Exploiting Common Neighbor Graph for Link Prediction

Hao Tian
Data Lab, EECS Department
Syracuse University
haotian@data.syr.edu

Reza Zafarani
Data Lab, EECS Department
Syracuse University
reza@data.syr.edu

ABSTRACT

Link prediction aims to predict whether two nodes in a network are likely to get connected. Motivated by its applications, e.g., in friend or product recommendation, link prediction has been extensively studied over the years. Most link prediction methods are designed based on specific assumptions that may or may not hold in different networks, leading to link prediction methods that are not generalizable. Here, we address this problem by proposing general link prediction methods that can capture network-specific patterns. Most link prediction methods rely on computing similarities between between nodes. By learning a γ-decaying model, the proposed methods can measure the pairwise similarities between nodes more accurately, even when only using common neighbor information, which is often used by current techniques.

CCS CONCEPTS

• **Mathematics of computing → Graph algorithms**; • **Information systems → Social networks**.

KEYWORDS

Common Neighbors, Common Neighbor Graph, Link Prediction

ACM Reference Format:
Hao Tian and Reza Zafarani. 2020. Exploiting Common Neighbor Graph for Link Prediction. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3417464

1 INTRODUCTION

Networks have become widespread and are growing in size across various disciplines. Network-based problems have been extensively studied in recent years with examples such as representation learning [26], community detection [8], visualization [5], and the like. Akin to how new friendships are formed between people in real world, many networks evolve over time with new edges appearing, motivating researchers to predict future links [16]. Link prediction methods aim to solve this problem, and are widely used in social networks [15], biological networks [2], and dynamic networks such as e-mail or communication networks [19].

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3417464

Most link prediction methods have shown remarkable performance in social networks research. Justified by theoretical findings in social sciences and extensive empirical studies, researchers have developed many *similarity-based link prediction* methods that can measure how likely a link will appear between two nodes given their similarity. This abstract similarity can be as simple as the number of common neighbors between two nodes [15], or the *Jaccard similarity* [10] between their *one-hop neighborhoods* (i.e., friends). Such similarity-based link prediction methods are widely used in recommender systems for social/product networks [22], or when seeking collaborators in collaboration networks [21].

However, new links appear for various reasons within graphs, and some reasons are indeed exogenous to the graphs. Even within a specific social network, it is difficult to determine the importance of different factors (i.e., similarity measures) such as the number of common neighbors, degrees, path lengths, and the like on the formation of new links. Hence, one often faces the elusive problem of selecting the "best" similarity measure for link prediction methods, which is often addressed by testing various similarity measures. One solution to this problem is to build a supervised combination of various similarity measures. An alternative solution is to directly learn the link formation patterns from the network. For instance, Zhang and Chen [27] propose a neural network link prediction method, which operates by encoding various neighborhoods in the graph as inputs to train a neural network. Due to higher computation cost, both solutions cannot completely replace similarity-based methods, which are fast, when dealing with large-scale social networks.

The Present Work. We introduce links prediction methods that are fast, but also generalize across graphs. To the best of our knowledge, we introduce the first of such link prediction methods. Our work is inspired by the recent study of Dong et al. [7], which studies the impact of the common neighbors on link existence across multiple networks. Instead of designing a new similarity measure, we explore the reasons behind the success of past similarity measures. In particular, recently Zhang and Chen [27] proposed a γ-decaying model that can explain how the overlap between the k-hop neighborhoods of two nodes becomes exponentially less important in forming a link between them as k increases. This result shows that the common neighbors of two nodes (i.e., $k = 1$) are the most important when designing link prediction similarity measures, which explains the success of many link prediction similarity measures. This paper takes this result even further and shows that this decaying phenomenon not only exists with respect to the nodes shared within the k-hop neighborhoods, but also with respect to how these nodes are connected. Most importantly, we show that *the way common neighbors of two nodes are connected has a major impact on the formation of a new link between them.* Dong et al. [7] had similar findings. Overall, we make the following contributions:

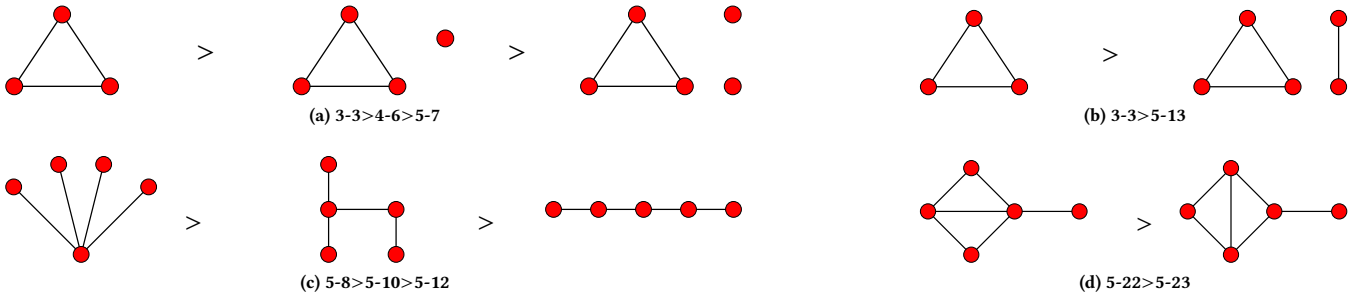

Figure 1: Comparing Link Formation Probabilities (> symbol) for Different Common Neighbor Graphs (the first number n is the number of nodes and the second is the graph index (see footnote) among graphs with n nodes). Here, only common neighbors are shown and the two nodes who share these common neighbors (are connected to all of them) are not shown.

(1) For the first time, we quantify the relation between how common neighbors are connected and link formation probabilities;
(2) We explain such relationships by a γ-decaying model;
(3) We incorporate our findings into two (one supervised and one unsupervised) new link prediction methods; and
(4) We evaluate the developed methods and show that they can significantly outperform current link prediction measures.

2 RELATED WORK

Among link prediction methods, similarity-based methods have a relatively longer history and are widely used in different applications. These methods measure the chance of two nodes getting linked based on their 'similarity', which can be measured using network information. Similarity-based method (or indices) can be further divided into local and global methods.

Local methods measure the similarity of two nodes based on their neighborhood information, usually one-hop (friends) or two-hop (friends-of-friends) neighborhoods. The most basic and well-studied measure is the *Common Neighbor* index, which is basically the number of neighbors shared among two nodes. The measure is also used in many other measures. In their classical paper, Liben-Nowell and Kleinberg [15] show that the common neighbor index performs surprisingly well for link prediction in social networks. The *Adamic-Adar Index* [3] is another well-established local similarity measure, which penalizes common neighbors by their degrees. Besides these, there are many other local similarity-based methods developed for various contexts and objectives. To name a few, *Resource Allocation Index* [28] models connections as the resource transmitted from one node to another through common neighbors, and *Preferential Attachment Index* [4] assumes link formation between two nodes relies on their probability of getting connected in scale-free networks. Classical similarity measures can also be used as local methods for link prediction, e.g., *Jaccard Index* [10] and *Mutual Information* [23].

On the contrary, global methods use the topology information of the whole network to score each potential link, often by using information on paths. *Katz Index* [12] and *Global Leicht-Holme-Newman Index* [14] count a weighted total of all possible paths between two nodes. Random walk based methods (e.g., *PageRank* [20] and *SimRank* [11]) measure similarity using random walk visiting probabilities starting from pairs of nodes. While global methods have performed quite well on link prediction, the complexity of exploring the whole graph hinders their use on large networks. As nicely pointed out by Bliss et al. [6], all link prediction similarity measures are heuristics based on some assumptions, where each similarity may have some level of impact on the formation of new links.

