PageRank Revisited: On the Relationship between Node Degrees and Node Significances in Different Applications *

Jung Hyun Kim
Arizona State University
Tempe, AZ 85287, USA
jkim294@asu.edu

K. Selçuk Candan
Arizona State University
Tempe, AZ 85287-8809
candan@asu.edu

Maria Luisa Sapino
University of Torino
I-10149 Torino, Italy
marialuisa.sapino@unito.it

ABSTRACT

Random-walk based techniques, such as PageRank, encode the structure of the graph in the form of a transition matrix of a stochastic process from which significances of the graph nodes can be inferred. Recommendation systems leverage such *node significance* measures to rank the objects in the database. Context-aware recommendation techniques complement the data graph with additional data that provide the *recommendation context*. However, despite their wide-spread use in many graph-based knowledge discovery and recommendation applications, conventional PageRank-based measures have various shortcomings. As we experimentally show in this paper, one such shortcoming is that PageRank scores are tightly coupled with the degrees of the graph nodes, whereas in many applications the relationship between the *significance* of the node and its degree in the underlying network may not be as implied by PageRank-based measures. In fact, as we also show in the paper, in certain applications, the *significance* of the node may be *negatively* correlated with the node degree and in such applications a naive application of PageRank may return poor results. To address these challenges, in this paper, we propose *degree decoupled PageRank (D2PR)* techniques to improve the effectiveness of PageRank based knowledge discovery and recommendation systems. These suitably penalize or (if needed) boost the transition strength based on the degree of a given node to adapt the node significances based on the network and application characteristics.

1. INTRODUCTION

In recent years, there has been considerable interest in measuring the *significance of a node in a graph* and *relatedness between two nodes in the graph*, as if measured accurately, these can be used for supporting many knowledge discovery, search, and recommendation tasks [1, 7, 9, 12, 26]. The *significance* of a node in a given graph often needs to reflect the topology of the graph. Measures like the *betweenness* measure [27] and the *centrality/cohesion* [5], help quantify how *significant* any node is on a given graph based on the underlying graph topology. The *betweenness* measure [27], for example, quantifies whether deleting the node would disconnect or disrupt the graph. *Centrality/cohesion* [5] measures quantify how close to a clique the given node and its neighbors are. Other *authority*, *prestige*, and *prominence* measures [1, 5, 6] quantify the significance of the node through eigen-analysis or random walks, which help measure how reachable a node is in the graph.

1.1 PageRank as a Measure of Significance

Since enumerating all paths among the graph nodes would require time exponential in the size of the graph, random-walk based techniques encode the structure of the network in the form of a transition matrix of a stochastic process from which the node significance can be inferred.PageRank [6] is one of the most widely-used random-walk based methods for measuring node significance and has been used in a variety of application domains, including web search, biology, and social networks. The basic thesis of PageRank is that a node is important if it is pointed to by other important nodes – it takes into account the connectivity of nodes in the graph by defining the score of the node $v_i \in V$ as the amount of time spent on v_i in a sufficiently long random walk on the graph. More specifically, given a graph $G(V, E)$, the PageRank scores are represented as $\vec{r}$, where

$$\vec{r} = \alpha \mathbf{T}_G \vec{r} + (1 - \alpha)\vec{t}$$

where $\mathbf{T}_G$ is a transition matrix corresponding to the graph G, $\vec{t}$ is a teleportation vector (such that $\vec{t}[i] = \frac{1}{\|V\|}$), and α is the residual probability (or equivalently, $(1 - \alpha)$ is the so-called teleportation probability). Unless the graph is weighted, the transition matrix, $\mathbf{T}_G$, is constructed such that for a node v with k (outgoing) neighbors, the transition probability from v to each of its (outgoing) neighbors will be $1/k$. If the graph is weighted, then the transition probabilities are adjusted in a way to account for the relative weights of the (outgoing) edges.

1.2 Tight Coupling of PageRank Scores of Nodes and their Degrees

Let us consider an undirected graph $G(V, E)$. There are two factors that contribute to the PageRank of a given node, $v \in V$:

- *Factor 1: Significance of Neighbors:* The more significant the neighbors of a node are, the higher its likelihood to be also significant.
- *Factor 2: Number of Neighbors (Degree of the Node) :* Even if the neighbors are not all significant, a large number of

*This work is supported by NSF Grants 1339835 "E-SDMS: Energy Simulation Data Management System Software", 1318788 "Data Management for Real-Time Data Driven Epidemic Spread Simulations", 1518939 "RAPID: Understanding the Evolution Patterns of the Ebola Outbreak in West-Africa and Supporting Real-Time Decision Making and Hypothesis Testing through Large Scale Simulations", and 1430144 "Fraud Detection via Visual Analytics: An Infrastructure to Support Complex Financial Patterns (CFP) based Real-Time Services Delivery".

Data Set	Listener Graph (Friendship edges, Last.fm)	Article Graph (co-author edges, DBLP)	Movie Graph (co-contributor edges, DBLP)
Correlation between PageRank and Degree	0.988	0.997	0.848

Table 1: Spearman's rank correlation between the node degree ranks and the node ranks' based on PageRank scores for various data graphs (see Section 4 for details of the data sets)

neighbors would imply that the node, v, is well-connected and, thus, likely to be structurally important.

In theory, these two factors should complement each other. In practice, however, the PageRank formulation described above implies that there is a very tight coupling between the degrees of the nodes in the graph and their PageRank scores (see Table 1).

1.2.1 Problem I: When a Large Node Degree Does Not Indicate High Node Significance

In this paper, we highlight (and experimentally show) that, in many applications, node degree and node significance are in fact inversely related and that the tight-coupling between node degrees and PageRank scores might be counter-productive in generating accurate recommendations.

