Link Prediction in Weighted Networks: Analyzing heterogeneous data in Foursquare Social Network

Francisco Izaguirre
fizaguir@stanford.edu
Thawsitt Naing
thawsitt@cs.stanford.edu

Stanford University

Abstract. Link Prediction is a classic network problem in which we predict future links in the network using the current state of the network. In our reaction paper, we review and critique important research papers on link prediction. Then, we propose a project to predict links in Foursquare Social Network using a variety of novel approaches.

Keywords: Link Prediction · Social Networks · Weighted Networks.

1 Reaction Paper

1.1 Link Prediction Problem for Social Networks

Summary: In this paper, Liben-Nowell and Kleinberg [1] formalizes the problem of link prediction. Given a social network $G = \langle V, E \rangle$ in which each edge $e = \langle u, v \rangle \in E$ represents an interaction between u and v, we want to accurately predict future edges based on the current state of the network. The authors perform different link prediction algorithms on co-authorship networks. First, the network is divided into training and testing sub-graphs. All link prediction methods in this paper are score-based: all pairs of node $\langle u, v \rangle$ are assigned a score(x, y) based on input graph G. Then, the scores are sorted in descending order such that pairs with higher scores are more likely to be connected than those with lower scores. Each method is evaluated based on how much it outperforms baseline–a random predictor which simply randomly selects two nodes that are not connected in the training sub-graph.

Critique: This paper provides a good theoretical foundation of link prediction. It surveys different types of similarity-based link prediction algorithms which will be a good starting point for our project. However, the paper is limited in the type of dataset it uses because (1) it only considers academic collaboration networks and does not mention how to generalize it to other types of social networks and (2) it does not consider additional information of the network such as edge weights or different node types. In our project, we plan to consider a

network with different types of nodes as well as compare the results on both unweighted and weighted networks. The paper discusses 2 directions to explore. (1) There is much room for improvement in link prediction as even the best method considered in this paper is correct on only about 16% of its predictions. The paper challenges the readers to find ways to better take advantage of the information in the training data. (2) Another path is to improve the efficiency of proximity-based algorithms on very large networks. The authors hint that fast algorithms for approximating the distribution of node-to-node distances may be the answer. Since our dataset is very large (2.1 M users and 1.1M venues) and diverse (check-ins, ratings, user-friendships, locations), we definitely plan to tackle the future challenges raised by this paper.

1.2 Link Prediction in Weighted Networks: A Weighted Mutual Information Model

Summary: In this paper, Boyao Zhu et al. [2] focus on resolving knowledge discovery and data mining for undirected and weighted networks, where weights are applied to appreciate the distinguishment of edges from different nodes. Their model resulted in a high prediction accuracy for unweighted networks and weighted networks. The extent of their research even demonstrates the potential to predict edge weights.

Critique: The proposed Weighted Mutual Information (WMI) model combines both structural features and link weights to offer improved feature calculations for Weighted Common Neighbor (WCN), Weighted Adamic-Adar (WAA), and Weighted Resource Allocation (WRA). Interestingly, the paper states that emphasizing the role of weak ties, referred as pure WMI-based indices, results in even higher prediction precision. By comparing model accuracies, the authors demonstrate that weak ties play a significant role for predicting edges. This link back to the question raised by the first paper [1] about discovering new information from the input network. The paper also suggests that the influence of new information offered by weak ties has significant influence over solely accounting for edge weight, which may or may not proportionally suggest the strength of a link ties. We find this difficult to believe given they're benchmarks and will explore this more by using Louvain's algorithm for community detection and further investigating the influence that communities have in link prediction. We are also interested in further investigating how distinct sets of common neighbors with equal cardinality can enhance the accuracy of predicting links.

1.3 Link Prediction in Social Networks: the State-of-the-Art

Summary: In this paper, Peng et al. [3] review and analyze state-of-the-art link prediction techniques for dynamic social networks, attempting to categorize link prediction problems and link prediction techniques. Our link detection experiments will use a few techniques from their catalog to measure node and topological features to measure the likelihood of an edge.

Critique: The various number of proposed metrics for evaluating link evolution in a social network are limited to mostly homogeneous networks. We will attempt to expand some link prediction models for heterogeneous social networks. Specifically, we are interested in refining the following metrics for heterogeneous network applications: Node-based metric: common check-ins and ratings to generate users' similarity. Neighbor-based metrics: Refine models for nd Common Neighbors, Jaccard Coefficient, Adamic-Adar Coefficient, and Resource Allocation (with the influence of Zhu and Xias's research). Path-based metrics: Local Path, Relation Strength Similarity, and FriendLink. However, we found it difficult to abide by their categorization model with so little expansion on how external information (weights, attributes, knowledge repository, etc) can influence the resolution of our link prediction.

