The Predictive Value of Young and Old Links in a Social Network

Hung-Hsuan Chen[†], David J. Miller[‡], C. Lee Giles[†§]
[†]Computer Science and Engineering, [‡]Electrical Engineering, [§]Information Sciences and Technology
The Pennsylvania State University, University Park, PA 16802, USA
hhchen@psu.edu, djmiller@engr.psu.edu, giles@ist.psu.edu

ABSTRACT

Recent studies show that vertex similarity measures are good at predicting link formation over the near term, but are less effective in predicting over the long term. This indicates that, generally, as links age, their degree of influence diminishes. However, few papers have systematically studied this phenomenon. In this paper, we apply a supervised learning approach to study age as a factor for link formation. Experiments on several real-world datasets show that younger links are more informative than older ones in predicting the formation of new links. Since older links become less useful, it might be appropriate to remove them when studying network evolution. Several previously observed network properties and network evolution phenomena, such as "the number of edges grows super-linearly in the number of nodes" and "the diameter is decreasing as the network grows", may need to be reconsidered under a dynamic network model where old, inactive links are removed.

Categories and Subject Descriptors

G.2.2 [**Mathematics of Computing**]: Discrete Mathematics—*Graph Theory*; E.1 [**Data Structures**]: Graphs and Networks; G.3 [**Mathematics of Computing**]: Probability and Statistics—*Correlation and Regression Analysis*

General Terms

Theory, Algorithm

Keywords

Vertex Similarity, Network Revolution, Link Prediction, Link Analysis, Densification Power-law, Graph Theory

1. INTRODUCTION

Several interesting statistical properties of networks and network evolution have been observed in real-world networks. For example, real networks usually have high clustering coefficients, small average shortest path lengths, and power-law degree distributions. As a network evolves, i.e., as more nodes join the network and more links are established, the number of edges usually grows faster than the number of nodes, with the densification following a power-law pattern (we use the terms "node" and "vertex", and "edge" and "link" interchangeably). Even more surprisingly, network diameters often shrink as networks grow. Several network generating models, which capture networks that follow some of these properties, have been proposed.

These statistical observations and discoveries mostly assume that links will always exist once they are established. In several real networks, however, this assumption seems to be naïve. For example, a Ph.D. student usually collaborates with the faculty and other graduate students at the same university. However, she/he may exclusively work with a different group of scholars after graduating and moving to another institute. When modeling coauthoring behavior via a coauthorship network, the links between an individual and "old" colleagues could become less prominent and even gradually die out (become inactive). Most previous studies do not consider link age as a factor influencing network evolution and new link formation. Thus, conclusions reached by these studies may no longer hold if, to accurately reflect the current "active" network, old (inactive) links are removed.

In this paper, we study how link age influences the evolution of a network at the finest granularity, i.e., the impact of a link's age on the formation of new links. Although we are interested in studying the relationship between link age and new link formation on many different types of networks, most available data sets have no age information associated with links. This is probably the reason why there is little previous work systematically studying the age factor. In this work, we analyze two networks with age values on edges: a coauthorship network among computer scientists and a co-starring network among actors. We observe the relationship between the existing links' ages and the formation of new links, and quantify the relative influential power of young links and old links via the parameters of a logistic regression classifier learned by gradient ascent on the training set data log-likelihood. Although one could apply more sophisticated models such as kernel-based support vector machines or ensemble techniques to achieve higher link prediction accuracy, these models are not used here because the influence of individual parameters/features is less easy to infer in such models.

Two main conclusions are highlighted here. First, on the application domains considered here, the active periods of links are usually short. Second, by modeling the link predic-

DBSocial '13 New York, NY USA

tion problem as a logistic regression classification problem, young links are shown to be more informative than old links in predicting the formation of new links.

The rest of the paper is organized as follows. In Section 2, we review previous work on the statistics, evolution properties, and generating models of networks. Several link prediction methods and vertex similarity measures are also reviewed. Section 3 introduces the two network domains considered in our experiments. Section 4 studies typical active link periods for the two target networks. Section 5 applies supervised learning to quantify the relative influential power of young links and old links. Finally, discussions and future work appear in Section 6.