EXAMPLE 1. *Consider, for example, a recommendation application where a movie graph, consisting of* `movie` *and* `actor` *nodes, is used for generating movie recommendations. In this application, the first factor (significance of neighbors) clearly has a positive contribution: a movie with good actors is likely to be a good movie and an actress playing in good movies is likely to be a good actress. On the other hand, the second factor (number of neighbors) may in fact be a negative contributor to node significance: the fact that an actor has played in a large number of movies may be a sign that he is a non-discriminating ('B movie') actor, whereas an actress with relatively fewer movies may be a more discriminating ('A movie') actress.*

As we see in Section 4, this observation turns out to be true in many applications, where (a) acquiring additional edges has a cost that is correlated with the significance of the neighbor (e.g. the effort one needs to invest to a high quality movie) and (b) each node has a limited budget (e.g. total effort an actor/actress can invest in his/her work).

1.2.2 Problem II: When PageRank Does Not Sufficiently Account for Contributions of Degrees

The mismatch between PageRank and node significance is not limited to the cases where node degrees are inversely related to the node significance. As we see in Section 4, there are other scenarios where PageRank may, in fact, fail to sufficiently account for the contribution of the node degrees to their significances.

1.3 PageRank Revisited: De-coupling Node Significance from Node Degrees

As we discussed above, one key shortcoming of the conventional PageRank scores is that they are often tightly coupled with the degrees of the graph nodes and in many applications the relationship between the *significance* of the node and its degree in the underlying network may not be as implied by PageRank-based measure: in certain applications, the *significance* of the node may be *negatively* correlated with the node degree, whereas in others PageRank may not be sufficient in accounting for degree contributions. Naturally, in such applications a naive application of PageRank in generating recommendations may return poor results.

To address these challenges, in this paper, we propose *degree decoupled PageRank (D2PR)* techniques to improve the effectiveness of PageRank based knowledge discovery and recommendation systems. These techniques suitably penalize or (if needed) boost[1] the transition strength based on the degree of a given node to adapt the node significances based on the network and application characteristics. This paper is organized as follows: Next, we discuss the related literature. In Sections 3, we introduce the proposed degree-decoupled PageRank techniques. We evaluate the proposed techniques in Section 4 and conclude in Section 5.

2. RELATED WORKS

2.1 Context-Sensitive PageRank

Path-length based definitions of node *relatedness*, such as those proposed by [4, 24] help capture the relatedness of a pair of nodes solely based on the properties of the nodes and edges on the *shortest* path between the pair. Random-walk based definitions, such as hitting distance [10, 21] and personalized page rank (PPR) score [1, 9, 16], of node relatedness further take into account the density of the edges: as in path-length based definitions, random-walk based definitions also recognize that a node is more related to another node if there are short paths between them; however, random walk-based definitions of relatedness also consider how well the given pair of nodes are connected.

In [7], authors construct a transition matrix, $\mathbf{T}_S$, where edges leading away from the seed nodes are weighted less than those edges leading towards the seed nodes. An alternative approach for contextualizing PageRank scores is to use the PPR techniques [1,9] discussed in the introduction. One key advantage of this teleportation vector modification based approach over modifying the transition matrix, as in [7], is that the term α can be used to directly control the *degree of seeding (or personalization)* of the PPR score. [10, 21] rely on a random walk hitting time based approach, where the hitting time is defined as the expected number of steps a random walk from the source vertex to the destination vertex will take. [17] leveraged these properties of PPR to develop locality-sensitive algorithms to rank nodes of graphs which are relative to a given set of seed nodes efficiently.

2.2 Improvements to the PageRank Function

Due to the obvious relationship between ranking and monetary rewards (e.g. through selling of advertisements on web search applications), there has been considerable effort in engineering (or manipulating) graphs in a way to maximize ranking scores of particular nodes. This is commonly referred to as *PageRank optimization*. One way to achieve this goal is carefully adding or removing certain links: If, for example, one or more colluding webmasters can add or remove edges, PageRank scores of target web pages or domains can be increased [23]. [20] established several bounds indicating to what extent the rank of the pages of a website can be changed and the authors derived an optimal referencing strategy to boost PageRank scores. A related, but opposite, problem is to protect the PageRank scores against negative links (which may indicate, for example, negative influence or distrust in a social network), artificial manipulation, and spam. [3], for example, focused on identifying spam pages and link farms and showed that better PageRank scores can be obtained after filtering spam pages and links. In [14], authors show that PPR algorithms that do not differentiate among the seed nodes may not properly rank nodes and present robust personalized PageRank (RPR) strategies, which are insensitive to noise in the set of seed nodes.

[1]In this context, *de-coupled* does not necessarily imply *de-correlated*. In fact, D2PR can boost correlation between node degree and PageRank if that is required by the application.

There are some efforts to change the impact of degrees on the PageRank computation. [2] proposed a way to boost the power of low-degree nodes in a network. The impact from nodes which are important but are not hubs is relatively small compared to other nodes which are less important with high degrees. To boost the low-degree important nodes for equal opportunity, the teleportation vector is modified with being proportional to the degrees of nodes. [11] boosted the degrees of nodes to reduce the expected cover time of the entire graph by the biassed random-walk.

3. DEGREE DE-COUPLED PAGERANK

The key difficulty of de-coupling node degrees from the PageRank scores is that the definition of the PageRank, based on random walk transitions, is inherently dependent on the number of transitions available from one node to the other. As we mentioned above, the more ways there are to reach into a node, the higher will be its PageRank score.

3.1 Desideratum

Therefore, to de-couple the PageRank score from node degrees, we need to modify the transition matrix. In particular, for each node v_i in the graph, we would like to be able to control the transition process with a *single parameter* (p), such that

- if $p \ll -1$, transitions from node v_i are $\sim 100\%$ towards the neighbor with the highest degree,
- if $p = -1$, transition probabilities from node v_i are proportional to the degrees of its neighbors,
- if $p = 0$, the transition probabilities mirror the standard PageRank probabilities (assuming undifferentiated neighbors),
- if $p = 1$, transition probabilities from node v_i are inversely proportional to the degrees of its neighbors,
- if $p \gg 1$, transitions from node v_i are $\sim 100\%$ towards the neighbor with the lowest degree.

In other words, the transition function should *de-couple* the transition process from node-degrees and *penalize* or *boost* the contributions of node degrees in the transition process, as needed.

3.2 Degree De-coupling Transition Matrix

In this subsection, we will consider degree de-coupling of the transition matrix as implied by the above desideratum.

