Capturing Missing Links in Social Networks Using Vertex Similarity

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

ABSTRACT

The vertex similarity measure is a useful tool to discover and justify the relationship of vertices in a complex network. We propose the Relation Strength Similarity (RSS), a new vertex similarity measure that utilizes the network topology to discover similar vertices. Compared to other vertex similarity measures, RSS has the following advantages. 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 a weighted network, in which the relation strength of two neighboring nodes can be explicitly expressed. Third, users could adjust the "discovery range" parameter for better performance based on their domain knowledge. Using coauthorship network as experimental data, our method outperforms other vertex similarity measures in terms of the ability to predict future coauthoring behavior among scholars.

Categories and Subject Descriptors

E.1 [**Data Structures**]: Graphs and networks; G.2.2 [**Discrete Mathematics**]: Graph Theory—*Graph algorithms*; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation, Measurement

Keywords

Coauthor network, Graph Theory, Link Analysis, Link Prediction, Information Retrieval, Web of Linked Data

1. INTRODUCTION

Complex network is a graph in which each vertex acts as an object and each edge corresponds to an interaction between two objects. Scientists can infer the nature of the vertices or relationship between vertices by the graph statistical mechanics, such as vertex degree, clustering coefficient, betweenness centrality, and shortest path length [3]. Among all the graph mechanics, one important graph measure is vertex similarity [15], 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 [6], coauthoring behavior inference [21], and knowledge capturing using representational components [7].

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 people 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.

In this paper, we focus on investigating topology based vertex similarity. We present Relation Strength Similarity (RSS), a new 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, the "discovery range" parameter can be adjusted based on user's domain knowledge about the network.

To evaluate and compare RSS with other network topology based vertex similarity measures, we use CiteSeerX[1] dataset to build coauthor networks for experiment. Ex-

[1]`http://citeseerx.ist.psu.edu/`

perimental results show that our method outperforms other vertex similarity measures in terms of the ability to predict the future coauthoring behavior.

The rest of the paper is organized as follows. In Section 2, we review previous work related to vertex feature based and network topology based vertex similarity measures. The calculation, analysis, and example of relation strength similarity measure are introduced 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 potential links on coauthorship network. Summary and future work appears in Section 5.

2. RELATED WORK

2.1 Vertex Feature Based Similarity

Vertex feature based vertex similarity measures utilize the distance between feature vectors to indicate the similarity among vertices. In the simplest case, all the features are treated equally important, and the inverse of Euclidean distance between the feature vectors can be a proxy of similarity. In more realistic cases, users usually need to analyze training data to produce a regression function. The distance between vertices is then determined by the regression function and their feature values. Research issues about feature based similarity include distance function definition [17], feature selection and dimensionality reduction [10], inductive bias problem [5], and so on. Vertex feature based measures are popular in several fields. In [18], the author defined the similarity between words by using the distribution pattern of words as features. In [6], Bilenko et al. proposed string-based similarity computation to identify distinct records referring the same entity. For a complete survey, please see [26].

2.2 Network Topology Based Similarity

While vertex feature based approaches focus on the intrinsic properties of the vertices, other measures exploit the network topology to determine the similarity among vertices. Several topology based approaches, such as Jaccard similarity [24] and cosine similarity [23], stands 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 a phenomenon 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 [22], and protein network [9]. Based on the empirical observation, Newman [20] 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 [28].

Instead of using local neighboring information, the global topology can also be used for vertex similarity calculation. Katz [14] 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 [13], Leicht-Holme-Newman (LHN) [15], and P-Rank [27] 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 some 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 relationship. 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 [21].

Although global topology based measures see a boarder picture of the whole network, they usually computationally expensive. Several approximations for global topology based measures are proposed in recent study. Gou et al. approximated LHN by clustering the social network into virtual nodes to reduce the graph size [11, 12]. Li et al. approximated SimRank by incremental updating [16], but this measure allows only link updating, i.e., it assumes that the total number of vertices in graph is fixed.

Table 1 lists and compares the characteristics of several well known vertex similarity measures. Only part of these works can only be employed on unweighted networks. Our proposed measure, RSS, is the only asymmetric measure.

3. RELATION STRENGTH VERTEX SIMILARITY

3.1 Relation Strength Similarity Calculation

Relation strength similarity permits users to explicitly assign the weights to every edge for initialization. If users have no clue about the relative importance of the edges, they could just 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 vertex A to vertex B is calculated as follows.

