Investigating Link Inference in Partially Observable Networks: Friendship Ties and Interaction

Mehwish Nasim, Raphaël Charbey, Christophe Prieur, and Ulrik Brandes

***Abstract*— While privacy preserving mechanisms, such as hiding one's friends list, may be available to withhold personal information on online social networking sites, it is not obvious whether to which degree a user's social behavior renders such an attempt futile. In this paper, we study the impact of additional interaction information on the inference of links between nodes in partially covert networks. This investigation is based on the assumption that interaction might be a proxy for connectivity patterns in online social networks. For this purpose, we use data collected from 586 Facebook profiles consisting of friendship ties (conceptualized as the network) and comments on wall posts (serving as interaction information) by a total of 64 000 users. The link-inference problem is formulated as a binary classification problem using a comprehensive set of features and multiple supervised learning algorithms. Our results suggest that interactions reiterate the information contained in friendship ties sufficiently well to serve as a proxy when the majority of a network is unobserved.**

***Index Terms*— Data privacy, Facebook, supervised learning.**

I. Introduction

A. Background and Motivation

IN THE first quarter of 2015, Facebook had 1.44 billion active users, which makes it the most actively used social networking site. Facebook allows its users various interaction options, such as sending messages, participating in events, sharing pictures, videos, status posts, and so on.

Issues related to privacy of shared content and personal information have received significant attention not only in the research community but also in the media. The privacy settings for content shared on Facebook match users' expectations only 37% of the time; if incorrect, then the content is usually exposed to more users than expected [23]. Facebook also provides privacy mechanisms to protect friendship information such as hiding friends from public or other contacts.

In case friend's information is hidden, we propose that a malicious contact who has access to a user's "timeline" may not only infer user's active friends but also the connections between them. Facebook, as well as the third parties, can create applications and games which Facebook users can add to their profiles. These apps have access to users' profile data such as list of friends, wall posts, demographic information, and so on [36]. Due to privacy concerns related to the use of data obtained by the third-party applications, Facebook restricted access to several data fields in its Application programming interface (API). As of April 30, 2015, friend list now only returns friends who also use the same app/game. Our question is hence: given these privacy mechanisms, is it still possible for a third-party app (or a bot) to determine the personal network of a user who is using that application?

We demonstrate that interaction information can boost the prediction of unobserved (missing or hidden) relations in partially observed networks. Information on ties in online social networking sites can thus be obtained even when it can be hidden or withheld (e.g., on Facebook [1], Google+ [3], and Flickr [2]). While such information can potentially be used for product and social recommendation systems, it undermines the privacy mechanisms that users of such systems rely on.

Simmel's [31] theory of social circles posits that a person has several personality aspects owing to membership in different social groups. It applies to real-life social groups (workplace colleagues, school mates, family members, organization members, and so on [13]) as well as online circles [24]. Social roles tend to reinforce each other and they may result in pressure to conform to them. Blau [8] conceived social structure as a space in which positions are determined by the individuals' characteristics. Homophily, then, is the assumption that individuals farther apart in this space are less likely to interact with each other. It is argued that social constraints may impede intergroup relations, which means that when homophily is correlated along multiple social constraints within communities, it can lead to limited intergroup interaction.

Both theories suggest that interactions are likely to take place among individuals who are also connected. In fact, repeated interaction is often conceived as a prerequisite for the establishment of social ties. Hence, we attempt to discover unobserved links in online social networks with the help of interaction via participation in joint discussions, but without considering the content of these discussions. Taking a small personal network as an example, we illustrate the problem at hand in Fig. 1. Nodes in white are the ones whose network information is available (e.g., visible circles membership in Google+ or friends on Facebook). Nodes in gray are the ones whose network information is not visible. In the original

Manuscript received September 11, 2015; revised March 30, 2016 and September 6, 2016; accepted October 2, 2016. Date of publication November 17, 2016; date of current version December 2, 2016. This work was supported in part by the DEUTSCHE FORSCHUNGSGEMEINSCHAFT under Grant Br 2158/6-1 and in part by the French National Research Agency, through the Algopol Project under Grant ANR-12-CORD-013.