3.2.1 Undirected Unweighted Graphs

Let $G = (V, E)$ be an undirected and unweighted graph. Let α also be a given residual probability parameter, and $deg(v)$ be a function which returns the number of edges on the node v. We represent degree de-coupled PageRank (D2PR) scores in the form of a vector

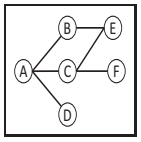
$$\vec{d} = \alpha \mathbf{T}_D \vec{d} + (1 - \alpha)\vec{t},$$

where $\vec{t}$ is the teleportation vector, such that $\vec{t}[i] = \frac{1}{\|V\|}$ for all i and $\mathbf{T}_D$ is a degree de-coupled transition matrix,

$$\mathbf{T}_D(j,i) = \frac{deg(v_j)^{-p}}{\sum_{v_k \in neighbor(v_i)} deg(v_k)^{-p}}, \quad (1)$$

where

- $\mathbf{T}_D(j,i)$ denotes the degree de-coupled transition probability from node v_i to node v_j over an edge $e_{ij} = [v_i \rightarrow v_j]$ when there exists at least one edge between two nodes,
- $neighbor(v_i)$ is the set of all neighbors of the source node, v_i, and



(a) A sample graph

Dest. v_j	$deg.$ (v_j)	Transition probability from A to its neighbors v_j		
		$p=0$	2	-2
B	2	0.33	0.18	0.29
C	3	0.33	0.08	0.64
D	1	0.33	0.74	0.07

(b) Transition probabilities from A

Figure 1: In conventional PageRank ($p = 0$), the transition probabilities from node $v_i = A$ to all its neighbors v_j are the same. In degree de-coupled PageRank (D2PR), the value of p can be used to penalize ($p > 0$) or boost ($p < 0$) transition probabilities based on the degree of the destination

node id	node degree	*Ranks of the graph nodes for different de-coupling weights (p)*				
		-4	-2	0	2	4
53608	883	1	1	69	5549	6793
351	739	2	12	425	1992	1935
...	...	...	...	...	...	...
79538	1	7661	7545	4149	195	182
79917	1	7793	7790	7522	2443	2043

Table 2: Ranks of graph nodes of different degrees on a sample graph for different de-coupling weights, p: as we see in this figure, when $p > 0$, high degree nodes are pushed down in the rankings (reducing the correlation between degree and rank), while when $p < 0$, they are pulled up (improving the correlation between degree and rank)

- $p \in \mathbb{R}$ is a degree de-coupling weight.

Intuitively, the numerator term, $deg(v_j)^{-p}$, ensures that the edge incoming to v_j is weighted by its degree: if $p > 0$, then its degree negatively impacts (reduces) transition probabilities into v_j, if $p < 0$ then its degree positively impacts (boosts[2]) transition probabilities into v_j, and if $p = 0$, we obtain the standard PageRank formulation without degree de-coupling. In other words, the transition function satisfies our desideratum of de-coupling the transition process from node-degrees and penalizing or boosting the contributions of node degrees on-demand. Note that, since all transitions from the node v_i are degree de-coupled individually based on the degrees of their destinations, the denominator term, $\sum_{v_k \in neighbor(v_i)} deg(v_k)^{-p}$, ensures that the transition probabilities from node v_i add up to 1.0. Note also that when there is no edge between node v_i and v_j, $\mathbf{T}_D(j,i) = 0$ and, consequently, the term $\mathbf{T}_D(j,i)$ is not affected by the degree de-coupling process.

EXAMPLE 2. *Figure 1 shows how the random walk probabilities are differentiated in a degree de-coupled transition matrix on a sample graph where a node A has three neighbors, B (with degree 2), C (with degree 3), and D (with degree 1). In conventional PageRank, the transition probabilities from node A to all its neighbor nodes are equal to 0.33. In degree de-coupled PageRank (D2PR), however, the value of p is used for explicitly accounting for the impact of node degree on the transition probabilities: When $p = 2$, the transition probabilities from A to its neighbors are 0.18, 0.08, and 0.74, which penalizes nodes which have larger degrees, whereas when $p = -2$, D2PR boosts the transition probabilities to large degree nodes leading to transition probabilities 0.29, 0.64, and 0.07, respectively.* ◇

This example shows that, in degree de-coupled PageRank (D2PR), as we also see in Table 2, the value of p can be used to penalize ($p > 0$) or boost ($p < 0$) transition probabilities based on the degree of the destination, v_j.

[2]In fact, a similar function was used in [11] to quickly locate nodes with higher degrees in a given graph.

3.2.2 Directed Unweighted Graphs

The semantics of degree de-coupling is slightly different in directed graphs. In particular, edges incoming to v_i often do not require a particular effort from v_i to establish and hence are often out of the control of v_i, but indicate a certain degree of *interestingness*, *usefulness*, or *authority* as perceived by others. The same is not true for edges outgoing from v_i; in particular, a vertex with a large number of outgoing edges may either indicate a potential *hub* or simply indicate a non-discerning connection maker. The distinction between these two situations gains importance especially in applications where establishing a new connection has a non-negligible cost to the source node and, thus, a large number of outgoing edges may indicate either (a) a very strong participant to the network or (b) a very poor participant with a large number of weak linkages.

Let $G = (V, E)$ be a directed graph and for the simplicity of the discussion, without any loss of generality, let us assume that G is unweighted. Let us also be given a residual probability parameter, α and let $outdeg(v)$ be a function which returns the number of outgoing edges from the node v. The degree de-coupled PageRank (D2PR) scores can be represented in the form of a vector $\vec{d}$, $\vec{d} = \alpha \mathbf{T}_D \vec{d} + (1-\alpha)\vec{t}$, where $\vec{t}$ is the teleportation vector, such that $\vec{t}[i] = \frac{1}{\|V\|}$ for all i and

$$\mathbf{T}_D(j,i) = \frac{outdeg(v_j)^{-p}}{\sum_{[v_i \to v_k] \in out\_edges(v_i)} outdeg(v_k)^{-p}},$$

where $\mathbf{T}_D(j,i)$ denotes the degree de-coupled transition probability from node v_i to node v_j over an edge $e_{ij} = [v_i \to v_j]$, $out\_edges(v_i)$ is the set of out-going edges from the source node, v_i, and $p \in \mathbb{R}$ is a degree de-coupling weight.

