Evolution of Friendship: a case study of MobiClique

JooYoung Lee
Network Science and Information Laboratory
Institute of Information Systems
Innopolis University, Innopolis, Russia
j.lee@innopolis.ru

Kontantin Lopatin
Innopolis University
Innopolis, Russia
k.lopatin@innopolis.ru

Rasheed Hussain
Institute of Information Systems
Innopolis University, Innopolis, Russia
r.hussain@innopolis.ru

Waqas Nawaz
College of Computer& Information Systems
Islamic University of Medina
Medina, Saudi Arabia
wnawaz@iu.edu.sa

ABSTRACT

Understanding the evolution of relationship among users, through generic interactions, is the key driving force to this study. We model the evolution of friendship in the social network of MobiClique using observations of interactions among users. MobiClique is a mobile ad-hoc network setting where Bluetooth enabled mobile devices communicate directly with each other as they meet opportunistically. We first apply existing topological methods to predict future friendship in MobiClique and then compare the results with the proposed interaction-based method. Our approach combines four types of user activity information to measure the similarity between users at any specific time. We also define the temporal accuracy evaluation metric and show that interaction data with temporal information is a good indicator to predict temporal social ties. The experimental evaluation suggests that the well-known static topological metrics do not perform well in ad-hoc network scenario. The results suggest that to accurately predict evolution of friendship, or topology of the network, it is necessary to utilise some interaction information.

CCS CONCEPTS

•Networks → Online social networks; Online social networks; •Information systems → Social networks;

KEYWORDS

social network anlaysis, community detection, link prediction

ACM Reference format:
JooYoung Lee, Kontantin Lopatin, Rasheed Hussain, and Waqas Nawaz. 2016. Evolution of Friendship: a case study of MobiClique. In *Proceedings of ACM International Conference on Computing Frontiers, Sienna, Italy, May 2017 (CF'17),* 4 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

CF'17, Sienna, Italy

DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Link prediction problem aims to find missing links or predict new links from an observed network. Link prediction helps to understand the evolution of social networks and has a power to complete the current observed graphs. As it involves many other areas of study in social network analysis, such as ranking, link prediction has many important applications. Recommender systems is one of the applications where link prediction is utilised. For example, if new connections are predicted, common interests for online shopping sites and new collaborators citation networks could be found.

Traditionally, in the literature, the link prediction methods have been focused on the analysis of topological similarity to foresee future link appearances. There are many popular metrics based on neighbours and paths such as *common neighbours* and *preferential attachment* [7].

Despite the fact that social networks are highly dynamic and volatile, link prediction metrics have not paid much attention to the temporal aspects of networks [2]. In this paper, we propose an interaction-based method to predict future links in the network of MobiClique [3, 4].

Contributions The contributions of this study are as as follows. *First*, we define how to measure temporal accuracy of link prediction. The evaluation of prediction methods is crucial to prove quality of results. However, it is difficult to evaluate link prediction strategies since the evaluation processes rely upon predefined thresholds [5]. In fixed threshold metrics, the precision and recall for top-N predictions are widely used. We measure the accuracy of the prediction at every time stamp so that it captures the true accuracy at the moment. *Second*, we formulate an interaction-based temporal link prediction measure that combines real activities of network users associated with time and static measures. *Third*, we apply time decaying function to count the number of edges that are correctly predicted (true positives). Since, predictions are made at each time step, it is unfair to ignore the links that are previously predicted but not appeared on time. We discount the count for number of links, which appear later than the prediction, according to a logistic curve.

2 RELATED WORK

Many previous works in link prediction use topological information in a given graph to find missing links between nodes [8]. We apply six of existing similarity indices to MobiClique network: common

neighbours, Adamic-Adar index, Resource allocation index, Jaccard index, Preferential attachment index and Simrank index. These metrics take information only about adjacent vertices and degrees of nodes, and no information about "real world" properties such as interests of users and their communication patterns.

There are other types of link prediction techniques such as node-based, path-based, random walked-based, social theory based, learning based [6], tensor based [7], etc. In these models, if we consider the time of link appearing then it become a problem of temporal link prediction. Social network with time can be organised as a third-order tensor or multidimensional array, therefore, matrix and tensor-based techniques can be used.

