Discovering Missing Links in Networks Using Vertex Similarity Measures

Hung-Hsuan Chen†, Liang Gou‡, Xiaolong (Luke) Zhang‡, C. Lee Giles†‡
†Computer Science and Engineering, ‡Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802, USA
hhchen@psu.edu, {lug129, lzhang, giles}@ist.psu.edu

ABSTRACT

Vertex similarity measure is a useful tool to discover the hidden relationships of vertices in a complex network. We introduce relation strength similarity (RSS), a vertex similarity measure that could better capture potential relationships of real world network structure. RSS is unique in that is is an asymmetric measure which could be used for a more general purpose social network analysis; allows users to explicitly specify the relation strength between neighboring vertices for initialization; and offers a discovery range parameter could be adjusted by users for extended network degree search. To show the potential of vertex similarity measures and the superiority of RSS over other measures, we conduct experiments on two real networks, a biological network and a coauthorship network. Experimental results show that RSS is better in discovering the hidden relationships of the networks.

Categories and Subject Descriptors

G.2.2 [**Discrete Mathematics**]: Graph Theory—*Graph algorithms*; E.1 [**Data Structures**]: Graphs and networks

General Terms

Algorithms, Experimentation, Measurement

Keywords

Link Analysis, Link Prediction, Information Retrieval, Web of Linked Data, Social Network, Complex Network

1. INTRODUCTION

A complex network is a graph with non-trivial topological features that occur in actual real world graphs and in which each vertex acts as a complex object and each edge corresponds to an interaction between two objects (we use the terms "node" and "vertex" interchangeably, and "network" and "graph" interchangeably). The nature of the vertices or relationship between vertices can be inferred by the graph statistics and measures, such as vertex degree, clustering coefficient, betweenness centrality, and shortest path length [3]. Among all the graph measures, one important measure is vertex similarity [16], which measures how similar two vertices are. Vertex similarity measure can be applied in several applications, such as potential web linking information discovery [1], duplicate object identification [5], coauthoring behavior inference [20], and knowledge capturing using representational components [8].

The vertex similarity problem can be categorized into two classes: vertex feature based similarity and network topology based similarity. Vertex feature based methods measure the similarity of two vertices based on their attributes. For example, two users might be interested in similar topics because they are both at the same age. Topology based methods, on the other hand, measure the similarity of two nodes based on the topology of the graph. For example, we could intuitively say that two nodes are more similar if they have more common neighbors.

Here, we investigate topology based vertex similarity. We introduce Relation Strength Similarity (RSS) [6, 7], a vertex similarity measure that has the following characteristics. First, it is an asymmetric metric which allows the measure to be used in more general social network applications. Second, it can be employed on weighted networks, in which the relationship strength between two nodes can be explicitly expressed using edge weights. Third, we propose a "discovery range" parameter that can be adjusted based on user's domain knowledge about the network to explore higher relationships between nodes.

To evaluate and compare RSS with other network topology based vertex similarity measures, we conduct experiments on two real networks: a human disease network and a coauthorship network of Computer Scientists. The human disease network is provided by Diseasome[1], which contains $1,284$ different human diseases. The human disease network experiment demonstrates the power of vertex similarity measures to capture the potential links. The coauthorship network is built by a subset of CiteSeer^{X}[2] dataset, which consists of over $1,300,000$ computer science related documents and over $300,000$ unique authors. This experiment shows that the vertex similarity analysis helps predict network evolution. Details of converting these information into a graph will be addressed later. Experimental results show that our method outperforms other vertex similarity

[1] http://diseasome.eu/

[2] http://citeseerx.ist.psu.edu/

SAC'12 March 25-29, 2012, Riva del Garda, Italy.

measures in both networks.

The rest of the paper is organized as follows. In Section 2, we introduce several well-known vertex similarity measures as the background. The relation strength similarity measure is addressed and analyzed in Section 3. In Section 4, we evaluate and compare the performance of relation strength similarity with other topology based vertex similarity measures in terms of their ability to predict the potential links on networks. Summary and future work appear in Section 5.

