Spring 2019

Link Predictions for Social Networks in an Online Health Community

Shangguan Wang
University of Iowa

LINK PREDICTIONS FOR SOCIAL NETWORKS IN AN ONLINE HEALTH COMMUNITY

by

Shangguan Wang

A thesis submitted in partial fulfillment of the requirements
for graduation with Honors in the Business Analytics and Information Systems

_____

Kang Zhao
Thesis Mentor

Spring 2019

All requirements for graduation with Honors in the
Business Analytics and Information Systems have been completed.

_____

Jennifer A. Blair
Business Analytics and Information Systems Honors Advisor

Link Predictions for Social Networks in an Online Health Community

by

Shangguan Wang

A thesis submitted in partial fulfillment of the requirements for graduation with Honors in the Tippie College of Business

Kang Zhao
Thesis Advisor

Spring 2019

Thesis received in partial satisfaction of the requirements to graduate with Honors in the Major

Prof. John Murry
Tippie College of Business Honors Director

Link Predictions for Social Networks in an Online Health Community

Abstract

Online social networks are ubiquitous in our daily life and offer different ways for people to interact with each other. Link prediction aims at predicting future social connections or interactions between two users within a social network. This project implements different link prediction algorithms and evaluates their performance for online social networks among users of an online health community for smoking cessations. The social networks were based on users' online interactions via four communication channels: blog comments, message boards, group discussions and private messages. The outcome of this study will help to provide insights into the design of recommender systems in such online social networks, and to improve user experience and engagement in online health communities.

Keywords: link prediction, social networks, online health communities

Introduction

Online social networks have become an integral part of our daily life and offer users a chance to initiate relationship with strangers through different communication channels such as direct message and public discussion. Online health communities (OHCs) are forms of social networks that allow patients and their families as well as medical experts to exchange medical information and related social support [1,7,8]. Nonetheless, designing an OHC can be difficult because of the various levels of members' medical background, the different stages of their health conditions, as well as the users' preferences of different communication channels. For example, a new user who just joined the community might participate more often in discussion forums and ask questions compared to an old user who can engage with other users through direct message and interact under blog posts.

In this paper, we will conduct link prediction on networks obtained from BecomeAnEX, one of the largest OHCs for smoking cessations. Link prediction aims at predicting potential future connections between two users based on whom they have interacted with in the past [2]. By identifying emerging interactions in the community based on current network structure, friend and content recommendations can then be improved to enhance user experiences. The outcome of this study can provide insights into users' interaction patterns in such online social networks and improve the efficiency of recommender systems.

Related work

A social network can be represented in the form of $G <N, E>$ where N indicates users in the form of nodes and E represents edges or interactions between a pair of users. According to Liben-Nowell et al. [2], link prediction problem involves two time periods, training period t_0 and testing period t' where t_0 is when the interactions represented by edges between node *u* and *v* have not happened yet and t' is the time interval where the interactions have happened. There could be multiple interactions between N_u and N_v and we only consider first connection in our study. Prediction methods are based on estimating the proximity between N_u and N_v and rank node pairs with scores from high to low. We will introduce an array of proximity measures below.

Common neighbor

Newman [3] has designed the common neighbor method which calculates the number of intersection neighbors between two users. It follows the theory of triadic closure which means there is higher tendency for two parties to form connections if they share more common connections in a social network.

$$CN(x, y) = |N(x) \cap N(y)|$$

Resource allocation index

Resource allocation index measures the fraction of resource that a node can sent to another through their common neighbors [4]. The assumption of this method is that the sender node will distribute its resources evenly through their neighbors. In other words, it penalizes those pairs who share common neighbors whom they themselves have a high degree.

$$RA(x, y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{N(u)}$$

Adam Adar

Adam Adar is similar to resource allocation index. The only difference is that it log-transformed the denominator [5].

