You are who knows you:
Predicting links between non-members of Facebook

Emőke-Ágnes Horvát, Michael Hanselmann, Fred A. Hamprecht, Katharina A. Zweig

Interdisciplinary Center for Scientific Computing (IWR)
Heidelberg Collaboratory for Image Processing (HCI)
Marsilius Kolleg
University of Heidelberg, Germany
Technical University of Kaiserslautern, Germany

Abstract

Could online social networks like Facebook be used to infer relationships between non-members? We show that the combination of relationships between members and their e-mail contacts to *non*-members provides enough information to deduce a substantial proportion of the relationships between non-members. Using structural features we are able to predict relationship patterns that are stable over independent social networks of the same type. Our findings are not specific to Facebook and can be applied to other platforms involving online invitations.

Keywords: online social network, privacy, link prediction, machine learning, random forest classifier

1 Introduction

Inference of user attributes and link prediction in online social networks is a challenging task that has attracted the attention of many researchers in the past few years. They showed that characteristics of a given user, such as its political preference or its sexual orientation, can be accurately inferred based on the attributes of its friends [3, 7, 9, 14]. Also, previously unrevealed or future relationships have been predicted with high precision using both supervised [6, 13] and unsupervised [5] learning methods. As an extension of these insights into the transparency of the members of such platforms, we asked a novel question: How many of the relationships between *non*-members can online social networks infer [2]? Besides revealing the potential of predicting relationships *outside* the network, we present a challenging link prediction task where learning and testing were performed on distinct Facebook networks.

2 Ground truth imputation

Given an online social network, a society is divided into a fraction ρ of members and $1-\rho$ of non-members (see Figure 1). Members are linked through mutually confirmed friendship relationships. Furthermore, aiming at expanding their circle of friends, a fraction α of the platform members import their whole e-mail address-book, thereby sharing also their contacts to non-members. Based on the seemingly innocuous combination of these two information types, we infer links between the non-members.

Data comprising the whole social network is unattainable (i.e., friendships via online social networks, e-mail contacts to non-members, and acquaintances between them). Thus, in order to provide the ground truth for a machine learning approach, we impute the missing information. We assume that the available real Facebook friendship networks of students from five different US universities (`UNC`, `Princeton`,

