [6]. Usually, the weights of the edges represent the strength of the interactions (or correlations) between the nodes. Therefore, distance values should be calculated from the weight values in order to calculate the shortest paths. For example, if $w(u, v)$ is the weight of interactions between two nodes u and v, then the weight on each edge can be translated to distance path using $1 - w(u, v)$ or $-log(w(u, v))$ [6]. We expect these intuitions and observations to be helpful in designing the proposed algorithm.

3. THE PROPOSED MODEL

In this section, we present the proposed algorithm, DiffRank, for finding the differential hubs. By considering connectivity and centrality, our algorithm can capture both local and global differences between two homogenous networks under a unified framework. We also provide some theoretical analysis of the proposed algorithm.

3.1 Local Structure Measure

Differential connectivity measures the local differences between two networks, G^A and G^B. Rather than just comparing the degree of a given node in both networks as in [12], we consider the actual weights of all edges in computing the differential connectivity. Moreover, we integrate the rank of each neighbor to weight the differential connectivity. The differential rank of each node will be propagated in the network. If two nodes are connected, then the propagation of the differential score between them is proportional to the weight of the edge connecting them.

Definition 1. (**Differential Connectivity**) Given two networks, G^A and G^B, we define the differential connectivity of node v, at iteration i as:

$$\Delta C^i(v) = \sum_{u=1}^{N} \frac{|w^A(u, v) - w^B(u, v)|.\pi_u^i}{\sum_{z=1}^{N} |w^A(u, z) - w^B(u, z)|} \quad (1)$$

π_v^i is the rank of node v at iteration i. It is initialized to $\frac{1}{N}$, and will be updated as explained later in this section. If a given node has the same set of edges in both networks, with the same weights, then the differential connectivity of that node will be 0. On the other hand, when a node has different sets of edges, it will get a high value for the differential connectivity. In addition to the number of edges and their weights, differential connectivity of a node also depends on the differential scores of the neighbors it is connected to. Each node is initialized with a uniform score $\pi_u = \frac{1}{N}$.

3.2 Global Structure Measure

Comparing the values of BC may not detect the topological changes between the two networks. For example, the shaded node in Figure 2 has the same value for BC in both networks. However, the shortest paths that pass through that node are different between the two networks. Therefore, we propose to consider the actual shortest paths to compare the centrality role of a node between two networks.

Definition 2. (**Differential Betweenness Centrality**) Let SP_k^v be a binary $N \times N$ matrix, such that $SP_k^v(s, t) = 1$ if one of the shortest paths from s to t passes through the node v in network $k = \{A, B\}$, where $s \neq t$, and it is 0 otherwise. We define differential betweenness centrality of a node v as follows:

$$\Delta BC(v) = \sum_{s=1}^{N} \sum_{t=1}^{N} (|SP_A^v(s, t) - SP_B^v(s, t)|) \quad (2)$$

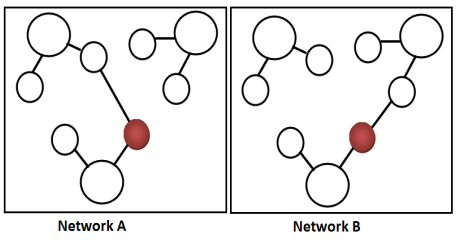


Figure 2: An illustration of differential betweenness centrality. The shaded node has the same value of betweenness centrality in both networks. However, the shortest paths that pass through that node are different between the two networks.

3.3 The DiffRank Algorithm

In addition to connectivity, centrality is important because changes in the central nodes could significantly alter the interconnection and the topology of the network. Therefore, we integrate both differential connectivity and differential betweenness centrality in the proposed algorithm.

We propose DiffRank algorithm that optimizes an objective which is a linear combination of differential connectivity and differential betweenness centrality (parameterized by λ) within a PageRank-style framework [15], such that the rank of each node v is computed as follows:

$$\pi_v^i = (1-\lambda).\frac{\Delta BC(v)}{\sum_{u=1}^{N}\Delta BC(u)} + \lambda.\Delta C^i(v) \quad (3)$$

The parameter λ controls the trade-off between differential connectivity and differential betweenness centrality. It can be assigned any value in the range $[0,1]$. When $\lambda = 0$, the ranking depends only on the differential betweenness centrality, and when $\lambda = 1$, the ranking depends only on the differential connectivity. Any other value of λ combines both terms in the ranking.