Other than similarity scores, which are unsupervised, there are also supervised link prediction methods. By formulating link prediction as a binary classification problem, traditional classifiers such as Bayesian classifiers or deep learning methods can be used to train a link prediction model, where features are graph properties [9].

3 INFLUENCE OF COMMON NEIGHBOR GRAPH ON FORMATION OF NEW LINKS

Common neighbors play an important role in most heuristics used for predicting links in social network. These heuristics either directly use the number of common neighbors (e.g. *Common Neighbor* [15] and *Adamic-Adar* [3]), or are indirectly influenced by the common neighbors (e.g. *Katz* [12] and *SimRank* [11]). Through extensive experiments on various social networks (we skip the details for brevity), we discover that common neighbors contain much more information than their sole counts. In particular, we demonstrate how the induced subgraph formed by common neighbors of two nodes, which we denote as the *Common Neighbor Graph* influences new link formations. We present our findings using one of our timestamped datasets, *Facebook-links* [1, 25] as an example, where we analyze the impact of the structure of the common neighbor graph on probabilities of new links appearing. We observe similar results in other datasets.

Figure 1 presents some typical observations[1] of new links appearing in a common neighbor graph, which we summarize as

(1) Increase in the number of common neighbors (the size of the common neighbor graph) in general increases the probability of new links forming between two unconnected nodes;
(2) Isolated components in common neighbor graph have a negative effect on link formation (Figure 1 a, b); and
(3) High eccentricity (maximum distance of one node to others) also lowers probabilities of forming new links (Figure 1 c, d).

These insights confirm and extend those given by Dong et al. [7].

[1]Graph index is available at https://www.graphclasses.org/smallgraphs.html

4 PROPOSED METHODS

In this section, we use the observations in Section 3 and combine them with the foundation of current link prediction methods to design novel and generalizable link prediction techniques. While current link prediction techniques are many, recent findings have shown that current link prediction heuristics can often be unified in terms of an influence decaying model controlled by some decay parameter [27]. We briefly review this influence decay model first. Next, we propose a conceptual overview of our link prediction approach and present two link prediction techniques: (1) a simple unsupervised approach which involves influence decay, is fast, but not tunable for different networks and (2) a supervised model, whose parameters can be properly trained for different networks.

4.1 The Influence Decaying Model

With various empirical studies demonstrating the success of link prediction methods, there has been some recent attempts in terms of identifying theories that can explain how these methods capture the nature of social networks. One general explanation is given by the γ-decaying heuristic. The heuristic states that (1) most current similarity measures can be unified in terms of an influence function decaying on the orders of neighborhood expansions, i.e., 1-hop, 2-hop, ..., n-hop neighborhoods, and (2) high-order link prediction heuristics, which use neighborhood information on more number of hops, can be accurately approximated using small orders of neighborhood information. A γ-decaying heuristic H for nodes (x, y) is defined as

$$H(x, y) = \eta \sum_{l=1}^{\infty} \gamma^l f(x, y, l), \tag{1}$$

where l is the heuristic order, f is a non-negative function of x, y, l, value γ is a positive decay factor, and η is a positive constant or a positive function of γ.

A γ-decaying heuristic could model many common link prediction similarity measures such as *Common Neighbor [15], Katz [12], PageRank [24]*, and the like (see [27] for details). This influence decaying function has been used to model the impact of neighborhood expansions on link candidates. However, as shown by our observations in Section 3, there exists another decaying pattern just within a one-hop neighborhood (common neighbors), where there is a decay in link formation probabilities as distances between common neighbors increase. From a computational perspective, this observation is much more useful as computing various measures for common neighbors tends to be fast, even for large graphs.

