OPEN ACCESS Freely available online

PLoS one

One Plus One Makes Three (for Social Networks)

Emöke-Ágnes Horvát[1], Michael Hanselmann[2], Fred A. Hamprecht[2,3], Katharina A. Zweig[1,3]*

1 Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg, Heidelberg, Germany, **2** Heidelberg Collaboratory for Image Processing (HCI), University of Heidelberg, Heidelberg, Germany, **3** Marsilius Kolleg, University of Heidelberg, Heidelberg, Germany

Abstract

Members of social network platforms often choose to reveal private information, and thus sacrifice some of their privacy, in exchange for the manifold opportunities and amenities offered by such platforms. In this article, we show that the seemingly innocuous combination of knowledge of confirmed contacts between members on the one hand and their email contacts to non-members on the other hand provides enough information to deduce a substantial proportion of relationships between *non*-members. Using machine learning we achieve an area under the (receiver operating characteristic) curve (AUC) of at least 0.85 for predicting whether two non-members known by the same member are connected or not, even for conservative estimates of the overall proportion of members, and the proportion of members disclosing their contacts.

Citation: Horvát E-Á, Hanselmann M, Hamprecht FA, Zweig KA (2012) One Plus One Makes Three (for Social Networks). PLoS ONE 7(4): e34740. doi:10.1371/journal.pone.0034740

Editor: Sergio Gómez, Universitat Rovira i Virgili, Spain

Received November 15, 2011; **Accepted** March 5, 2012; **Published** April 6, 2012

Funding: EAH and KAZ are supported by the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences (http://www.mathcomp.uni-heidelberg.de/), 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 (http://www.marsilius-kolleg.uni-heidelberg.de/), University of Heidelberg. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: nina@ninasnet.de

Introduction

Some individuals prefer to keep intimate details such as their political preferences or sexual orientation private. Recent results suggest that such details can nonetheless be inferred with high probability if a sufficient number of confirmed contacts in a social network chooses to reveal *their* details [1–4]. As a consequence, some of the more circumspect choose to stay away from social network platforms such as Facebook in the belief that this will help protect their privacy. In this article, we show that such an assumption is no longer valid: with the help of machine learning, social network operators can make predictions regarding the acquaintance or lack thereof between two *non*-members with a high rate of success. To our knowledge these are the first results on the potential of social network platforms to infer relationships between *non*-members.

Inference of undisclosed, unobserved, or future contacts (the ''edges'') between people or agents (the ''nodes'') is known as the ''link prediction'' problem [5–7]. It is a difficult problem mainly because the imbalance between possible and realized future edges is extremely high in most cases [8,9]. In contrast, the prediction of some properties of given links, e.g. the sign of the weight on a link [10] is simpler because the problem is typically more balanced. Link prediction was mostly approached with unsupervised [11] and recently also with supervised learning methods [12–14]. Inference was done both using solely structural measures based on the network topology [15,16] but also by additionally taking into account the nodes' attributes [17–19]. The most common setting of the link prediction problem is, given an evolving network at an early stage, to predict newly acquired edges at a later stage. The success of link prediction has usually been estimated by cross-validation within the same network [20,21]. This typically implies a dependence between training and test data and, hence, an overly optimistic estimate of the accuracy of an algorithm. To our knowledge, we present the first link prediction work where learning and testing are performed on entirely independent networks.

Methods

The Problem

All members of society can be seen as nodes in an unobservable social graph. This latent social graph is dynamic and extremely complex, with edges of widely differing quality (two people may be kindred, or engaged, or work together, they may like or dislike each other, etc.). From the point of view of a social network platform like Facebook, the set of all people can be divided into a fraction $0 \leq \rho \leq 1$ of members and $1-\rho$ of non-members. The multi-faceted relationships between people are much simplified, in an extreme case into mere binary form: two members may declare a ''friendship'' which is then represented by an edge in the set E_I. In reality, social networks typically have access to more information that allows to estimate the quality of an edge (its strength, its asymmetry, etc.) especially if they integrate a messaging service. Additionally, a fraction $0 \leq \alpha \leq 1$ of all platform members may also share their contacts to non-members, e.g., by uploading their email address book (Figure 1). Social network platforms then have direct access to two different sets of relationships: on the one hand, the mutually confirmed contacts between platform members (E_I); and on the other hand, their members' unilateral declarations of their acquaintance with non-members (E_B). The edges in both E_I and E_B are an abstraction and a subset of the edges in the latent social graph. The central

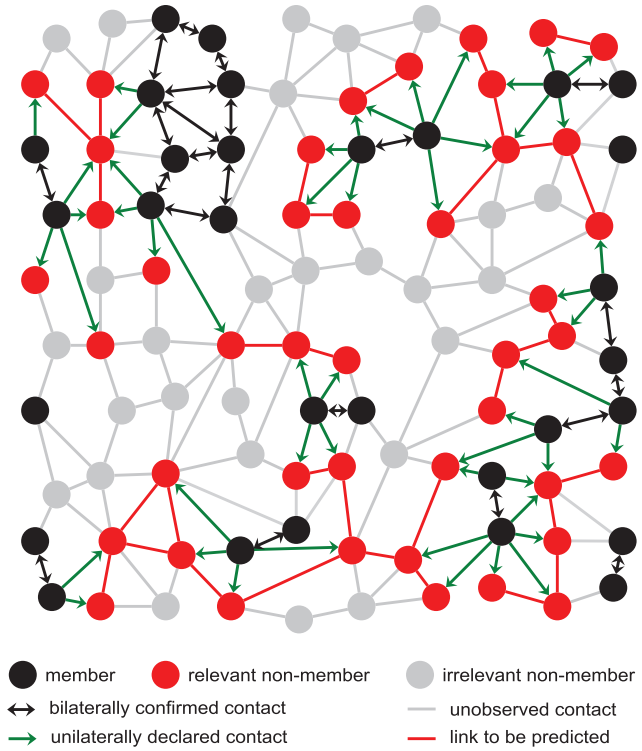

Figure 1. Definitions and examples. Any social network platform divides society into two sets: the set of members M (black nodes) and of non-members $\overline{M}$. In our toy example 30 of 100 individuals, i.e. a fraction of $\rho=0.3$, are members. The relevant subset $\overline{M_R}$ of non-members (red nodes) that are in contact with at least one member is distinguished from other non-members (gray nodes). 15 of the 30 members, i.e., a fraction of $\alpha=0.5$, have disclosed their outside social contacts. The knowledge of the set of edges E_I between members (black, bi-directed) and the set of edges E_B (green) to non-members is enough to infer a substantial fraction of edges between non-members (red edges).
doi:10.1371/journal.pone.0034740.g001