The integration of BC in the ranking formula adds significant global topological information to the differential analysis of networks. It was shown that BC is not significantly correlated with PageRank [8]. This means that BC can measure different prospectives compared to what PageRank can measure [8]. Therefore, integrating both connectivity and centrality enables DiffRank to capture the changes in both the local and global topological structures.

3.4 Preservation and Convergence

To begin with, all the nodes are initialized to $\frac{1}{N}$ (uniform distribution), so that the sum of the rankings is 1, $\sum_{v=1}^{N}\pi_v^i = 1$. The rankings will be updated in each iteration. There is no need to normalize after each step since the sum of the rankings is preserved to unity.

LEMMA 1. *The sum of the node ranks (Π_Δ) obtained by DiffRank are preserved to unity.*

PROOF. Let us assume that the algorithm is in the iteration i and the $\sum_{v=1}^{N}\pi_v^i = 1$, now we will show that the sum of ranking is preserved for the next iteration $(i+1)$:

$$\begin{aligned}
\sum_{v=1}^{N}\pi_v^{i+1} &= \sum_{v=1}^{N}\left(\frac{(1-\lambda).\Delta BC(v)}{\sum_{u=1}^{N}\Delta BC(u)} + \lambda.\sum_{u=1}^{N}\Delta DC^i(v)\right)\\
&= (1-\lambda).\left(\frac{\sum_{v=1}^{N}\Delta BC(v)}{\sum_{u=1}^{N}\Delta BC(u)}\right)\\
&+ \lambda.\left(\sum_{v=1}^{N}\sum_{u=1}^{N}\frac{|w^A(u,v)-w^B(u,v)|.\pi_u^i}{\sum_{z=1}^{N}|w^A(u,z)-w^B(u,z)|}\right)\\
&= (1-\lambda)\\
&+ \lambda.\left(\sum_{u=1}^{N}\pi_u^i\frac{\sum_{v=1}^{N}|w^A(u,v)-w^B(u,v)|}{\sum_{z=1}^{N}|w^A(u,z)-w^B(u,z)|}\right)\\
&= (1-\lambda) + \lambda.\sum_{u=1}^{N}\pi_u^i\\
&= (1-\lambda)+\lambda = 1
\end{aligned}$$

□

One issue that needs to be resolved is handling the sinks (or isolated nodes). These nodes will be assigned uniform weighted edges to each other node in the network in order to ensure the convergence of the DiffRank algorithm [17].

THEOREM 1. *The result from the DiffRank model converges to a unique rank vector Π_Δ.*

PROOF. Let us define $M^{N\times N}$ as a square matrix, such that

$$M_{uv} = \frac{|w^A(u,v)-w^B(u,v)|}{\sum_{z=1}^{N}|w^A(u,z)-w^B(u,z)|}$$

We replace all rows with zeros by $\frac{1}{N}$. Now, M is considered to be a stochastic matrix in which the sum of each row is 1: $\sum_{v=1}^{N}M_{uv} = 1$, $1 \le u \le N$. Let P denote a vector of length N, such that

$$P_v = \frac{\Delta BC(v)}{\sum_{u=1}^{N}\Delta BC(u)}$$

then we will have $\sum_{v=1}^{N}P_v = 1$. Finally we define a new matrix M' as follows:

$$M' = \lambda.M + (1-\lambda).P^T$$

. The combination of the stochastic matrix M, and the vector P reduces the effect of the isolated nodes $\lambda \in [0,1]$. The rank vector, Π_Δ, can be computed by solving the following eigenvector problem:

$${\Pi_\Delta}^T M' = {\Pi_\Delta}^T$$

Since M' is a stochastic matrix, the DiffRank model is reduced to a personalized PageRank model for which a unique solution is guaranteed [17, 15]. □

3.5 Network-Specific Analysis

In some applications, we are interested in identifying the differential nodes of a specific network. For example, in gene networks, it is important to find the genes that are rewired in the cancer cells. For this purpose, we can modify some definitions based on the particular network of interest. To find the differential nodes in network B, the differential connectivity (ΔC) can be redefined as follows:

$$\Delta C'^i(v) = \sum_{u=1}^{N}\frac{max(w_B(u,v)-w_A(u,v),0).\pi_u^i}{\sum_{z=1}^{N}max(w_B(u,z)-w_A(u,z),0)} \quad (4)$$

This new definition excludes any edge in the network of interest if the corresponding edge in the other network has a higher weight. Similarly, the new definition of differential betweenness centrality, ΔBC, includes the unique shortest paths that are in the network of interest and excludes the unique shortest paths in the other network.