A Conceptual Framework for Link Prediction. We propose a general conceptual model to incorporate our discussion and exploit common neighbor graph information for link prediction. We posit that the link formation probability between two unlinked nodes can be defined in terms of two component: (1) an *individual* component, which models the impact that different nodes in the graph can have on connecting two unlinked nodes. Past link prediction measures, such as the number of common neighbors, can be considered an example of this component, and (2) a *community* component, which measures communities potentially formed as a result of this new link formation, modeling our observations in Section 3. Hence, a similarity measure to model this conceptual framework will be

$$Sim(u, v) = \text{individual component} + \text{community component}. \tag{2}$$

Using this conceptual model, we propose two link prediction techniques, where one is unsupervised and simple, and the other is supervised; hence, requires training, but is more accurate.

4.2 An Unsupervised Approach

To introduce an unsupervised approach, we basically define a simple similarity measure that captures the decaying phenomenon and our empirical observations. Same as the *Common Neighbor* [15] measure, which only has an individual component that measures the number of common neighbors, our measure only involves a community component capturing connections between common neighbors. Our measure is inspired by closeness centrality:

Definition 4.1. For node i, its closeness centrality $C(i)$ is $C(i) = \sum_j \frac{1}{d(i,j)}$, where $d(i, j)$ is the distance from node i to node j.

To design our measure, we add a decaying function to pairwise geodesic distances of common neighbors,

$$Sim(u, v) = \begin{cases} \frac{\sum_{i,j \in |CN(u,v)|}^{i \neq j} 1/d(i,j)}{|CN(u,v)|-1}, & \text{if } |CN(u, v)| > 1; \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $CN(u, v)$ denotes the common neighbors of u and v, and $d(i, j)$ is the distance between i and j. As the decaying factor cannot be learned from actual graphs in an unsupervised fashion, we simply select $1/x$ as the decay function f in Eq. (1) (similar to closeness centrality), while any other decay functions could also be used. We normalize this similarity by dividing it by the number of common neighbors, which ensures its scale is at the same level (at most 1/2) of the number of common neighbors. Note that n common neighbors have $\binom{n}{2} = n(n-1)/2$ pairwise distances. So, for each common neighbor, we count its community influence as the average distance from it to other $(|CN(u, v)| - 1)$ common neighbors.

We ignore the individual component when designing the unsupervised similarity measure for the following two reasons: (1) it is difficult to balance the relationship between the individual component and community component without looking at the graphs; and (2) we have already included the individual component to some extent by scaling the measure to be within the number of common neighbors scale. Our goal was to ensure the measure is as simple as possible, yet capturing the community aspect of link prediction. Our performance results in Section 5 show that this simple measure significantly outperforms current similarity measures.

4.3 A Supervised Approach

With the understanding that different graphs have various traits, we define a complete adaptive link prediction model, defined as:

$$Sim(u, v) = \begin{cases} \alpha \frac{|CN(u,v)|}{2} + (1-\alpha) \frac{\sum_{i,j \in |CN(u,v)|}^{i \neq j} \gamma^{d(i,j)}}{|CN(u,v)|-1}, & \text{if } |CN(u, v)| > 1; \\ \alpha \frac{|CN(u,v)|}{2} = \alpha/2, & \text{otherwise,} \end{cases} \tag{4}$$

where $\alpha \in [0, 1]$ is a weight that balances the contributions of individual/community components, and $\gamma \in (0, 1)$ is a decay parameter. These two parameters have to be learned from training data. A smaller α indicates higher community influence, and when u and v have only one common neighbor, we only have individual

Graphs	CN	AA	RA	J	Katz	Proposed Methods	
						Unsupervised	Supervised
Facebook-links [1]	0.5713	0.5928	0.5657	0.5460	0.5691	**0.6594**	**0.6594**
Bitcoin-Alpha [13]	0.5862	0.6017	0.5869	0.3802	0.6515	**0.6640**	0.6587
Flickr [18]	0.5621	0.5737	0.5848	0.5223	0.5507	**0.5964**	0.5698
Youtube [17]	0.6665	0.6738	0.6608	0.4781	0.6899	**0.7084**	0.7083

Table 1: The performance of the proposed methods (AUC values) compared to common techniques. The proposed techniques significantly improve other measures.

influence. We can prove that $\alpha = 0.5$ ensures equal contribution when number of common neighbors is greater than one.