question of this article is to what extent the acquaintance of two people who are *both* non-members can be predicted.

For the very reason that the latent social graph is fundamentally unobservable, both we and a social network operator with similar aims as described here needs to impute the missing information to admit a machine learning procedure. The approach we choose is to use the *observed* part of a social network – say, the Facebook network of all students at a given university – and presume it represents the *complete* (and unobservable) social graph of a hypothetical community. In other words, the edges in this social graph are considered the ground truth. We then proceed to partition this community into a set of members and non-members by a number of member recruitment models outlined below which represent a broad range of potential strategies by which people choose to become members. Finally, we predict the existence or otherwise of an edge between any two non-members and evaluate the accuracy of these predictions with respect to the ground truth.

Ground Truth Imputation

In line with what would be available to a social network operator, we use real social networks, in this case real-world Facebook friendship networks representing the students from five different US universities [22]. Figure 2 shows a comparison of their number n of members, their average degree (the average number of friends a member has), their density $2m/(n(n-1))$ where m is the number of friends, and their average clustering coefficient (the average probability that two friends of a member are friends themselves [23]).

The ground truth imputation comprises three steps. In the first step the platform penetration percentage ρ is modeled. The percentage of members in a given population varies strongly with the type of social network platform and the social community of interest. According to Facebook, it had more than 800 million active users in November 2011 [24], while the number of internet users worldwide is estimated at over 2 billion [25]. Thus, roughly 40% of all internet users are active Facebook members. Around 25% of all Facebook users are US citizens. Assuming that each Facebook account represents one individual, we can estimate that over 70% of all North American internet users are members of Facebook. For certain social strata the percentage of Facebook users is known exactly: one study showed that already back in 2005 over 60% of the undergraduates of the Carnegie Mellon University were members of the platform [26]. Later, in 2009 around 80% of all interviewed students of the University of Illinois of Chicago [27] and 85% of a polled contemporary Canadian sample [28] were Facebook members. The platform penetration parameter $0\leq\rho\leq1$ thus reflects different membership densities and allows to model different social network platforms and their acceptance in different communities.

Given a real Facebook friendship network and a choice of ρ, the second step of the ground truth imputation is to partition the nodes of the network into members and non-members. For this we need models for how people choose to become members of a platform which we call *member recruitment models*. An analysis of the evolution of online social networks [29] suggests that a network platform recruits its members through a mixture of online mediated invitations by friends who already are members, and independent decisions by individuals who are not yet friends of a member. Since the actual member recruitment process is unknown and probably also depends on the group of people that is considered (e.g. college students vs. employees), we have emulated the growth of social network platforms using processes ranging from strongly dependent to purely independent decisions. All models start with labeling a node chosen uniformly at random as the first member. Strongly dependent decisions are modeled by processes in which only people who know at least one member will join the network. In a breadth first search (BFS) model all friends of the first member are labeled as members after which all *their* friends are labeled and so on. In a depth first search (DFS) model a randomly chosen friend of the first member joins the platform after which a randomly chosen friend of the new member joins and so on recursively. Less dependent decisions are modeled by a random walk (RW) which is restarted from a new node as soon as a friend of a new member is chosen which is already a member. The ego networks selection (EN) model joins the independent decision of some randomly chosen seed members with the dependent decision of their direct friends. Purely random selection of members (RS) is based entirely on independent decisions modeled by the random selection of a set of members. These member recruitment models are described in more detail in Text S1. For an analysis of the structural properties of the partitions obtained with different member recruitment models see Figures S1, S2, and S3. Figure 3 shows the resulting partitions of a toy graph under all five models. We show below that our main findings are robust with respect to the specific choice of the member recruitment model.

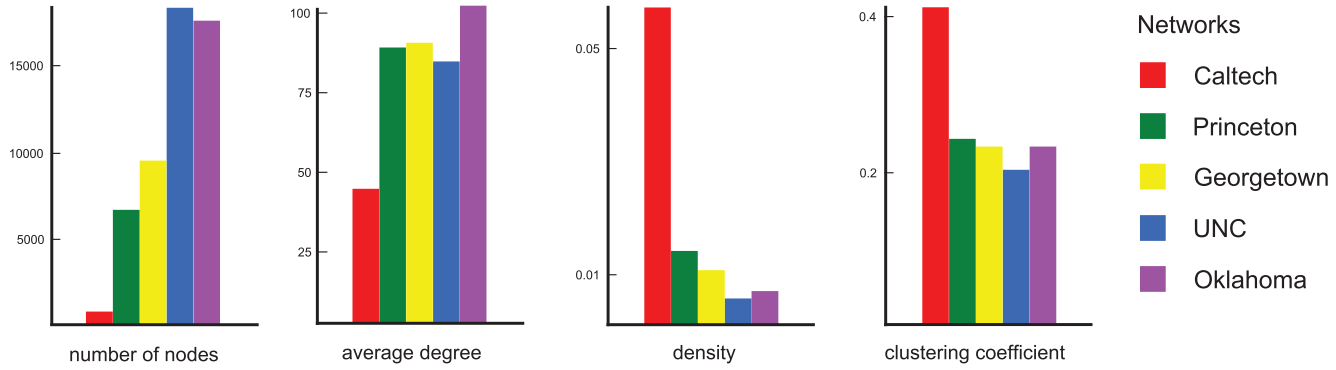

Figure 2. Comparison of basic network analytic statistics of the five data sets obtained from Traud et al. [22].
doi:10.1371/journal.pone.0034740.g002

In the third step, a so-called disclosure parameter $0 \leq \alpha \leq 1$ is chosen to model the probability with which a member opens her email address book to the platform, e.g. through revealing her email contacts. We consider that a member who revealed her email contact list shared thereby *all* of her contacts justified by the feature of the platform allowing for easy automatic uploading of the entire email address book. α governs the fraction of connections between the member and non-member sets. As such, it is a key ingredient of the ground truth imputation.