$$\Delta BC'(v) = \sum_{s=1}^{N}\sum_{t=1}^{N}max(SP_B^v(s,t)-SP_A^v(s,t),0) \quad (5)$$

Then the second version of DiffRank is modified as follows:

$$\pi_v^i = (1-\lambda).\frac{\Delta BC'(v)}{\sum_{u=1}^{N}\Delta BC'(u)} + \lambda.\Delta C'^i(v) \quad (6)$$

With these two version of DiffRank, we can solve the following problems:

1. Find the differential hubs from two networks; this can be solved by the first version of DiffRank.

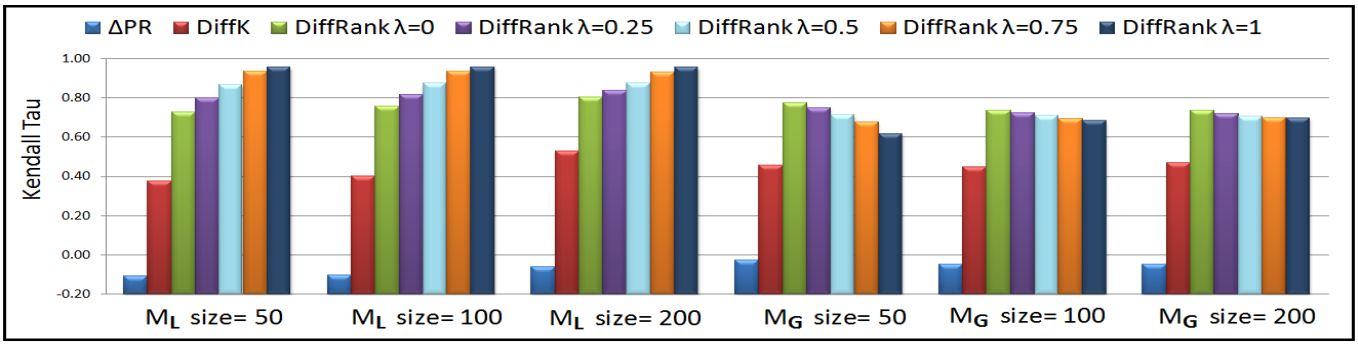


Figure 3: Results on Simulated networks with different sizes comparing with baseline methods with different λ values in DiffRank.

2. Find the network-specific differential hubs from two networks which are active in a particular network; this can be solved by the second version of DiffRank.

3.6 Scalability

Finding the shortest paths is the most time-consuming computation in the proposed model. Using the traditional Dijkstra's algorithm, computing the shortest paths between two nodes needs $O(m + nlog(n))$ where m is the number of links, and n is the number of nodes in the graph and solving all-pairs shortest paths requires $O(nm + n^2logn)$ time and $O(n^2)$ space [14]. However, Recent methods have been proposed to reduce the computational overhead by using approximation methods [14], which helps in efficiently applying DiffRank on large-scale networks.

4. EXPERIMENTS ON SIMULATED DATA

In this section, we will first describe the baseline methods for comparison and the evaluation measures used to evaluate the proposed algorithm. Then, we present a simulator that is used to generate scale-free networks based on two different models: random and evolving models.

4.1 Baseline Methods

To the best of our knowledge, ranking differential hubs in two networks has not been studied before, and there is no standard metric that can be used to capture the differences between two networks. As a baseline method, we used the difference between the scores given by the PageRank algorithm [22] in the two networks. We denote this measure as ΔPR, and define it as:

$$\Delta PR(v) = |PR^A(v) - PR^B(v)| \tag{7}$$

Where $PR^K(v)$ is the score for the node v obtained by applying PageRank on network k. In addition, we applied the differential connectivity score, DiffK [12]. Given the i^{th} node, $k^A(i)$ and $k^B(i)$ are the connectivity of the i^{th} node in networks A and B, respectively, DiffK is defined as follows:

$$DiffK(v) = |K^A(v) - K^B(v)| \tag{8}$$

where $K^A(v) = \frac{k^A(v)}{max(k^A)}$ and $K^B(v) = \frac{k^B(v)}{max(k^B)}$. When we refer to DiffRank, ΔPR and DiffK algorithms, we refer to the ranking of all the nodes based on their scores given by Equation (3), Equation (7) and Equation (8), respectively.