Proof. When individual influence and maximum community influence (a complete common neighbor graph: $d(i,j) = 1$) is equal:

$$\alpha \frac{|CN(u,v)|}{2} = \max_{d(i,j),\gamma} (1-\alpha) \frac{\sum_{i,j \in |CN(u,v)|}^{i \neq j} \gamma^{d(i,j)}}{|CN(u,v)| - 1}, \quad (5)$$

which can be solved as $\alpha = 0.5$. □

5 CASE STUDY: LINK PREDICTION

We evaluate the performance of the proposed methods on multiple time-stamped real-world social networks (friendship networks or trust networks), compared with various well-known link prediction measures: Common Neighbor (CN), Adamic-Adar (AA), Resource Allocation (RA), Jaccard (J), and Katz. Calculating pairwise similarities for the whole graph can be extremely time consuming, especially for global method like Katz. When link prediction is performed on graph $G = (V, E)$, we have $\binom{|V|}{2} - |E|$ potential link candidates. To save time, we first split the edges in order of timestamps to 90% for training and 10% as testing. Then, we include all edges in test set as positive class, while selecting same amount of missing edges as negative class.[2]

If we select negative class at random, the performances of most methods are similar since majority of negative candidate pairs have no similarity with each other. To test the ability of measuring similarities of 'similar' pairs, we select negative class by the criteria of 'at least one common neighbor'. we learn parameters α and γ by grid search. The results are in Table 1, clearly showing that proposed methods outperform all others.

6 CONCLUSION AND FUTURE WORKS

In this paper, for the first time, we quantify the influence of structural information of common neighbors on formation of new links. We find that number of common neighbors, number of isolated components, and distances between nodes within common neighbor graph impact link formation probabilities. We propose two new link prediction measures (supervised and unsupervised) that incorporate these observations. Our experimental results on real-world data show that the developed methods can outperform existing techniques, while being simple to implement.

REFERENCES

[1] 2017. Facebook friendships network dataset – KONECT. http://konect.uni-koblenz.de/networks/facebook-wosn-links

[2] 2018. Link Prediction Potentials for Biological Networks. *Int. J. Data Min. Bioinformatics* 20, 2 (Jan. 2018), 161–184. https://doi.org/10.1504/IJDMB.2018.093684

[3] Eytan Adamic and Lada A. Adar. 2003. Friends and neighbors on the web. 3 (July 2003), 211–230.

[4] Albert-László Barabási and Réka Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286, 5439 (1999), 509–512. https://doi.org/10.1126/science.286.5439.509 arXiv:https://science.sciencemag.org/content/286/5439/509.full.pdf

[5] Fabian Beck, Michael Burch, Stephan Diehl, and Daniel Weiskopf. 2017. A Taxonomy and Survey of Dynamic Graph Visualization. *Computer Graphics Forum* 36, 1 (2017), 133–159. https://doi.org/10.1111/cgf.12791 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.12791

[6] Catherine A. Bliss, Morgan R. Frank, Christopher M. Danforth, and Peter Sheridan Dodds. 2013. An Evolutionary Algorithm Approach to Link Prediction in Dynamic Social Networks. *arXiv e-prints*, Article arXiv:1304.6257 (April 2013), arXiv:1304.6257 pages. arXiv:physics.soc-ph/1304.6257

[7] Yuxiao Dong, Reid A. Johnson, Jian Xu, and Nitesh V. Chawla. 2017. Structural Diversity and Homophily: A Study Across More Than One Hundred Big Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 807–816. https://doi.org/10.1145/3097983.3098116

[8] Steve Harenberg, Gonzalo Bello, L. Gjeltema, Stephen Ranshous, Jitendra Harlalka, Ramona Seay, Kanchana Padmanabhan, and Nagiza Samatova. 2014. Community detection in large-scale networks: a survey and empirical evaluation. *WIREs Computational Statistics* 6, 6 (2014), 426–439. https://doi.org/10.1002/wics.1319 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1319

[9] Mohammad Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. 2006. Link Prediction Using Supervised Learning. (01 2006).