M. Nasim and U. Brandes are with the University of Konstanz, 78457 Konstanz, Germany (e-mail: mehwish.nasim@uni-konstanz.de; ulrik.brandes@uni-konstanz.de).

R. Charbey and C. Prieur are with Telecom ParisTech, 75013 Paris, France (e-mail: raphael.charbey@telecom-paristech.fr; christophe.prieur@telecom-paristech.fr).

Digital Object Identifier 10.1109/TCSS.2016.2618998

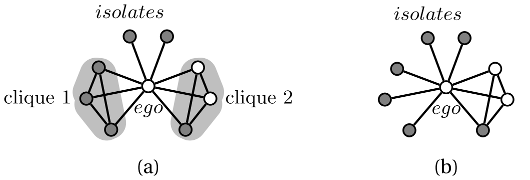

Fig. 1. Personal network example. Friendship links of gray nodes are hidden, whereas, those of white nodes are not. (a) Actual. (b) Observed.

network, a link between a white node and a gray node can be inferred by the information acquired from the white node (assuming there are indirect links). However, the situation becomes challenging when one has to infer links between any two gray nodes, i.e., where the network information is not visible.

B. Related Work

Studies on link prediction have focused on properties such as existing network structure, actor attributes, and interaction patterns to deduce information about the users. In this section, we will discuss various link-prediction methods found in the literature.

1) Network Structure: Liben-Nowell and Kleinberg's [21] pioneering work on link prediction considered a set of features limited to the topological properties of the network. The approach is generic and can be applied to any social network graph. Following the foundations set in [21], various network topology-based approaches have been devised to predict links in social networks [4], [16], [19]. Taking this paper a step further, Cukierski [11] looked into the accuracy of graph-based features for supervised link prediction. They tested multiple network-based features on a data set from Flickr. They found that the best classification results for a supervised link-prediction problem were achieved through a combination of a large number of features, which could capture various aspects of the graph structure. Horvát *et al.* [16] showed that the combination of knowledge of confirmed contacts between members on a social network and their email contacts to nonmembers provides enough information to deduce a substantial proportion of relationships between nonmembers. Leroy *et al.* [18] used interest groups on Flickr to infer hidden links. In a similar study, Zheleva *et al.* [39] showed that when there are tightly knit family circles, the accuracy of link-prediction models can be improved. They made use of the family circle features based on the structural equivalence of family members.

A major limitation of topology-based features is that missing network information may add noise in the training sets and eventually affect the performance of the classifier.

2) Attributes: Both social influence and social selection suggest that network structure and node attribute information should be reinforcing concepts. Several studies have, therefore, looked into improving link prediction with attribute information [15], [38]. Mislove *et al.* [25] inferred attributes of users in combination with the social network graph on the premise that groups in the network are formed around users who share certain attributes. Link-prediction methods utilizing attribute information first appeared in the relational learning community. Taskar *et al.* [35] addressed the problem of link prediction in relational domains. They have focused on the task of collective link classification, where they simultaneously predict and classify an entire set of links in a link graph. They use topological properties and attributes of nodes to define a single probabilistic model over the entire link graph. Modeling link graphs has numerous other applications, including: analyzing communities of people, identifying people who may play certain key roles, and also predicting current and future interactions. Yang *et al.* [37] proposed jointly predicting links and propagating node interests (e.g., music interest). They showed that the interest and friendship information are highly relevant for suggesting friends in social networks.

3) Interaction: Event-based network data consist of sets of events over a period of time. Examples of such networks are e-mail traffic, coauthorship events, telephone calls, and so on. Madadhain *et al.* [28] looked at the problem of temporal link prediction for coauthorship and e-mail networks. They argued that using techniques from data mining and machine learning can yield scalable robust algorithms for predictive modeling. Temporal interactions from call logs have been used to predict links and community structure in social networks [12], [14], [32]. Lee *et al.* [17] demonstrated (using social vector clocks) that link predictors can be made more effective if they operate directly on longitudinal dyadic communication data. Nasim *et al.* [27] provided an insight on the joint commenting behavior of Facebook users. Their results suggested that just like offline behavior people tend to conform to group dynamics when they are on online social networking sites, and their commenting behavior is a result of social influence. Romero *et al.* [29] studied the interplay between network and topical structures on Twitter and suggested that users who followed same tweets were more likely to have links between them. In order to predict links on Facebook, Backstorm and Leskovec [6] used edge creation time, edge initiator, probability of communication and profile observation in a one week period, and number of common friends as prediction features. Their method showed 11% relative improvement in Prec@20 as compared with random walk with restart and logistic regression (LR) (using node + network features). An interesting finding in this paper is that the most important features for Facebook are the ones related to time. We acknowledge that extracting good features that describe the network structure and connectivity patterns between the pair of nodes under consideration is a challenging task.