2. VERTEX SIMILARITY MEASURES

We introduce several well-known topology based vertex similarity measures, i.e., the similarity score between two vertices determined by the topology of the network. Several topology based approaches, such as Jaccard similarity [23] and cosine similarity [22], is based on the intuition that two vertices are more similar if they share more common neighbors. Adamic and Adar [2] refined the measures by assigning more weights to the vertices with fewer degrees. However, Adamic-Adar's measure cannot be normalized because in theory the maximum similarity values between two nodes could be infinity. Preferential attachment [4] is based on the likelihood that a high degree node is more likely to acquire new links. The phenomenon was observed in several large scale networks, such as World Wide Web [4], citation network [21], and protein network [10]. Based on empirical observation, Newman [19] proposed that the probability of a new edge established between two vertices is proportional to the product of their degree. Zhou et al. did a comprehensive empirically study on the local topology based similarities [25].

Instead of using local neighborhood information, the global topology can also be used for vertex similarity calculation. Katz [15] proposed a measure based on the total number of simple paths between vertices with lower weights to longer paths. Instead of calculating all the simple paths, the measure can be directly calculated by $(I - aC)^{-1} - I$, where I is the identity matrix, a is a parameter to decide the importance ratio between direct neighbors and indirect neighbors, and C is the adjacent matrix. Several other global topology based measures, such as SimRank [14], Leicht-Holme-Newman (LHN) [16], and P-Rank [24] defined the similarity measures recursively: two vertices are similar if their immediate neighbors in the network are themselves similar. Although these methods bear some similarity to each other, they have an important difference. SimRank and LHN regard two vertices similar if they are referenced by similar vertices, whereas P-Rank considers both in-link and out-link relationships. In addition, SimRank and P-Rank includes only paths of even length, which could make a substantial difference for the final similarity score. Several of these methods were compared in [20].

Although global topology based measures offer a boarder picture of the whole network, they are usually computationally very expensive. Several approximations for global topology based measures are have been proposed. Gou et al. approximated LHN by clustering the social network into virtual nodes to reduce the graph size [12, 13]. Li et al. approximated SimRank by incremental updating [17], but this measure allows only link updating, i.e., it assumes that the total number of vertices in a graph is fixed.

3. RELATION STRENGTH SIMILARITY

Among all the introduced measures in last section, only part of them can be used with unweighted networks. Moreover, none of them is an asymmetric measure. Our Relation Strength Similarity [6, 7] is an asymmetric measure that can be applied on a weighted network.

3.1 Relation Strength Similarity Calculation

Algorithm 1: Calculating the RSS score from a start vertex to all other vertices

```
Input: N: the target network, a: start vertex; r: discovery range
Output: S(a,*): RSS score from a to all the other vertices
if a not in N then
    return ERROR;
end
valid_paths ← GetValidPaths(a,r);
foreach p in valid_paths do
    b ← the end vertex of path p;
    Calculate $R^*_p(a,b)$ by Equation 2;
    $S(a,b) \leftarrow S(a,b) + R^*_p(a,b)$;
end
```

Relation strength similarity permits users to explicitly assign the weights to every edge for initialization. If users are unsure about the relative importance of the edges, they can naïvely assign the same weight to all of them. Relation strength similarity is calculated based on *relation strength*, a normalized edge weighting score defining the relative degree of similarity between neighboring vertices. The relation strength from a vertex A to another vertex B is calculated as follows.

$$R(A,B) := \begin{cases} \dfrac{\alpha_{AB}}{\sum_{\forall X \in N(A)} \alpha_{AX}} & \text{if } A \text{ and } B \text{ are adjacent} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where α_{AB} can be explicitly specified by users based on known conditions or their best knowledge, and $N(A)$ is the set of A's neighboring vertices. The value of relation strength is normalized between 0 and 1.

For any two vertices A and C, if A could reach C through a simple path p_m, we define the *generalized relation strength* from A to C through path p_m as

$$R^*_{p_m}(A,C) := \prod_{k=1}^{K-1} R(B_k, B_{k+1}), \tag{2}$$

where B_1 is vertex A, B_K is vertex C, path p_m is formed by K vertices B_1, B_2, …, B_{K-1}, and B_K.

