Predicting Recent Links in FOAF Networks

Hung-Hsuan Chen[1], Liang Gou[2], Xiaolong (Luke) Zhang[2], and C. Lee Giles[1,2]

[1] Computer Science and Engineering
[2] Information Sciences and Technology
Pennsylvania State University
hhchen@psu.edu, {lug129,lzhang,giles}@ist.psu.edu

Abstract. For social networks, prediction of new links or edges can be important for many reasons, in particular for understanding future network growth. Recent work has shown that graph vertex similarity measures are good at predicting graph link formation for the near future, but are less effective in predicting further out. This could imply that recent links can be more important than older links in link prediction. To see if this is indeed the case, we apply a new relation strength similarity (RSS) measure on a coauthorship network constructed from a subset of the CiteSeer^X dataset to study the power of recency. We choose RSS because it is one of the few similarity measures designed for weighted networks and easily models FOAF networks. By assigning different weights to the links according to authors coauthoring history, we show that recency is helpful in predicting the formation of new links.

Keywords: Social Network Analysis, Graph Analysis, Vertex Similarity, Coauthor Network Analysis, Relation Strength Similarity, RSS.

1 Introduction

A network is a set of vertices connected by links that formally modeled the relationship between objects. Using the terminology of graph theory, the objects are usually called nodes or vertices, and the links are usually called edges or links (we will use these terms interchangeably in the paper). It has been shown that several vertex properties and the relationship between vertices can be inferred by statistical measures, such as degree, average path length, and clustering coefficient of the vertices.

One measure that has recently attracted much attention is vertex similarity, which measures how similar two vertices are. Vertex similarity measures is usually used to predict the missing or future links of the networks based on the idea that two vertices tend to have a link connection if they are more similar. Recent studies have shown that vertex similarity measures are good at predicting the links of the near future, but they are less effective for the further future link prediction [3,5]. This implies that recent links could be more important indicators than the old links in the link formation process. However, there has been no systematic study of the effect of the recency factor for missing or future link prediction. We address this question with an empirical study and use a subset

S.J. Yang, A.M. Greenberg, and M. Endsley (Eds.): SBP 2012, LNCS 7227, pp. 156–163, 2012.

of the CiteSeer^{X}[1] dataset to build a coauthorship network. Applying the relation strength similarity as the vertex similarity measure to explore the FOAF network, we show that the performance of link prediction can be improved by assigning more weights to the new links than the old links.

The rest of the paper is organized as follows. In Section 2, we introduce previous work about link formation, link prediction, and vertex similarity measures. Section 3 introduces the new relation strength similarity measure, and the integration of a recency factor. The experiments described in Section 4 demonstrate the power of recency in terms of its ability to predict future collaboration behavior. Discussions and future work appear in Section 5.

2 Related Work

Sociologists have long studied the question - what influences people to make friends? Studies have shown that people sharing mutual friends will be more likely to have these mutual friends become friends in the future [14]. This phenomena, called "triadic closure" [12], has been observed in several types of networks such as coauthorship networks [3,4], social networks [10], and information networks [15]. Based on this observation, local structure based vertex similarity measures such as Jaccard similarity [18], cosine similarity [17], and Adamic-Adar's measure [1] have been suggested as an important measure. The principle behind these methods is that two non-adjacent nodes are more likely to connect if they share more common neighbors. Furthermore, these types of measure are usually computationally efficient.

Instead of using the local structure information such as the number of mutual friends, a global structure has been suggested as influential. Zhou introduced such a global structure by suggesting that two vertices n_i and n_j are similar if the average distance from n_i to any other node n_k is closer to that from n_j to n_k for $i \neq j \neq k$ [21]. Others defined the similarity index recursively: two vertices are similar if their corresponding neighbors are similar [7,9]. Katz proposed an index based on the number and the lengths of the simple paths between two vertices [8]. Although global structure based methods consider the complete graph, empirical studies showed that the global structure based methods are worse than the local structure ones in predicting the missing links [13,3,5].