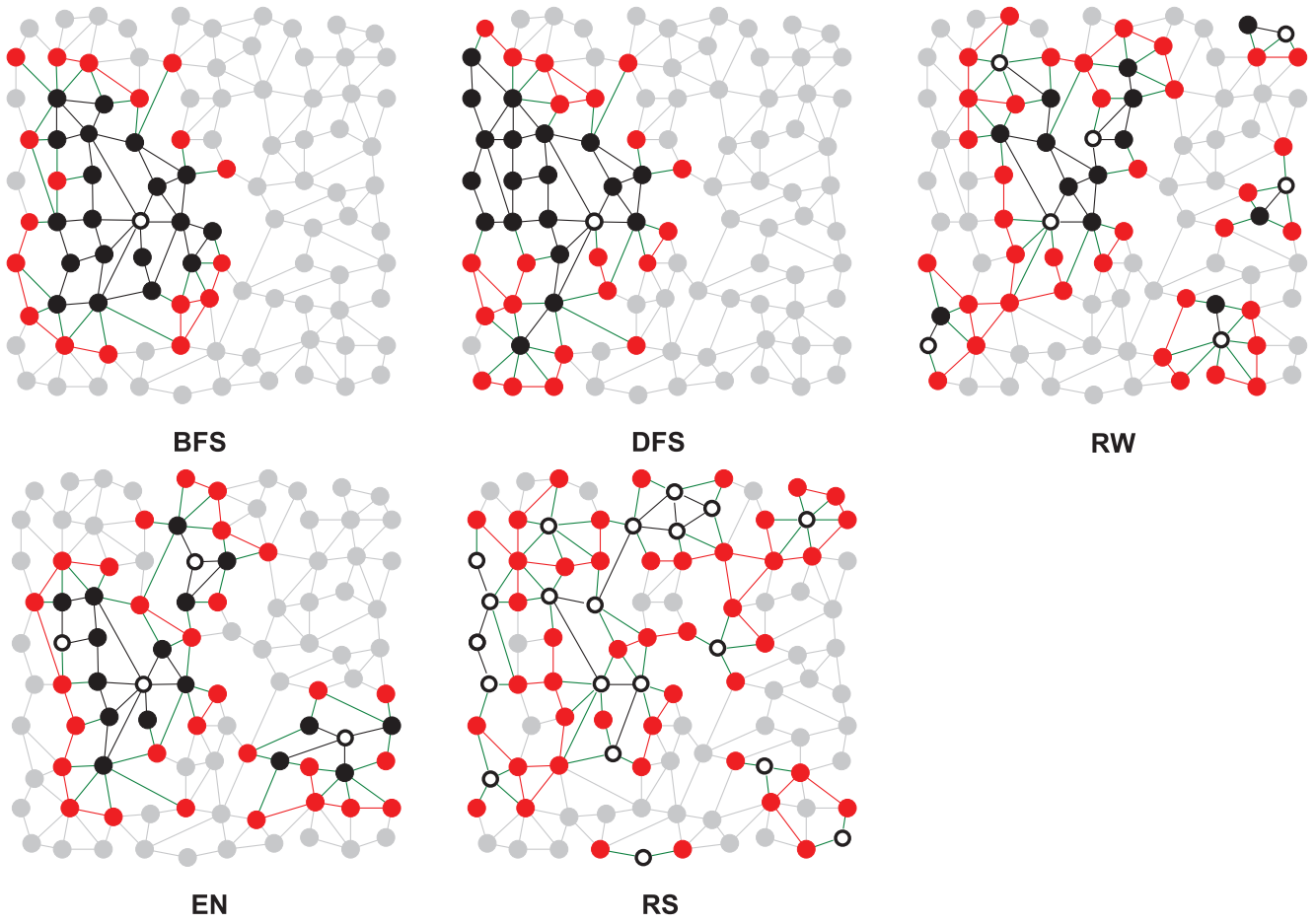

Figure 3. Membership propagation in a toy example according to different propagation models. Note that real social networks exhibit more long-range edges. Examples for the platform penetration value $\rho = 0.2$ show the nodes from which the propagation started (black nodes with white core). Other members are marked black and relevant non-members red; for ease of reading arrows are not displayed, but black edges are bidirectional while green edges point from black to red nodes. With BFS and DFS the network is explored starting from one node (denoted by a white circle); with RW and EN there are more nodes from which the propagation is launched; and finally, for RS all selected nodes can be seen as starting nodes.
doi:10.1371/journal.pone.0034740.g003

Given an underlying graph $G=(V,E)$ with nodes V and edges E, the simulated member recruitment model results in a subset of nodes $M \subset V$ that are considered members, and a set of non-members $\overline{M}=V\backslash M$. We will only focus on the set $\overline{M_R} \subset \overline{M}=\{y \in \overline{M} | (xy) \in E_B,\ x \in M\}$ of *relevant* non-members whose email address has been disclosed by at least one member (see Figure 1). These node sets induce the edge sets E_I and E_B as defined above, and additionally the edges between non-members, which are not directly accessible to the platform, and are at the core of our interest and prediction efforts. Let $G'=(M \cup \overline{M_R}, E_I \cup E_B)$ denote the new graph containing all the structural information that is assumed to be known by a given social network platform (black and red nodes, and black and green edges in Figure 1).