2. RELATED WORK

Network generation and evolution can be studied at both macro and micro levels. At the macro level, scientists have observed several statistical properties that commonly exist in different kinds of networks. They have proposed models which, when simulated, at least partially match these properties. A simulated network generated by the Watts-Strogatz model [21] has short average path length and high clustering coefficient. The Barabási-Albert model [2] generates a network with short average path length and power-law degree distribution. The Community Guided Attachment model [13] leads the network to follow a densification power law. The Fire-Forest model [13] generates a network that has heavy-tailed in-degree, high clustering coefficient, and also follows a densification power law.

At the micro level, the node-to-node link formation can be inferred by vertex similarity measures [4, 16]. It has been shown that these measures are good at predicting link formation over the near term but are less effective in predicting over the long term [4, 5]. It is also very interesting that simple local structure based similarity measures, such as cosine similarity or triadic closure based measures, usually better predict future link formation than global structure based measures such as SimRank [4, 11]. The link prediction problem can also be treated as a supervised classification problem in which the labels represent presence or absence of links and the features can be both topological (such as the shortest distance between a pair of nodes or the clustering index) and non-topological (such as the intrinsic properties of the nodes) [1]. Recently, Cukierski showed that using a large number of network structure based features for link prediction is promising [7]. Surveys of the link prediction problem can be found in [10, 17].

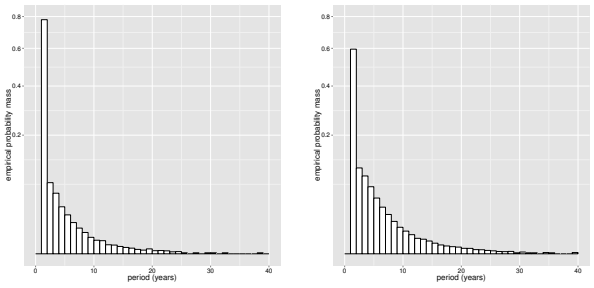
Our work is motivated by supervised link prediction [1] and the recently discovered observation that "younger links seem to be more influential in future link prediction" [5]. However, there are substantial differences between our current work and these publications. Unlike [5] which suggests a relatively ad-hoc aging model, here we apply logistic regression to quantify the relative importance of old links and young links. Different from [1], we specially focus on studying link age instead of general topological features or intrinsic features of nodes.

3. DATASET DESCRIPTION

This section describes the two data domains used in our study. The first network, a coauthorship network among computer scientists, was compiled using the DBLP Computer Science Bibliography[1]. Each node in the coauthorship network represents one author, with two nodes connected if the two authors coauthored at least one paper. At the time we crawled the data, DBLP had collected more than 2 million papers written by more than 1 million authors. Although the author names are disambiguated using coauthoring information [15], incorrect attributions still occur. In this work, we directly used the disambiguated result from DBLP, since author name disambiguation is outside the scope of this study.

The second network, a co-starring network among actors, was generated using the IMDB movie database[2]. Each node in the network represents an actor, and two nodes are connected if they co-star in a movie. The dataset we have is a collection containing movies before and inclusive of year 2007. Nevertheless, it is still a good example of a social network that evolves over time.

Although a variety of network datasets have been collected and shared[3][4], we selected the two networks used in our study because the ages of the links can be inferred from the interaction history between the nodes.



(a) Empirical probability mass function on coauthor periods for 10,000 randomly selected authors and all their coauthors.

(b) Empirical probability mass function on coauthor periods of 10,000 randomly selected authors who published at least 5 papers with coauthors who also published at least 5 papers.

Figure 1: The empirical probability mass function of coauthoring periods (y axis is on a square root scale).