EXAMPLE 3. *Figure 2 (a) in Section 4 provides an example illustrating the correlations between the degree de-coupled PageRank (D2PR) scores and external evidence for different values of p for some application: here, the higher the correlation, the better resulting ranking reflects the application semantics. As we see in this example, which we will investigate in greater detail in Section 4, the optimal de-coupling weight is not always $p = 0$ as implied by the conventional PageRank measure. In this particular case, for example, the correlation between D2PR and external evidence of significance is maximized when the de-coupling weight, p, is equal to 0.5, implying that in this application a moderate degree of penalization based on the node degrees is needed to align PageRank scores and application semantics.* ◇

3.2.3 Weighted Graphs

Once again, the semantics of degree de-coupling need to be reconsidered for weighted graphs. Let $G = (V, E, w)$ be a directed, weighted graph, where $w(e)$ is a function which returns the weight of the edge associated with edge e. It is important to note that, in such a graph, the weight of an edge can 1) indicate the strength of the connection between two nodes (thus positively contributing to the significance of the destination node); and at the same time and 2) contribute to the degree of a node as a multiplier (thus positively or negatively contributing to the node significance depending on the degree-sensitivity of the application). In other words, given an edge $e_{ij} = [v_i \to v_j]$, from node v_i to node v_j, the transition probability from v_i to v_j can be written as

$$\mathbf{T}(j,i) = \beta \mathbf{T}_{conn\_strength}(j,i) + (1-\beta)\mathbf{T}_D(j,i),$$

where

$$\mathbf{T}_{conn\_strength}(j,i) = \frac{w(v_i \to v_j)}{\sum_{[v_i \to v_h] \in out\_edges(v_i)} w(v_i \to v_h)},$$

accounts for the connection strength (as in the conventional PageRank) whereas $\mathbf{T}_D$ is a degree de-coupled transition matrix,

$$\mathbf{T}_D(j,i) = \frac{\Theta(v_j)^{-p}}{\sum_{[v_i \to v_k] \in out\_edges(v_i)} \Theta(v_k)^{-p}},$$

such that, $\mathbf{T}_D(j,i)$ denotes the degree de-coupled transition probability from node v_i to node v_j over an edge $e_{ij} = [v_i \to v_j]$, $p \in \mathbb{R}$ is a degree de-coupling weight, and

$$\Theta(v) = \sum_{[v \to v_h] \in out\_edges(v)} w(v \to v_h).$$

Note that, above, β controls whether accounting for the connection strength or degree de-coupling is more critical in a given application. In Section 4, we will study the impact of degree de-coupling in weighted graphs for different scenarios.

4. CASE STUDIES

In this section, we present case studies assessing the effectiveness of the degree de-coupling process and the relationship between the degree de-coupling weight p and recommendation accuracy for different data graphs.

4.1 Setup

For all experiments, the degree de-coupling weight, p, is varied between -4 and 4 with increments of 0.5. The residual probability, α, is varied between 0.5 and 0.9, with default value chosen as 0.85. We also varied the β parameter, which controls whether accounting for the connection strength or degree de-coupling is more critical in a given application, between 0.0 and 1.0, with the default value set to 0 (indicating full decoupling).

4.1.1 Datasets

Four real data sets are used for the experiments. Each data set is used to create two distinct data graphs and corresponding ratings data. Table 3 provides further details about the various graphs created using these four data sets. These recommendation tasks based on these data graphs are detailed below:

• For the **IMDB** [15] data set, we created (a) a *movie-movie* graph, where movie nodes are connected by an edge if they share common contributors, such as actors, directors, writers, composers, editors, cosmetic designers, and producers and (b) an *actor-actor* graph based on whether two actors played in the same movie. **Applications:** For this data set, we consider applications where movies are rated by the users: thus, we merged the IMDB data with the MovieLens 10M [22] data (based on movie names) to identify user ratings (between 1 and 5) for the movies in the graph. We consider the (a) *average user rating* as the significance of the movies in the movie-movie graph and (b) *average user rating of the movies played in* as the significance of the actors in the actor-actor graph.

• For the **DBLP** [26] data set, we constructed (a) an *article-article* graph where scientific articles were connected to each other if they shared a co-author and (b) an *author-author* graph based on co-authorship. **Applications:** (a) In the article-article graph, the *number of citations* to an article is used to indicate its significance. Similarly, (b) in the author-author graph, *average number of citations* to an author's papers is used as his/her significance.

• For the **Last.fm** [18], we constructed (a) a *listener-listener* graph, where the nodes are Last.FM listeners and undirected edges reflect friendship information among these listeners. We also constructed (b) an *artist-artist* graph based on shared listeners. **Applications:** (a) In the listener-listener graph, we considered the *total listening*

Data	Graph	# of nodes	# of edge	Average node degree	Standard deviation of node degrees	Median standard deviation of neighbors' node degrees
IMDB	movie-movie	191,602	4,465,272	23.30	51.86	2.89
	actor-actor	32,208	2,493,574	77.42	67.15	114.41
DBLP	article-article	8,808	951,798	108.06	171.25	309.92
	author-author	47,252	310,250	6.57	8.89	6.39
Last.fm	listener-listener	1,892	25,434	13.44	17.31	22.37
	artist-artist	17,626	2,640,150	149.79	299.66	998.53
Epinions	commenter-commenter	6,703	2,395,176	425.05	438.97	609.39
	product-product	13,384	2,355,460	175.99	224.12	202.78

Table 3: Data sets and data graphs

activity of a given listener as his/her significance. (b) In the artist-artist graph, the *number of times an artist has been listened* is considered as his/her significance.

• For the **Epinions** [25]: We constructed (a) a *commenter-commenter* graph based on the products on which two individuals both commented and (b) a *product-product* graph based on shared commenters. **Applications:** (a) For the nodes on the commenter-commenter graph, the *number of trusts* the commenter received from others is used as his/her commenter significance. (b) For each product in the product-product graph, its *average rating by the commenters* is used as its node significance.