$$AA(x,y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log |N(u)|}$$

Preferential attachment

Preferential attachment calculates product of the number of neighbors of x and the number of neighbors of y [6]. It assumes that the richer will get richer. In other words, nodes with higher degrees are more easily assessible to each other.

$$PA(x,y) = |N(x)| \times |N(y)|$$

Data

The dataset is from BecomeAnEX, a social media platform for smoking cessation. The whole aggregated network consists of 4 communication channels: blog comments (BC), message board (MB), group discussion (GD), and private messages (PM). BC, MB, and GD are public networks where every user can see activities posted while PM is a private network that only two users can see the content [9]. Together there are 88,321 unique users and 1,012,456 observations from year 2010 to year 2015 after removing self-loops and banned users. Because of some users tend to send welcome messages to new users, un-reciprocal ties in private message connections are also removed. Each observation represents an undirected connection between two users, meaning the connection can be either from A to B or B to A. Table 1 shows the summary statistics of the aggregated network: the density of the network is 0.001, which is typical for sparse social networks. The average clustering coefficient measures how tight-knit nodes are and social networks usually has a higher clustering coefficient compared to other networks. Most of the

nodes or users are in the largest connected components and it takes an average of 3 hops for two not-yet-connected nodes to be connected.

Number of nodes	10,174
Number of edges	53,447
Density	0.001
Average clustering coefficient	0.528
Fraction of nodes in LCC	98.25%
Average shortest path length in LCC	3.029

Table 1: Summary Statistics of the aggregated network

Methods

We picked week 50 to week 70 out of the 5-year time span because this is a period with a large and stable number of posts as shown in the highlighted red box of Figure 1.

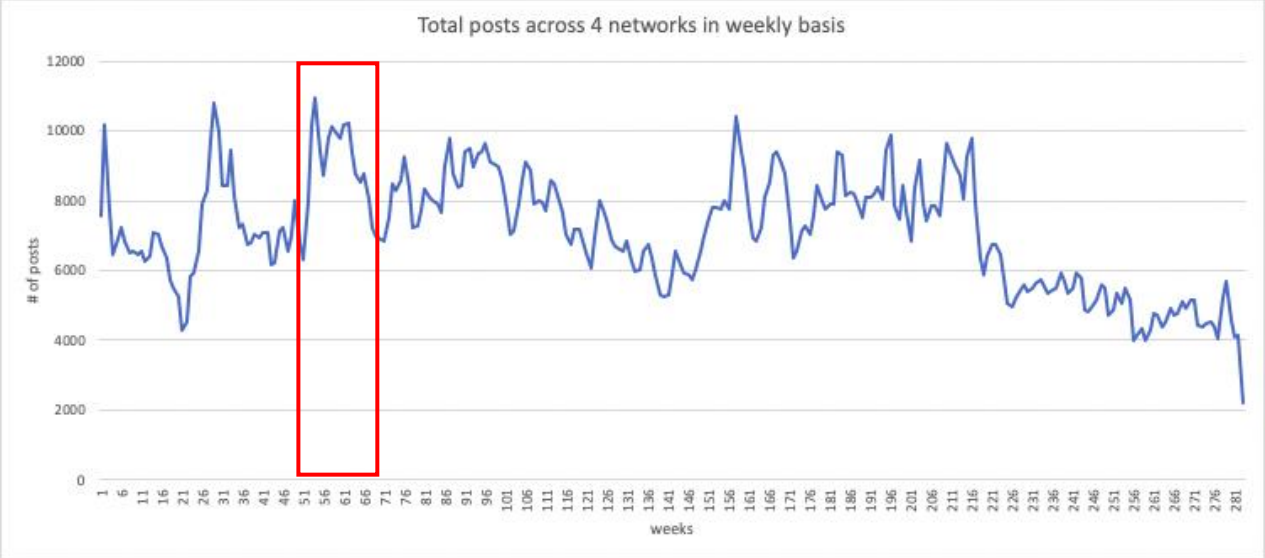

Figure 1: Count of the number of posts every week in 5 years

We split the data into 20 pairs of training and testing sets and construct undirected network using one-week time window. Our goal is to predict currently disconnected user pairs that may get connected next week. For each week, we construct a training network $G_{train} = <V_{train}, E_{train}>$, in which V represents all nodes in week w and E represents all current edges in week w. For every run we predict $G_{test} = <V_{test}, E_{test}>$. There are 2 conditions that have to be met prior to prediction:

1. $V_{train} = V_{test}$

   We first remove new nodes that appeared during week w+1 because newly arrived nodes are not even connected to the network yet.