The above equation requires computing all the paths between two vertices. So far, an exhaustive search is still the only way to solve the problem [18]. To make the calculation tractable, we propose a new *discovery range* parameter, r, to control the maximum degree of separation for an generalized relation strength calculation, i.e., we only look for paths at most r hops away. Thus, Equation 2 becomes

$$R^*_{p_m}(A,C) := \begin{cases} \prod_{k=1}^{K} R(B_k, B_{k+1}) & \text{if } K \leq r. \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

This way the discovery range for the social network can be based on the domain knowledge of the problem of interest. In our experiments, as discussed in Section 4, we found that even with a small discovery range RSS still outperforms other vertex similarity measures.

Assuming that there are M distinct simple paths p_1, p_2, $\ldots$, p_M from A to C with path length shorter than discovery range r, the relation strength similarity from a vertex A to another vertex C is defined as the summation of all the generalized relation strengths, as defined in Equation 4.

$$S(A,C) := \sum_{m=1}^{M} R^{*}_{p_m}(A,C). \qquad (4)$$

The procedure of calculating the RSS for two given vertices is shown in Algorithm 1. The GetValidPaths(a,b,r) function at line 4 returns all the simple paths with lengths no longer than r between vertices a and b. In practice, we use depth-first search to get these paths.

3.2 Analysis of Relation Strength Similarity

Here we first show that the value of RSS is always between 0 and 1. Next, we study and compare several characteristics of RSS with other similarity measures. We explain why introducing a discovery range parameter is a good idea. Finally, we finish the section by analyzing the time complexity of RSS.

Although normalization seems to be a straightforward step in defining a new measure, several vertex similarity measures, such as Adamic-Adar [2], preferential attachment [19], and Katz [15], cannot be normalized because their maximum possible value could be infinity. We show that the value of RSS is always between 0 and 1 by rewriting Equation 4 as follows.

$$\begin{aligned} S(A,C) &:= \sum_{m=1}^{M} R^{*}_{p_m}(A,C) && (5)\\ &= \sum_{m=1}^{M} \left[\prod_{k=1}^{K-1} R(B_k^{(m)}, B_{k+1}^{(m)}) \right] && (6)\\ &\leq \sum_{m=1}^{M} R(A, B_2^{(m)}) && (7)\\ &\leq \sum_{\forall X \in N(A)} \frac{\alpha_{AX}}{\sum_{\forall X \in N(A)} \alpha_{AX}} && (8)\\ &= 1, && (9) \end{aligned}$$

where $B_1^{(m)}, B_2^{(m)}, \ldots B_K^{(m)}$ form p_m, the m^{th} path between A and C, $A = B_1^{(1)} = B_1^{(2)} = \ldots = B_1^{(M)}$ since $B_1^{(m)}$ is the starting vertex of path p_m, $C = B_K^{(1)} = B_K^{(2)} = \ldots = B_K^{(M)}$ since $B_K^{(m)}$ is the ending vertex of path p_m, and $N(A)$ is the set of neighboring vertices of A.

Equation 7 holds because the generalized relation strength of any two vertices through a simple path p_m is less or equal to the relation strength of any two adjacent vertices along p_m by Equation 2. If C is a neighboring vertex of A, Equation 8 applies since vertices C, $B_2^{(1)}$, $B_2^{(2)}$, $\ldots$, $B_2^{(M)}$ form a subset of $N(A)$ and therefore $\sum_{X \in \{C, B_2^{(1)}, \ldots, B_2^{(M)}\}} R(A,X) \leq \sum_{\forall X \in N(A)} R(A,X)$. If C is not adjacent to A, $R(A,C)$ becomes 0 by Equation 1 and contributes nothing to the final measure. Equation 8 still applies because vertices $B_2^{(1)}$, $B_2^{(2)}$, $\ldots$, $B_2^{(M)}$ form a subset of $N(A)$.

