Interactions Around Social Networks Matter: Predicting the social network from associated interaction networks

Mohammed Abufouda and Katharina Anna Zweig
University of Kaiserslautern
Gottlieb-Daimler-Str. 48/672
67663 Kaiserslautern, Germany
email: {abufouda, zweig}@cs.uni-kl.de

***Abstract*—Tie formation in social networks is driven by different motives that are not always apparent in the social network itself. These motives differ from one social network to another, depending on, e.g., the network's purpose, such as advice seeking or collaboration, and the effort it costs to establish a friendship relationship. A common factor that exists in almost all social networks is homophily: the tendency of social network members to connect to similar members. In this work, we look at the tie formation process in social networks from a different perspective where we consider not only a social network SN, but also a set of associated interaction networks $\mathcal{G}_n$ around it. We show, based on 6 social networks and in total 20 different associated interaction networks, that it is possible to predict the entire social network's structure to a satisfactory extent, only by knowing the structure of these interaction networks. As social networks are based on a voluntary relationship while some of the interaction relationships are at most semi-controllable for most members, e.g., being together in a team, this seems to indicate that whom we choose as a friend is also determined by whom we interact with.**

***Keywords*—Network prediction, Multiple networks, Social homophily.**

I. Introduction

Tie formation in social networks is mainly voluntary and guided by personal motives. However, it is undoubtedly influenced by notions of homophily [1] which might induce relationships between persons based on external and not entirely controllable factors. Homophily in social networks is described as the tendency to connect with similar people. This similarity is based on internal factors like having the same hobby or being engaged in the same political party. Nevertheless, this similarity can also be influenced by external, semi-controllable factors such as being in the same company. Based on that premise, considering only the social network in a given complex system is not enough to understand the existing ties among its members and to predict their future ties. Thus, to provide a comprehensive and an informative view of a social network SN, it is important to consider all possible and available information about the members and their interactions in other environments. These interactions are represented by additional networks G_i, *interaction networks*, on the same set of actors of the social network SN. To show that networks G_i are really informative with respect to the social network they accompany, we give an evidence that they at least partly drive the process of tie formation in the SN. In this research we show that harnessing the information in associated networks G_i makes it possible to predict the link structure of the social network SN, using a very broad data set.

Motivating example: Consider a social coding platform like *github.com* where members are software developers. In addition to providing the functionalities to share their work, the developers can also accept each other as friends to build a social network SN. Also, there are several interaction networks which might influence the tie formation in the SN. These interaction networks include *The collaboration development network* G_1: vertices represent developers and a directed edge appears between two developers when one of them has committed to another developer's software repository at least once. *The watcher network* G_2: vertices represent developers and a directed edge appears between two developers when one of them is watching the software repository of the other developer. *The fork network* G_3: vertices represent developers and a directed edge appears between two developers when one of them forks a repository of the other developer. *The pull requests network* G_4: vertices represent developers and a directed edge appears between two developers when one of them sends a pull request to the other developer.

One way to analyze tie formation is to build a model that *predicts* links in a given social network. The closer the predicted link structure is to the real network's structure, the more convincing is the idea that the model captures the main motivations for tie formation. So far, link prediction approaches have assumed that the information given in a social network at time t is enough to deduce future tie formation at a time $t' > t$. Here, we test how much the ties in the social network can be predicted by those found in the surrounding networks G_i as described above without using any information in the social network itself. Our work is based on the link prediction problem initially proposed by Liben-Nowell et al. [2], namely to predict the formation of new links between actors in a time interval t based on the already existing network structure in the same social network in an earlier time interval. Here, we use the following variant of the link prediction problem: given a set of associated networks G_i and a social network SN of the same actors at any point of time t, predict the network structure of the SN at time t without using information from the SN itself. This prediction helps not only in revealing latent ties among the members of SN but also in providing information regarding the correlation between the SN and G_n.