4.2 Measures

In this section, our goal is to observe the impact of different D2PR degree de-coupling weights on the relationship between D2PR rankings and application specific significance measures for the above data sets[3]. We also aim to verify whether de-coupling weights can also be used to improve recommendation accuracies.

In order to measure the relationship between the degree de-coupled PageRank (D2PR) scores and the application-specific node significance, we used Spearman's rank correlation,

$$\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}},$$

which measures the agreement between the D2PR ranks of the nodes in the graph and their application-specific significances. Here, x are rankings by D2PR and y are significances for an application and $\bar{x}$ and $\bar{y}$ are averages of two values.

4.3 Impact of De-Coupling in Different Applications (Unweighted Graphs)

In this subsection, we present results that aim to assess D2PR under the settings described above. For these experiments, the residual probability, α, and the parameter, β, are set to the default values, 0.85 and 0, respectively. In these experiments, we consider only unweighted graphs (we will study the weighted graphs and the impact of parameter β later in Section 4.5).

Figures 2 through 4 include charts showing the Spearman's correlations between the D2PR ranks and application specific node significances for different values of p and for different data graphs. These figures clearly illustrate that different data graphs require different degrees of de-coupling[4] to best match the application specific node significance criterion.

4.3.1 Application Group A: When Degree Penalization Helps

The *actor-actor* (based on common movies) and *commenter-commenter* (based on common products) graphs have highest correlation at $p = 0.5$, with the correlations dropping significantly when the degrees are over-penalized (i.e., when $p \gg 0.5$). The Epinions *product-product* graph (based on common commenters, Figure 2(c)) also provides the highest correlations with $p > 0$, but behaves somewhat differently from the other two cases: the correlations stabilize and do not deteriorate significantly when degrees are over-penalized, indicating that the need for degree penalization is especially critical in this case: this is due to the fact that, the larger the number of comments a product has, the more likely it is that the comments are negative (Figure 5). In fact, we see that, among the three graphs, this is the only graph where the traditional PageRank (with $p = 0$) leads to **negative correlations** between node ranks and node significances.

These results indicate that actors who have had many co-actors, commenters who commented on products also commented by many others, or products which received comments from individuals who also commented on many other products are not good candidates for transition during random walk. This aligns with our expectation that, in applications where each new movie role or comment requires additional effort, high degree may indicate lower per-movie or per-comment effort and, hence, lower significance.

4.3.2 Application Group B: When Conventional PageRank is Ideal

Figure 3 shows that, for *movie-movie* (based on common actors) and *author-author* (based on common articles) graphs, the peak correlation is at $p = 0$ indicating that the conventional PageRank which gives positive weight to node degree, is appropriate.

This perhaps indicates that movies with a lot of actors tend to be big-budget products and that authors with a large number of co-authors tend to be experts with whom others want to collaborate. Note that, in these applications, additional boosting, with $p < 0$, negatively affects the correlation, indicating that the relationship between node degree and significance is not very strong (Figure 5). The quick change when $p < 0$ is because, as we see in Table 3, median standard deviations of neighbors' degrees are low; i.e., degrees of neighbors of a node are comparable: there is no dominant contributor to $\mathbf{T}_D(j, i)$ in Equation 1 (Section 3) and, thus, the transition probabilities are sensitive to changes in p, when $p < 0$.

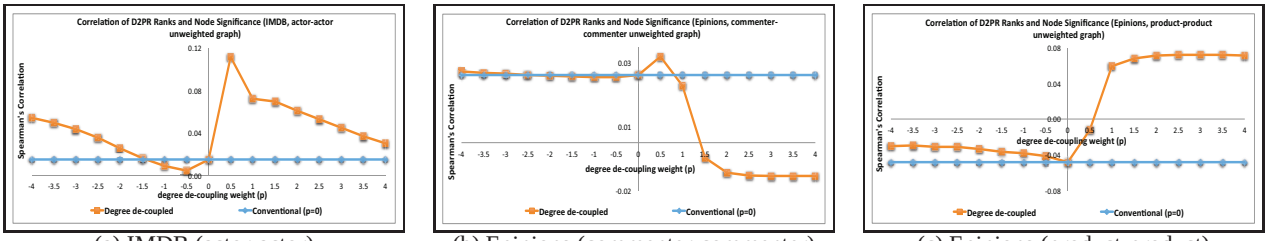
4.3.3 Application Group C: When Degree Boosting Helps

Figure 4 shows that there are scenarios where additional boosting based node degrees provides some benefits. The *article-article* (based on common authors), *listener-listener* (based on common artists), and *artist-artist* (based on common listeners) graphs reach their peaks around $p \sim -1$, indicating that these also benefit from large node degrees though improvements over $p = 0$ are slight.

A significant difference between applications in Group B and Group C is that, for $p < 0$, the correlation curve is more or less stable. This is because, as we see in Table 3, in these graphs median standard deviations of neighbors' degrees are high: in other words, for each node, there is a dominant neighbor with a high degree and this neighbor has the highest contribution to $\mathbf{T}_D(j, i)$; thus, the rankings are not very sensitive to p, when $p < 0$.

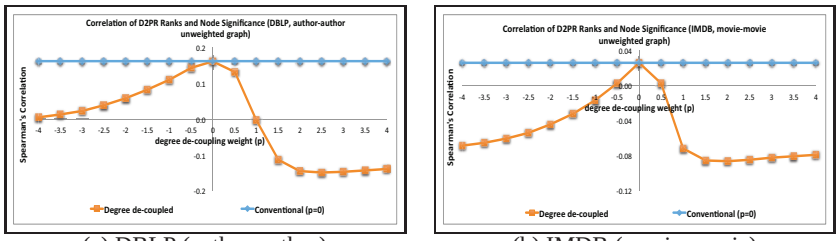
[3]In this paper, we are not proposing a new PageRank computation mechanism. Because of this (and since the focus is not improving scalability of PR), we do not report execution times and compare our results with other PageRank computation mechanisms.