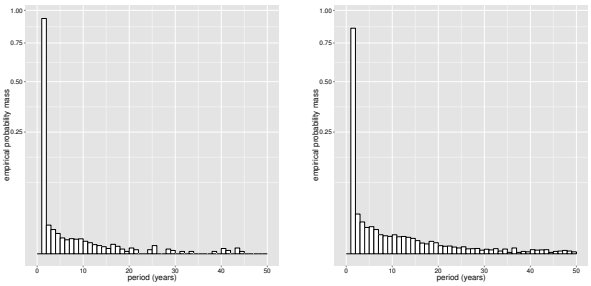
Table 1: Statistical summaries of coauthoring periods in years (Q1, Q2, Q3: the first, second, and third quantiles).

min-pub-num	min	Q1	Q2	mean	Q3	max
1	1	1	1	1.683	1	38
5	1	1	1	2.458	3	40

4. TYPICAL ACTIVE PERIODS OF LINKS

In this section, we study the typical active periods of links. We calculate the active period of a link by the time dif-

[1] http://www.informatik.uni-trier.de/~ley/db/
[2] http://www.imdb.com/
[3] http://snap.stanford.edu/data/
[4] http://www-personal.umich.edu/~mejn/netdata/



(a) Empirical probability mass function on co-starring periods for 10,000 randomly selected actors and all their co-starring actors.

(b) Empirical probability mass function on co-starring periods of 10,000 randomly selected actors who acted in at least 5 films with co-starring actors who also acted in at least 5 films.

Figure 2: The empirical probability mass function of co-starring periods (y axis is on a square root scale).

Table 2: Statistical summaries of co-starring periods in years (Q1, Q2, Q3: the first, second, and third quantiles).

min-movie-num	min	Q1	Q2	mean	Q3	max
1	1	1	1	1.433	1	44
5	1	1	1	2.056	1	49

ference between the first and last interactions between two connected nodes.

To study the typical active periods of links in a coauthorship network, we randomly selected 10,000 nodes as seeds. For each seed node, we compiled the active period with all of the seed's neighbors. The active period of a link is set as 1 plus the difference between the latest and the initial coauthoring years. The empirical probability mass function of active periods of links in the coauthorship network is shown in Figure 1(a). Note that the y axis is on a square root scale so that the small density bars can be seen more clearly. As shown, more than 80% of the coauthoring relationships end within 3 years. In a similar manner, we set the active period of a link in the co-starring network as 1 plus the difference between the latest and earliest release dates of their co-starring movies. The empirical probability mass function is shown in Figure 2(a). The typical active periods for the co-starring network is even shorter: more than 95% of the co-starring behaviors end within 3 years. This shows that most of the links age-out relatively quickly.

One might suspect that such a skewed distribution stems from the large number of authors who published only 1 or 2 papers, and from actors who performed in only 1 or 2 films. To eliminate this confounding factor from the coauthorship network, we randomly selected 10,000 authors who published at least 5 papers and picked all their coauthors who also published at least 5 papers. The associated empirical probability mass function of the coauthoring period is presented in Figure 1(b) and shows that approximately 80% of the collaboration periods are still shorter than 3 years. In a similar manner, we randomly selected 10,000 actors who performed in at least 5 films and picked all their co-starring actors who also participated in at least 5 films; the results are shown in Figure 2(b). Table 1 and Table 2 list statistics of the collaboration periods and the co-starring periods, respectively. Thus, even if we intentionally select nodes that are more actively interacting with others, the typical active periods of links are still short.

5. DIMINISHING INFLUENCE OF LINKS

In this section, we further study how a link gradually, over time, loses its influence in determining which new links will form in the network.

5.1 Predictive Model

Motivated by [1], we express this question as a supervised learning problem, where the labels represent presence or absence of links. We apply a logistic regression classifier as the model predicting new links for two reasons. First, this model can be easily updated to take in new data using an online gradient descent method. Although in our experiments the size of the training dataset will be fixed, in practice the network evolves over time, with newly formed links providing new supervising information, which can be exploited to online adapt the classifier, making it both more accurate and more up-to-date. Second, and most significantly for our purposes here, the predictive power of individual features in logistic regression models can be directly inferred from the magnitude of the learned coefficients. Thus, the influence of young and old links (and their associated features) is easily quantitated. Although applying ensemble supervised learning classifiers such as bagging and boosting [3, 19] (or kernel-based SVMs) may improve link prediction performance [1], it is difficult to infer the relative importance of individual features in these methods. Since our main target is to understand the relative importance of young and old links instead of purely pursuing high prediction accuracy, logistic regression is a better model choice here.