Actual applications have been developed using vertex similarity measures. CollabSeer[2], an academical collaborator recommendation system, analyzed both the researchers' research interests and the structure FOAF coauthorship network [4]. The "Don't forget Bob!" and "Got the wrong Bob?"[3] are two Gmail Lab features helping users to identify the right mail receivers by analyzing the email network [16].

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

[2] `http://collabseer.ist.psu.edu/`

[3] `http://gmailblog.blogspot.com/2011/04/dont-forget-bob-and-got-wrong-bob.html`

Another research topic related to this work is network evolution. Erdős-Rényi graph (ER graph) [6] is probably the simplest random graph generating model. It has been studied for decades thus several properties were carefully analyzed [11]. However, observations showed that several characteristics of real networks don't fit ER model. Other network generating models, such as the Watts-Strogatz graph (WS graph) and Barabási-Albert model (BA model) [20,2], were proposed to fit the the real network better.

3 RSS and Recency Factor

We propose to study the effect of recency on the formation of new edges. To do this, we assign different weights to different edges based partially on their ages. The weighted network is used to compute the similarity score between the vertices. Previous studies showed that the local structure based vertex similarity measures are more computationally efficient and better than global structure based measures in terms of link prediction [3]. However, most of the local structure based similarity measures, such as Jaccard similarity, cosine similarity, and Adamic Adar similarity, consider only the number of mutual friends between two vertices; thus, the edge weights cannot be integrated into these models. Therefore, we use relation strength similarity (RSS), a similarity measure that is designed for weighted networks. The discovery range parameter of RSS can be adjusted so it becomes a local structure based similarity measure, which is computationally efficient. In this section, we first briefly introduce RSS and then integrate into RSS a recency factor.

Relation strength similarity was proposed and analyzed in Chen [3,4,5]. Given a network, RSS is calculated based on the following intuitions: two non-neighboring vertices v_i and v_j are more similar if 1) the path length (number of hops) between v_i and v_j is shorter; 2) the number of distinct paths between v_i and v_j is larger; and 3) the relation strength of the neighbor vertices along the paths from v_i to v_j is larger.

We construct the coauthorship network as a weighted network as follows. Each identical author is regarded as a vertex in the graph. Two vertices v_i and v_j are connected if the two authors have previously coauthored. The edge weight is assigned as the number of coauthored papers. The relation strength is defined as the normalized weight [3] as follows.

$$R(v_i, v_j) := \frac{n_{ij}}{n_i}, \tag{1}$$

where n_{ij} is the number of v_i and v_j's coauthored papers, and n_i is the number of v_i's publications.

A researcher's research interests may vary over time. Most likely, a recent publication is more representative of a researcher's latest interests. Thus, new collaborators should be better for inferring a researcher's future collaboration preferences. Previous work showed that local structure based similarity measures are better in predicting coauthoring behavior in the near future [3,5]. This could imply that the new collaborators are more important than the old collaborators. To introduce a recency factor to the graph, we model the edge weights as an exponential decay function over time with a half life time of T_h. Let $n_{i,j}(t)$ denote the number of coauthored papers between v_i and v_j at year t. The decay rate λ of the exponential decay function can be derived by Equation 2.

$$\begin{aligned} & n_{i,j}(t+T_h) = \tfrac{1}{2} n_{i,j}(t) \\ \Rightarrow\ & n_{i,j}(t) \exp(-\lambda T_h) = \tfrac{1}{2} n_{i,j}(t) \\ \Rightarrow\ & \lambda = \tfrac{\ln 2}{T_h}. \end{aligned} \tag{2}$$

Let's assume author v_i and v_j coauthored $n_{i,j}^{(1)}, n_{i,j}^{(2)}, \ldots, n_{i,j}^{(K)}$ papers in year $y_1, y_2, \ldots, y_K$ respectively. The edge weight at time t_{now} is defined as

$$\begin{aligned} n_{i,j}(t_{now}) &= \textstyle\sum_{k=1}^{K} n_{i,j}^{(k)} \exp(-\lambda(t_{now} - t_k)) \\ &= \textstyle\sum_{k=1}^{K} n_{i,j}^{(k)} \exp\left(\frac{-\ln 2}{T_h}(t_{now} - t_k)\right). \end{aligned} \tag{3}$$