4.2 Evaluation Measures

Since there is no standard measure for comparing two networks, we developed two evaluation measures:

Local structure measure (M_L): This measure depends on comparing the edges of each node to find the differential nodes. It is a local measure which is defined as follows:

$$M_L(v) = \sum_{u=1}^{N} [w^A(u,v) - w^B(u,v)]^2 \tag{9}$$

Global structure measure (M_G): This measure captures the global changes in the networks, and it uses the shortest paths in the computation as follows: We define $dist(u, v, G^K)$ to be the distance between the nodes u and v in graph G^K computed through the shortest path between them, and we define $G_z^{K'}$ to be the same as G^K except that all the edges for node z are removed. Then, we define $\Delta_z dist(u, v, G^K) = (dist(u, v, G^K) - dist(u, v, G_z^{K'}))^2$. Finally, M_G is defined as follows:

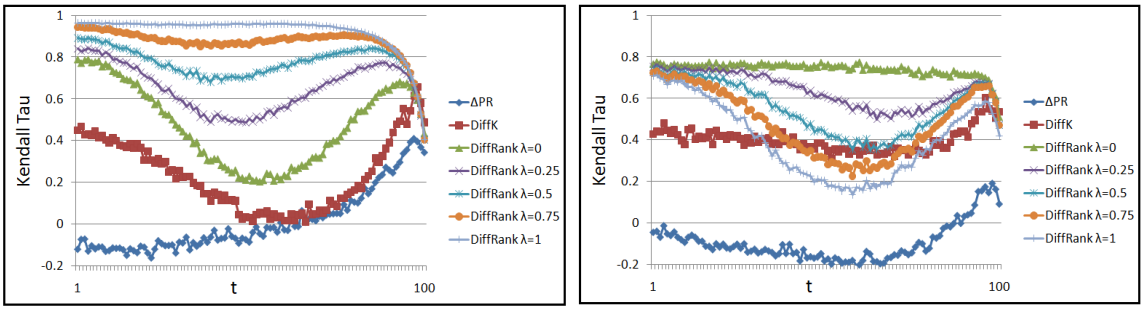
$$M_G(z) = \sum_{u=1}^{N}\sum_{v=1}^{N} [\Delta_z dist(u,v,G^A) - \Delta_z dist(u,v,G^B)]^2 \tag{10}$$

M_G measures the importance of each node to all other nodes in the network. We used $Kendall\ Tau$ statistic [18] to measure the correlation between the evaluation measures and the ranking algorithms.

4.3 Random Model

We developed a simulator to generate synthetic differential scale-free networks. First, we start with a small network as a seed, then we follow the preferential attachment rule [3, 20] in adding new nodes. The probability for any node to be connected with another node is proportional to its degree and equals to $\frac{d(v)}{\sum_{u=1}^{n} d(u)}$ where n is the number of existing nodes in the network [3]. To generate two differential networks of size n, we start with the same seed for each network of size m, and then generate the remaining $n - m$ nodes for each network separately to obtain two networks with different sets of edges.

Figure 3 shows the results on simulated data of different sizes: 50, 100 and 200. These results are the average of 10 runs. As shown in this figure, our proposed method outperforms the other methods in all of the cases. When using the local measure M_L, as the values of λ increase from 0



(a) Evaluation based on M_L (b) Evaluation based on M_G

Figure 4: Results of ΔPR, DiffK and DiffRank on the data generated using the evolving model.

to 1 results are improved. Since M_L depends only on the differences in the local structures, differential connectivity better fits such cases. On the other hand, when using M_G measure, as the values of λ decrease from 1 to 0 better results are obtained. Since M_G depends on the differences in the global structures between two networks. Hence, we can conclude that parameter λ plays a significant role in optimizing both the local and global measures. Based on our experiment analysis, we recommend to set its value to 0.75. We used this value in all of the experiments, unless stated otherwise.

4.4 Evolving Model

In the evolving model, we start with the same seed of m nodes in each network; then, we add n more nodes in each network such that the last $m - t$ are added separately in each network while the first $m+t$ nodes in the two networks will have the same edges, $1 \leq t \leq n$. We used $m = 5$ and $n = 100$ in this model. We compute all the rankings for each value of t. The results are shown in Figure 4. These are the average of 10 runs.

The results are close to what we obtained in the first model. As the value of t increases from 1 to 100 the differences between the two networks decrease. For example, when $t = 1$, only the seed nodes have the same connections between the two networks, and the remaining 100 nodes have different connections. When the value of $t = 100$, all the nodes have the same connections in both networks except the last node. When the value of t is small and the there are big differences between the two networks, DiffRank outperforms the other methods in majority of the cases.