The conditional probability of the logistic regression model is shown in Equation 1:

$$P(y = 1|\theta_0, \mathbf{\Theta}, \mathbf{X}') = \frac{1}{1 + \exp(-(\theta_0 + \mathbf{\Theta}^T \mathbf{X}'))}, \quad (1)$$

where $y = 1$ represents link presence, $\mathbf{X}' = [x'_1, x'_2, \ldots, x'_n]^T$ are the n re-scaled features, and θ_0 and $\mathbf{\Theta} = [\theta_1, \theta_2, \ldots, \theta_n]^T$ are the coefficients to be learned. A feature x_i with observed values $\mathcal{X}_i = \{v_1, v_2, \ldots, v_m\}$ is re-scaled to $\mathcal{X}_{i'}$ by Equation 2, with the range of the new values $[0, 1]$.

$$\mathcal{X}_{i'} = \left\{ \frac{v_1 - \min(\mathcal{X}_i)}{\max(\mathcal{X}_i) - \min(\mathcal{X}_i)}, \frac{v_2 - \min(\mathcal{X}_i)}{\max(\mathcal{X}_i) - \min(\mathcal{X}_i)}, \ldots, \frac{v_m - \min(\mathcal{X}_i)}{\max(\mathcal{X}_i) - \min(\mathcal{X}_i)} \right\} \quad (2)$$

By this re-scaling, quantitative evaluation of the predictive power of individual features simply reduces to evaluation of the magnitudes of the learned coefficients in $\mathbf{\Theta}$. Specifically, letting *odds* represents the ratio of the probability that the edge exists to the probability that the edge does not exist, the learned coefficients represent the change in the log odds of edge existence for a unit change in the feature. As a result, the relative influential power of two

Table 3: A list of networks used for generating training features and labels.

	year	purpose
G_1	2000-2002	generate training features for DBLP
G_2	2003	generate training labels for DBLP
H_1	1997-1999	generate training features for IMDB
H_2	2000	generate training labels for IMDB

features x_i and x_j is estimated by the ratio of the exponential of the learned coefficients $\exp(\theta_i) : \exp(\theta_j)$.

5.2 Link Feature Derivation

In this section, we explain in detail the creation of the training network, testing network, and the selected features for the new link prediction problem in a coauthorship network. The information required for co-starring network link prediction is created through a similar process, which will be briefly introduced at the end of the section.

To obtain the training and testing features of the coauthorship network, we selected from the DBLP dataset publications over three consecutive years y, $y+1$, and $y+2$ in the following seven conferences: ICDE, ICML, KDD, SIGIR, SIGMOD, VLDB, and WWW. The authors of these papers and all their coauthors are compiled to create a coauthorship network G_1. We limited to three consecutive years to generate the training features because 80% of the coauthoring periods are no longer than 3 years.

Specifically, six features are generated for each pair of nodes i and j in G_1 by considering both local topology structure and the ages of the links. The six features $x_1, x_2, \ldots x_6$ are listed below:

$$x_1 = \sum_{\forall k \in MutualNeighbor(i,j)} Weight(i,k,t=y), \quad (3)$$

$$x_2 = \sum_{\forall k \in MutualNeighbor(i,j)} Weight(i,k,t=y+1), \quad (4)$$

$$x_3 = \sum_{\forall k \in MutualNeighbor(i,j)} Weight(i,k,t=y+2), \quad (5)$$

$$x_4 = \sum_{\forall k \in MutualNeighbor(i,j)} Weight(j,k,t=y), \quad (6)$$

$$x_5 = \sum_{\forall k \in MutualNeighbor(i,j)} Weight(j,k,t=y+1), \quad (7)$$

$$x_6 = \sum_{\forall k \in MutualNeighbor(i,j)} Weight(j,k,t=y+2), \quad (8)$$

where the $MutualNeighbor(i,j)$ function returns the set of mutual neighbors (coauthors) of i and j and $Weight(i,k,t)$ returns the number of coauthored papers between i and k in year t.

Note that we do not include the full coauthoring history of i and j because we are interested in how link age influences the formation of "new" links, i.e., links that are not present in G_1.

We only pick the above six features because they are critical to our main research question: what is the relationship between edges' ages and the formation of new links. While it is shown that performance of link prediction can be improved by introducing a large number of features [7], most of these features have nothing to do with links' ages. Introducing many features may improve precision but will make it difficult to isolate the effect of link age.