Compared to other vertex similarity measures, a significant advantage of RSS is its asymmetry, i.e., $S(A,B)$ may not equal $S(B,A)$. This is because $R(A,B)$, the relation strength from A to B, may not be the same as $R(B,A)$, the relation strength from B to A, as shown in Equation 1. The asymmetric property is true for several real world scenarios where a social actor knows someone but is not known by that actor. Most of previous vertex similarity measures [2, 14, 15, 16, 19, 22, 23, 24] are symmetric by nature.

In addition, unlike many other similarity methods, RSS can be used on weighted graphs. Much previous work has treated neighboring vertices equally important and the edges have only binary values [2, 16, 19, 22, 23] and neglect the fact that neighboring vertices may have different strengths. Different from these approaches, the initial setting of our method allows users to explicitly specify the known relation strength between objects based on domain knowledge. Consider the coauthorship network for example where the weights of edges could be used to represent the number of coauthored papers between two authors. For a gene promoter network, weighted edges could stand for bp-sharing between promoters.

Users can adjust the discovery range of how complex a relationship between nodes can be explored to n degrees of separation. Compared with previous work [2, 22, 23], the local topology based measures are too restrictive in the sense that they only look for vertices with two degrees of separation. The global topology based measures [14, 15, 16, 24] are not computationally feasible for the large or dynamic networks. Our algorithm allows a user to control the discovery range to achieve balance between the two. Although introducing a discovery range parameter disregards the effect of long paths between vertices, the approximation is reasonable because once the path length is too long, the product form in Equation 2 would make R^{*}_{pm} very small, and therefore contributes little to the final similarity measure score (Equation 4).

Let's consider the required time complexity of RSS with discovery range r for a network with n vertices, e edges, and average degree d. Referring to Algorithm 1, the first three lines are to check whether both the start and end vertices are in the network. It requires $O(n)$ to examine through all the vertices. The GetValidPath(a,b,r) function at line 4 is essentially a depth-first search algorithm with the early termination condition: disregard a path when the length is longer than r. To get all the valid paths from the starting vertex a requires $O(d^r)$. Line 5 to line 9 calculate the RSS score $S(a,*)$ by looping through all the valid paths, which requires $O(dd^r) = O(d^{r+1})$ time in average. Thus, the time complexity to calculate all the RSS scores from a vertex a is $O(n) + O(d^r) + O(d^{r+1}) \sim O(d^{r+1})$. Since there are n vertices in the network, the time complexity to compute the RSS between any two vertices in the network is $O(nd^{r+1}) \sim O(n)$ because in practice $d \ll n$ and $r \ll n$.

4. EXPERIMENTS

Evaluating similarity measures is difficult because vertex similarity results usually lack interpretability [9]. We conduct two experiment to compare RSS with other measures. In the first experiment, we aim to use known topology of

Table 1: Statistical measures of the training network for the human disease network.

Statistical Measure	Value
Number of Nodes	867
Number of Edges	$1,231$
Average Degree	2.84
Average Clustering Coefficient	0.37
Average Shortest Path Length	7.83
Diameter	19

the network to capture the exist but unknown relationship among the vertices. We use Diseasome dataset to build a human disease network for this experiment. In the second experiment, the snapshot of a growing network is utilized to infer how the network evolves over time. A subset of the CiteSeer^X dataset is used to build a coauthorship network for experiments.

4.1 Capturing Unknown Relationship of Networks

The human disease network information is built from Diseasome, a bipartite graph with two disjoint sets of vertices [11]. One set contains all known genetic disorders, and the other set includes all known disease genes in the human genome. A disorder and a gene are connected if the mutation of the gene would cause the disorder.

4.1.1 Experimental Setup

We use Diseasome to build the human disease network (HDN), which contains $1,284$ vertices and $1,527$ edges. Each vertex represents a human disease. An edge attaches two vertices if there are one or more genes that are implicated in both. Edge weights correspond to the number of common genes between the two disorders.