5. EXPERIMENTS ON REAL DATA

In this section, we illustrate the performance of DiffRank algorithm on the *DBLP* dataset and a lung cancer datasets. For the *DBLP* dataset, we performed two experiments. One based on topic variation, and the other one based on temporal variation. For the lung cancer dataset, we focused on the phenotypic variation.

5.1 Results for Topic Variation

In the first experiment we used the *Arnetminer* **DBLP** dataset[1]. In this experiment, we included all the authors who have at least two papers from 2000 to 2010 in database conferences (DM) and data mining (DM) conferences.

[1]http://www.arnetminer.org/DBLP_Citation

Table 2: Results of the network-specific DiffRank algorithm on the DM network.

diffRank	Author	ΔPR	diffK
1	Philip S. Yu	5180.5	33
2	Christos Faloutsos	5156	81
3	Zheng Chen	3067.5	77
4	Jiawei Han	5187	329
5	Qiang Yang	4395	36
6	Wei Fan	4823	106
7	Heikki Mannila	4290.5	65
8	Jun Yan	2	205
9	Vipin Kumar	4483	76
10	Eamonn J. Keogh	4903	247
11	Huan Liu	1878	70
12	Chris H. Q. Ding	1	147
13	Hui Xiong	3245	71
14	Tao Li	1415.5	100
15	Bing Liu	3301.5	108

- **DB**: {*ICDE, VLDB, SIGMOD, PODS, EDBT* }.
- **DM**: { *KDD, SDM, ICDM, PKDD, PAKDD* }.

Two co-authorship networks were constructed from this dataset. The first network was constructed from the papers published in the DB conferences, while the second network was constructed from the papers published in the DM conferences. The nodes are the authors and the edges represent collaborations. The weight of each edge represents the number of papers written by the two linked authors together. Each network contains 5188 nodes. The number of edges in the DB network is 24916, and the number of edges in the DM network is 12932.

Table 1 shows the results of applying ΔPR, DiffK and DiffRank on the DB and DM networks. As discussed earlier, there are several factors that affect the ranking such as connectivity, centrality and the rank of the neighbors. For simplicity, we provide some examples of the interesting results by explaining the number of links only. It should be noted that though we are just mentioning the number of links, our approach optimizes for the differential propagation network rather than just based on the number of links.

The top ranked author, *Divesh Srivastava*, has published with 97 authors, 90 of them published in the DB conferences only. Similarly, for *Beng Chin Ooi* and *Gerhard Weikum*, most of the connections are in the DB network. While considering top ranked authors such as *Philip S. Yu* and *Jiawei Han*, who published in both DB and DM conferences, they were top ranked because they collaborated with a different

Table 1: Results of ΔPR, DiffK and DiffRank on the DB and DM networks

Rank	Top 10 based on Δ PR			Top 10 based on DiffK			Top 10 based on DiffRank		
	Author	DiffK	DiffRank	Author	ΔPR	DiffRank	Author	ΔPR	DiffK
1	Chris H. Q. Ding	147	106	Beng Chin Ooi	2926	3	Divesh Srivastava	4227.5	3
2	Jun Yan	205	68	Gerhard Weikum	4821	7	Philip S. Yu	5180.5	33
3	Hillol Kargupta	147	179	Divesh Srivastava	4227.5	1	Beng Chin Ooi	2926	1
4	Pang-Ning Tan	205	140	Michael J. Carey	73	17	Christos Faloutsos	5156	81
5	Zhi-Hua Zhou	83	194	Ioana Manolescu	76	26	Jiawei Han	5187	329
6	Jieping Ye	101	125	Alon Y. Halevy	67	13	Nick Koudas	4711	13
7	Takashi Washio	147	183	Donald Kossmann	69	15	Gerhard Weikum	4821	2
8	Changshui Zhang	250	223	Daniela Florescu	95	38	Surajit Chaudhuri	63	10
9	Jing Gao	224	136	R. Ramakrishnan	4689	11	Jeffrey Xu Yu	5185	542
10	Joydeep Ghosh	224	254	Surajit Chaudhuri	63	8	Kian-Lee Tan	3546.5	18

Table 3: Results of ΔPR, DiffK and DiffRank two DM networks from two mutually exclusive time intervals.