[10] Paul Jaccard. 1901. Etude de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles* 37 (01 1901), 547–579. https://doi.org/10.5169/seals-266450

[11] Glen Jeh and Jennifer Widom. 2002. SimRank: A Measure of Structural-Context Similarity. In *In KDD*. 538–543.

[12] Leo Katz. 1953. A new status index derived from sociometric analysis. *Psychometrika* 18, 1 (01 March 1953), 39–43. https://doi.org/10.1007/BF02289026

[13] Srijan Kumar, Francesca Spezzano, VS Subrahmanian, and Christos Faloutsos. 2016. Edge weight prediction in weighted signed networks. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 221–230.

[14] E. A. Leicht, Petter Holme, and M. E. J. Newman. 2006. Vertex similarity in networks. *Physical Review E* 73, 2 (Feb 2006). https://doi.org/10.1103/physreve.73.026120

[15] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology* 58, 7 (2007), 1019–1031. https://doi.org/10.1002/asi.20591 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20591

[16] Víctor Martínez, Fernando Berzal, and Juan-Carlos Cubero. 2016. A Survey of Link Prediction in Complex Networks. *ACM Comput. Surv.* 49, 4, Article Article 69 (Dec. 2016), 33 pages. https://doi.org/10.1145/3012704

[17] Alan Mislove. 2009. *Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems*. Ph.D. Dissertation. Rice University, Department of Computer Science.

[18] Alan Mislove, Hema Swetha Koppula, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. 2008. Growth of the Flickr Social Network. In *Proceedings of the 1st ACM SIGCOMM Workshop on Social Networks (WOSN'08)*.

[19] Joshua O'Madadhain, Jon Hutchins, and Padhraic Smyth. 2005. Prediction and Ranking Algorithms for Event-Based Network Data. *SIGKDD Explor. Newsl.* 7, 2 (Dec. 2005), 23–30. https://doi.org/10.1145/1117454.1117458

[20] L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The PageRank citation ranking: Bringing order to the Web. In *Proceedings of the 7th International World Wide Web Conference*. Brisbane, Australia, 161–172.

[21] Milen Pavlov and Ryutaro Ichise. 2007. Finding Experts by Link Prediction in Co-Authorship Networks. In *Proceedings of the 2nd International Conference on Finding Experts on the Web with Semantics - Volume 290 (FEWS'07)*. CEUR-WS.org, Aachen, DEU, 42–55.

[22] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. *Collaborative Filtering Recommender Systems*. Springer-Verlag, Berlin, Heidelberg, 291–324.

[23] Fei Tan, Yongxiang Xia, and Boyao Zhu. 2014. Link Prediction in Complex Networks: A Mutual Information Perspective. *CoRR* abs/1405.4341 (2014). arXiv:1405.4341 http://arxiv.org/abs/1405.4341

[24] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. 2006. Fast Random Walk with Restart and Its Applications. *Sixth International Conference on Data Mining (ICDM'06)* (2006), 613–622.

[25] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. 2009. On the Evolution of User Interaction in Facebook. In *Proc. Workshop on Online Social Networks*. 37–42.

[26] Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. 2018. Network Representation Learning: A Survey. *CoRR* abs/1801.05852 (2018). arXiv:1801.05852 http://arxiv.org/abs/1801.05852

[27] Muhan Zhang and Yixin Chen. 2018. Link Prediction Based on Graph Neural Networks. *CoRR* abs/1802.09691 (2018). arXiv:1802.09691 http://arxiv.org/abs/1802.09691

[28] Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. 2009. Predicting missing links via local information. *The European Physical Journal B* 71 (2009), 623–630.

[2] All codes can be found at https://github.com/ttt78952/cn_structure