Table 1: A comparison of different topology based vertex similarity measures. (n: number of vertices, K: maximum number of iteration, d: average degree (including in-degree and out-degree), r: discovery range, assuming $d << n$, $K << n$, $r << n$)

Name	Time Complexity	Topology Range	Asymmetric?	Weighted Network?	Ref.
RSS	$O\left(nd^r\right) \sim O\left(n\right)$	adjustable	✓	✓	this paper
Jaccard	$O\left(nd^3\right) \sim O\left(n\right)$	local			[23]
cosine	$O\left(nd^3\right) \sim O\left(n\right)$	local			[24]
Adamic-Adar	$O\left(nd^3\right) \sim O\left(n\right)$	local			[2]
Pref. Attach.	$O\left(n^2d\right) \sim O\left(n^2\right)$	local			[20]
Katz	$O\left(n^3\right)$	global		✓	[14]
SimRank	$O\left(Kn^2d^2\right) \sim O\left(n^2\right)$	global		✓	[13]
LHN	$O\left(n^3\right)$	global			[15]
P-Rank	$O\left(Kn^2d^2\right) \sim O\left(n^2\right)$	global		✓	[27]

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

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 *indirect 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 [19]. To make the calculation tractable, we conceive a new *discovery range* parameter, r, to control the maximum degree of separation for indirect relation strength calculation, i.e., we only look for paths at most r hops away. Thus, Equation 2 becomes this.

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

Users could adjust the discovery range based on their domain knowledge. In our experiment, 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, ..., p_M from A to C with path length shorter than discovery range r, the relation strength similarity from vertex A to vertex C is defined as the summation of the relation strength and all the indirect relation strengths, as defined in Equation 4.

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

3.2 Analysis of Relation Strength Similarity

In this section, we first show that the value of RSS is always between 0 and 1. Next, we study and compare several characteristic of the RSS with other similarity measures. Finally, we explain why introducing discovery range is reasonable.

Although normalization seems to be a straightforward step in defining a new measure, several vertex similarity measures, such as Adamic-Adar [2], preferential attachment [20], and Katz [14], cannot be normalized because their maximum possible value could be infinity by their nature. We show the value of RSS is always between 0 and 1 by rewriting Equation 4 as follows.

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

where $B_1^{(m)}, B_2^{(m)}, \ldots B_K^{(m)}$ form p_m, the m^{th} path be-

tween 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, $C = B_{K+1}^{(1)} = B_{K+1}^{(2)} = \ldots = B_{K+1}^{(M)}$ since $B_{K+1}^{(m)}$ is the ending vertex of path p_m, and $N(A)$ is the set of neighboring vertices of A.

Equation 7 holds because the indirect 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, $B_2^{(1)}$, $B_2^{(2)}$, ..., $B_2^{(M)}$ 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 $B_2^{(1)}$, $B_2^{(2)}$, ..., $B_2^{(M)}$ form a subset of $N(A)$.

Compared to other vertex similarity measures, the first advantage of RSS is asymmetric, 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 closer to the real world scenario. We will illustrate a real life example in Section 3.3 to help readers understand more about the powerfulness of asymmetricity. Most of previous vertex similarity measures [2, 13, 14, 15, 20, 23, 24, 27] are symmetric by their nature.

In addition, RSS can be employed on weighted graph. Several previous works treat neighboring vertices equally important in the initial setting [2, 15, 20, 23, 24]. They neglect the fact that neighboring vertices may still have different strength of relation. Different from these approaches, the initial setting of our method allows users to explicitly specify the known relation strength between objects based on application. Take coauthorship network for example, the weights of edges could be used to represent the number of coauthored papers between two authors. For gene promoter network, weighted edges could stand for bp-sharing between promoters.

Finally, users could adjust the discovery range by their domain knowledge. Compared with previous work [2, 23, 24], the local topology based measures are too restrictive in the sense that they only look for vertices with two degree of separation. The global topology based measures [13, 14, 15, 27] are not computationally feasible for large or dynamic networks. Our algorithm allows users to control the discovery range to achieve balance between the two. Although introducing 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 (Equation 4).

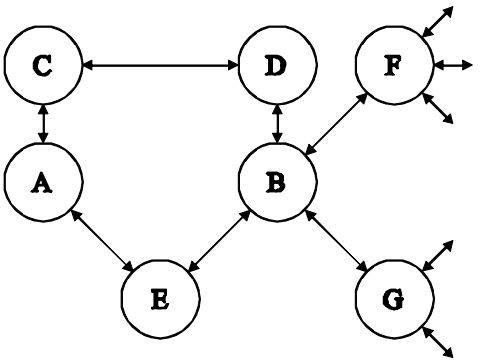


Figure 1: Relation strength similarity example

3.3 A RSS Example

Let's consider a real world scenario happening in academic circle. A young faculty usually has fewer connections with other researchers compared to a senior faculty. Therefore, each connection for the young faculty is relatively important. In addition, a young faculty is usually more eager to establish connections with senior faculties, whereas a senior faculty might be less interested in forming new links, since he or she has have several connections.