Rank	Top 10 based on ΔPR			Top 10 based on DiffK			Top 10 based on DiffRank		
	Author	DiffK	DiffRank	Author	ΔPR	DiffRank	Author	ΔPR	DiffK
1	Honghua Dai	81	110	Vipin Kumar	2271.5	10	Philip S. Yu	2433.5	2384
2	Jiong Yang	5	60	Hongjun Lu	3	32	Jiawei Han	2431	131.5
3	Hongjun Lu	2	32	Zhi-Hua Zhou	2108	4	Christos Faloutsos	2424.5	6
4	Yuchang Lu	22	445	Jimeng Sun	1040.5	8	Zhi-Hua Zhou	2108	3
5	William Perrizo	208	1776	Jiong Yang	2	60	Heikki Mannila	2424.5	86
6	Chidanand Apte	15.5	149	Christos Faloutsos	2424.5	3	Jian Pei	2417	690
7	Djamel A. Zighed	121.5	343	Ke Wang	2368.5	31	S. Papadimitriou	2379	926
8	Ricardo Vilalta	39	341	Wensi Xi	32	386	Jimeng Sun	1040.5	4
9	Ron Kohavi	81	128	Sheng Ma	484	39	Wei Fan	2384.5	962
10	C. Ratanamahatana	51	450	Xindong Wu	2221.5	23	Vipin Kumar	2271.5	1

set of authors in each field. Therefore, the connections for such authors are different between DB and DM networks.

The results in Table 1 include authors who have many papers either in DB or DM or in both of them. However, we performed network-specific differential analysis using the second version of DiffRank, defined in Equation(6), by looking at the differential authors who publish in DM conferences, and we reported the top 15 authors in Table 2. These authors are highly active in DM conferences, and they have different sets of links compared to their collaborations in the DB conferences. For example, *Jun Yan* and *Chris H. Q. Ding* published only in DM conferences.

We also compared the results in Table 1 to the results obtained from the Topic Affinity Analysis (TAP) algorithm [24]. The TAP results include a list of representative authors for the DB and the DM topics. We found the number of common authors between TAP and ΔPR is 0, between TAP and DiffK is 1 and between TAP and DiffRank is 6. In addition, 7 authors appeared in both Table 2 and [24]. Therefore, the results reported by DiffRank can also be considered as representative for the topic of interest.

5.2 Results for Temporal Variation

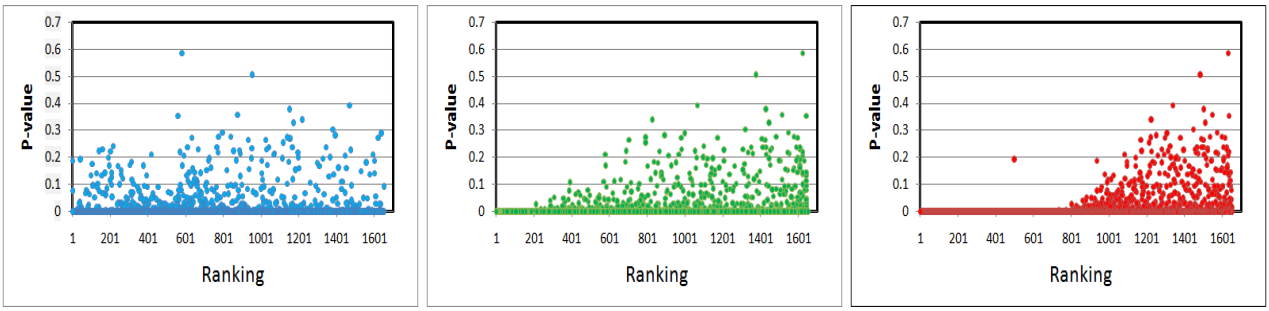
In this experiment, we consider the authors who published in the *DM* conferences. We constructed two networks, the first network included the authors who published during the time interval 2000-2005, and the second network included the authors who published during the time interval 2006-2010. The number of nodes in each network is 2434. The number of edges in the first network is 6906, and the number of edges in the second network is 13736. Table 3 shows the results of ΔPR, DiffK and DiffRank on the two networks described. In this kind of analysis, interesting results can be obtained about authors who started publishing in the second time interval. For example *Zhi-Hua Zhou* has 4 collaborations in the first time interval, but he has 29 collaborations from 2006 to 2010. Similarly *Jimeng Sun* has only 2 collaborations in the first time interval, but he has 25 in the second time interval. Such results help in identifying evolving authors in a specific topic of interest. Merely, looking at the results (qualitatively), one can see that our approach yields more prominent authors in the data mining community compared to the other two methods.