[4]Degree penalization or degree-based boosting



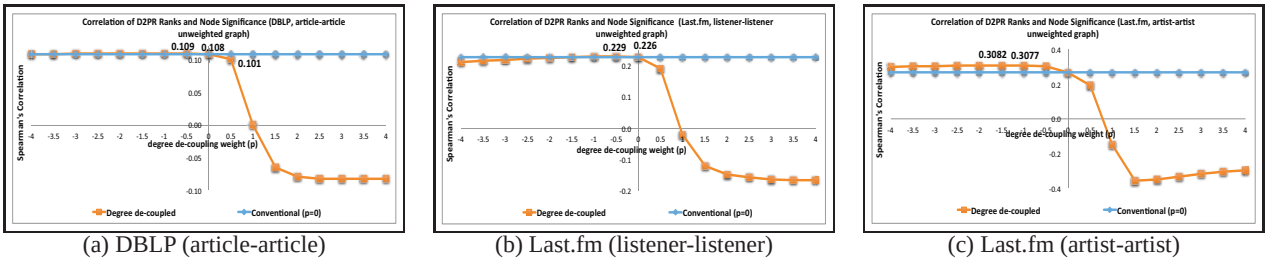
(a) IMDB (actor-actor) (b) Epinions (commenter-commenter) (c) Epinions (product-product)

Figure 2: Application Group A: $p > 0$ is optimal (i.e., node degrees need to be penalized)



(a) DBLP (author-author) (b) IMDB (movie-movie)

Figure 3: Application Group B: $p = 0$ is optimal



(a) DBLP (article-article) (b) Last.fm (listener-listener) (c) Last.fm (artist-artist)

Figure 4: Application Group C: $p < 0$ is optimal (i.e., node degrees need to be boosted)

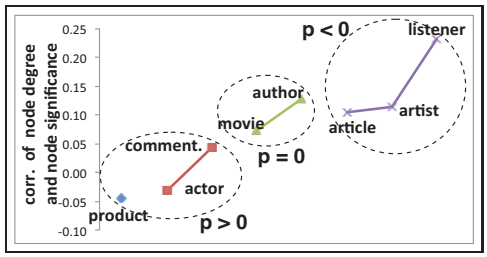


Figure 5: Correlations between node degrees and application specific significances for different data graphs (each color group is a distinct pattern in Figures 2 through 4).

4.3.4 *Summary: Correlations between Node Degrees and Application Specific Significances*

The experiments reported above show that degree de-coupling is important as different applications, even on the same data set, may associate different semantics to node degrees and the conventional PageRank scores are too tightly coupled with node degrees to be effective in all scenarios. Figure 5, which plots correlations between node degrees and application specific significances for different data graphs, re-confirms that the ideal value of the p is related to the usefulness of the node degree in capturing the application specific definition of node significance.

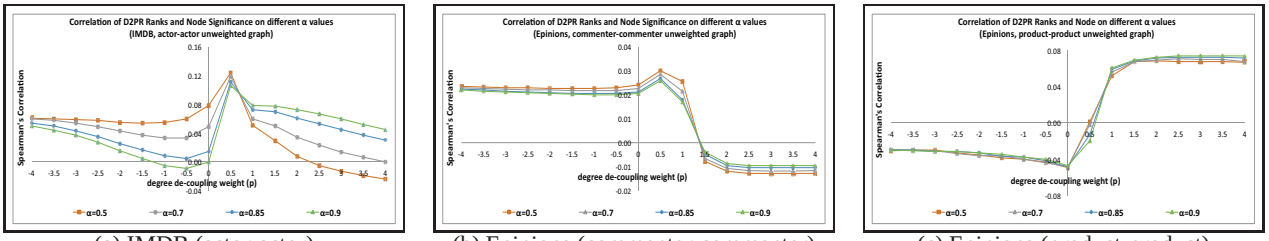
4.4 Relationship between α and p

In Figures 6 through 8, we investigate the relationship between the value α and the degree de-coupling parameter p for different application types. Here we use the default value, 0, for the parameter β and present the results for unweighted graphs (the results for the weighted graphs are similar).

First thing to notice in these figures is that the grouping of the applications (into those where, respectively, $p > 0$, $p = 0$, or $p < 0$ is useful) is preserved when different values of α are considered.

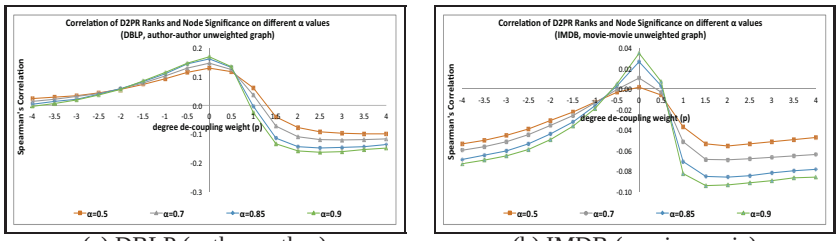
Figure 6 studies the impact of the value of α in *application group A, where degree penalization helps ($p > 0$)*. As we see here, for the IMDB *actor-actor* (Figure 6(a)) and Epinions *commenter-commenter* (Figure 6(b)) graphs, having a lower value of α (i.e., lower probability of forward movement during the random walk) provides the highest possible correlations between D2PR ranks and node significance (with the optimal value of p being ~ 0.5 independent of the value of α). This indicates that in these graphs, it is not necessary to traverse far during the random walk. Interestingly, though, when degrees are over-penalized (i.e., $p \gg 0$), smaller values of α start leading to worse correlations, indicating that (while not being optimal) severe penalization of node degrees helps make random traversals more useful than random jumps. As we have already observed in Figure 2(c), the Epinions *product-product* graph (Figure 6(c)) behaves somewhat differently from the other two cases where degree penalization ($p > 0$) leads to larger correlations: in this case, unlike the other two graphs, the highest possible correlations between D2PR ranks and node significance are obtained for large values of α, indicating that this application benefits from longer random walks (though the differences among the correlations for different α values are very small).