Before continuing the scenario, let's take a look at the example illustrated in Figure 1. To simplify the explanation, we assume all the edge weights equal 1, and all the links are reciprocal.

We want to calculate the relation strength similarity from vertex A to vertex B. By Equation 1, we know the $R(A, C)$, relation strength from A to C equals $1/2$, since A has 2 equally important adjacent vertices. Similarly, we could get $R(C, D) = R(D, B) = R(A, E) = R(E, B) = 1/2$. Because path $A - C - D - B$ and path $A - E - B$ are the only two simple paths from A to B, by Equation 4 we get $S(A, B)$ be $R(A, C) \cdot R(C, D) \cdot R(D, B) + R(A, E) \cdot R(E, B)$, which is 0.375. Using similar steps, one can verify that $S(B, A)$ is 0.1875, which is smaller than $S(A, B)$.

Let's go back to the academic scenario. The young faculty's character is like vertex A in the graph. Compared with a senior faculty (vertex B), the relation strength of A to A's neighbors is $1/2$, which is twice as important as B to B's neighbors $(1/4)$. In addition, RSS also displays that the young faculty A would be more eager in getting in touch with the senior faculty B than the other way around.

Compared with other similarity measures, the local topology based measures considers only path $A - E - B$ and

Table 2: Statistical mechanics of the training graph.

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

fails to consider a longer path $A - C - D - B$ in the example. Moreover, most of the similarity measures would determine the young faculty and the senior faculty have equal motivation to establish connection with each other because the symmetric nature of these measures. Our proposed RSS measure successfully explains the real world asymmetric situation.

4. EXPERIMENTS

Evaluating similarity measures is difficult because vertex similarity result is usually lack of interpretability [8]. To compare RSS with other measures, we use CiteSeerX dataset to build the coauthorship network and study the performance of different measures in terms of their ability to predict future collaboration behaviors. To eliminate the author ambiguity problem, we use random forest learning [25] to disambiguate the authors with similar names.

4.1 Experimental Setup

We retrieve the papers published between 1995 and 1997 by CiteSeerX dataset and build a training coauthorship network, G_0, by the authors of the papers. The statistical mechanics of the training network is shown in Table 2.

To generate the testing network, we build a coauthorship network by authors who has publications between 1998 and 2000. The authors who has 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 network 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

$$R(A, B) := \frac{n_{AB}}{n_A}, \tag{10}$$

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

Table 3: Prediction accuracy by determining top 500 similar node pairs will connect in the testing graph

	G_1	G_2	G_3
Random Select	0.004%	0.002%	0.001%
Jaccard	0.604%	0.097%	0.001%
Adamic-Adar	0.498%	0.001%	0.0003%
SimRank	0.310%	0.001%	0.0004%
RSS ($r = 2$)	4.001%	0.701%	0.001%
RSS ($r = 3$)	3.793%	0.626%	0.0007%

Table 4: Prediction accuracy by determining top $5,000$ similar node pairs will connect in the testing graph

	G_1	G_2	G_3
Random Select	0.004%	0.002%	0.001%
Jaccard	0.776%	0.167%	0.046%
Adamic-Adar	0.650%	0.068%	0.042%
SimRank	1.140%	0.201%	0.084%
RSS ($r = 2$)	1.983%	0.547%	0.099%
RSS ($r = 3$)	2.445%	0.658%	0.145%