C. Contribution

We take Facebook as an example of an online social networking site and investigate whether link predictors for online social networks (e.g., Facebook) can be made more accurate if they operate on appropriate communication data. This paper does not focus on developing a sophisticated link-prediction algorithm, rather it aims at identifying appropriate features that can help predict links when information about friendship ties is missing. The uniqueness of this paper comes from the fact that it is inspired from the sociological theory behind friendship formation, which claims that individuals organize relations around points of common interest known as "foci" [13]; this entails that interactions carry important information about the presence of friendship ties. The individuals perceive themselves as the representatives of different social circles [31], and due to the effects of social influence or social selection, they participate in similar activities. When this concept is extended to online social networks, it is observed that people who are friends with each other share similar interests. For instance, people from the same community involve in the same discussions (users commenting on the same posts), e.g., on Facebook [27]. Traditional link-prediction schemes may miss such useful information depicted in the social behavior.

II. Problem Formulation

Given observed mutual friendships among the alters of an ego, we want to infer additional, unobserved, mutual friendships by including information on interaction among them. Formally, we assume, for each ego profile, an undirected graph $G = (V, E \cup E')$ of vertices V representing the alters, and edge sets E and E' representing observed and unobserved friendship ties. We propose to design a binary classifier, which uses set of features X with cardinality $|X| = p + r$ extracted from the given network $G = (V, E)$ and interactions of alters in order to infer links that are likely to be present in the graph G. The features are computed for all dyads (all node pairs). Input to the classifier is p network features $(X_1, \ldots, X_p \in X)$ and r interaction features $(X_{p+1}, \ldots, X_{p+r} \in X)$. The kth feature for node pair i and j is denoted by X_k^{ij}. The outcome, i.e., the class label $(Y_{ij}|X_1^{ij}, \ldots, X_{p+r}^{ij})$, is conditioned on features $(X_1^{ij}, \ldots, X_{p+r}^{ij} \in X)$ and, therefore, $Y_{ij} = 1$ if there is a link between node pair i and j and $Y_{ij} = 0$ otherwise.

III. Data Set

Existing data sets in the public domain contain limited information for the purpose of this paper. Most publicly available data sets (e.g., data sets from Facebook, Flickr, and Google+) contain network and attributes information but are devoid of any interaction history or contain very limited information about the interaction between nodes in the network. The Stanford Network Analysis Project [20] hosts many such data sets. Data sets that contain interaction information, for example the data sets used in [26] and [33], are not available. For this paper, we used an original Facebook data set provided by the Algopol project[1] as part of a survey among volunteer participants selected by a poll institute to make a representative panel of French Facebook users [7]. A Facebook application collected data from Facebook profiles. Data were completely anonymized before analysis. The application downloaded the mutual friendship graph, wall posts, and attributes of ego of his/her friends.

A. Data Statistics

The data set contains 586 ego profiles, where 38% of egos are male, 60% female, and remaining 2% did not report the gender. There are 64 000 friends in total, of which 52% are females, 46% males, and 2% did not report the gender. A total number of comments are about 0.6 million.

B. Network Statistics

Statistics are averaged over all egos. Number of friends: 102; number of edges: 390; number of isolates in the original networks: 9; number of communities (using Louvain clustering [9]): 6; size of the largest connected component: 66; clustering coefficient: 0.5; and density of networks: 0.12, and 73% commenters were friends who commented more than twice in the overall commenting history.