2. $(V_{train} \cap V_{test}) = \emptyset$

   We then exclude test edges that already existed during the training period to make sure test edges only include newly formed edges among the same sets of training nodes.

We can then apply 4 neighborhood-based unsupervised methods. All 4 methods are based on the assumption that two users are more likely to connect if they are closer (or structurally similar) to each other in the network.

Results

To evaluate the performance of each link prediction method, we rank all node pairs that are disconnected during the training period by their proximity scores in the descending order, and select the top K pairs as prediction of new edges during the testing period. We tried K from 10 to 100 with an increment of 10. Then we calculate the average weekly precision at K. We specifically used precision instead of recall because for recommendation tasks, we care less about false negatives.

$$Average\ Weekly\ Precision\ at\ K = \frac{Number\ of\ correct\ predictions\ out\ of\ k\ (TP)}{K}$$

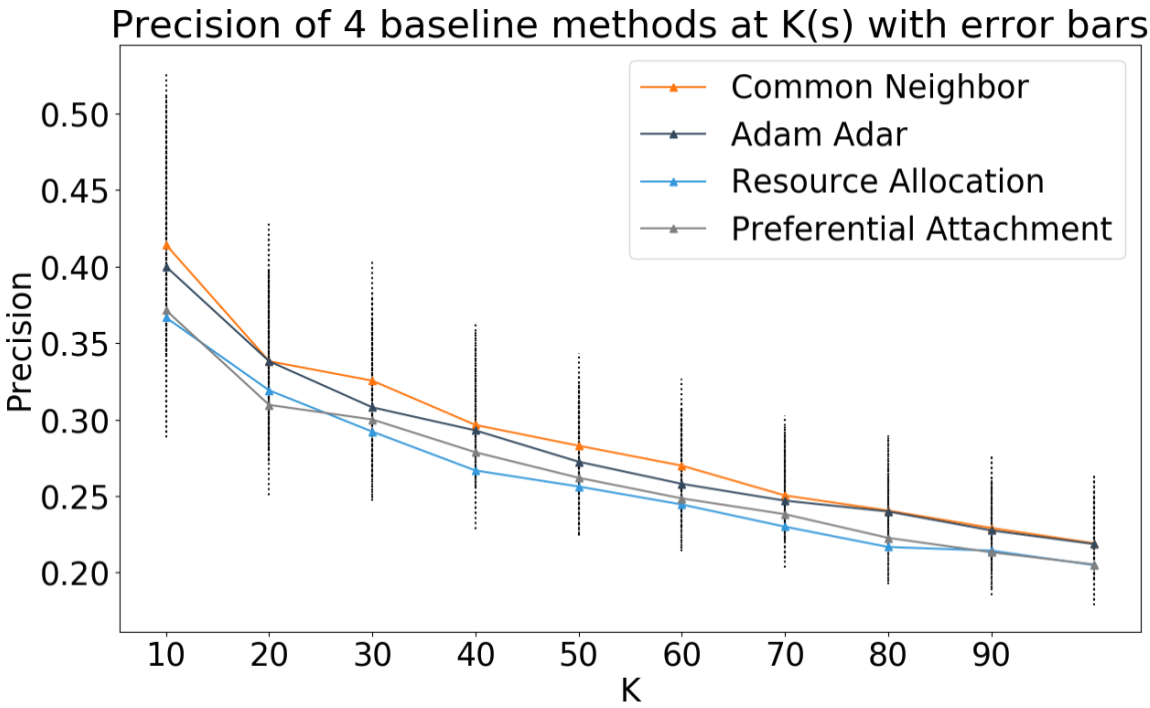

Figure 2 Precision of 4 baseline methods at K(s) with error bars

According to the results in Figure 2, Common neighbor performs the best followed by Adam Adar, Preferential Attachment and Resource Allocation Index. The best precision is 0.414 at K = 10, meaning that if we predict 10 new links, more than 4 of them will actually occur. Naturally, as we increase K, we bring more false positives. The vertical dotted lines at each K mean the 95% confidence interval of the mean precision across 20 weeks.

Conclusion and Future Work

Link prediction based on social networks is challenging given the dynamic nature of social networks and user churn behaviors [10]. Throughout this work, we have shown that link prediction based solely on user interactions can yield satisfying performance. We implemented four different link prediction methods on weekly undirected graphs constructed from the online health community. Common Neighbor, the simplest method outperforms the rest. We believe the outcome of this study will help to provide insights into the design of recommender systems in

such online social networks, and to improve user experience and engagement in online health communities.

We only used neighborhood-based unsupervised link prediction algorithms in this study. Future works can consider more sophisticated methods such as network embeddings and methods based on supervised learning [11]. Moreover, our current setup does not consider different communication channels. In the future, we would like to distinguish different communication channels and apply link prediction in a multi-relational network setting [12].