Figure 7 shows that the pattern is different for *application group B, where conventional PageRank is ideal ($p = 0$)*: in this case, having a larger value of α (i.e., larger probability of forward movement during the random walk) provides the highest correlations between ranks and significance. Interestingly, in these applications, when



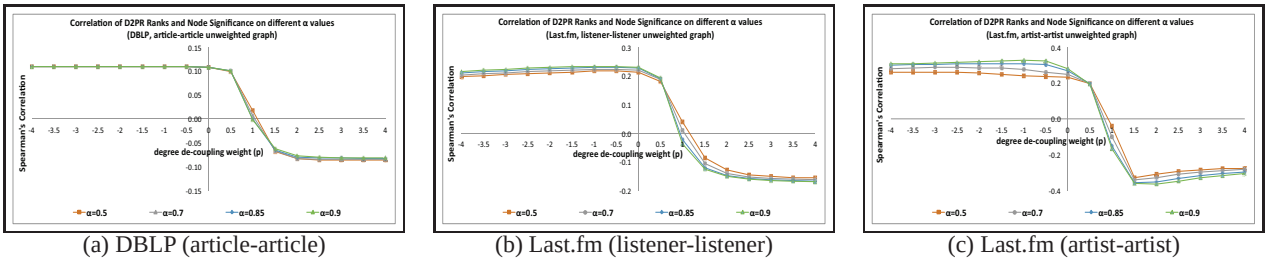
(a) IMDB (actor-actor) (b) Epinions (commenter-commenter) (c) Epinions (product-product)

Figure 6: Relationship between p and α, for application group A, where $p > 0$ is optimal (i.e., degrees need to be penalized)



(a) DBLP (author-author) (b) IMDB (movie-movie)

Figure 7: Relationship between p and α, for application group B, where $p = 0$ is optimal



(a) DBLP (article-article) (b) Last.fm (listener-listener) (c) Last.fm (artist-artist)

Figure 8: Relationship between p and α, for application group C, where $p < 0$ is optimal (i.e., node degrees need to be boosted)

$p \ll 0$ or $p \gg 0$, higher probabilities of random walk traversal (i.e., larger α) stop being beneficial and lower values of α lead to larger correlations. This re-confirms that, for these applications, $p \sim 0$ leverages the random walk traversal the best.

As we see in Figure 8, in *application group C, where degree boosting helps ($p < 0$)*, it is also the case that larger values of α (i.e., larger probabilities of forward transitions during the random walk) provides the highest correlations between node ranks and significance. On the other hand, in these applications, $p \sim 0.5$ serves as a balance point where the value of α stops being relevant; in fact, for $p > 0.5$ the higher values of α stops being beneficial and lower values of α lead to larger correlations. This re-confirms that smaller values of p (which provides degree boosting) help leverage the random walk traversal the best.

4.5 Relationship between β and p in Weighted Graphs

Finally, in Figures 9 through 11, we investigate the relationship between the value β (which controls whether accounting for the connection strength or degree de-coupling is more critical in a given application) and the degree de-coupling parameter p for different application types. Here we use the default value, 0.85, for the parameter α and present the results for weighted graphs:

Figure 9 depicts the impact of the value of the parameter β in *application group A, where degree penalization helps ($p > 0$)*. As we see here, for all three weighted graphs, performing degree penalization (i.e., $\beta < 1.0$) provides better rank-significance correlation than relying solely on the connection strength (i.e., $\beta = 1.0$). Note that the value of β impacts the optimal value of degree penalization parameter p: the more weight is given to connection strength (i.e., the greater β is), the larger is the optimal value of p.

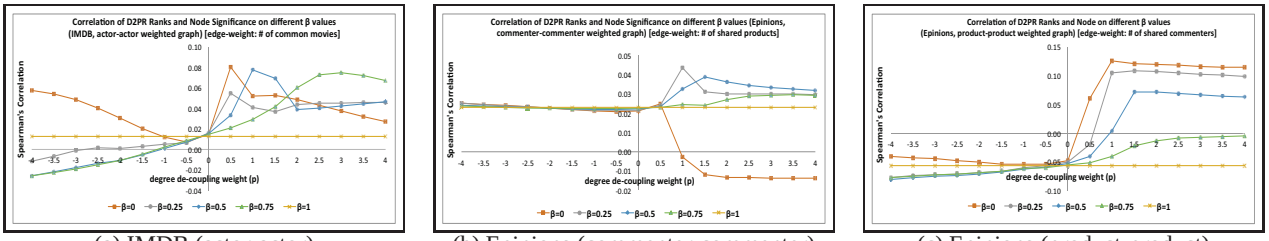
Figure 10 shows that, for *applications in group B, where $p \sim 0$ is ideal*, when the connection strength is given significantly more weight than degree de-coupling (i.e., $\beta \sim 0$), we observe high rank-significance correlations. Interestingly however, for the *movie-movie* graph (where the edge weights denote common actors) the highest correlations are obtained not with $p = 0$, but with $p = 0.5$ and $\beta = 0.75$, indicating that degree penalization is actually beneficial in this case: movies that share large numbers of actors with other movies are likely to be B-movies, which are not good candidates for transitions during the random walk.

Figure 11 shows that in *application group C, where degree boosting ($p < 0$) helps*, giving more weight to connection strength (i.e., $\beta \sim 1.0$) is a good, but not necessarily the best strategy. In fact, in these graphs, the highest overall correlations are obtained with $\beta = 0$ or $\beta = 0.25$, indicating that degree de-coupling is beneficial also in these cases. Interestingly, (unlike the case with the unweighted *listener-listener* graph, where the best correlation was obtained when $p < 0$) for the weighted version of the *listener-listener* graph (where edge weights denote the number of shared friends), when $\beta = 0$ through 0.5, $p = 0$ provides the highest correlation and when $\beta = 0.75$, $p = 0.5$ provides the highest correlation – these indicate that listeners who have large numbers of shared friends with others are good candidates for random walk.

Note that a key observation from the above results is that the conventional PageRank, based on connection strength (i.e., $\beta = 1.0$), is not always the best strategy for the applications considered.