The ground truth imputation is thus determined by the choice of the percentage of individuals ρ deciding to become members of the social network, the member recruitment model (BFS, DFS, RW, EN, RS), and the choice of α, the propensity of members to disclose their contacts with non-members.

Feature Extraction

To predict whether two non-members v,w are connected, we compute 15 topological graph features of the network around v and w in G' on which the prediction is based. We deduce the features from relational knowledge because the data at hand is anonymized and therefore no node attributes are available. The exact choice of features is rooted in the known structural properties of (online) social networks [30,31]. The intuition that two people sharing common friends are likely to be friends themselves motivates including a feature that counts the absolute number of common neighbors v and w have. However, the absolute number of common neighbors might be misleading if v has just a few neighbors, while w has many. Thus, we add three normalized versions of the number of common neighbors where the normalization is done by the smaller degree, the larger degree and the number of nodes which are neighboring at least one of the two nodes (the so-called *Jaccard coefficient*). The typically high assortativity (measuring the likelihood for nodes to connect to other nodes with similar degrees [32]) and the significant local clustering [33] of nodes in online social networks justifies focusing on the average degree and the clustering coefficient of the common neighbours of v and w. The community structure of social networks [34] leads us to construct several features that reflect the interconnectedness of the member side neighbors of the two nodes as illustrated by Figure 4. Finally, we count the absolute number of distinct paths between v and w in G' with exactly three edges. For a precise description of the features see Text S2.

For each pair of non-members these scalars are stored in a 15 dimensional feature vector. A feature vector relating to two connected non-members is called a positive sample, and one describing unconnected non-members is a negative sample. Based on this vector, supervised machine learning is used to predict which pairs of non-members are connected (acquainted) and which ones are not.

The Prediction Algorithm

Supervised learning requires a training set on which the classifier's parameters are adjusted. Its performance is then evaluated on an independent test set. We restrict our predictions to those pairs of non-members with at least one common neighbor among the members. In this respect we follow similar approaches which restricted link predictions to pairs of nodes with a maximum distance of two [9,12]. Our focus is thus on predicting whether two non-member friends of a member are friends themselves, i.e. whether a pair of non-members is contained as an edge or not. We

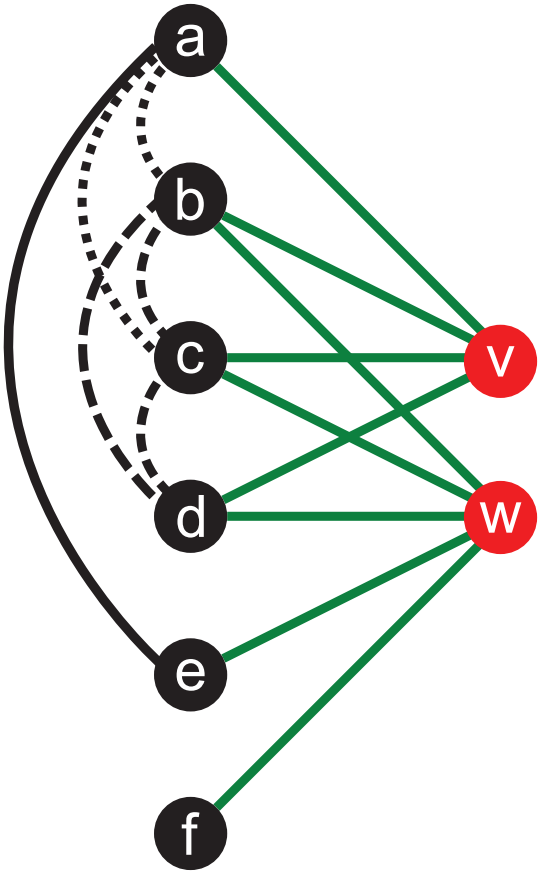

Figure 4. Features based on different edge sets between the exclusive, joint, and common neighborhoods of v and w. All left-hand nodes belong to the joint neighborhood of v and w. a is exclusive to v, while e,f are exclusive to w, and b,c,d are common neighbors of both. Our features comprise the absolute number of edges between common neighbors (black, dashed edges), exclusive neighbors (black, straight edge), joint neighborhood (all black edges between nodes a,b,c,d,e), and an exclusive and a common neighbor (black, dotted edges). For each of them we also added their normalized value. Normalization was done by the number of possible edges between the neighbors they have.
doi:10.1371/journal.pone.0034740.g004

employ the random forest classifier [35], an ensemble of decision trees that has previously been used for link prediction in dynamic networks [9,11,12]. For a more detailed description see Text S3.

Once the random forest has been trained it can be applied to the test set, and edges with a probability higher than some threshold are predicted to exist. This prediction can then be compared to the ground truth.

Accuracy Measures for Prediction

A good classification result is characterized by a high sensitivity (probability of predicting an edge that truly exists) and high specificity (probability of predicting the absence of an edge that truly doesn't exist). In the following we use two classic accuracy measures for the link prediction problem, the AUC and the PPV_k which combine sensitivity and specificity [13,36]: Varying the threshold allows to trade-off sensitivity vs. specificity. The receiver operating characteristic (ROC curve) shows the *sensitivity* against $1-$*specificity* plot. The area under this curve (AUC) is a scalar performance measure that aggregates the prediction accuracy over all possible settings of this threshold. A perfect predictor achieves

an AUC of 1 while random guessing in a two-class problem yields a value of 0.5.