Instead of conducting the expensive biological experiments to verify the results, we imitate the supervised learning technique by separating the known information into training and testing data set to show the potential of vertex similarity measures. Specifically, for the $1,527$ known links in the HDN, each link has a probability p to be included in the training network and $(1-p)$ in the testing network $(0 < p < 1)$. The expected numbers of links in the training network and testing network are $1,527p$ and $1,527(1-p)$ respectively. In addition, among the $1,284$ vertices in HDN, 417 of them are singletons, i.e., the the vertices have no links attached to it. The singletons are removed because the similarity score between a singleton and any other vertices is always zero by vertex similarity measures. Thus, the training network of HDN contains 867 vertices.

We apply the vertex similarity measures on the training network to get the similarity scores of each non-neighbor vertex pair. The potential links are predicted by claiming the top-n most similar pairs should be connected. The correctness of the prediction is validated by the testing network. The procedure is repeated 20 times independently. Table 1 shows the important statistical measures of the training network for one of the twenty trials.

Unlike the coin flip guessing problem which has 50% precision by naïve random guessing, link prediction is much harder because the precision of a random guess is very low.

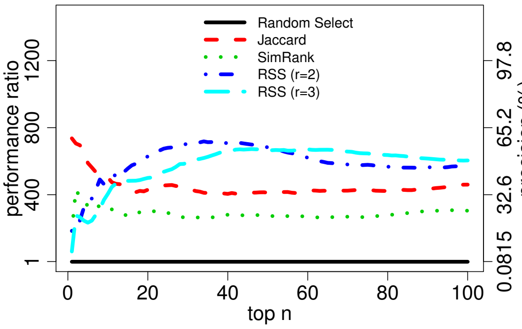


Figure 1: Average performance ratio and precision of various vertex similarity measures for HDN. (baseline measure: random select)

When the training network contains $p = 80\%$ of the edges of the original network, the training network of HDN would have 867 vertices and $1,222$ edges. Randomly picking two vertices and claiming the two should be connected gives $\binom{867}{2} = 375,411$ possible combinations. Since $1,222$ of them are already connected in the training network, there are $375,411 - 1,222 = 374,189$ non-neighbor pairs. Only $1,527(1-p) = 305$ of them are the correct pairs. Thus, the precision of a naïve random pick for the HDN is only $305/374,189 = 0.0815\%$.

To demonstrate the effectiveness of vertex similarity measures, we show both the precision and performance ratio for each measure. The performance ratio $P(S_m, n)$ is defined in Equation 10 as:

$$P(S_m, n) := \frac{Prec(S_m, n)}{Prec(S_r, n)}, \qquad (10)$$

where S_m is the given vertex similarity measure, S_r is a naïve random select measure, $Prec(S_m, n)$ is the precision of S_m by claiming the top-n similar vertex pairs should be connected. A larger performance ratio score is preferred.

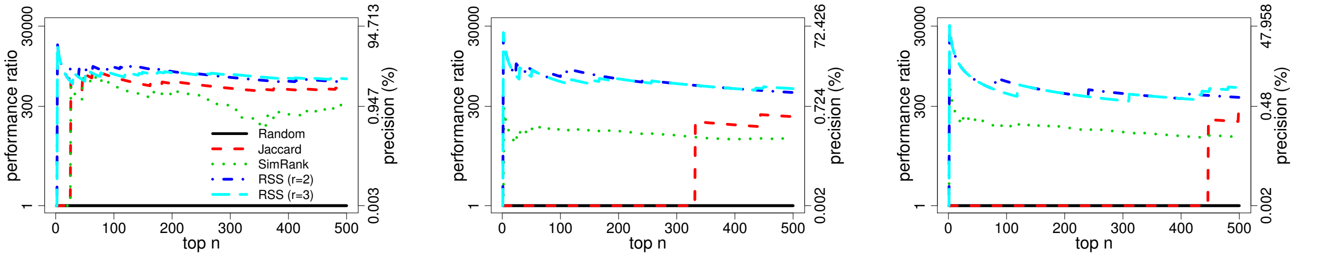
4.1.2 Experimental Results

Figure 1 shows the average precision of 20 independent trials for different vertex similarity, including the baseline method random select, one local topology based approach Jaccard similarity, one global topology based approach SimRank, RSS with discovery range 2, and RSS with discovery range 3.