Table 4: Results of DiffRank on lung cancer dataset.

diffRank	Gene Symbol	ΔPR	DiffK
1	CSF1	1911	1406
2	AFF3	1960.5	1413
3	CLDN14	1950.5	1397.5
4	RAB32	1058	1152
5	RBL1	1375	1237
6	PPP1CB	173	843
7	PNRC1	109.5	1085
8	PAX7	1254.5	1475
9	NRTN	1906.5	1618
10	HBE1	743.5	539
11	RUNX2	1525.5	1500
12	POU2F3	1828	1373
13	ZFP36L1	869	615
14	PAX8	1881	1494
15	PPP2R5D	1828	1100

5.3 Results for Phenotypic Variation

The proposed work was originally motivated by the biological application described in the introduction section. We provide more quantitative results in this application. In this experiment, we applied DiffRank algorithm on lung cancer dataset obtained from [9]. This dataset is comprised of 1975 genes and 169 samples: 102 are cancer samples, and 67 normal samples. We constructed the gene networks using Mutual Information (MI) as described in [4]. The first net-



(a) ΔPR (b) DiffK (c) DiffRank

Figure 5: The distributions of P-value obtained by applying the MDA test on the lung cancer dataset. The Y-axis shows the values of the MDA test, and the X-axis shows the gene ranking obtained by ΔPR, DiffK and DiffRank algorithms.

work, constructed from the normal samples, contains 655316 edges, and the second network, constructed from the cancer samples, contains 146117 edges.

To statistically test the differential connectivity of genes, we used the mean absolute distance (MDA) test that was proposed in [13]:

$$D(g) = \frac{1}{N-1} \sum_{g' \neq g} |MI^A(g, g') - MI^B(g, g')| \qquad (11)$$

where $MI^k(g, g')$ is the mutual information of the genes g and $g^{'}$ in network $k = \{A, B\}$. The p-value of this test was computed by permutating the samples, and then computing the MDA statistic for each pair of genes. If C_1 is the number of samples for the first condition, and C_2 is the number of samples for the second condition. The permutation is performed for the $C_1 + C_2$ conditions. The first C_1 samples will be considered as the samples for the first condition, and the last C_2 samples will be considered as the samples for the second condition [13]. The approximate p-value of each gene is calculated as:

$$P - value(g) = \frac{1}{P} \sum_{p=1}^{P} (D_p \geq D_{observed}) \qquad (12)$$

where P is the number of permutations, D_p is the value of the MDA test on the permutation p and $D_{observed}$) is the value of the MDA test on the Original data . The distributions of p-values obtained from each algorithm are shown in Figure 5. If a gene has lower p-value, it implies that the gene is significant and more differential than other genes with higher p-values. As shown in this Figure, the top differential nodes obtained using our method are statistically significant. From this figure, we can see that the top differential genes obtained from the other two baseline approaches have higher p-values and many of them are not statistically significant with p-values greater than 0.05.

6. RELATED WORK

To the best of our knowledge, DiffRank is the first algorithm to rank the nodes of two networks based on their differential behavior and to identify the differential hubs. Community evolution has been studied in several papers such as [2] and [25] to discover evolving groups in social networks. An event-based framework for analyzing the evolution of dynamic graphs was proposed in [1]. In our proposed algorithm, we are interesting in studying the individual behaviour change in two networks to identify the differential hubs.

Topic Affinity Analysis (TAP) was proposed in [24] to differentiate the social influences from different topics. This algorithms takes as input an existing network structure and topic distribution, and it aims to find topic-level social influence graph [24]. This algorithm assumes one network structure for different topics, while in this paper, each topic will be represented in a different network.

In the biological domain, there are some differential measures that have been proposed to measure the differences between two gene networks. Examples of such measures include comparing the distribution of small subgraphs, referred to as *graphlets* [23] and DiffK, used in this paper for comparisons, was proposed in [12] to compare the genes in two networks based on their degrees. Most of these methods depend on pair-wise comparisons of nodes in two networks However, our proposed algorithm captures local and global changes between two networks to obtain a deeper insight into disease networks.