While the AUC measures the accuracy over the full range of possible thresholds, the PPV_k is based on a specific threshold: let k denote the number of positive samples in the test set, i.e., the non-member pairs connected by an edge, and let all samples of the test set, i.e., all non-member pairs having at least one common member friend, be ordered non-increasingly by their prediction value. The PPV_k introduced by [11] is defined as the percentage of correctly classified positive samples among the first k samples in the ranking, and is thus also equal to the sensitivity achieved by predicting these k samples to be edges. It can be shown that the specificity is linearly dependent on PPV_k and thus both measures are captured by it. The higher the value the more the number of positive samples is enriched among the k highest-ranked samples. Note that the PPV_k should always be at least as large as the overall fraction of positive edges among all edges. Otherwise, the prediction algorithm performs worse than a naive algorithm in which k samples are drawn uniformly at random from all samples and predicted to be edges.

Training Schemes

All prior work on link prediction with high imbalance between possible and realized future edges that we are aware of uses cross-validation where training and testing are achieved within a single network. In reality, however, one would want an approach that can be trained once and then allows to predict edges in independent networks (cross-prediction). In general, cross-validation within a single network tends to be overly optimistic in its results since training and test samples may be dependent. It is thus more meaningful to test the cross-prediction performance of a prediction algorithm. Accordingly, we have devised the following training scheme (see Text S3):

1. cross-prediction $4\rightarrow1$: The classifier was trained on subsets of the samples from four of the five data sets and the test set consisted of all samples from the fifth data set. This is an instance of cross-validation across networks, rather than within a network. To stress this difference, we use the term cross-prediction in the following. To be able to compare classification results obtained on different data sets and for varying parameters ρ and α, we subsampled the available training data in the following way: First, 1200 positive samples and $(1/\eta-1)\cdot1200$ negative samples were selected at random. Here, η denotes the fraction of positive samples among all samples. To obtain balanced training sets when constructing individual trees of the random forest, we then randomly selected 1200 samples from both classes from the reduced set. We applied this procedure to each of the four training data sets to obtain a total of 4800 positive samples and the corresponding number of negative samples. Since especially the Caltech data set did not always provide enough samples, we oversampled the available data whenever necessary (that is by sampling with replacement). Each tree was trained with a balanced data set of 4800 samples from both classes.

2. cross-prediction $1\rightarrow1$: For each pair of data sets, the classifier was trained on one and evaluated on the other. The same sampling scheme was used as above and the trees were trained with 1200 samples per class.

The $4\rightarrow1$ training scheme is less prone to overfitting by training on four different networks, while the $1\rightarrow1$ scheme was used to find out whether a single known network contains enough information to obtain high-quality predictions.

Results

For all member recruitment models, the ratio between the number of positive and negative samples is very small and lies between 0.0002 and 0.03 for four out of five of the university networks. This imbalance of the two classes is in the typical range of imbalances for various link prediction problems [9,11,36,37] and seems to determine the hardness of the problem.

As argued above, the prediction accuracy of an intra-network cross-validation should upper-bound the prediction accuracy of an inter-network cross-prediction approach. Thus, as a base line, Figure 5 visualizes the prediction accuracy as measured by the AUC, using different member recruitment models in conjunction with cross-validation on the UNC data set. The general pattern is that the prediction accuracy increases with ρ and α. In other words, the greater the percentage of members and the higher their propensity to share their email contacts, the easier it is to predict the network between non-members. One exception to this pattern is the BFS model whose prediction accuracy shows a maximum for $\rho\sim0.5$. The behavior of the AUC and the PPV_k are very consistent over all member recruitment models for all university data sets, implying that the exact model of the member recruitment model is not crucial (see Figures S4 and S5).

The $4\rightarrow1$ cross-prediction training scheme results in AUC values of at least 0.85 for all combinations with $\rho\geq0.5$ and $\alpha\geq0.4$ in the case of UNC, Princeton, Georgetown and Oklahoma, for all but the BFS member recruitment model. Similarly, the PPV_k is at least 0.4 for the same range of ρ and α, on UNC, Georgetown and Oklahoma, and for all but the BFS and the DFS member recruitment models. That means, if for each data point we select the k samples with the highest prediction values, at least 40% of them indeed represent two non-members that know each other. Figure 6 shows for each combination of ρ and α and all five data sets the minimal (lower triangle) and maximal (upper triangle)

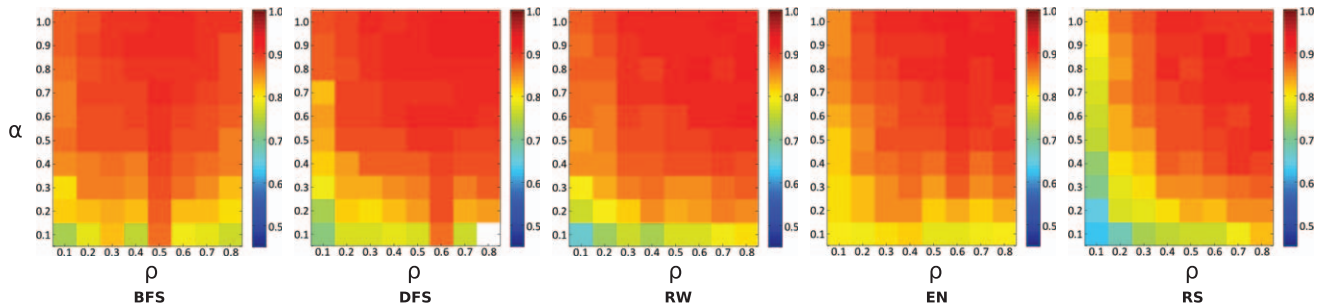

Figure 5. Prediction accuracy (AUC) of samples based on all member recruitment models in the cross-validation training scheme applied to UNC data. The white square denotes a data point where there was not enough data to make the prediction.
doi:10.1371/journal.pone.0034740.g005