II. Related work

In their seminal work, Liben-Nowell and Kleinberg modeled and addressed the link prediction problem in complex networks by providing a set of proximity measures as predictors in an unsupervised machine learning approach [2]. The authors used different co-authorship data sets to predict future coauthor-relationships based on a set of proximity measures [2]. These are still the main proximity measures used in later work by a number of researchers, particularly those who employed machine learning techniques. Al Hassan et al. were the first to apply supervised machine learning to predict ties in co-authorship networks [3], which is still an active field of research for predicting all kinds of social relationships [4], [5], [6]. Researchers also started to use more than one relationship to predict the network structure of a complex network. For example, Lu et al. used references, co-authorship, and co-citation information in time interval t to predict the formation of new co-authorship-relationships in a later time interval [7].

All of the aforementioned work followed the same paradigm for link prediction, namely dividing the social network (and possibly additional networks) into two independent temporal snapshots for training and testing. Here, we aim to identify the influence of a **single** interaction network on the social network's structure without using any information from the social network structure itself. Thus, our work differs significantly from the related work as we predict the ties of the **entire** social network, not only the structure of newly added ties. This enables insights regarding the influence of semi-controllable interaction networks and the voluntarily build structure in a given social network.

III. The Proposed Method

Our approach is based on a re-definition of the classical link prediction problem to incorporate multi-networks as training sets to predict the ties in the SN using machine learning classification algorithms. Figure 1 shows the general approach where the G_i networks are used to build a features data model (FDM), i.e., a set of topological features for each pair of nodes (v, w) together with ground truth, that is used to train a classifier that predicts the SN. The following subsections describe the construction of the FDM in detail. In general, the FDM contains *node-dependent features* which are described in the following.

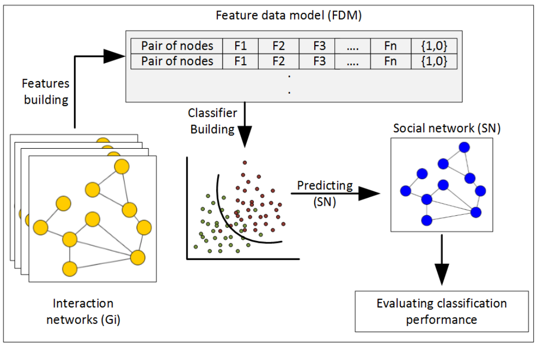
Feature data model (FDM)
Pair of nodes | F1 | F2 | F3 | .... | Fn | {1,0}
Pair of nodes | F1 | F2 | F3 | .... | Fn | {1,0}
Features building
Classifier Building
Predicting (SN)
Social network (SN)
Interaction networks (Gi)
Evaluating classification performance

Fig. 1: An abstract view of how to predict the link structure of a social network using associated interaction networks.

A. Node-dependent features

To predict whether a pair of nodes (v, w) is connected in the SN, neighborhood measures of *v* and *w* in G_i are calculated to provide the features of the FDM. These features are:

- *Cooccurrence* (*coocc*): For each pair of nodes v and w, the cooccurrence is defined as the number of their common neighbors. $coocc(v, w) = |\Gamma(v) \cap \Gamma(w)|$
- *Resource Allocation* ($\mathcal{RA}$): This measure was proposed by Zhou et al. [8] and showed a slightly better performance than $coocc$ in link prediction. This measure assumes that each node has a given amount of resources that is distributed equally among its neighbors. This concept is adapted by incorporating two nodes v and w: $\mathcal{RA}(v, w) = \sum_{z \in \Gamma(v) \cap \Gamma(w)} \frac{1}{|\Gamma(z)|}$
- *Adamic-Adar coefficient* ($\mathcal{AAC}$): The Adamic-Adar coefficient is defined as [9]: $\mathcal{AAC}(v, w) = \sum_{z \in \Gamma(v) \cap \Gamma(w)} \frac{1}{log|\Gamma(z)|}$
- *Jaccard Index* ($\mathcal{JI}$): This measure was first proposed in information retrieval field [10] as a method to characterize the similarity of two sets. $\mathcal{JI}(v, w) = \frac{|\Gamma(v) \cap \Gamma(w)|}{|\Gamma(v) \cup \Gamma(w)|}$
- *Preferential Attachment* ($\mathcal{PA}$): In collaboration networks, Newman showed that the probability of collaboration between any two nodes(authors) v and w is correlated to the product of $deg(v)$ and $deg(w)$ [11]: $\mathcal{PA}(v, w) = |\Gamma(v)|.|\Gamma(w)|$