7. CONCLUSIONS AND FUTURE WORK

In this work, we propose the novel problem of finding the differential hubs in homogenous networks. Given two networks with same nodes but different edges, we could find and mine the differential hubs that are responsible for the differences between the two networks. We make several key observations about how the local and global measures mutually influence the ability to identify the differential nodes, and propose a novel algorithm, called DiffRank, for mining the top K differential hubs in the two networks. Comprehensive experimental studies on real-world datasets and synthetically generated datasets showed that our approach outperforms the baselines. In the future, we will study how to automatically set the value for the parameter λ based on the structure of the networks.

Our approach can potentially enable informative analysis on various real-world applications. This work opens the door to several interesting directions for future work. One interesting future research is to further explore the problem of differential networking analysis in heterogenous or multimode networks. Another interesting research directions is to integrate the concepts of influential nodes [26] and effectors [19] in the differential analysis of multiple social networks.

8. REFERENCES

[1] S. Asur, S. Parthasarathy, and D. Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 913–921, New York, NY, USA, 2007.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 44–54, New York, NY, USA, 2006.

[3] A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999.

[4] A. J. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 418–429, 2000.

[5] C. Chen, X. Yan, F. Zhu, J. Han, and P. S. Yu. Graph olap: Towards online analytical processing on graphs. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 103–112, Washington, DC, USA, 2008.

[6] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1029–1038, New York, NY, USA, 2010.

[7] A. de la Fuente. From 'differential expression' to 'differential networking' identification of dysfunctional regulatory networks in diseases. *Trends in Genetics*, 26(7):326–333, July 2010.

[8] Y. Ding, E. Yan, A. Frazho, and J. Caverlee. Pagerank for ranking authors in co-citation networks. *J. Am. Soc. Inf. Sci. Technol.*, 60:2229–2243, November 2009.

[9] G. Fang, R. Kuang, G. Pandey, M. Steinbach, C. L. Myers, and V. Kumar. Subspace differential coexpression analysis: problem definition and a general approach. *Pacific Symposium on Biocomputing*, pages 145–156, 2010.

[10] M. Francesconi, D. Remondini, N. Neretti, J. Sedivy, L. Cooper, E. Verondini, L. Milanesi, and G. Castellani. Reconstructing networks of pathways via significance analysis of their intersections. *BMC Bioinformatics*, 9(Suppl 4):S9, 2008.

[11] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, March 1977.

[12] T. Fuller, A. Ghazalpour, J. Aten, T. Drake, A. Lusis, and S. Horvath. Weighted gene coexpression network analysis strategies applied to mouse weight. *Mammalian Genome*, 18:463–472, 2007.

[13] R. Gill, S. Datta, and S. Datta. A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics*, 11(1):95, 2010.

[14] A. Gubichev, S. Bedathur, S. Seufert, and G. Weikum. Fast and accurate estimation of shortest paths in large graphs. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 499–508, New York, NY, USA, 2010.

[15] T. H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15:784–796, 2003.

[16] G. C. Kennedy, H. Matsuzaki, S. Dong, W. M. Liu, J. Huang, G. Liu, X. Su, M. Cao, W. Chen, J. Zhang, W. Liu, G. Yang, X. Di, T. Ryder, Z. He, U. Surti, M. S. Phillips, B. M. T. Jacino, S. P. Fodor, and K. W. Jones. Large-scale genotyping of complex DNA. *Nature Biotechnology*, 21(10):1233–7, 2003.

[17] A. N. Langville and C. D. Meyer. Deeper inside pagerank. *Internet Mathematics*, 1:2004, 2004.

[18] M. Lapata. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484, 2006.

[19] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila. Finding effectors in social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1059–1068, New York, NY, USA, 2010.

[20] Q. Mei, J. Guo, and D. Radev. Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1009–1018, New York, NY, USA, 2010.

[21] O. Odibat, C. K. Reddy, and C. N. Giroux. Differential biclustering for gene expression analysis. In *Proceedings of the ACM Conference on Bioinformatics and Computational Biology (BCB)*, pages 275–284. ACM, 2010.

[22] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[23] N. Prđulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[24] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 807–816, New York, NY, USA, 2009.

[25] L. Tang, H. Liu, J. Zhang, and Z. Nazeri. Community evolution in dynamic multi-mode networks. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 677–685, New York, NY, USA, 2008.

[26] Y. Wang, G. Cong, G. Song, and K. Xie. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1039–1048, New York, NY, USA, 2010. ACM.

[27] Z. Wen and C.-Y. Lin. On the quality of inferring interests from social neighbors. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 373–382, New York, NY, USA, 2010.