Instead of equation 1, the new relation strength considering both the number of coauthored papers and the recency factor between v_i and v_j is defined as Equation 4.

$$R(v_i, v_j) := \frac{n_{i,j}(t_{now})}{\sum_{\forall k \in N(v_i)} n_{i,k}(t_{now})}, \tag{4}$$

where $N(v_i)$ returns all the neighbors of v_i.

RSS uses the relation strength between neighbor vertices as the foundation to calculate the similarity score between non-neighboring vertices. Assume v_i can arrive v_j through path p_m, which is formed by vertices $v_i(= u_1)$, u_2, u_3, $\ldots$, u_{K-1}, $v_j(= u_K)$. The general relation strength from v_i to v_j through p_m is defined in Equation 5.

$$R^*_{p_m}(v_i, v_j) := \begin{cases} \prod_{k=1}^{K-1} R(u_k, u_{k+1}) & \text{if } K \le r \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where r is the discovery range parameter controlling the maximum degree of separation for collaborator recommendation. The parameter plays a tradeoff between the computation efficiency and relation discovery range.

Assuming there are M distinct paths from v_i to v_j, the relation strength similarity is calculated by Equation 6.

$$S(v_i, v_j) := \sum_{1}^{M} R^*_{p_m}(v_i, v_j). \tag{6}$$

Since RSS guarantees the vertex similarity measures within 0 and 1 as long as the relation strength is normalized [4], it is easy to integrate RSS with other scoring.

Table 1. Information and statistical measures of training networks G_1, G_2, and G_3

	G_1	G_2	G_3
Year	1995 – 1997	1999 – 2001	2003 – 2005
Number of Vertices	1,019	2,556	2,198
Number of Edges	2,286	5,308	4,303
Average Degree	4.49	4.15	3.92
Average Clustering Coefficient	0.55	0.55	0.54
Average Shortest Path Length	13.14	14.44	14.19
Diameter	37	45	40

4 Experiments

To show the power of recency, we use a subset of the CiteSeer^X dataset to build coauthorship networks and study the performance before and after introducing the recency factor in terms of their ability to predict future collaboration behavior. To eliminate the author ambiguity problem, random forest learning methods [19] are used to disambiguate different authors with similar names and authors whose names have several variations.

4.1 Experiment Setup

We retrieve the authors who published at least 5 papers between 1995 and 1997 from CiteSeer^X dataset and build a coauthorship network among the authors. The giant component of the network is called network G_1. The same process is performed from 1999 to 2001 and from 2003 to 2005 to generate two more networks G_2 and G_3. The networks G_1, G_2, and G_3 are the training networks because they are used to calculate the the similarity scores between non-neighboring vertices. The information and the statistical measures of the training networks are shown in Table 1.

We create a testing network H_1 from the coauthorship network of the authors who have publications in 1998. The authors who have publications in 1998 but not in interval $[1995, 1997]$ are disregarded since they are not presented in the training network. The edges that already appeared in $[1995, 1997]$ are also disregarded because we are only interested in predicting new collaboration behavior. By similar manner, we created two more testing networks H_2 of year 2002 and H_3 of year 2006. The information and statistical measures of the testing networks are shown in Table 2. Note that the average shortest path length and the diameter are not shown because each of H_1, H_2, and H_3 is not a connected component.