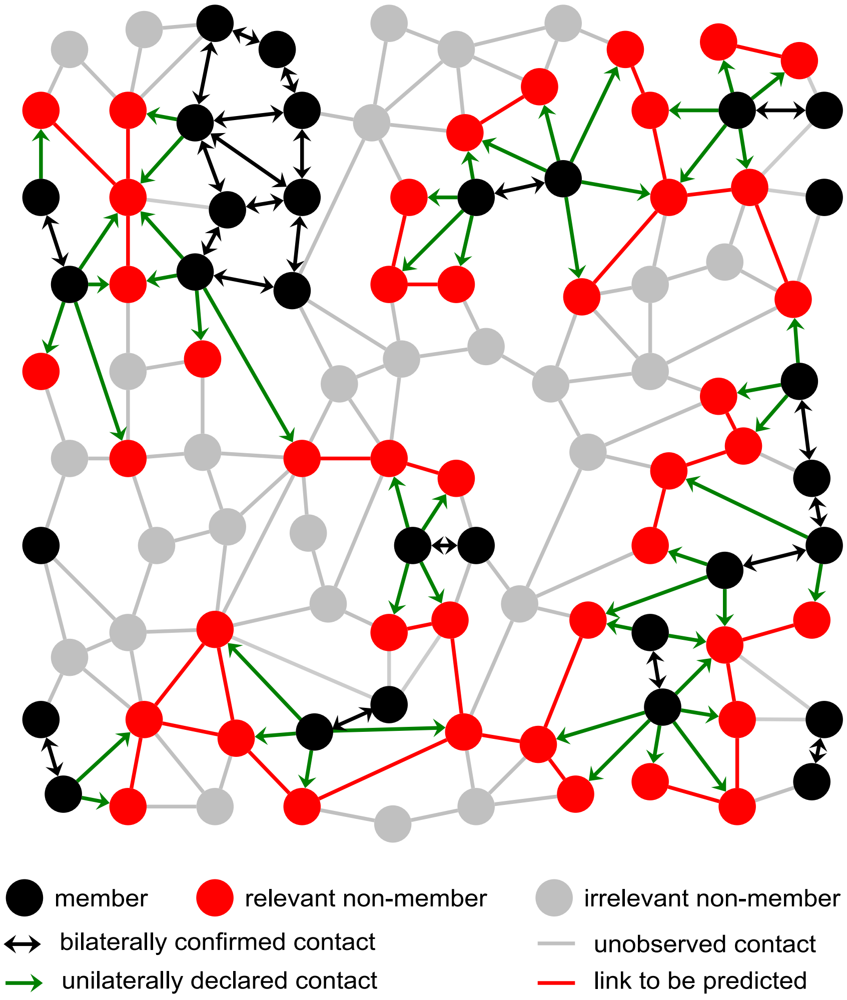

Figure 1: Division of the social network into members (black nodes) and non-members. In the depicted toy example, a fraction of $\rho = 0.3$ (30 out of 100) individuals are members. The relevant subset of non-members consists of those who are in contact with at least one member (red nodes). A fraction of $\alpha = 0.5$ (15 out of 30) members have disclosed their e-mail contacts to non-members. The edges between members (black, bi-directed arrows) and to non-members (green arrows) are used to predict edges between non-members (red lines). For clarity, the values of ρ and α are exaggerated and the weak ties between individuals are omitted. Figure reprinted from [2] with permission.

Georgetown, Oklahoma, and Caltech) represent their complete online social network [11]. We then partition the students into members and non-members. We do not have a clear understanding of how people decide upon joining online social networks. On one hand, a recent analysis of the growth of Facebook showed that the probability of a non-member joining the platform increases with the structural diversity of its acquaintances who are members, i.e., with the number of connected components in its Facebook neighbourhood [12]. On the other hand, there is indication that platforms recruit their members through a mixture of online mediated invitations by friends who are already members and through independent decisions by individuals who are not yet friends of a member [4]. In line with the latter investigation, we cover a wide range of possible mechanisms: we use a series of different member recruitment models and show that our results are stable for all of them. The considered models are the following: *1)* the *breadth-first search* model (BFS): once we have a starting member, all her friends become members followed by all their friends and so on, *2)* the *depth first search* model (DFS): a randomly chosen friend of a member joins, followed by a randomly chosen friend of the new member and so on, *3)* we start a *random walk* (RW) from a member and restart the walk as soon as we would choose someone who is already a member, *4)* the *ego networks selection* model (EN): we select the members randomly and together with them, all their friends join the platform, and *5)* the *random selection* model (RS): people decide independently from their friends whether to become a member or not, i.e., we choose each member randomly from the remaining non-members. Accordingly, the ground truth imputation for our inference problem consists in fixing the fraction of members ρ, partitioning the community into members and non-members by using one of the member recruitment models above described, and finally choosing the disclosure parameter α, thereby controlling for the percentage of contacts that are made public. Having devised the ground truth, we use the following approach.

3 Link prediction

The available Facebook networks are anonymized. In the absence of user attributes, we base our predictions solely on topological graph features. For each pair of non-members, we compute a set of features deduced from the known structural properties of (online) social networks [10, 8]. For example, based on the recognition that people sharing a friend are usually friends themselves, we include a feature that counts the number of neighbours two non-members share. Other features weight this number in several ways (e.g., by the popularity of the common neighbour) or count the number of paths of length 3 between the two non-members. We use the feature vectors to employ a standard supervised learning method called random forest classifier [1]. We adjust the parameters of the classifier on a training set and then apply it to a test set. We predict that those pairs of non-members are linked for which the edge probability determined by the algorithm is higher than some threshold value. In a final step, we validate our predictions by comparing them with the ground truth. We use two measures to quantify the accuracy of the algorithm:

1) the *Area Under the Curve* (AUC) which is a standard machine learning measure that quantifies the probability that the classifier algorithm assigns higher prediction values to true positives than to true negatives. Thus, a perfect classifier has an AUC of 1 while random guessing results in an AUC of 0.5.

2) the *positive predicted value* of the k top-ranked predictions (PPV_k), introduced by [5], is defined as the percentage of correctly classified edges among the first k pairs in the ranking and is thus equal to the sensitivity achieved by predicting these k samples to be edges.

Instead of training and testing within the same network, we assure the independence of these two sets by learning and testing on different networks. We do so by devising two training schemes. *1)* In the $4 \rightarrow 1$ cross-prediction scenario the classifier is trained on samples from *four* data sets and tested on samples from a fifth data set. With this scheme we avoid overfitting. *2)* In the $1 \rightarrow 1$ cross-prediction setting the classifier is trained on *one* university data set and evaluated on another. The goal here is to evaluate whether a single network contains enough characteristic patterns to obtain high-quality predictions for an entirely different network.