Supervised learning is another alternative solution to link prediction problem [1]. There are many classification algorithms that can be used. The most efficient and popular methods are decision tree, bagging, multilayer perceptron, k-nearest neighbours. They give a very high accuracy by using few features, i.e., from 4 to 9 features in different datasets. All of aforementioned approaches either do not support temporal aspects or only rely on structural information.

3 METHODOLOGY

In this section, we define the problem and our proposed solution along with the evaluation measure and MobiClique description.

3.1 Temporal link prediction

In this section, we formally describe our problem setting. Given an undirected and unweighted graph $G^t = (V^t, E^t)$ where V^t is the set of vertices and E^t is the set of edges in G^t, we find $G^{t'} = (V^{t'}, E^{t'})$ where $t < t'$. In other words, given a current graph G^t, we predict a future graph $G^{t'}$. In many other link prediction metrics, the future time of the prediction is not specified, i.e. t' is some time in the future or the latest time of given datasets. In this case study, we predict $G^{t'}$ for each time stamp, i.e., every 1 hour, to find reliable accuracy of predictions over time.

3.2 Interaction-based Similarity Measure

We define a similarity among users based on their static preferences and heterogeneous temporal interactions, i.e. interaction-based similarity measure. In order to achieve this, we need to specify and acquire temporal interactions along with the static preferences of users.

3.2.1 Temporal Interaction Information: MobiClique. MobiClique is an application used to acquire temporal interactions among users in terms of their groupings, proximity, messaging, and static profile information. We utilise four types of user interactions or behaviours to predict future links.

- **Interest group** MobiClique collected lists of Facebook groups of participants to initialise profiles. During the experiment, users were allowed to join any existing group or create new groups at any time. Hence, interests groups of participants were changing over time.
- **Proximity** For every 120 seconds, devices of participants performed discovering for other devices. The trace records of all the nearby Bluetooth devices are reported by the periodic Bluetooth device.
- **Messaging** MobiClique allowed messaging between friends or among members of an interest group. There are three types of messages; unicast (type=U), multicast (type=M), or broadcast (type=B). We considered only unicast messages between two users since other types do not contribute much to form new links.
- **Users profiles** The profiles consist of information about the institute, the city and the country of each participant. We assume that this information does not change overtime and regarded as non-temporal. We understand that in reality profile information may change but it is not frequent.

Participants logged on to their Facebook accounts to initialise profiles, but they had a possibility to hide any information before it was recorded. Hence, the dataset does not contain the full information about friends and groups of participants.

3.2.2 Collaborative Similarity towards Link Prediction. The basic idea for link predication in this study is that if two users have strong or frequent interactions between them then there is a high probability that both are friendship on social platform, e.g. Facebook. These interactions need to be consistent over different time intervals. For instance, if two users have exchanged many messages between them in a time span of 24 hours whereas the time interval is an hour then there is low probability for their friendship in future.

We define similarity measure based on two different aspects of temporal interactions among users and static profile information of each user, as discussed in Section 3.2.1. We describe each measure separately in the following discussion and explain the collaborative measure at the end.

- *Common interest*: Let the common interest based similarity between an arbitrary pair of users a and b at time t be defined as $C^t_{a,b} = \frac{|G^t_a \cap G^t_b|}{|G^t_a \cup G^t_b|}$, where G^t_a is the set of interest groups that a is subscribing at time t.
- *Proximity*: Let the proximity similarity between two users a and b be defined as $P^t_{a,b} = \frac{\Sigma_M d^m_{a,b}}{|M|}$, where $d^m_{a,b}$ is the duration of an instance of the meeting $m \in M$.
- *Messaging*: There are two modes of message sharing among users, 1) *multicase* (broadcast) where the message is shared among many (all) users, 2) *unicase* messages that shared with specific users. Intuitively, first category of messages do not significantly contribute to promote friendships among users, therefore, we use unicast only to measure messaging based similarity. And this similarity $M^t_{a,b}$ is defined as the number of messages exchanged between user a and b before time t.
- *Profiling*: In real life, people working in some institution know each other because of occasional gatherings or events. Therefore, user profiles would be a healthy indicator for actual relationship. We define the profile similarity, $N_{a,b}$, between two users a and b as the number of identical attributes, e.g. institute, city, and country. Profile information does not change while data acquiring phase in MobiClique and therefore it is considered as a static or non-temporal measure.