To test the power of recency, we assign different values to the half life time parameter T_h in the calculation. Specifically, we assign T_h to be 0.5, 1.0, 1.5, 2.0, and ∞ (years). When $T_h = \infty$, the model considers only the number of coauthored papers between two authors. We use RSS with discovery parameter

Table 2. Information and statistical measures of testing networks H_1, H_2, and H_3

	H_1	H_2	H_3
Year	1998	2002	2006
Number of Vertices	656	$1,613$	$1,205$
Number of Edges	$1,255$	$2,991$	$2,034$
Average Degree	3.83	3.71	3.38
Average Clustering Coefficient	0.54	0.52	0.52

$r = 2$ to calculate the similarity scores between vertices, and claim the top-n similar non-neighboring vertices will connect. The vertex similarity scores calculated from G_1, G_2, and G_3 are used to predict the links in H_1, H_2, and H_3 respectively.

4.2 Experimental Results

Previous studies showed that the precision of link prediction is usually very low [3,5,13]. This is because the sparsity of the links makes a naïve random guess very unlikely to be correct.

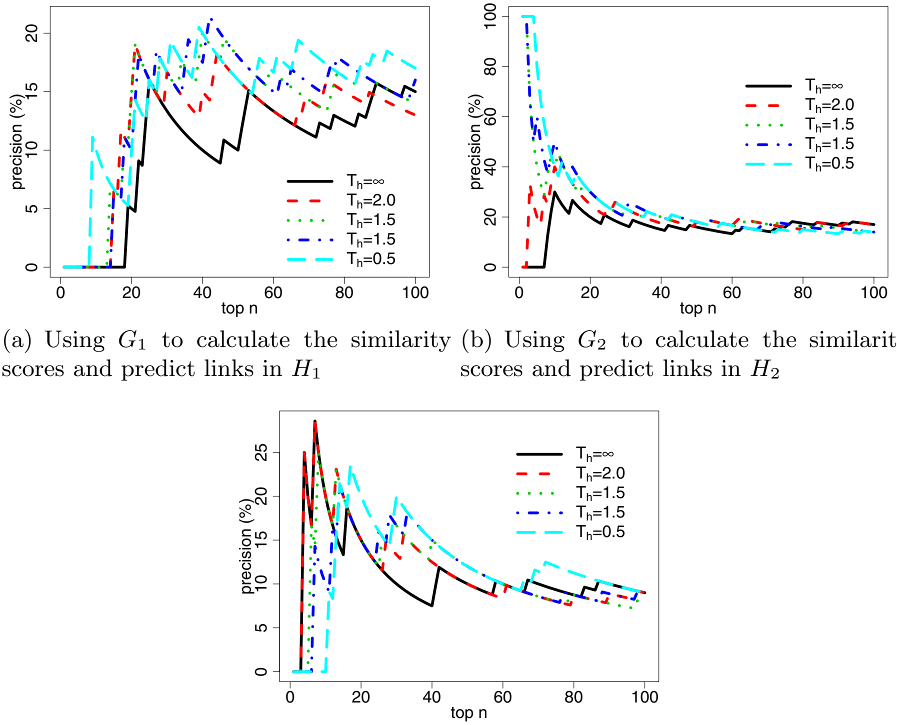
As mentioned in last section, we claim the top-n similar non-neighboring vertices to be connected. Different n will cause different precision. To be fair, we show the precisions of different n (from 1 to 100) in Figure 1.

The five different lines in each sub-graph represent $T_h = \infty$ (years), $T_h = 2.0$ (years), $T_h = 1.5$ (years), $T_h = 1.0$ (years), and $T_h = 0.5$ (years) respectively. The lower the value of the half life parameter T_h, the more important the recent edges are. In general, a smaller half life parameter yields better precision in all three experiments. This means the recent edges do play a more important role in future link formation.

5 Discussion and Future Work

Although the evolution of networks has been well studied, most work only considers how the network grows. One interesting topic rarely discussed is whether the nodes or edges degenerate and therefore gradually lose their influence over time? Probably because most of the available network datasets don't contain such information, little has been done to explore this question.