Acknowledgement

This study is Shangguan Wang's honor thesis. This work is supported by a fellowship from the Iowa Center for Research by Undergraduates (ICRU). The author wants to express her sincere gratitude to Dr. Kang Zhao for his gracious support and encouragement along the way.
The author would like to thank Xi Wang for providing the dataset and Sulyun Lee for her collaboration.

References

[1] Coulson, N. S., Buchanan, H., & Aubeeluck, A. (2007). Social support in cyberspace: A content analysis of communication within a Huntington's disease online support group. *Patient Education and Counseling*, *68*(2), 173–178. https://doi.org/10.1016/j.pec.2007.06.002

[2] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[3] M. E. J. Newman. Clustering and preferential attachment in growing networks. Physical Review Letters E, 64(025102), 2001.

[4] T. Zhou, L. Lu, Y.-C. Zhang. *Predicting missing links via local information*. Eur. Phys. J. B 71 (2009) 623.

[5] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[6] A Papadimitriou, P Symeonidis, and Y Manolopoulos. Friendlink: Link prediction in social networks via bounded local path traversal. In *Computational Aspects of Social Networks (CASoN), 2011 International Conference on, pages* 66–71. IEEE, 2011.

[7] Klemm, P., Bunnel, D., Cullen, M., Soneji, R., Gibbons, P., & Holecek, A. (2003). Online cancer support groups: A review of the research literature. *Computers Informatics Nursing*, *21*, 136–142.

[8] Zhao, K., Greer, G. E., Qiu, B., Caragea, C., Mitra, P., Wu, D., … Yen, J. (2011). *Finding influential users of an online health community: a new metric based on sentiment influence*. Presented at the The 21st Annual Workshop on Information Technologies and Systems (WITS'11), Shanghai, China.

[9] Zhao, K., Wang, X., Cha, S., Cohn, A. M., Papandonatos, G. D., Amato, M. S., … Graham, A. L. (2016). A Multirelational Social Network Analysis of an Online Health Community for Smoking Cessation. *Journal of Medical Internet Research*, *18*(8), e233.

[10] Wang, X., Zhao, K., & Street, N. (2017). Analyzing and Predicting User Participations in Online Health Communities: A Social Support Perspective. *Journal of Medical Internet*

[11] Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online Learning of Social Representations. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 701–710.

[12] Davis, D., Lichtenwalter, R., & Chawla, N. V. (2012). Supervised methods for multi-relational link prediction. *Social Network Analysis and Mining*, *3*(2), 127–141.