As shown, Jaccard similarity is good when n is smaller than 10. When n is between 11 and 100, RSS outperforms all other measures for both $r = 2$ and $r = 3$. While SimRank considers the global topology, it seems to have no advantage over other methods. However, even the worst SimRank measure is more than 300 times better than random select in most cases. This demonstrates the potential of vertex similarity measures as the non-expensive indicators for the genetic diseases sharing common genes.

4.2 Network Evolution

4.2.1 Experimental Setup



(a) Performance ratio of similarity measures in G_1 (between 1998 and 2000)

(b) Performance ratio of similarity measures in G_2 (between 2001 and 2003)

(c) Performance ratio of similarity measures in G_3 (between 2004 and 2006)

Figure 2: The performance ratio of different similarity measures for top-n returns. (baseline measure: random select similarity)

Table 2: Statistical measures of the training network for the coauthorship network.

Statistical Measure	Value
Number of Nodes	$26,082$
Number of Edges	$59,742$
Average Degree	4.58
Average Clustering Coefficient	0.48
Average Shortest Path Length	10.99
Diameter	36

We retrieve the papers published between 1995 and 1997 from the CiteSeer^X dataset and build a training set of a coauthorship network, G_0, from the authors of the papers. The statistical measures of the training network is shown in Table 2.

To generate the testing network, we build a coauthorship network from authors who have publications between 1998 and 2000. The authors who have publications in interval $[1998, 2000]$ but not in $[1995, 1997]$ are disregarded since they are not presented in the training network. We repeat the same procedure to produce two more testing coauthorship networks in interval $[2001, 2003]$ and interval $[2004, 2006]$. The three testing coauthorship networks are labeled as G_1, G_2, and G_3 respectively.

We use the number of coauthored papers as the weight of each edge. Therefore, the relation strength from author A to author B becomes

$$R(A, B) := \frac{n_{AB}}{n_A}, \qquad (11)$$

where n_{AB} is the number of A and B's coauthored papers, n_A is number of A's published papers.

We calculate different vertex similarity measures among vertices on the training network G_0 and use the information to infer future collaboration behavior.

Similar to the last experiment, we rank all the node pairs by their similarity scores from the highest to the lowest and claim the top-n node pairs as the authors who will collaborate in the future. Compared with the test network of actual collaborations that occurred, we could calculate each similarity measure's precision, which is used as a proxy of the performance of all the similarity measures. In addition, for this experiment we only care about new collaboration behavior. For two authors who have publications in the training network, their collaboration behavior in the testing network is excluded in the performance evaluation.

4.2.2 *Experimental Results*

As shown in Figure 2, two RSS results (with discovery range equals 2 and 3 respectively) both outperform the local structure based Jaccard similarity and the global structure based SimRank. Note that the y-axis is in logarithmic scale for better visualization.

Figure 2(a) shows the link prediction results for the test graph G_1. The two RSS results are slightly better than Jaccard similarity in general. Even the worst SimRank is more than 100 times better than random select. This shows the potential of vertex similarity measures as a powerful tool to predict the network evolution.

While G_1 is the coauthoring behavior of the near future, G_2 and G_3 represent a further future. Thus, the coauthoring behavior in G_2 and G_3 should be less predictable. RSS shows the superiority in the sense of the stable performance from the near future G_1 to the further future G_2 and G_3. Jaccard similarity performs well in the near future G_1, but the performance is barely satisfactory in the further future G_2 and G_3. This is because Jaccard similarity can only look for nodes at most two hops away. Although the new collaborating behavior of the near future shrink the distance between an author and other non-neighboring people, the training network G_0 cannot be aware of these updates. On the other hand, the global topology based similarity SimRank performs steady compared to Jaccard similarity in the further future G_2 and G_3 because SimRank considers the global topology of the network.