All these four aspects are computed independently and normalised in the range of $[0, 1]$. One of the trivial approach for integration is to use linear aggregation for these factors. Consequently, the similarity between an arbitrary pair of users a and b is defined as follows:

$$S^t_{a,b} = \alpha_1 C^t_{a,b} + \alpha_2 P^t_{a,b} + \alpha_3 M^t_{a,b} + \alpha_4 N_{a,b} \quad (1)$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are constants.

To predict a link based on the similarity metric defined above, $S^t_{a,b}$, we also define a temporal threshold $\theta^t = min_{a,b \in F^t}(S^t_{a,b})$, where F^t is the set of existing and predicted links before t. If $S^t_{a,b} > \theta^t$, then we predict that, at time t, a new link appears between a and b.

3.3 Accuracy Evaluation Measure

We define the temporal accuracy of the prediction. For each time step t, let L^t be the set of predicted links at time t and E^t be the set of newly appeared links before t. Then, let $(FN)^t = E^t - L^t$ be the set of unpredicted links (False Negative). Similarly, let $(FP)^t = L^t - E^t$ be the set of wrongly predicted links (False Positive) and $(TP)^t = (L^t \cap E^t) - (L^{t''} \cap E^{t''})$ be the set of correctly predicted links (True Positive), where t'' is the previous time step. The accuracy of the prediction at time t is defined as follows.

$$(Acc)^t = \frac{|(TP)^t| + |(TN)^t|}{|(FN)^t| + |(TP)^t| + |(FP)^t| + |(TN)^t|}$$

$$|(TP)^t| = \sum_i l^t_i$$

$$l^t_i = \begin{cases} 1 & \text{if } t \geq t^p_i \\ F(d) & \text{otherwise} \end{cases}$$

where, $l^t_i \in (TP)^t$, t^p_i is the predicted time of link i's appearance, $d = |t - t^p_i|$, and $F(d) = \frac{1}{1+e^d}$. We compute accuracies of predictions for each individual metrics as well as the combination of all. The metrics shown are *interest group*, $C^t_{a,b}$, *individual messaging*, $M^t_{a,b}$, *proximity*, $P^t_{a,b}$, *profiles*, $N_{a,b}$, and the combination, $S^t_{a,b}$.

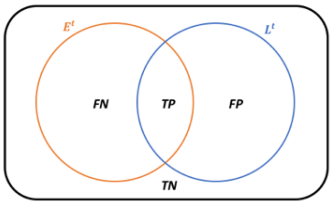

Figure 1: Relationship of the edge sets.

For any given time t, we can plot the relationship of given sets as shown in Figure 1. Our definition of accuracy, $(ACC)^t$, coincides with Rand index.

4 EXPERIMENTS

In this section, we explain our experimental design and show the results of the prediction in comparison with other well known link prediction methods.

4.1 Dataset

We employ the MobiClique dataset to apply our proposed method, which was collected at *SIGCOMM 2009* conference in Barcelona, Spain. During the first two days, 76 participants used MobiClique application and different types of interaction data was collected. The duration of the entire dataset is about 62 hours. The time interval for our experiments is one hour. We apply different methods to predict links in MobiClique dataset and then estimated the accuracies for each method. Accuracy of the prediction is computed for each time interval independently.

Ground Truth: To evaluate the accuracy of prediction results, we use information about friendship of users. Initially, friends information was collected from Facebook accounts of participants in MobiClique application. During the experiments, users could discover and add new users as their friends.

4.2 Comparison: Topology-based methods

We use five static link prediction methods, based on the topological properties of a network, for comparison with our proposed method. We use two notations for the definitions below. $\Gamma(a)$ is the neighbours of user a and deg_b is the degree of node b.