To get the training labels, we create a coauthorship network G_2 for year $y+3$. Note that the authors who appear in year $y+3$ but not in the period $[y, y+2]$ are disregarded since their features cannot be obtained from G_1. By setting y to year 2000, the network G_2 (representing the coauthorship network of the year 2003) contains 460 nodes and 610 edges, among which 245 of these edges have already appeared in G_1 (representing the coauthorship network for the period $[2000, 2002]$). When training a model to predict the formation of "new" edges, these 245 edges should be disregarded, i.e., they are neither positive nor negative instances. Thus, the training examples contain $610 - 245 = 365$ positive instances and $\binom{460}{2} - 610 = 104,960$ negative instances. A logistic regression classifier C_1 is trained by using the labels derived from G_2 and the features derived from G_1.

In addition, we create three other logistic regression classifiers, using different sets of features, for comparison. The classifiers C_2, C_3, and C_4 predict coauthoring events in year $y+3$ using only the features derived from the coauthor network in year $y+2$, $y+1$, and y respectively, i.e., C_2 uses features x_3 and x_6, C_3 uses features x_2 and x_5, and C_4 uses features x_1 and x_4. The learned classifiers will be compared with C_1.

We select other consecutive 4-year periods to generate the testing data. The first three years are used to generate the testing features, with the fourth year used to obtain the testing labels. The testing features are fed to C_1, C_2, C_3, and C_4 for link prediction.

The training data, testing data, and features for the co-starring network are generated in a similar manner. First, we select $1,000$ actors of the highest degrees between 1991 and 2007 as set A. Next, we create a co-starring network H_1 by the co-starring behavior among actors in A between 1997 and 1999. Six features, as listed from Equation 3 to Equation 8, are generated for each pair of nodes in H_1, where $Weight(i,k,t)$ returns the number of co-starring movies between i and k in year t. Again, we use only the co-starring network of three consecutive years because more than 90% of co-starring behaviors are active for less than 3 years. The co-starring behaviors among actors in A in the year 2000 are used to generate the training labels.

Table 3 shows networks used for generating the training features and labels for the coauthorship network and the co-starring network.

5.3 Experimental Results

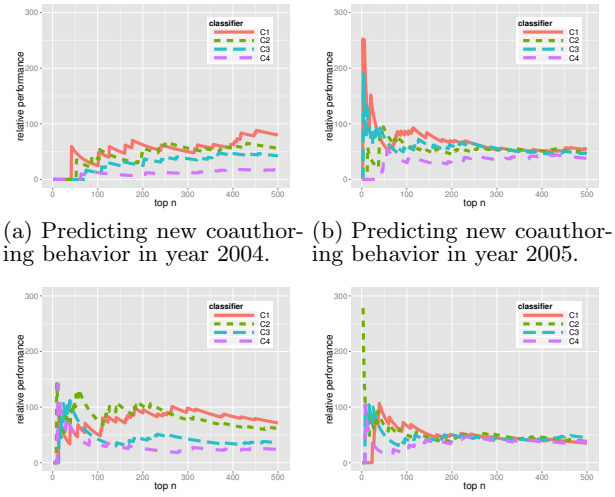
The learned coefficients of the classifier C_1 on the coauthorship network are listed in Table 4. As shown, the later coauthoring behaviors (essentially, the younger links) play a more important predictive role than the earlier coauthoring events (the older links). We can roughly think of the links of the 1st, 2nd, and 3rd years as old, mid-age, and young links. From the learned coefficients, the average influential power of the old, mid-age, and young links to the odds of new link formation are $\exp(0.7540) : \exp(3.0010) : \exp(4.0239)$. Thus, the young links are almost 3 times more informative than the mid-age links and 26 times more informative than

Table 4: Coefficients of the learned logistic regression classifier when using the coauthorship network over the years 2000-2002 to generate training features and the coauthorship network of 2003 to generate the training labels.