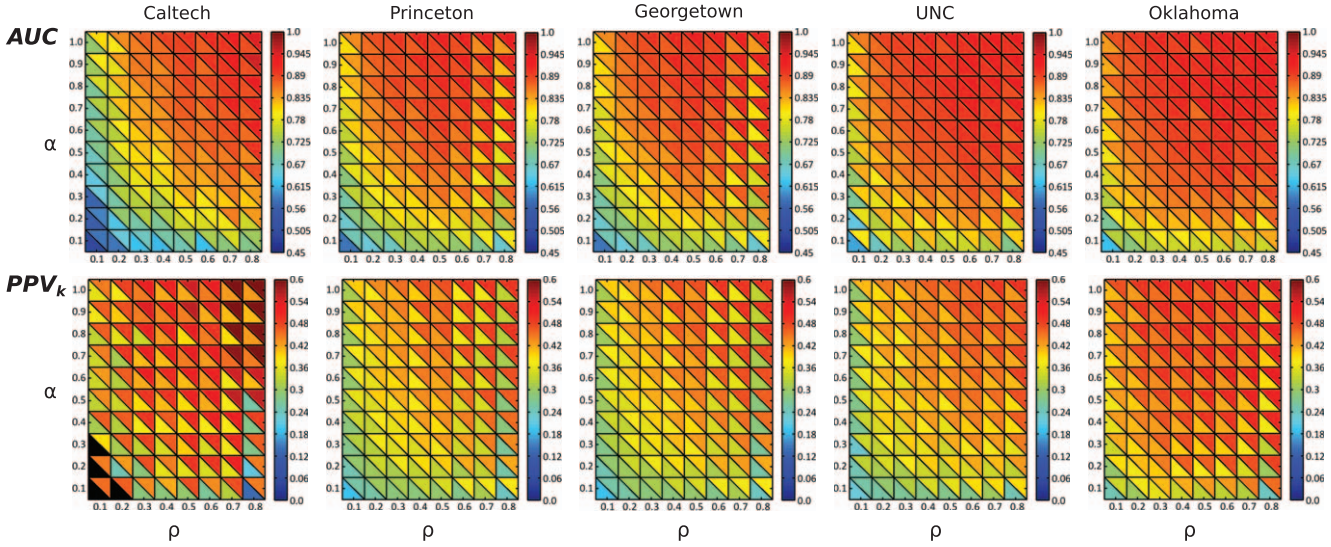

Figure 6. 4 → 1 cross-prediction accuracy. Minimal (lower triangle) and maximal (upper triangle) prediction accuracy for all five member recruitment models are shown as a function of platform penetration ρ and the disclosure parameter α. Upper row: AUC; lower row: PPV_k; black triangles denote data points where PPV_k was smaller than the according fraction of positive samples among all samples.
doi:10.1371/journal.pone.0034740.g006

AUC and PPV_k value over all member recruitment models. For the complete results refer to Figures S6 and S7. It can clearly be seen again that the differences depending on the member recruitment model are small in most cases. The one part where the difference is most pronounced is for $\rho \geq 0.6$ for Georgetown and Princeton, where the BFS member recruitment model results in notably worse predictions than the other member recruitment models.

The $4 \rightarrow 1$ setting is more robust against overfitting since it learns from four independent networks. To see if one can reliably predict with even fewer training data, we also evaluated in the $1 \rightarrow 1$ setting how good the predictions remain if the random forest is trained on only one network at $\rho = \alpha = 0.5$. For the case of Facebook in particular, the estimates of ρ and α appear conservative, because of the pervasiveness that this platform has achieved, and the simplicity with which members can upload their full email address book. At the time of writing, Facebook still asks novice members, upon the creation of their account, to supply the password of their email provider to parse all their email contacts. The choice of $\alpha = 0.5$ is supported by an informal survey among 100 members of Facebook. Additionally, for the particular case of Facebook, the number of members whose email contacts are known can be expected to further grow with the integrated email service that has recently been introduced. Figure 7 shows the prediction accuracy in the $1 \rightarrow 1$ training scheme. On the diagonal, the cross-validation value is plotted as a reference and it can be seen that the AUC values in this cross-prediction training scheme are around the same value as those of the cross-validation within the same network. It can also 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. Based on the network statistics shown in Figure 2, it can be seen that UNC and Oklahoma are the two largest networks while Caltech is the smallest, with around half the average degree of the other four networks. At the same time, its clustering coefficient is almost twice as large as that of the others. Thus, the prediction accuracy as measured by AUC is almost as good in $1 \rightarrow 1$-cross prediction as in within-network cross-validation, provided the network used for training is structurally not too different from the one to be predicted.

In summary, we achieve an $AUC \geq 0.85$ and a PPV_k of more than 40% in a cross-prediction training scheme for realistic parameter values of $\rho \geq 0.5$ and $\alpha \geq 0.4$. The high AUC value implies that the prediction is considerably better than random guessing. The $PPV_k > 0.4$ means that, if there is an estimate on the number k of edges to expect between all pairs of non-members

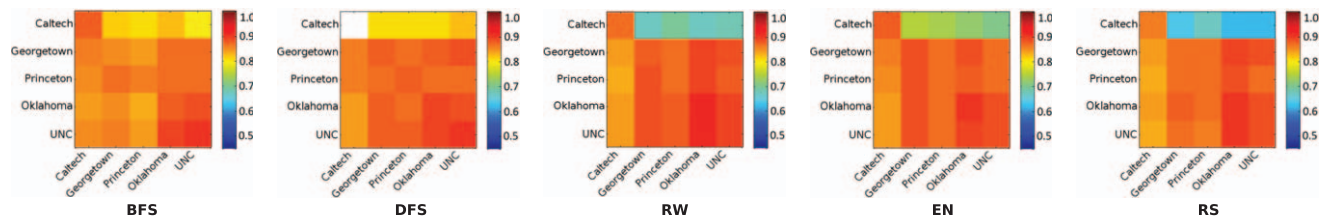

Figure 7. 1 → 1 cross-prediction accuracy. AUC values for each of the five member recruitment models at $\rho = \alpha = 0.5$. The y and x-axis show on which network the random forest was trained and tested, respectively. The white field indicates that there were too few edge samples to reasonably train the classifier.
doi:10.1371/journal.pone.0034740.g007

that share a member friend, then 40% of the predicted edges are indeed edges and of these edges between non-members 40% are predicted. Note that the restriction to predicting only connections between non-members sharing a member friend leaves us with at least 75% of all befriended non-members (in almost all cases this percentage is above 85%, Table S1).