4.2.1 Common neighbors. A common neighbor measure counts the number of neighbours that users a and b have in common, $S_{a,b} = \Gamma(a) \cap \Gamma(b)$. If two users have large number of common neighbors then they are probably friends to each other.

4.2.2 Jaccard Coefficient. Jaccard Coefficient is defined as the size of the intersection divided by the size of the union of the sets, $S_{a,b} = \frac{|\Gamma(a) \cap \Gamma(b)|}{|\Gamma(a) \cup \Gamma(b)|}$. This can be regarded as the normalised version of common neighbours approach. However, the semantics for friendship remains same.

4.2.3 Preferential attachment. Preferential attachment means that the more connected a node is, the more likely it is to have new links. To compute the similarity based on the preferential attachment, the product of the number of neighbors of two users a and b is used. Formally, $S_{a,b} = |\Gamma(a)| \times |\Gamma(b)|$. This measure has a serious issue with the high degree vertices, which is addressed in the subsequent measures.

4.2.4 Adamic-Adar. Adamic-Adar measure is defined as inverted sum of degrees of common neighbours for given two vertices, $S_{a,b} = \sum_{c \in \Gamma(a) \cap \Gamma(b)} \frac{1}{log(deg_c)}$.

4.2.5 Resource allocation. Resource allocation metric is similar to Adamic-Adar. Both suppress the contribution of the high-degree common neighbors, but resource allocation metric punishes more heavily, $S_{a,b} = \sum_{c \in \Gamma(a) \cap \Gamma(b)} \frac{1}{deg_c}$.

In Fig. 2, we analyze the accuracies of both topological and our proposed measures. It is evident that interaction-based measure performs better than all the topological ones. We can also notice that the common neighbors index measure has highest accuracy among all other topological measures. Adamic-Adar and Resource allocation metrics behave almost in the same way at different timestep due to their similar nature as explained in the above definitions.

The accuracy decreases over time for all topological metrics, because of their inability of predicting new edges. In other words,

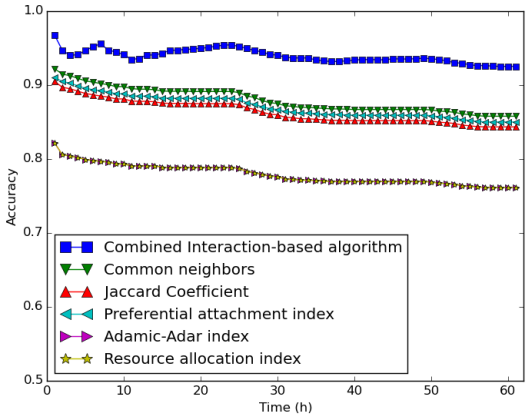

Figure 2: Accuracy of predictions based on different metrics (Topological).

we compute the accuracy of the prediction for every timestep, and these static metrics keep predicting new links based on the previously predicted network, not the original network. Therefore, the accuracy of static metrics decreases over time.

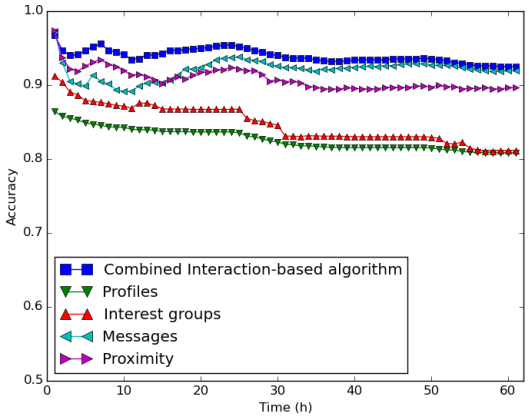

Figure 3: Accuracy of predictions based on proposed metrics (Interaction-based).