IV. Methodology

We model the mutual friendship information collected from each ego profile as an undirected graph $G = (V, E)$, consisting of the set of V vertices and set of E edges. An edge between two vertices indicates a friendship tie. Facebook provides various interaction options to users. One of them is posting on one's wall. Friends can then write comments on that post. For each ego profile, we can represent the set of posts $\{p_1, \ldots, p_n\} \in P$ and the commenters $\{c_1, \ldots, c_m\} \in C \subseteq V$, as a bipartite graph (a two-mode network, and the derived one-mode network of commenters). In this paper, we are looking at the structural features of comments received on wall posts of an ego. Content of the posts and comments is not analyzed.

A. Data Sampling

In order to hide some links, we sample our network in such a way that the links incident on certain nodes is removed. We call those links hidden or missing links. This is to create a situation similar to the one where a certain percentage of ego's friends have made the friends list hidden, as discussed in Section I (please refer to Fig. 1). We removed edges incident to nodes until a certain percentage of nodes (V^*) are isolates. We have restricted $V^*$to 50%, 75%, 90%, and 100%. Please note that the features are recomputed for each case after removing the edges (both in training and test sets).

Next, we compute the features listed in Table I for all pairs of nodes $u \neq v \in V$ (network structure in the top part and discussion participation in the bottom part).

B. Network

All but one (Katz similarity) are local features based only on neighborhoods $N(u)$ and $N(v)$. Note that the number of common neighbors equals the number of triads closed when u and v are linked. These features represent the state of the art for predicting links in social networks in both theoretical [30] and empirical studies [5], [6], [11], [34].

C. Discussion

Six network features are transferred to the interaction space, where $S(u)$ and $S(v)$ represent the discussions in which the alters participate. Note that we do not make use of labels and thus do not discriminate discussions by, say, topic.

[1] http://algopol.huma-num.fr/grant #ANR-12-CORD-013

TABLE I

FEATURES FOR PREDICTING A TIE BETWEEN ALTERS u AND v

Name	Formula	Description
Common neighbors	$\lvert N(u) \cap N(v)\rvert$	Size of intersection of the two neighborhoods
Common neighbors (min-normalized)	$\frac{\lvert N(u)\cap N(v)\rvert}{\min\{deg(u),deg(v)\}}$	Number of common neighbors normalized by smaller neighborhood
Common neighbors (max-normalized)	$\frac{\lvert N(u)\cap N(v)\rvert}{\max\{deg(u),deg(v)\}}$	Number of common neighbors normalized by larger neighborhood
Jaccard similarity	$\frac{\lvert N(u)\cap N(v)\rvert}{\lvert N(u)\cup N(v)\rvert}$	Common neighbors normalized by joint neighborhood size
Adamic-Adar	$\sum_{k\in N(u)\cap N(v)} \frac{1}{log(deg(k))}$	Similarity based on degrees of common neighbors
Preferential attachment	$\lvert N(u)\rvert \cdot \lvert N(v)\rvert$	Product of neighborhood sizes
Katz similarity	$\sum_{l=1}^{\infty} (\alpha A)^l$	Walks from u to v weighted by length, where A is the adjacency matrix
Cosine similarity	$\frac{\lvert N(u)\cap N(v)\rvert}{\lvert N(u)\rvert\lvert N(v)\rvert}$	Common neighbors normalized by preferential attachment
Joint discussions	$\lvert S(u) \cap S(v)\rvert$	Number of discussions both alters participated in
Joint discussions (min-normalized)	$\frac{\lvert S(u)\cap S(v)\rvert}{\min\{\lvert S(u)\rvert,\lvert S(v)\rvert\}}$	Joint discussions normalized by smaller participation set
Joint discussions (max-normalized)	$\frac{\lvert S(u)\cap S(v)\rvert}{\max\{\lvert S(u)\rvert,\lvert S(v)\rvert\}}$	Joint discussions normalized by larger participation set
Jaccard (discussions)	$\frac{\lvert S(u)\cap S(v)\rvert}{\lvert S(u)\cup S(v)\rvert}$	Joint discussions normalized by combined set of discussions
Smallest discussion size	$\min_{d\in S(u)\cap S(v)} \lvert U(d)\rvert$	Smallest number of participants in joint discussions
Largest discussion size	$\max_{d\in S(u)\cap S(v)} \lvert U(d)\rvert$	Largest number of participants in joint discussions