4 Results

Imputing the ground truth required introducing two parameters (the membership ρ and the disclosure α) as well as a member recruitment model (BFS, DFS, RW, EN, or RS). We investigate the prediction accuracies for a wide range of their combinations using two measures (AUC and PPV_k) and two training schemes.

First, we examine the performance of our algorithm with $4 \rightarrow 1$ cross-prediction for each combination of ρ and α, all member recruitment models and all five university data sets (see Figure 2). Based on the minimal (lower triangle) and maximal (upper triangle) AUC and PPV_k values, we see that the differences between the member recruitment models are small in most cases. The AUC values are above 0.85 for all combinations with $\rho \geq 0.5$ and $\alpha \geq 0.4$ in the case of `UNC`, `Princeton`, `Georgetown` and `Oklahoma`, for all member recruitment models except the BFS. This implies that in most cases the prediction is considerably better than random guessing. The PPV_k is at least 0.4 for the same range of ρ and α and in the case of `UNC`, `Georgetown`, and `Oklahoma`, and for all member recruitment models except the BFS and the DFS. A value of 0.4 means that when selecting the k samples with the highest prediction values, at least 40% of them indeed represent two non-members that know each other. To interpret this value correctly, we have to note that our data set shows a striking imbalance which makes prediction difficult. While there is a huge number of node pairs that could be linked by an edge, there are only a few pairs which are truly linked. More precisely, depending on the chosen member recruitment model and on the ρ membership and α disclosure parameters, the ratio between the number of edges and non-edges lies between 0.0002 and 0.03 for four out of five university networks.

Second, in the $1 \rightarrow 1$ cross-prediction setting, we evaluate how reliable the predictions are if the random forest is trained on only one network at $\rho = \alpha = 0.5$. Given the coverage of Facebook and the heavy usage of the friend finder application by both novice members and experienced users of the platform, these estimates of ρ and α are rather conservative. Figure 3 shows the corresponding prediction accuracy. On the diagonal, we plot as reference the prediction accuracy when we train and test on the same network, while the off-diagonal elements correspond to the cross-prediction case. It can be seen that some data sets are easy to predict, namely `Oklahoma` and `UNC`, while `Caltech` is hard to predict based on any of the four other data sets. Furthermore, if the classifier is trained on `Caltech` data, the predictions are consistently the worst among all cross-predictions. The intuition behind this observation is that `Caltech` is a clear outlier among the used data sets because it is by far the smallest and the densest.

5 Conclusions

Our work reveals the potential that social network platforms have in predicting links between non-members based only on the connection patterns of the befriended members and their e-mail contacts to non-members. Accordingly, people without a profile on an online social network – such as Facebook – are not immune against data mining based on data available to the given platform. This finding is based solely on topological features, i.e. we used purely contact data and no user attributes. If we had access to more comprehensive data including details about the members like their age, location, or occupation, then our inference could be considerably more accurate.