2 Project Proposal

Our project aims to perform link prediction on Foursquare Social Network.

2.1 Dataset

We will analyze the FourSquare dataset provided by Sarwat, Levandoski, Eldawy, and Mokbel, all of which was extracted from the Foursquare application through their public Application Program Interface (API). Note that user anonymization was applied to users and geolocations. The data are contained in five files, users.dat, venues.dat, check-ins.dat, socialgraph.dat, and ratings.dat.

More concretely, the data consists of the following five tables:

1. **Users**: Consists of a list of 2,153,471 unique user IDs and their geolocation (longitude, latitude). The user IDs are associated with the network social graph, check-ins, and ratings.
2. **Social Graph**: A list of 27,098,490 edges (connections) that exist between users. Each social connection consists of two users (friends) represented by two unique IDs.
3. **Venues**: Consists of a list of 1,143,092 venues that are associated to ratings from users and user check-ins.
4. **Check-ins**: A list of 1,021,970 check-ins (visits) of users at venues. Each check-in has a unique id, the user who checked-in, and the venue, the geolocation logged when checking in, and the time the user checked-in.
5. **Ratings**: A list of 2,809,581 implicit ratings that quantifies how much a user likes a specific venue, from 1 to 5 where 1 represent strong dislike and 5 represents a strong appreciation.

2.2 Problem Description

As explained in 2.1, the diversity of the dataset makes this a very interesting network to analyze. Given the current state of the Foursquare social network, how accurately can we predict future edges between users and venues? Does the accuracy increase when we consider the edge weights in the network (obtained from the ratings graph)? What happens if we consider additional information such as the location of the venues and the social graph of users? Furthermore, can we predict the weight of future edges? These are the questions we wish to answer at the end of our research project.

2.3 Methodology

Similarity-based metrics on unweighted graph: First, we will use the common link prediction algorithms such as Jaccards Coefficient, Adamic/Adar, Katz-measure and PageRank on our network and compare the results. From these experiments, we hope to gain valuable insights as well as useful results to benchmark against future improvements.

Weighted metrics: Next, we plan to use the metrics suggested in paper [2] to extend our similarity-based metrics to weighted networks. We will use the values from ratings network to generate edge weights for our user-venue network. For example, if a user u has been to venue v and gives it a rating of 3, then the u-v edge will have a weight of 3. We will also explore the importance of weak ties in predicting missing links by introducing a free parameter α which control the relative contributions of weak ties to the similarity measures.

Heterogeneous network features: Finally, we plan to take advantage of the diversity of the dataset by considering additional information such as the friendships between users, and the location of the venues.

2.4 Evaluation

As with all social networks, our Foursquare network is sparse. Therefore, simply measuring the accuracy of our predictions does not provide meaningful insights. Instead, as suggested by Liben-Nowell et al [1], we will compare the results from our link prediction methods against a random predictor baseline which simply randomly selects pairs of (user, venue) that are not connected in the training dataset.

2.5 Potential Challenges

Currently, these are the challenges that come to mind:

- Deriving an application methods for homogeneous techniques for measure features, like Jaccard Coefficient, Local Path, Relation Strength Similarity, and FriendLink.

– Apart from link prediction, we want to also predict a weight for the suggested link.
– Computation time. Having over 20 million edges between without our social graph alone, we hope that network computations and reductions do not take too long to compute. We do not expect this to completely hinder us, but its a concern worth noting.

References

1. Liben-Nowell, David, and Jon Kleinberg. 'The Link Prediction Problem for Social Networks.' Proceedings of the Twelfth International Conference on Information and Knowledge Management - CIKM 03, 2003. doi:10.1145/956958.956972.
2. Zhu, Boyao, and Yongxiang Xia. 'Link Prediction in Weighted Networks: A Weighted Mutual Information Model.' Plos One 11, no. 2 (May 2016). doi:10.1371/journal.pone.0148265.
3. Wang, Peng, Baowen Xu, Yurong Wu, and Xiaoyu Zhou. 'Link Prediction in Social Networks: the State-of-the-Art.' Science China Information Sciences 58, no. 1 (March 2014): 138. doi:10.1007/s11432-014-5237-y.