TABLE II

AVERAGE AREA UNDER THE ROC CURVE (AUCROC) AND AREA UNDER THE PRECISION RECALL CURVE (AUCPR) FOR INDIVIDUAL NETWORK FEATURES USING LR AND TENFOLD CROSS VALIDATION ON UNBALANCED DATA FOR DIFFERENT PERCENTAGES OF ISOLATES IN THE NETWORK

	AUCROC			AUCPR		
Network Features	50%	75%	90%	50%	75%	90%
neighbors	0.859	0.514	0.501	0.643	0.192	0.165
– min-norm.	0.791	**0.519**	**0.509**	0.544	**0.198**	**0.191**
– max-norm.	0.770	0.504	0.502	0.533	0.174	0.176
Jaccard	0.840	0.514	0.501	0.557	0.188	0.165
Adamic-Adar	**0.865**	0.514	0.501	**0.666**	0.194	0.165
preferential	0.761	0.515	0.503	0.429	0.198	0.174
Katz	0.501	NA	NA	0.174	NA	NA
Cosine	0.780	0.512	0.503	0.378	0.194	0.172

TABLE III

AVERAGE AUCROC AND AUCPR FOR INDIVIDUAL DISCUSSION FEATURES (WITHOUT ANY NETWORK STRUCTURE) USING LR AND TENFOLD CROSS VALIDATION ON UNBALANCED DATA

Discussion Features	AUCROC	AUCPR
Joint discussions	**0.608**	**0.284**
– min-normalized	0.606	0.262
– max-normalized	0.607	0.273
Jaccard (discussions)	0.607	0.275
Smallest discussion size	0.598	0.250
Largest discussion size	0.603	0.260
Naïve (random)	0.506	0.173
Naïve (all 1s)	0.500	0.163

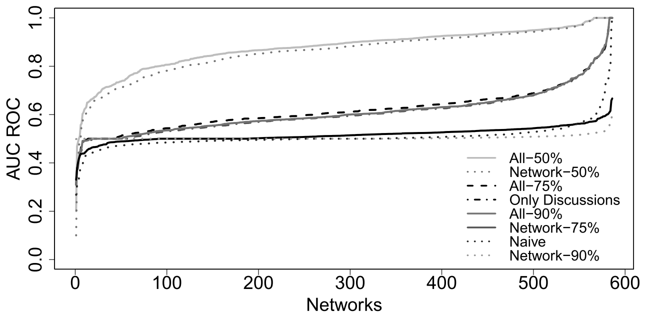

Fig. 2. Performance relative to percentage of hidden neighborhoods.

We have formulated the link-inference problem as a supervised binary classification problem. Supervised learning methods are apt at dealing with data sets, which have greater class imbalance (e.g., online friendship networks) [22]. We divide the data into training and test sets. The model is trained on the training sample and is used to classify the class labels in the test sample. We use a tenfold cross validation. For every fold, we train the algorithm on 90% of the networks and test the performance on the remaining 10%. We have ensured mutual exclusivity between data points as well as individual networks that are used in training and test samples in the experiments. We then apply a learning model and test the performance of our features using: LR, linear discriminant analysis (LDA), and support vector machines (SVMs) and found the performance to be invariant of the underlying model. For a social network data set with nearly 2 million data points, LR is the preferred classification algorithm. It measures the relationship between the categorical-dependent variable and one or more independent variables by estimating probabilities. It is very scalable (unlike SVM) and does not assume that the features are normally distributed (unlike LDA). Because of the sparsity of the networks, we also experimented with balancing our data but there was no significant difference in the average performance of the classifier. For brevity, we, therefore, only report results using LR with tenfold cross validation on the original imbalanced data.