	Year 2000	Year 2001	Year 2002
i-k	$\theta_1 = 0.5181$	$\theta_2 = 2.6930$	$\theta_3 = 3.4582$
j-k	$\theta_4 = 0.9899$	$\theta_5 = 3.3089$	$\theta_6 = 4.5895$
Average	0.7540	3.0010	4.0239

Table 5: Coefficients of the learned logistic regression classifier when using the costar network over the years 1997-1999 to generate training features and the coauthorship network of 2000 to generate the training labels.

	Year 1997	Year 1998	Year 1999
i-k	$\theta_1 = 0.1193$	$\theta_2 = 1.4216$	$\theta_3 = 6.8277$
j-k	$\theta_4 = 0.3026$	$\theta_5 = 0.8970$	$\theta_6 = 6.3317$
Average	0.2110	1.1593	6.5797



(a) Predicting new coauthoring behavior in year 2004.

(b) Predicting new coauthoring behavior in year 2005.

(c) Predicting new coauthoring behavior in year 2006.

(d) Predicting new coauthoring behavior in year 2007.

Figure 3: Comparison of the relative performance at n of different classifiers for coauthorship network.

the old links. The influential power of young links in co-starring network is even more prominent: the young links are 226 times more informative than mid-age links and 583 times more informative than the old links, as shown in Table 5. Since typical active periods of links in the co-starring network are usually shorter than links in the coauthorship network, as shown in Figure 1 and Figure 2, it is not surprising that the young links in the co-starring network are more influential than the young links in the coauthorship network.

Next, we show the performance of each classifier on predicting future active links. Because the two classes (link presence and absence) are highly imbalanced, successfully predicting a positive instance is very challenging. For example, when predicting new links in the coauthorship network for year 2004 using features derived from the period $[2001, 2003]$, the true positive rate of a naïve random guess is only 0.2406629%. To make the metric more meaningful, we show the "relative performance at n" for each classifier. This is defined as the true positive rate over the first n predictions for a classifier divided by the true positive rate from naïve random guessing.

We used each classifier to predict new links of the coauthorship network in years 2004, 2005, 2006, and 2007. The accuracy measure being compared is the "relative performance at n", with n ranging from 1 to 500, since different values of n affect the performance. As shown in Figure 3, the classifiers that consider the younger links (C_1 and C_2) generally perform better than the one considering only the older links (C_3 and C_4). Figure 4 shows the relative performance at n for predicting new links of co-starring network in year 2001, 2002, 2003, and 2004. Again, the classifiers considering younger links generally perform better.

6. DISCUSSION AND FUTURE WORK

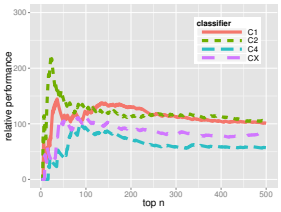
A study on Facebook, the currently most dominant social medium, shows that the average U.S. user was friends with around 214 other U.S. users in May 2011 [20]. This number is almost 1.5 times larger than the famous Dunbar's Number, which suggested that the cognitive limit to the number of individuals with whom any one person can maintain stable relationships is on average 147.8 [8]. This result implies that, on average, approximately one third of the relationships in an ego-network of a Facebook user are unstable. Although this does not necessarily mean the old links are more fragile, it does suggest that a large portion of links might gradually become inactive and, to stabilize the network, these links could be removed.

To study the influential power of old links and young links, we separated links by their age, and studied how the age of a link influences the formation of new links. Using one coauthorship network and one co-starring network as domains, we found that younger links are more informative in predicting the formation of new links. As far as we know, this is the first paper to quantitate the relative importance of links of different ages.

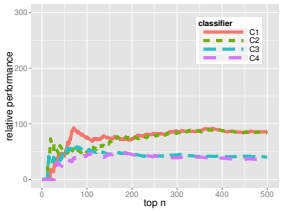
Since as links become older their predictive power generally diminishes, it might be most appropriate to disregard or remove "sufficiently old links". This leads to the following research questions that are rarely discussed in previous studies.

First, network evolution theory should consider not only link/node addition, but also link/node removal as well. Although we quantify the relative importance of young, mid-age, and old links, how to appropriately age-out old links (or even old nodes) is still an open question. We are interested in designing a network generation model to address this problem.