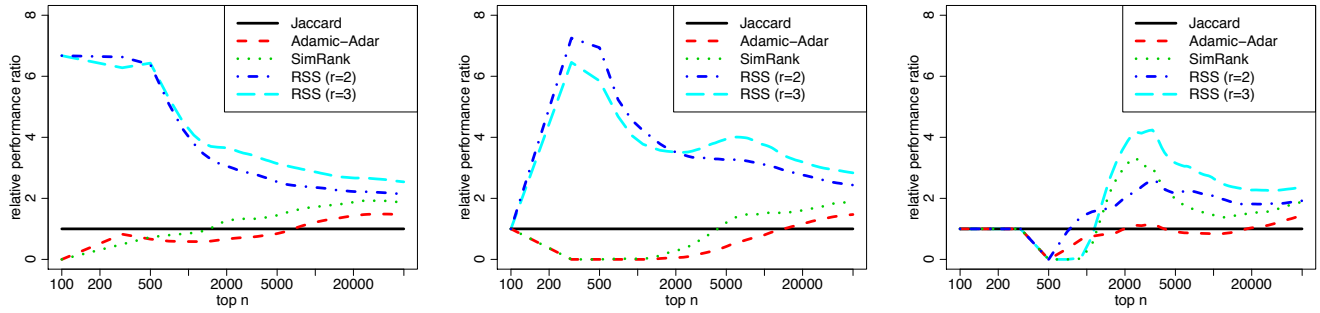
4.2 Evaluation Method

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

For each similarity measure, we rank all the node pairs by their similarity scores from the highest to the lowest. We determine the top-n node pairs as the authors who will collaborate in the future. Comparing the judgment with the testing graph, we could calculate each similarity measure's success rate, which is used as a proxy of the performance of the measure. Since coauthoring behavior is reciprocal, it does not make sense to include both $S(A, B)$ and $S(B, A)$ in the ordering list. Therefore, we only include the larger one of the two. In addition, we only care about new collaboration behaviors in the experiment. For two authors who have publications in the training network, their collaboration behavior in the testing network is excluded in the performance evaluation.

We compare two local topology based similarity measures (Jaccard similarity and Adamic-Adar similarity) and one global topology based similarity measure (SimRank) with our RSS measure by setting discovery range (r) to 2 and 3.

4.3 Experimental Results

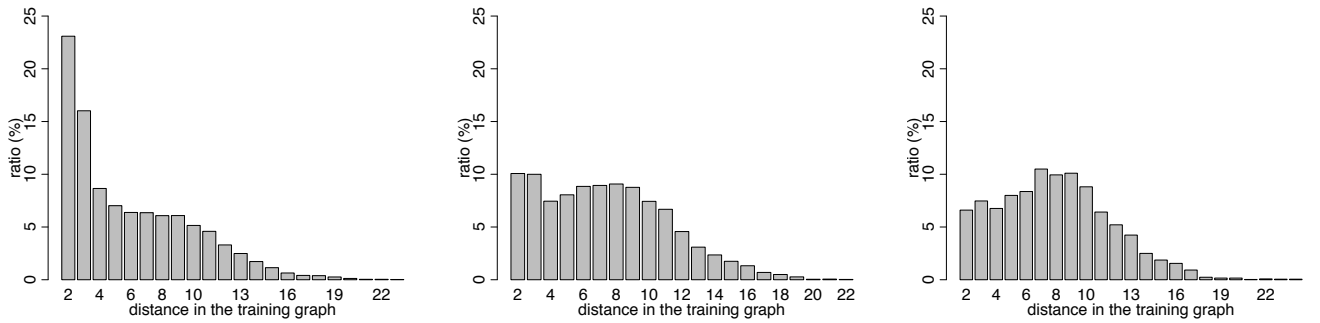


(a) Performance ratio of similarity measures in G_1 (between 1998 and 2000)

(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: Relative performance ratio of different similarity measures for top-n returns. (reference measure: Jaccard similarity)



(a) Authors first collaborate in G_1 (between 1998 and 2000).

(b) Authors first collaborate in G_2 (between 2001 and 2003).

(c) Authors first collaborate in G_3 (between 2004 and 2006).

Figure 3: The distance distribution of two vertices in G_0 before their first collaboration.

In addition to the network topology issues, there are a lot of reasons for two authors to (or not to) collaborate. For example, a PhD student generally works with his or her adviser and well connected with part of the faculties in the department. After graduating, the student usually changes to another institution and starts to get in touch with new faces. This kind of transition behavior cannot be inferred from network topology alone. Therefore, the accuracy of predicting new links by only topology based similarity measures is relatively low. A similar result was also reported in Nowell and Kleinburg's paper [21].

To make the comparison more meaningful, we show the accuracy of random select measure, which randomly pick up two non-adjacent vertices in the training network. As shown in Table 3, by determining the top 500 similar node pairs will connect, Jaccard similarity is more than 150 times better than random select in testing network G_1, which represents the coauthoring behavior between 1998 and 2000. For G_2 (the coauthoring behavior between 2001 and 2003), Jaccard similarity is 48 times better than random select. Although Jaccard similarity has similar performance with random select in G_3, it is 46 times better than random select when we determine top $5,000$ similar node pair will connect, as shown in Table 4. The predicting accuracy of other similarity measures is also listed.