V. RESULTS

A. Evaluation Metric

There are different ways to measure the performance of a classifier for a link-prediction problem. We have used two evaluation metrics to judge the performance of our classifier: 1) precision recall and 2) receiver operating characteristic. We calculate the area under the curve (AUC) for each

TABLE IV

AVERAGE AUC (AND STANDARD DEVIATION) FOR ROC AND PR CURVES OF ALL NETWORK FEATURES WITH AND WITHOUT DISCUSSION FEATURES

	AUCROC				AUCPR			
Features	50%	75%	90%	100%	50%	75%	90%	100%
All network features	0.859 (0.099)	0.516 (0.03)	0.503 (0.009)	-	0.647 (0.177)	0.198 (0.130)	0.171 (0.122)	-
All network features + all discussion features	***0.875*** (0.090)	***0.619*** (0.087)	***0.619*** (0.087)	0.606 (0.087)	***0.661*** (0.175)	***0.308*** (0.185)	***0.286*** (0.189)	0.284 (0.189)

metric. The range of AUC is between 0 and 1, AUC value of 1 being the ideal case. The two evaluation mechanisms (ROC and PR) calculate different measures. If the intention is to find out how many "true positives" are successfully detected by the classifier, then ROC is the better choice. However, if one is interested in the precision of the classifier, then PR is the better choice. We use commenting behavior to predict friendship links by a linear combination of best-performing network and discussion features.

1) Individual Performance of Features: The individual performance of network and discussion features is reported in Tables II and III. The ranking was checked and confirmed using information-gain criteria [10].

The selection methods rank Adamic–Adar similarity, number of common neighbors and Jaccard similarity the highest among the network features, and common discussion size among discussion features.

While Katz similarity actually performs best on an almost complete network, its performance degrades rapidly with the number of excluded neighborhoods. Sarkar *et al.* [30] argue that in social networks longer walks are more relevant if short paths (and two-hop paths via common neighbors in particular) are rare. The coefficient performs poorly in networks containing 50% or more isolates, whereas local features based on triadic closure still do well. This is in line with previous empirical results on Online social network (OSNs) such as Facebook or Flickr [5], [6], [34].

In case no network information is available, no network but only discussion features are applicable (Table III) and we also report on two naïve models for comparison. The first one is a random predictor based on the density of ties in the original network and the second one assumes a complete mutual friendship graph. All discussion features perform better than naïve prediction on average, with higher AUC values on 543 out of 586 networks.

If two friends are repeatedly in discussions, which have a fewer number of participants, then this is a good indicator of friendship prediction (smallest common discussion size). Larger discussions indicate posts with global interests. The estimate coefficients for the size of smallest common discussion and the size of largest common discussion are negative. This signifies that smaller the size of a discussion, the more exclusive it would be, which is indicative of the fact that participants involved in the discussion share friendship ties and they may point that the discussion is of interest to a specific group. Note that we did not look at the topological links between all discussion participants, since our focus is the structural features of discussions.

2) Combined Performance of Features: Fig. 2 shows the distribution of AUCROC for linear combination of network and discussion features. The accompanying Table IV reports the average AUCROC and AUCPR and the deviation from mean. When 50% of the nodes are isolates, there is a slightly higher AUC for some networks when both network and discussion features are utilized as compared with network-only features, and discussion features are not adding a lot of value. For more sparse networks, the importance of discussion features gets evident. Discussion features do add significant information as demonstrated in the case when 90% of nodes are isolates and we see more than 20% relative improvement in AUCROC. We noticed that with the addition of a single discussion feature, the AUC improves by 0.101 points. Adding further discussions information results in additional performance improvement. This shows that when the network is partially available then discussions act as a proxy to detect friendship ties. This strengthens our assumption that even on online social networking sites, people involve more in intragroup interaction.

VI. CONCLUSION

We posed the question whether it is possible for a third-party application such as Candy Crush or a bot to determine the personal network of a user who is using that application. We find that if as little as 10% of a user's friends have installed the application, it can reveal a significant portion of the user's personal network. We utilized the stylized fact that individuals act as members of multiple social groups. Members of the same group tend to participate in similar activities. Our results are based on multiple network and interaction features as well as multiple classification algorithms, and they suggest that in the absence of network structure, interaction information may be used as a proxy for friendship ties and thereby improve the performance of link prediction.

Privacy is a concern for many users of online social networks. Our findings suggest, however, that network ties are reflected in social behavior. Privacy preserving mechanisms, i.e., hiding friends lists (Facebook) or circles (Google+), may not serve the purpose if interaction information is visible to a third-party application. When there is no network information available, interactions provide substantial information about the unobserved ties. While we do not know the algorithm used to determine the content of the "News feed" in Facebook, filtering based on shared attributes should boost within-group discussion. This paper thus suggests that filtering mechanisms based on homophily counteract the privacy goals expressed by hiding one's friends list. If users are commenting on posts because they are shown only posts written by their friends, then inference of connections could be made less likely to be compromised by showing users a mix of posts in their news feeds. However, calculating the right mix requires further knowledge about the tradeoffs between the quality of users' online experience and privacy.