One final remark concerns the variability of prediction accuracy among the member recruitment models. We found that the prediction accuracy is rather independent of how individuals decide to become a member of an online social network platform, but that some percentage of independent decisions like in the *RW*, *EN*, and *RS* model helps the platform to explore the latent social network more efficiently.

Discussion

The above results give us a good indication of how accurately one can predict links between non-members of a social network platform, based only on information extracted from the friendship and email contact information of their members. Additionally, the quite high prediction accuracy achieved in inter-network cross-prediction as compared with intra-network cross-validation is astonishing. As always, machine learning can only work if the training data is representative of the observations for which predictions are required in the future. This expectation is borne out by our experiments which show that cross-prediction works best if training and test set are similar in terms of basic network statistics.

Note that we used purely topological or structural indicators, i.e., we only exploited the presence or absence of edges in the computation of features for prediction. Social network platform operators, however, typically have access to much more detailed information on nodes such as the age, sex and (approximate) location of their members; and if they provide messaging services they can infer the quality of an acquaintance from its communication pattern. Including such information into the features will likely improve prediction accuracy.

The ground truth imputation on which the results are based relies on three important modeling decisions which need to be discussed. First, we imply that non-members are similar to members in terms of the revealed network characteristics. Two studies from 2006 and 2009 indicate that there are indeed statistically significant differences between members and non-members among university students: these differences concern age, ethnicity, and gender, but not important social factors such as life satisfaction, social trust, or privacy concerns [38,39]. Although the sociability of members and non-members was not directly assessed, these studies give no indication that members and non-member differ significantly in the structure of their contact networks. Second, since the contact network between members and non-members of no social network platform is available, we take the known Facebook friendship network as a proxy for the structure of the email contact network between members and non-members. This is justified by the fact that both belong to the large set of social networks with scale-free degree distribution, high clustering coefficient, small-world behavior, and a positive assortativity [31,40,41]. Third, we only take into account pure member recruitment models which might not be realistic apiece. The surprising result that the choice of member recruitment models does not alter the main conclusions shows that the analysis of the pure models does not constrain the approach. Even for BFS, which makes good predictions hardest, good results are obtained. This implies that however individuals decide to join an online social network, the unilateral declaration by members of their contacts with non-members allows social network platforms to gain substantial insight into the relationships of non-members. This increase in coverage thanks to link prediction will be most successful if the individual members' decisions to join the network exhibit some independence - a knowledge that could be exploited by the platform when elaborating new recruitment strategies.

Altogether, our results indicate that knowledge of the social network between members of a platform along with part of the contacts to non-members is sufficient to infer a substantial part of the network between non-members. This perhaps surprising finding has been afforded by leveraging relational information provided by members in a unilateral, non-consensual manner. Ultimately, it evokes the question of the ownership and exploitation of relational data in the information age.

Supporting Information

Figure S1 Some structural properties of the networks resulting from applying different member recruitment models on the Oklahoma **data set in dependence of ρ for $\alpha = 1$.**
(EPS)

Figure S2 The coverage of all member recruitment models in dependence of ρ for the networks of the five universities. For all member recruitment models and all data, $\alpha = 0.5$, i.e., half of the members have shared their email contacts.
(EPS)

Figure S3 The coverage of all member recruitment models in dependence of α for the networks of the five universities. For all member recruitment models and all data, $\rho = 0.5$, i.e., half of the nodes are selected members.
(EPS)

Figure S4 Shown are the *AUC* values for all $\rho - \alpha$ combinations in all five data sets. The random forest was trained on 90% of the samples within the same data set and tested on the remaining 10% (9/10-cross validation). The procedure was repeated 10 times with randomly chosen training sets, and the values of all runs were averaged. White squares denote data points where there were not enough edge samples to do the learning.
(EPS)

Figure S5 Shown are the PPV_k values for all $\rho - \alpha$ combinations in all five data sets. The random forest was trained on 90% of the samples within the same data set and tested on the remaining 10% (9/10-cross validation). The procedure was repeated 10 times with randomly chosen training sets, and the values of all runs were averaged. White squares denote data points where there were not enough edge samples to do the learning; black squares denote data points where PPV_k was smaller than the according fraction of positive samples among all samples.
(EPS)

Figure S6 Shown are the *AUC* values for all $\rho - \alpha$ combinations in all five data sets. To predict samples from any of the data sets, the random forest was trained on 1200 samples each of the other four data sets (4→1 prediction). White squares denote data points where there were not enough edge samples to do the learning.
(EPS)

Figure S7 Shown are the PPV_k values for all $\rho - \alpha$ combinations in all five data sets. To predict samples from any of the data sets, the random forest was trained on 1200 samples each of the other four data sets (4→1 prediction). White squares denote data points where there were not enough edge

samples to do the learning; black squares denote data points where PPV_k was smaller than the according fraction of positive samples among all samples.
(EPS)

Table S1 Percentage of edges between two non-members that commonly know at least one member of all edges between two non-members. Shown is the average of ten runs for $\rho = 0.5$ and $\alpha = 0.5$.
(PDF)

Text S1 Details about the experimental setting.
(PDF)

Text S2 Details regarding feature extraction.
(PDF)

Text S3 More information on the learning procedure with random forests.
(PDF)

Author Contributions

Conceived and designed the experiments: EAH MH FAH KAZ. Performed the experiments: EAH MH. Analyzed the data: EAH MH. Wrote the paper: EAH MH FAH KAZ. Defined the research question: FAH. Coordinated the writing of the article: KAZ.