5. CONCLUSION AND FUTURE WORKS

We introduce relation strength similarity, an asymmetric vertex similarity measure that can be applied on weighted networks. The performance is measured in terms of 1) their ability to capture the hidden relationship among vertices, and 2) the power of predicting network evolution. Conducting experiments on the human disease networks and the coauthorship networks, we discover the followings. First, vertex similarity measures has the potential of capturing the network's missing links, which are used to represent the hidden or unknown relationship among objects. Second, RSS is a stable and superior vertex similarity measure compared to the local topology based Jaccard similarity and global topology based SimRank. Third, while the local topology based measures are good at predicting the new established relationship of near future, global topology based measures are better the further in the time we go.

For future work, we plan to investigate the influence of new links and old links in terms of their ability to discover the missing links and a deeper investigation of the role of distance to non-neighbor vertices. We speculate that RRS could be used to predict and capture other social network knowledge such as what friends will share information with others and better predict information spreading in social media.

6. REFERENCES

[1] S. Adafre and M. de Rijke. Discovering missing links in wikipedia. In *Proceedings of the 3rd International Workshop on Link Discovery*, pages 90–97. ACM, 2005.

[2] L. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[3] R. Albert and A. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[4] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[5] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg. Adaptive name matching in information integration. *Intelligent Systems, IEEE*, 18(5):16–23, 2005.

[6] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles. Capturing missing edges in social networks using vertex similarity. In *Proceedings of the 6th ACM International Conference on Knowledge Capture*, pages 195–196. ACM, 2011.

[7] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles. Collabseer: A search engine for collaboration discovery. In *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 231–240. ACM, 2011.

[8] P. Clark, J. Thompson, K. Barker, B. Porter, V. Chaudhri, A. Rodriguez, J. Thoméré, S. Mishra, Y. Gil, P. Hayes, et al. Knowledge entry as the graphical assembly of components. In *Proceedings of the 1st International Conference on Knowledge Capture*, page 29. ACM, 2001.

[9] C. Desrosiers and G. Karypis. Enhancing link-based similarity through the use of non-numerical labels and prior information. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, pages 26–33. ACM, 2010.

[10] E. Eisenberg and E. Levanon. Preferential attachment in the protein network evolution. *Physical Review Letters*, 91(13):138701, 2003.

[11] K. Goh, M. Cusick, D. Valle, B. Childs, M. Vidal, and A. Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685, 2007.

[12] L. Gou, H.-H. Chen, J. Kim, X. Zhang, and C. L. Giles. Sndocrank: a social network-based video search ranking framework. In *Proceedings of the International Conference on Multimedia Information Retrieval*, pages 367–376. ACM, 2010.

[13] L. Gou, X. Zhang, H.-H. Chen, J. Kim, and C. L. Giles. Social network document ranking. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, pages 313–322. ACM, 2010.

[14] G. Jeh and J. Widom. SimRank: A measure of structural-context similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 538–543. ACM, 2002.

[15] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[16] E. Leicht, P. Holme, and M. Newman. Vertex similarity in networks. *Physical Review E*, 73(2):26120, 2006.

[17] C. Li, J. Han, G. He, X. Jin, Y. Sun, Y. Yu, and T. Wu. Fast computation of simrank for static and dynamic information networks. In *Proceedings of the 13th International Conference on Extending Database Technology*, pages 465–476. ACM, 2010.

[18] M. Migliore, V. Martorana, and F. Sciortino. An algorithm to find all paths between two nodes in a graph. *Journal of Computational Physics*, 87(1):231–236, 1990.

[19] M. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):25102, 2001.

[20] D. Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM03: Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pages 556–559, 2003.

[21] D. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976.

[22] G. Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer.* 1989.

[23] P. Tan, M. Steinbach, V. Kumar, et al. *Introduction to data mining.* Pearson Addison Wesley Boston, 2006.

[24] P. Zhao, J. Han, and Y. Sun. P-Rank: a comprehensive structural similarity measure over information networks. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management*, pages 553–562. ACM, 2009.

[25] T. Zhou, L. Lü, and Y.-C. Zhang. Predicting missing links via local information. *The European Physical Journal B-Condensed Matter and Complex Systems*, 71(4):623–630, 2009.