Capturing Missing Edges 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
University Park, PA 16802, USA
hhchen@psu.edu, {lug129, lzhang, giles}@ist.psu.edu

ABSTRACT

We introduce the graph vertex similarity measure, Relation Strength Similarity (RSS) [2], that utilizes a network's topology to discover and capture similar vertices. The RSS has the advantage that it is asymmetric; can be used in a weighted network; and has an adjustable "discovery range" parameter that enables exploration of friend of friend connections in a social network. To evaluate RSS we perform experiments on a coauthorship network from the CiteSeerX database. Our method significantly outperforms other vertex similarity measures in terms of the ability to predict future coauthoring behavior among authors in the CiteSeerX database for the near future 0 to 4 years out and reasonably so for 4 to 6 years out.

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

Coauthor Network,Vertex Similarity, Link Analysis, Link Prediction, Information Retrieval, Web of Linked Data

1. VERTEX SIMILARITY INTRODUCTION

Among all graph measures, an important one is vertex similarity, which is a measure of the similarity between vertices. Vertex similarity calculation can be classified into local and global structure based approaches. Local structure based approaches, such as Jaccard similarity [4], are calculated based on the intuition that two vertices are more similar if they share more common friends. Adamic and Adar [1] refined these measures by assigning more weight to vertices with fewer degrees because these are better discriminators. Although local structure based approaches are computationally efficient, they fail to consider all orders of edges such as neighbors neighbors neighbors. Global structure based measures define the similarity recursively: two vertices are similar if their immediate neighbors in the network are themselves similar. SimRank [3] is the most well known of the global based measures. However, SimRank is a symmetric measure, i.e., the similarity of vertex A to vertex B is commutative .

K-CAP'11, June 26–29, 2011, Banff, Alberta, Canada.
ACM 978-1-4503-0396-5/11/06.

2. RSS CALCULATION

Our proposed Relation Strength Similarity (RSS) is a asymmetric vertex similarity measure that can be used on a weighted network. RSS of vertices explicitly assigns the weights to every edge for initialization. RSS is calculated from a normalized edge weighting score based on the relative degree of similarity between neighboring vertices. The relation strength R from vertex A to vertex B is:

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

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.

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

$$R^*_{p_m}(A,C) := \begin{cases} \prod_{k=1}^{K} R(B_k, B_{k+1}) & \text{if } K \leq r, \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where r is the discovery range parameter that control the maximum degree of separation, i.e., we only look for paths at most r hops away. The discovery range for a social network can be based on a network's domain knowledge. In our experiments, 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 3,

$$S(A,C) := \sum_{m=1}^{M} R^*_{p_m}(A,C). \quad (3)$$

3. A RSS EXAMPLE

Let's consider a real world scenario. A young researcher usually has fewer connections with other researchers com-

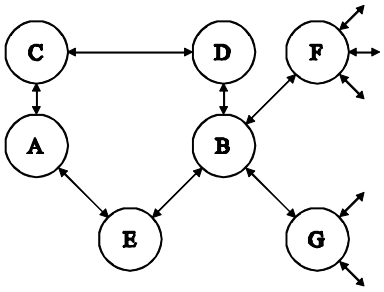

Figure 1: Relation strength similarity example.

pared to a senior researcher. Therefore, each potential research connection for a young researcher is relatively more important. In addition, a young researcher is usually eager to establish connections with strong collaborators, whereas a senior collaborator might be less interested in forming new links, since he or she already has have several connections and collaborators.

To explain this scenario, consider the example illustrated in Figure 1. To simplify, 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 3 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 $S(A, B)$.

For the scenario previously discussed, the young researcher could be considered as vertex A and the senior researcher (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 for this case implies that the young researcher A may be more eager in getting making contact with the senior researcher B than the other way around.

4. EXPERIMENT AND CONCLUSIONS

To test out our hypothesis and the value of RSS, we consider a coauthorship network. We use the CiteSeerX dataset to build a coauthorship network and study the performance of different measures in terms of their ability to predict future collaboration behavior. Specifically, the papers published between 1995 and 1997 are used to build a training coauthorship network, G_0. The training network contains $26,082$ vertices and $59,742$ edges. 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. The number of coauthored papers is used as the weight of each edge. Therefore, the relation strength from author A to author B becomes $R(A, B) := \frac{n_{AB}}{n_A}$, where n_{AB} is the number of A and B's coauthored papers, n_A is number of A's published papers.

Table 1: Improvement ratio over random selection specifying the linked top 1000 similar vertex pairs.

	G_1	G_2	G_3
Random Select	0.004%	0.002%	0.001%
Jaccard	221	116	46
Adamic-Adar	125	108	50
SimRank	91	83	75
RSS ($r = 2$)	498	**428**	95
RSS ($r = 3$)	**598**	399	**98**

We compare RSS with two local structure based measures (Jaccard similarity and Adamic-Adar similarity) and one global structure based measure (SimRank) against randomly selecting any author as a possible collaborator with another with the percentage the likelihood two authors will collaborate in the future. The RSS outperforms all the vertex similarity measures. As shown in Table 1, by determining the top 1000 similar vertex pairs will connect, RSS with discovery range 2 is 500 times better than random select in testing network G_1 and more than 400 times better than the random select in G_2. Compared to G_1 and G_2, G_3 is less predictable because it represents a farther future. However, RSS still much better; it is nearly 100 times better than random select. Increasing the discovery range of RSS is helpful in predicting near future (G_1), but the advantage is less obvious in predicting a further future (G_2 and G_3).

An observation is that SimRank seems to have no apparent advantage over Jaccard and Adamic-Adar for G_1 and G_2 even though SimRank considers global topology. This is because coauthors tend to work with those who are near their social circle. For the testing network G_3, the majority of the collaborators are of hop distance 7 to 9 in G_0. Since local topology based similarity measures (Jaccard and Adamic-Adar) can only look for vertices 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 are better predictors of collaborators further in the future. Future work would be the investigation of robustness of RSS to link noise and temporal changes and realistic measures of recommendations for collaborators.

5. REFERENCES

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

[2] H.-H. Chen, L. Gou, X. Zhang, , and C. L. Giles. Collabseer: A search engine for collaboration discovery. In *Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM, 2011.

[3] 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.

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