References

1. Jernigan C, Mistree B (2009) Gaydar: Facebook friendships expose sexual orientation. First Monday [Online] 14.
2. Lindamood J, Heatherly R, Kantarcioglu M, Thuraisingham B (2009) Inferring private information using social network data. In: Proceedings of the 18th International Conference on World Wide Web (WWW '09). pp 1145–1146.
3. Mislove A, Viswanath B, Gummadi KP, Druschel P (2010) 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 (WSDM '10). pp 251–260.
4. Zheleva E, Getoor L (2009) 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 (WWW '09). pp 531–540.
5. Getoor L, Diehl CP (2005) Link mining: a survey. ACM SIGKDD Explorations Newsletter 7: 3–12.
6. Dietterich TG, Domingos P, Getoor L, Muggleton S, Tadepalli P (2008) Structured machine learning: the next ten years. Machine Learning 73: 3–23.
7. Kolaczyk ED (2009) Statistical Analysis of Network Data: Methods and Models. New York: Springer.
8. Weiss GM (2004) Mining with rarity: a unifying framework. ACM SIGKDD Explorations Newsletter 6: 7–19.
9. Wang D, Pedreschi D, Song C, Giannotti F, Barabasi AL (2011) Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11). pp 1100–1108.
10. Leskovec J, Huttenlocher D, Kleinberg J (2010) Predicting positive and negative links in online social networks. In: Proceedings of the 19th International Conference on World Wide Web (WWW'10). pp 641–650.
11. Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology 58: 1019–1031.
12. Lichtenwalter RN, Lussier JT, Chawla NV (2010) New perspectives and methods in link prediction. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10). pp 243–252.
13. Wang C, Satuluri V, Parthasarathy S (2007) Local probabilistic models for link prediction. In: 7th IEEE International Conference on Data Mining (ICDM'07). pp 322–331.
14. Kashima H, Abe N (2006) A parametrized probabilistic model of network evolution for supervised link prediction. In: 6th IEEE International Conference on Data Mining (ICDM'06). pp 340–349.
15. Clauset A, Moore C, Newman MEJ (2008) Hierarchical structure and the prediction of missing links in networks. Nature 453: 98–101.
16. Redner S (2008) Networks: Teasing out the missing links. Nature 453: 47–48.
17. O'Madadhain J, Hutchins J, Smyth P (2005) Prediction and ranking algorithms for event-based network data. ACM SIGKDD Explorations Newsletter 7: 23–30.
18. Crandall DJ, Backstrom L, Cosley D, Suri S, Huttenlocher D, et al. (2010) Inferring social ties from geographic coincidences. Proceedings of the National Academy of Sciences 107: 22436–22441.
19. Hasan MA, Chaoji V, Salem S, Zaki M (2005) Link prediction using supervised learning. In: Workshop on Link Discovery: Issues, Approaches and Applications;LinkKDD-2005.
20. Zhou T, Lü L, Zhang YC (2009) Predicting missing links via local information. The European Physical Journal B 71: 623–630.
21. Liu W, Lü L (2010) Link prediction based on local random walk. Europhysics Letters 89: 58007.
22. Traud A, Kelsic E, Mucha P, Porter M (2011) Comparing community structure to characteristics in online collegiate social networks. SIAM Review 53: 526–543.
23. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393: 440–442.
24. Facebook statistics website. Available: http://www.facebook.com/press/info.php?statistics. Accessed 2011 Nov, 11.
25. Internet usage statistics website. Available: http://www.internetworldstats.com/stats.htm. Accessed 2012 Mar, 12.
26. Gross R, Acquisti A (2005) Information revelation and privacy in online social networks (The Facebook case). In: Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society (WPES'05). pp 71–80.
27. Pasek J, More E, Hargittai E (2009) Facebook and academic performance: Reconciling a media sensation with data. First Monday [Online] 14.
28. Ross C, Orr ES, Sisic M, Arseneault JM, Simmering MG, et al. (2009) Personality and motivations associated with Facebook use. Computers in Human Behavior 25: 578–586.
29. Kumar R, Novak J, Tomkins A (2006) Structure and evolution of online social networks. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06). pp 611–617.
30. Newman MEJ, Park J (2003) Why social networks are different from other types of networks. Physical Review E 68: 036122.
31. Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattarcharjee B (2007) Measurement and analysis of online social networks. In: Proceedings of the 7th ACM Sigcomm Conference on Internet Measurement (IMC'07). pp 29–42.
32. Newman MEJ (2002) Assortative mixing in networks. Physical Review Letters 89: 208701.
33. Adamic LA, Buyukkokten O, Adar E (2003) A social network caught in the web. First Monday 8.
34. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99: 7821–7826.
35. Breiman L (2001) Random forests. Machine Learning 45: 5–32.
36. Backstrom L, Leskovec J (2011) Supervised random walks: predicting and recommending links in social networks. In: Proceedings of the 4th ACM international conference on Web search and data mining (WSDM'11). pp 635–644.
37. Rattigan MJ, Jensen D (2005) The case for anomalous link discovery. ACM SIGKDD Explorations Newsletter 7: 41–47.
38. Acquisti A, Gross R (2006) Imagined communities: Awareness, information sharing, and privacy on the Facebook. In: Proceedings of 6th Workshop on Privacy Enhancing Technologies. pp 36–58.
39. Valenzuela S, Park N, Kee KF (2009) Is there social capital in a social network site?: Facebook use and college students' life satisfaction, trust, and participation. Journal of Computer-Mediated Communication 14: 875–901.
40. Adamic L, Adar E (2005) How to search a social network. Social Networks 27: 187–203.
41. Toivonena R, Kovanena L, Kiveläa M, Onnela JP, Saramakia J, et al. (2009) A comparative study of social network models: Network evolution models and nodal attribute models. Social Networks 31: 240–254.