In this paper, we try approach the recency problem by introducing a recency factor to the edges. We construct the coauthorship network and assign the initial weight of an edge to be proportional to the number of coauthored papers between two authors. The weight decays exponentially as time unfolds, and the weight can be strengthen again if the two authors recently have coauthored new papers. By integrating the recency factor, we show that future links can be better predicted. This demonstrates that recent links should be more representative than the older



(a) Using G_1 to calculate the similarity scores and predict links in H_1

(b) Using G_2 to calculate the similarity scores and predict links in H_2

(c) Using G_3 to calculate the similarity scores and predict links in H_3

Fig. 1. The accuracy of different half time values

links for the formation of future links, and implies that the links and nodes may gradually lose their influence and predictive power over time, i.e. they age.

For future work, the effect of recency and aging factor can be further investigated by various machine learning methods. A user survey can also show the effectiveness of link prediction. Since networks do age, it would be interesting to investigate not only the growth but also the degeneration of networks.

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

References

1. Adamic, L., Adar, E.: Friends and neighbors on the web. Social Networks 25(3), 211–230 (2003)
2. Barabási, A., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509 (1999)

3. Chen, H.-H., Gou, L., Zhang, X., Giles, C.L.: Capturing missing edges in social networks using vertex similarity. In: The 6th International Conference on Knowledge Capture. ACM (2011)
4. Chen, H.-H., Gou, L., Zhang, X., Giles, C.L.: Collabseer: A search engine for collaboration discovery. In: Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries. ACM (2011)
5. Chen, H.-H., Gou, L., Zhang, X., Giles, C.L.: Discovering missing links in networks using vertex similarity measures. In: The 27th ACM Symposium on Applied Computing. ACM (2012)
6. Erdös, P., Rényi, A.: On random graphs, i. Publicationes Mathematicae (Debrecen) 6, 290–297 (1959)
7. Jeh, G., Widom, J.: Simrank: a measure of structural-context similarity. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 538–543. ACM (2002)
8. Katz, L.: A new status index derived from sociometric analysis. Psychometrika 18(1), 39–43 (1953)
9. Leicht, E., Holme, P., Newman, M.: Vertex similarity in networks. Physical Review E 73(2), 026120 (2006)
10. Leskovec, J., Backstrom, L., Kumar, R., Tomkins, A.: Microscopic evolution of social networks. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 462–470. ACM (2008)
11. Newman, M.E.J.: Random graphs as models of networks. In: Handbook of Graphs and Networks, pp. 35–68 (2003)
12. Nguyen, V.-A., Leung, C.W.-K., Lim, E.-P.: Modeling Link Formation Behaviors in Dynamic Social Networks. In: Salerno, J., Yang, S.J., Nau, D., Chai, S.-K. (eds.) SBP 2011. LNCS, vol. 6589, pp. 349–357. Springer, Heidelberg (2011)
13. Nowell, D., Kleinberg, J.: The link prediction problem for social networks. In: CIKM 2003: Proceedings of the Twelfth International Conference on Information and Knowledge Management, pp. 556–559 (2003)
14. Rapoport, A.: Spread of information through a population with socio-structural bias: I. assumption of transitivity. Bulletin of Mathematical Biology 15(4), 523–533 (1953)
15. Romero, D., Kleinberg, J.: The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In: Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, pp. 138–145 (2010)
16. Roth, M., Ben-David, A., Deutscher, D., Flysher, G., Horn, I., Leichtberg, A., Leiser, N., Matias, Y., Merom, R.: Suggesting friends using the implicit social graph. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 233–242. ACM (2010)
17. Salton, G.: Automatic text processing: the transformation, analysis, and retrieval of information by computer (1989)
18. Tan, P., Steinbach, M., Kumar, V., et al.: Introduction to data mining. Pearson Addison Wesley, Boston (2006)
19. Treeratpituk, P., Giles, C.: Disambiguating authors in academic publications using random forests. In: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 39–48. ACM (2009)
20. Watts, D., Strogatz, S.: Collective dynamics of small-world networks. Nature 393(6684), 440–442 (1998)
21. Zhou, H.: Distance, dissimilarity index, and network community structure. Physical Review E 67(6), 061901 (2003)