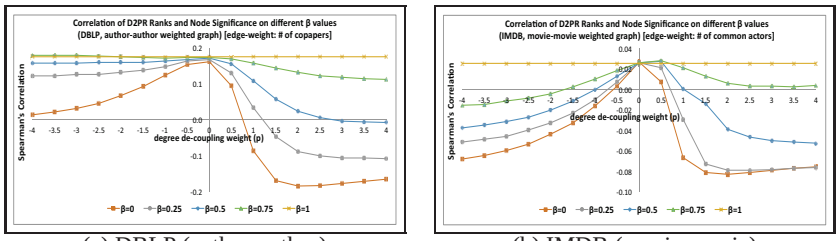
5. CONCLUSIONS

In this paper, we noted that in many applications the relation-



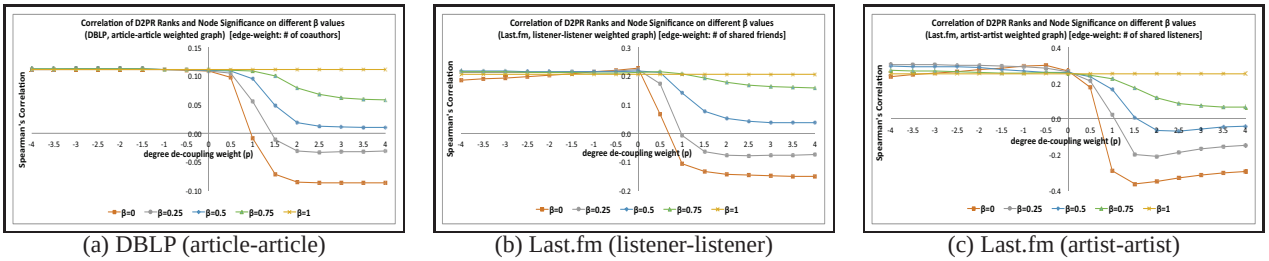
(a) IMDB (actor-actor) (b) Epinions (commenter-commenter) (c) Epinions (product-product)

Figure 9: Relationship between p and β, for application group A, where $p > 0$ is optimal (i.e., node degrees need to be penalized)



(a) DBLP (author-author) (b) IMDB (movie-movie)

Figure 10: Relationship between p and β, for application group B, where $p = 0$ is optimal



(a) DBLP (article-article) (b) Last.fm (listener-listener) (c) Last.fm (artist-artist)

Figure 11: Relationship between p and β, for application group C, where $p < 0$ is optimal (i.e., node degrees need to be boosted)

ship between the *significance* of the node and its degree in the underlying network may not be as strong (or as weak) as implied by PageRank-based measures. We proposed *degree de-coupled PageRank (D2PR)* to improve the effectiveness of PageRank based knowledge discovery and recommendation tasks. Evaluations on different data graphs and recommendation tasks have confirmed that degree de-coupling would be an effective way to match application specific node significances and improve recommendation accuracies using PageRank based approaches.

6. REFERENCES

[1] A. Balmin, V. Hristidis, and Y. Papakonstantinou. ObjectRank: Authority-based keyword search in databases. In VLDB'04.

[2] D. Banky, G. Ivan, V. Gromusz. Equal Opportunity for Low-Degree Network Nodes: A PageRank-Based Method for Protein Target Identification in Metabolic Graphs. PLoS One 8(1): e542-4, 2013.

[3] L. Becchetti *et al.* Using Rank Propagation and Probabilistic Counting for Link-Based Spam Detection. WebKDD, 2006.

[4] P. Boldi *et al.* HyperANF: Approximating the neighbourhood function of very large graphs on a budget. WWW, 2011.

[5] M.G.Borgatti, et al. Network measures of social capital. Connections 21(2):27-36, 1998.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 30, 1998.

[7] K.S. Candan and W.D. Li. Using random walks for mining web document associations. PAKDD'00, pp. 294-305, 2000.

[8] M. Chen., et al. Clustering via random walk hitting time on directed graphs. AAAI'08.

[9] S. Chakrabarti. Dynamic personalized pagerank in entity-relation graphs. In WWW'07, pp 571–580, 2007.

[10] M. Chen, J. Liu, and X. Tang. Clustering via random walk hitting time on directed graphs. AAAI'08, pp. 616–621, 2008.

[11] C. Cooper, *et al.* A fast algorithm to find all high degree vertices in graphs with a power law degree sequence. In WAW'12, 2012.

[12] O. Fercoq. PageRank optimization applied to spam detection. arXiv:1203.1457, 2012.

[13] T. H. Haveliwala. Topic-sensitive PageRank. WWW, 2002.

[14] S. Huang, X. Li, K.S Candan, M.L Sapino. "Can you really trust that seed?": Reducing the Impact of Seed Noise in Personalized PageRank. ASONAM'14, 2014.

[15] IMDB website: `http://www.imdb.com/`

[16] G. Jeh and J. Widom. Scaling personalized web search. Stanford Univ. Tech. Report. 2002.

[17] J.H Kim, K.S Candan, M.L Sapino. Locality-sensitive and Re-use Promoting Personalized PageRank Computations. Knowledge and Information Systems. 10.1007/s10115-015-0843-6, 2015

[18] `http://ir.ii.uam.es/hetrec2011/datasets.html`.

[19] A.N. Nikolakopoulos and J. Garofalakis, NCDawareRank: A Novel Ranking Method that Exploits the Decomposable Structure of the Web. WSDM, 2013.

[20] F. Mathieu and L. Viennot, Local aspects of the global ranking of web pages. In I2CS'06, pp. 1–10, 2006.

[21] Q. Mei, D. Zhou, and K. Church. Query suggestion using hitting time. CIKM'08, 2008.

[22] `http://grouplens.org/datasets/movielens`

[23] M. Olsen. Maximizing PageRank with New Backlinks. CIAC, 2010.

[24] C. Palmer, P. Gibbons, and C. Faloutsos. Anf: a fast and scalable tool for data mining in massive graphs. KDD, 2002.

[25] J. Tang, H. Gao, and H. Liu. mTrust: Discerning multi-faceted trust in a connected world. WSDM, 2012.

[26] J. Tang, *et al.* ArnetMiner: Extraction and Mining of Academic Social Networks. In SIGKDD'08, pp 990–998, 2008

[27] D.R. White, *et al.* Betweenness centrality measures for directed graphs. Social Networks, 16, 335-346,1994.