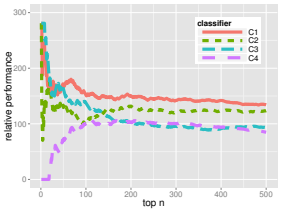
Second, we are interested in observing the statistical properties of networks and network evolutions when links/nodes do age-out. As a starting point, we could assume the age of each link to follow the distribution of observed coauthoring periods and examine both whether and how these statistical



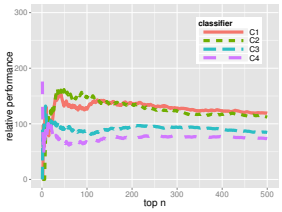
(a) Predicting new co-starring behavior in year 2001.



(b) Predicting new co-starring behavior in year 2002.



(c) Predicting new co-starring behavior in year 2003.



(d) Predicting new co-starring behavior in year 2004.

Figure 4: Comparison of the relative performance at n of different classifiers for co-starring network.

properties change when these links are removed.

Third, applications based on network structures [9, 12, 14, 18] may be influenced as well. For example, one popular research direction of social network based viral marketing is to identify the key influencers to advertise to, with the hope that these influencers can disseminate the information to a larger group of individuals. The top influencers are usually estimated via degree-based measures [6, 12, 14] without considering that links could become inactive, i.e., the interaction between the targeted influencers and others may break. Researchers may need to work on more generalized models to capture the effect of inactive links.

One caveat on our results is that, while old links appear in general to be less influential, the aging speed can be link or node dependent. In particular, some old links may still be powerful determinants. For example, consider researchers A and B mentored by the same advisor, many years apart. They may start collaborating at some point in the future, mainly because they are former students of the same advisor.

We believe one reason that previous studies pay less attention to link/node removal when studying network evolution is because such events are difficult to directly observe or to definitively ascertain (e.g., long periods of link inactivity do not preclude future activity). As a pioneering approach, this paper uses the coauthoring or co-starring period to obtain a proxy for the ages of the links. On the other hand, we are certainly desirous to obtain networks that do have explicit link or node removal behavior, as such networks could be used to further validate our work and its main observations.

References

[1] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *Workshop on Link Analysis, Counter-terrorism and Security*, 2006.

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

[3] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[4] H. Chen, L. Gou, X. Zhang, and C. Giles. Discovering missing links in networks using vertex similarity measures. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*. ACM, 2012.

[5] H. Chen, L. Gou, X. Zhang, and C. Giles. Predicting recent links in FOAF networks. *Social Computing, Behavioral-Cultural Modeling and Prediction*, 2012.

[6] H.-H. Chen, Y.-B. Ciou, and S.-D. Lin. Information propagation game: a tool to acquire humanplaying data for multiplayer influence maximization on social networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1524–1527. ACM, 2012.

[7] W. Cukierski, B. Hamner, and B. Yang. Graph-based features for supervised link prediction. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1237–1244. IEEE, 2011.

[8] R. Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16(4), 1993.

[9] W. Fan, J. Li, J. Luo, Z. Tan, X. Wang, and Y. Wu. Incremental graph pattern matching. In *Proceedings of the 2011 International Conference on Management of Data, SIGMOD*, volume 11, pages 925–936, 2011.

[10] M. Hasan and M. Zaki. A survey of link prediction in social networks. *Social Network Data Analytics*, pages 243–275, 2011.

[11] G. Jeh and J. Widom. SimRank: a measure of structural-context similarity. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining*. ACM, 2002.

[12] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[13] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining*. ACM, 2005.

[14] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429. ACM, 2007.

[15] M. Ley and P. Reuther. Maintaining an online bibliographical database: The problem of data quality. *Extraction et gestion des connaissances (EGC'2006), Actes des sixiémes journées Extraction et Gestion des Connaissances, Lille, France*, pages 17–20, 2006.

[16] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 2007.

[17] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

[18] D. O'Doherty, S. Jouili, and P. Van Roy. Towards trust inference from bipartite social networks. In *Proceedings of the 2nd ACM SIGMOD Workshop on Databases and Social Networks*, pages 13–18. ACM, 2012.

[19] R. Schapire. The boosting approach to machine learning: an overview. 2001.

[20] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the Facebook social graph. *arXiv Preprint arXiv:1111.4503*, 2011.

[21] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684), June 1998.