Since different number of returns may cause different accuracy, we show the relationship between the number of returns and the relative performance ratio. Using Jaccard similarity as the reference measure, the relative performance of other similarity measures with number of returns ranging from 10 to $30,000$ is displayed in Figure 2.

As shown in Figure 2(a) and Figure 2(b), Two RSS results (with discovery range equals 2 and 3 respectively) both outperform Jaccard similarity by a factor of 1 to 7, depending on the number of returns. The performance of Adamic-Adar and SimRank is similar to Jaccard similarity, although they tend to be better as the number of returns increases.

While G_1 and G_2 are the coauthoring behavior of near future, G_3 represents the farther future. Therefore, the coauthoring behavior in G_3 is less predictable, as shown in Figure 2(c). For the top 500 returns, all the similarity measures have no advantage over random select (refer the last column of Table 3). The similarity measures start to perform better as the number of returns

increases. As the number of returns reaches $5,000$, Jaccard similarity is 46 times better than random select. Our proposed RSS is 145 times better even discovery range is set to a small number 3.

An interesting discovery is that SimRank seems have no apparent advantage over Jaccard and Adamic-Adar for G_1 and G_2, although SimRank considers global topology. To further investigate the finding, we plot the distance distribution between two authors in training network (G_0) if the two authors have no coauthored paper in G_0 and have coauthoring behavior in G_1, G_2, or G_3. As shown in Figure 3(a), authors tend to work with people who are not far away in the coauthoring network. The collaborators within distance 4 account for 50% of the new established links. Although these new links would shrink the distance between an author and other non-neighboring people, the training graph G_0 cannot be aware of these updates. Therefore, the distribution in Figure 3(b) and Figure 3(c) looks like authors start to work with people of a larger distance as time goes by. For testing network G_3, the majority of the collaborators are of distance 7 to 9. Since local topology based similarity measures (Jaccard and Adamic-Adar) can only look for nodes at most two hops away, global topology based similarity (SimRank) starts to outperform these methods. This tells us that while local topology based measures are good at predicting near future collaborating behaviors, global topology based measures could be a better option if we want to predict connections of farther future.

5. CONCLUSION AND DISCUSSION

In this paper, we introduce relation strength similarity (RSS), a new vertex similarity measure that could better capture the potential relationship of real world structure context. RSS is unique in three aspects. First, it is an asymmetric measure which could be used for a more general purpose social network analysis. Second, it allows users to explicitly specify the relation strength between neighboring vertices for initialization. Third, the discovery range parameter could be adjusted by users based on their domain knowledge for computation efficiency and performance concern. We illustrate a real life example to demonstrate that RSS could better explain a scholar's degree of interest to collaborate with other scholars.

The vertex similarity measures are evaluated in terms of their ability to predict future collaboration behavior among scholars. Although the collaborating behavior cannot be interpreted by network structure alone, introducing vertex similarity measure can be a great help for future collaboration prediction. Experimental results show that RSS outperforms both local topology based similarity measures (Jaccard and Acadmic-Adar) and global topology based similarity (SimRank) by a factor of 1 to 7, even with small discovery range. We also discover that local topology based measures are good at predicting collaborations of near future, whereas global topology based method is a better option if we are asking for a long term prediction.

We plan to investigate the influence of new links and old links in terms of their ability to predict collaboration as future work.

6. ACKNOWLEDGMENTS

We gratefully acknowledge partial support from Alcatel-Lucent and NSF.

7. 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] J. Baxter. A model of inductive bias learning. *JAIR*, 12:149–198, 2000.

[6] 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.

[7] 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.

[8] 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.

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

[10] I. Fodor. A survey of dimension reduction techniques. *Livermore, CA: US DOE Office of Scientific and Technical Information*, 18, 2002.

[11] 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.

[12] 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.

[13] 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.

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

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

[16] 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.

[17] M. Li, X. Chen, X. Li, B. Ma, and P. Vitányi. The similarity metric. *Information Theory, IEEE Transactions on*, 50(12):3250–3264, 2004.

[18] D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics, 1998.

[19] 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.

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

[21] 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.

[22] 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.

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

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

[25] P. Treeratpituk and C. L. Giles. Disambiguating authors in academic publications using random forests. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 39–48. ACM, 2009.

[26] V. Vapnik. *The nature of statistical learning theory*. Springer Verlag, 2000.

[27] 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.

[28] 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.