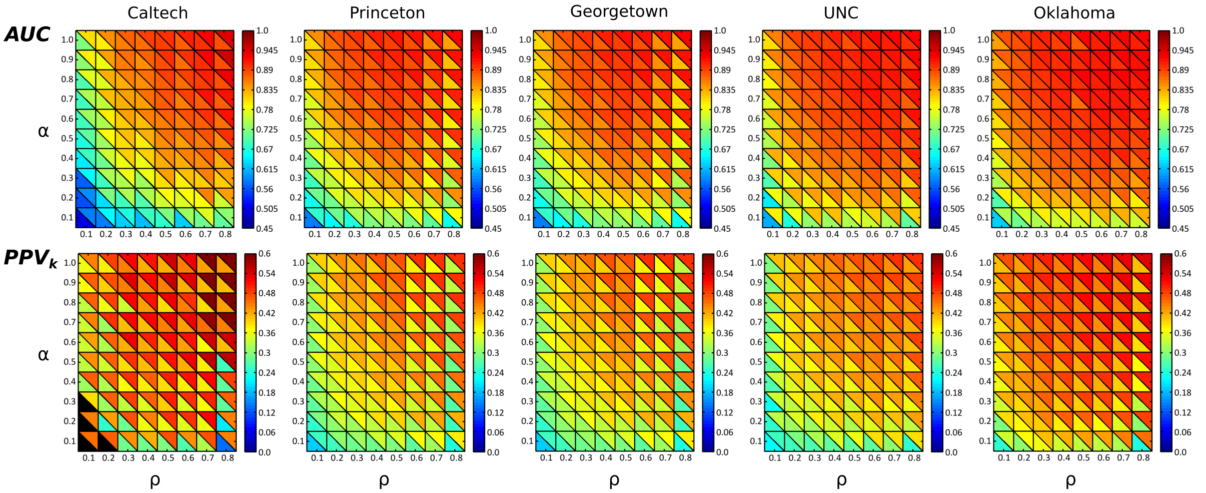

Figure 2: Minimal (lower triangle) and maximal (upper triangle) prediction accuracy in the 4 → 1 training scheme for all five member recruitment models as a function of the membership parameter ρ and the disclosure α. Upper row: AUC; lower row: PPV_k; black triangles denote data points where PPV_k was smaller than the according fraction of positive samples among all samples, i.e., it was worse than expected by chance. Figure reprinted from [2] with permission.

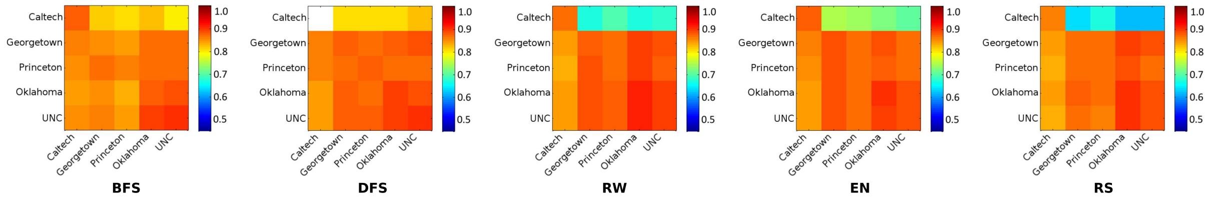

Figure 3: 1 → 1 cross-prediction accuracy: AUC values for each of the five member recruitment models when $\rho = \alpha = 0.5$. The y and x-axis show on which network the random forest was trained and tested, respectively. The white field indicates that there were too few edge samples to reasonably train the classifier. Figure reprinted from [2] with permission.

Acknowledgments.

EÁH and KAZ were supported by the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences, University of Heidelberg, Germany, which is funded by the German Excellence Initiative (GSC 220). FAH and KAZ were supported by a fellowship of the Marsilius Kolleg, University of Heidelberg, Germany.

References

[1] Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)

[2] Horvát, E.-Á., Hanselmann, M., Hamprecht, F., Zweig, K.A.: One plus one makes three (for social networks). PLoS ONE 7(4), e34740 (2012)

[3] Jernigan, C., Mistree, B.: Gaydar: Facebook friendships expose sexual orientation. First Monday [Online] 14 (2009)

[4] Kumar, R., Novak, J., Tomkins, A.: Structure and evolution of online social networks. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 611–617 (2006)

[5] Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology 58, 1019–1031 (2007)

[6] Lichtenwalter, R.N., Lussier, J.T., Chawla, N.V.: New perspectives and methods in link prediction. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 243–252 (2010)

[7] Lindamood, J., Heatherly, R., Kantarcioglu, M., Thuraisingham, B.: Inferring private information using social network data. In: Proceedings of the 18th International Conference on World Wide Web, pp. 1145–1146 (2009)

[8] Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattarcharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 7th ACM Sigcomm Conference on Internet Measurement, pp. 29–42 (2007)

[9] Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: inferring user profiles in online social networks. In: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, pp. 251–260 (2010)

[10] Newman, M.E.J., Park, J.: Why social networks are different from other types of networks. Physical Review E 68, 036122 (2003)

[11] Traud, A., Kelsic, E., Mucha, P., Porter, M.: Comparing community structure to characteristics in online collegiate social networks. SIAM Review 53, 526–543 (2011)

[12] Ugander, J., Backstrom, L., Marlow, C., Kleinberg, J.: Structural diversity in social contagion. Proceedings of the National Academy of Sciences 109(16): 5962–5966 (2012)

[13] Wang, C., Satuluri, V., Parthasarathy, S.: Local probabilistic models for link prediction. In: 7th IEEE International Conference on Data Mining, pp. 322–331 (2007)

[14] Zheleva, E., Getoor, L.: To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In: Proceedings of the 18th International Conference on World Wide Web, pp. 531–540 (2009)