References

[1] *Facebook*, accessed on May 1, 2015. [Online]. Available: http://www.facebook.com/

[2] *Flickr*, accessed on May 1, 2015. [Online]. Available: https://flickr.com/

[3] *Google+*, accessed on May 1, 2015. [Online]. Available: https://plus.google.com/

[4] L. A. Adamic and E. Adar, "Friends and neighbors on the Web," *Social Netw.*, vol. 25, no. 3, pp. 211–230, 2003.

[5] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer, "Friendship prediction and homophily in social media," *ACM Trans. Web*, vol. 6, no. 2, 2012, Art. no. 9.

[6] L. Backstrom and J. Leskovec, "Supervised random walks: Predicting and recommending links in social networks," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, 2011, pp. 635–644.

[7] I. Bastard, D. Cardon, G. Fouetillou, C. Prieur, and S. Raux, "Travail et travailleurs de la donnée," in *Big Data: Nouvelles Partitions de L'Information*, L. Calderan, P. Laurent, H. Lowinger, and J. Millet, Eds. De Boeck/ADBS, 2015.

[8] P. M. Blau, *Inequality Heterogeneity: A Primitive Theory Social Structure*. New York, NY, USA: The Free Press, 1977.

[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Statist. Mech. Theory Experim.*, vol. 2008, no. 10, p. 10008, 2008.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

[11] W. Cukierski, B. Hamner, and B. Yang, "Graph-based features for supervised link prediction," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul./Aug. 2011, pp. 1237–1244.

[12] N. Eagle, A. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 36, pp. 15274–15278, 2009.

[13] S. L. Feld, "The focused organization of social ties," *Amer. J. Sociol.*, vol. 86, no. 5, pp. 1015–1035, 1981.

[14] M. Goldberg, S. Kelley, M. Magdon-Ismail, K. Mertsalov, and A. Wallace, "Finding overlapping communities in social networks," in *Proc. IEEE 2nd Int. Conf. Social Comput. (SocialCom)*, Aug. 2010, pp. 104–113.

[15] N. Z. Gong *et al.*, "Joint link prediction and attribute inference using a social-attribute network," *ACM Trans. Intell. Syst. Technol. (TIST)*, vol. 5, no. 2, 2014, Art. no. 27.

[16] E.-Á. Horvát, M. Hanselmann, F. A. Hamprecht, and K. A. Zweig, "One plus one makes three (for social networks)," *PLoS ONE*, vol. 7, no. 4, p. e34740, 2012.

[17] C. Lee, B. Nick, U. Brandes, and P. Cunningham, "Link prediction with social vector clocks," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 784–792.

[18] V. Leroy, B. B. Cambazoglu, and F. Bonchi, "Cold start link prediction," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 393–402.

[19] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 641–650.

[20] J. Leskovec and A. Krevl. (Jun. 2014). *SNAP Datasets: Stanford Large Network Dataset Collection*. [Online]. Available: http://snap.stanford.edu/data

[21] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, 2007.

[22] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 243–252.

[23] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove, "Analyzing Facebook privacy settings: User expectations vs. reality," in *Proc. ACM SIGCOMM Conf. Internet Meas. Conf.*, 2011, pp. 61–70.

[24] J. Mcauley and J. Leskovec, "Discovering social circles in ego networks," *ACM Trans. Knowl. Discovery Data (TKDD)*, vol. 8, no. 1, 2014, Art. no. 4.

[25] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You are who you know: Inferring user profiles in online social networks," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining*, 2010, pp. 251–260.

[26] M. Mondal, Y. Liu, B. Viswanath, K. P. Gummadi, and A. Mislove, "Understanding and specifying social access control lists," in *Proc. Symp. Usable Privacy Secur. (SOUPS)*, 2014, pp. 271–283.