In Fig. 3, it is shown that messages and proximity give higher accuracy than profiles and interests. Also, the accuracy of messages and proximity increases little over time, while the accuracy of interest groups and profiles only decreases. Since profile information ($N_{a,b}$) does not change during experiment and we cannot predict new edges, the accuracy decreases over time. When users join new interest groups, the value of $C^t_{a,b}$ becomes higher and chance of predicting incorrect edges becomes higher. Therefore, the accuracy of interest groups also decreases. In case of messages and proximity measures, they are more reliable since new friends often communicate to each other by messaging ($M^t_{a,b}$) or through physical communication ($P^t_{a,b}$). Therefore proximity and messages have stable and increasing accuracy over time.

According to our observations on the metrics above, we combine them with different weights to compute the similarity of users in Equation (1). We find the appropriate weights for each term as $\alpha_1 = 2$, $\alpha_2 = 0.2$, $\alpha_3 = 0.2$, and $\alpha_4 = 2$. Also, the threshold θ^t is computed for each time interval as a minimum value of similarities of existing edges.

5 CONCLUSIONS AND FUTURE WORK

We proposed an interaction-based link prediction measure that predicts probable temporal social ties. We combined four different aspects of user interactions, namely grouping, profiling, messaging activities, and proximity, to measure the similarity among users. We have defined a lower-bound, based on the current minimum social ties between users, as a threshold to determine the links. Since, we can measure the similarity of existing users at any time, therefore, we use this threshold to predict a new link at any timestep.

The experimental results have shown the link prediction accuracy increased by 8%-16% compared to traditional structural based approaches. We have also noticed the superiority of proximity and messaging based measures over profile and interest group based measures, where former measures achieved 6%-8% higher accuracy and 3% lower mean square error. It has proven our hypothesis that more temporal information helps to predict links with higher accuracy. Thus, future social ties are much influenced by social temporal interactions than structural information.

The weights in interaction-based similarity are determined empirically, however, estimating these weights automatically is non-trivial and regarded as our future work. We are also interested to employ frequent temporal communication patterns to analyse and estimate the persistence of existing links.

REFERENCES

[1] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. 2006. Link prediction using supervised learning. In *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*.

[2] Mujtaba Jawed, Mehmet Kaya, and Reda Alhajj. 2015. Time Frame Based Link Prediction in Directed Citation Networks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM '15)*. ACM, New York, NY, USA, 1162–1168. DOI: https://doi.org/10.1145/2808797.2809323

[3] A-K Pietil"ainen, E. Oliver, J. LeBrun, G. Varghese, and C. Diot. 2009. MobiClique: Middleware for Mobile Social Networking. In *WOSN'09: Proceedings of ACM SIGCOMM Workshop on Online Social Networks*.

[4] Anna-Kaisa Pietiläinen, Earl Oliver, Jason LeBrun, George Varghese, and Christophe Diot. 2009. MobiClique: Middleware for Mobile Social Networking. In *Proceedings of the 2Nd ACM Workshop on Online Social Networks (WOSN '09)*. ACM, New York, NY, USA, 49–54. DOI: https://doi.org/10.1145/1592665.1592678

[5] Ben Taskar, Ming fai Wong, Pieter Abbeel, and Daphne Koller. 2004. Link Prediction in Relational Data. In *Advances in Neural Information Processing Systems 16*, S. Thrun, L. K. Saul, and B. Schölkopf (Eds.). MIT Press, 659–666. http://papers.nips.cc/paper/2465-link-prediction-in-relational-data.pdf

[6] Anna Tigunova, JooYoung Lee, and Sadegh Nobari. 2015. Location Prediction via Social Contents and Behaviors: Location-Aware Behavioral LDA. In *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. IEEE, 1131–1135.

[7] Peng Wang, Baowen Xu, Yurong Wu, and Xiaoyu Zhou. 2015. Link prediction in social networks: the state-of-the-art. *SCIENCE CHINA Information Sciences* 58, 1 (2015), 1–38. http://dblp.uni-trier.de/db/journals/chinaf/chinaf58.html#WangXWZ15

[8] Zhiqiang Wang, Jiye Liang, Ru Li, and Yuhua Qian. 2016. An Approach to Cold-Start Link Prediction: Establishing Connections between Non-Topological and Topological Information. *IEEE Transactions on Knowledge and Data Engineering* 28, 11 (2016), 2857–2870.