[27] M. Nasim, M. U. Ilyas, A. Rextin, and N. Nasim, "On commenting behavior of Facebook users," in *Proc. 24th ACM Conf. Hypertext Social Media*, 2013, pp. 179–183.

[28] J. O'Madadhain, J. Hutchins, and P. Smyth, "Prediction and ranking algorithms for event-based network data," *ACM SIGKDD Explorations Newslett.*, vol. 7, no. 2, pp. 23–30, 2005.

[29] D. M. Romero, C. Tan, and J. Ugander, "On the interplay between social and topical structure," in *Proc. 7th Int. AAAI Conf. Weblogs Social Media (ICWSM)*, 2013, pp. 1–10.

[30] P. Sarkar, D. Chakrabarti, and A. W. Moore, "Theoretical justification of popular link prediction heuristics," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, vol. 22. 2011, pp. 2722–2727.

[31] G. Simmel, *Soziologie. Untersuchungen über die Formen der Vergesellschaftung*. Berlin, Germany: Duncker & Humblot, 1908.

[32] L. Tabourier, A.-S. Libert, and R. Lambiotte, "Predicting links in ego-networks using temporal information," *EPJ Data Sci.*, vol. 5, no. 1, pp. 1–16, 2016.

[33] J. Tang, S. Wu, and J. Sun, "Confluence: Conformity influence in large social networks," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 347–355.

[34] F. Tarissan, M. Latapy, and C. Prieur, "Efficient measurement of complex networks using link queries," in *Proc. IEEE INFOCOM Workshops*, Apr. 2009, pp. 1–6.

[35] B. Taskar, M.-F. Wong, P. Abbeel, and D. Koller, "Link prediction in relational data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003.

[36] N. Wang, H. Xu, and J. Grossklags, "Third-party apps on Facebook: Privacy and the illusion of control," in *Proc. 5th ACM Symp. Comput. Human Interaction Manage. Inf. Technol.*, 2011, Art. no. 4.

[37] S.-H. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng, and H. Zha, "Like like alike: Joint friendship and interest propagation in social networks," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 537–546.

[38] Z. Yin, M. Gupta, T. Weninger, and J. Han, "A unified framework for link recommendation using random walks," in *Proc. Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2010, pp. 152–159.

[39] E. Zheleva, L. Getoor, J. Golbeck, and U. Kuter, "Using friendship ties and family circles for link prediction," in *Advances in Social Network Mining and Analysis*. Berlin, Germany: Springer, 2010, pp. 97–113.

Mehwish Nasim received the BS(CS) degree from the National University of Computer and Emerging Sciences, Islamabad, Pakistan, in 2005, the master's degree in computer software engineering from the National University of Sciences and Technology, Islamabad, in 2010, and the Ph.D. degree from the University of Konstanz, Germany, in 2016.

Her current research interests include network analysis, data mining, and machine learning.

Dr. Nasim is a member of the Algorithmics Group, University of Konstanz.

Raphaël Charbey is currently pursuing the Ph.D. degree in computer science with the Department of Social and Economic Sciences, Telecom ParisTech, Paris, France, under the supervision of C. Prieur and A. Casilli.

His current research interests include structural analysis of networks and data mining.

Christophe Prieur is currently a Researcher in sociology and algorithms. He is involved in sociability and on methods to study it from large volumes of data.

Dr. Prieur is a member of the Interdisciplinary Institute of Innovation, a joint laboratory of CNRS, and Telecom Paris Tech, a prominent engineering school in Paris.

Ulrik Brandes received the Diploma degree from RWTH Aachen University, Aachen, Germany, in 1994, the Ph.D. degree from the University of Konstanz, Konstanz, Germany, in 1999.

After his Habilitation in 2002, He became an Associate Professor with the University of Passau, Passau, Germany in the same year. Since 2003, he has been a Professor of Computer Science with the University of Konstanz. With a background in algorithmics, his current research interests include network analysis and visualization, with application to social networks in particular.

Dr. Brandes was a member of the Graph Drawing Steering Committee from 2007 to 2014, and is on the Editorial Board of the *Journal of Graph Algorithms and Applications*. He is a member of the board of directors of the International Network for Social Network Analysis, an Associate Editor of *Social Networks*, and an Area Editor of *Network Science*.