The previous similarity measures are for undirected networks. For directed networks, two versions for each measure are used by providing two versions of the neighborhood set Γ, the in-neighbors $\Gamma(v)_{in}$ and the out-neighbors $\Gamma(v)_{out}$. Based on this, an *in* and an *out* version of the above measures can be constructed. For example, the in-cooccurrence for two nodes v, w is: *coocc*$(v, w)_{in} = |\Gamma(v)_{in} \cap \Gamma(w)_{in}|$. Based on these node-dependent features, the FDM is constructed as described in the following.

B. Features data model ($FDM(\mathcal{G}_i)$)

The features data model ($FDM(\mathcal{G}_i)$) for a single network $\mathcal{G}_i$ contains $\frac{n(\mathcal{G}_i)(n(\mathcal{G}_i)-1)}{2}$ entries where $n(\mathcal{G}_i)$ is the number of nodes in $\mathcal{G}_i$. For each *v* and *w* $\in \mathcal{G}_i$, an entry $\mathcal{I}(v, w)$ is a tuple that contains: (1) the node-dependent features of v and w and (2) a binary classification, $\{1, 0\}$, that indicates whether there is an edge $e(v, w)$ in $\mathcal{G}_i$ or not. Having constructed the $FDM(\mathcal{G}_i)$, it can be used in a machine learning approach ψ*(*$FDM(\mathcal{G}_i)$*)* as network model of the network $(\mathcal{G}_i)$. There are a couple of out-of-the-box machine learning classifiers that can be used to predict the SN. In the experiment we use the logistic regression classifier.

IV. Data Sets and Evaluation Metrics

In this section we provide information about the data sets used to validate the method and the evaluation metrics.

A. Data sets description

Here we describe a variety of data sets used in the experiment. *Research Group* [12]: Includes the *Facebook* social network along with four associated interaction networks built between the employees of the research group. Relations in these other networks are co-working, co-author, going out to lunch, and leisure. *International Internet* [13]: Includes three different networks for 75 nations' internet relations. *Hyperlinks* is a directed network such that an edge exist between two nodes (countries) if there is a website in one of these countries' domains that points to a website from the other country's

domain's. We consider this network as the social network among countries. *Bandwidth* is a network among countries where edges represent the existence of an internet connection between two countries. In the *shared website* network, an edge appears between two countries if they share at least one common most-frequently visited website. The original hyperlinks network is directed (with reciprocity 0.92), while the other two networks are undirected. To overcome this problem, only the reciprocal edges in the original hyperlink network are considered. *Terrorists network* [14]: Includes the friendship network of 79 individuals together with information on associated interaction networks like trainings done together, meetings between them, places commonly visited by two persons, and business ties. *Github*: A social network of software developers with a set of associated interaction networks as compiled by Gousi et al. [15] and described earlier in Section I. *Brightkite* [16]: Is a location-based social network[1]. Originally, check-in is a bipartite network of actors and places where an actor can check-in to the software to let it know that he or she visited that place. We performed one-mode projection to construct the check-in network such that there is an edge between two persons if they were at the same place at least one time. *Law Firm* [17]: A social network of law firm partners with information on two other interaction networks: *co-working* and *seeking advice from*.
Table I shows network statistics for all of the networks we used. These statistics include the number of nodes n, the number of edges m, the clustering coefficient $cc(G)$ [18], and the network's density η.

TABLE I: Data set statistics.

dataset	Networks	n	m	$cc(G)$	$\eta(G)$
Research group	SN **facebook**	32	248	0.48	0.24
	G1 Work	60	338	0.34	0.1
	G2 Co-author	25	42	0.43	0.08
	G3 Lunch	60	386	0.57	0.01
	G4 Leisure	47	176	0.34	0.08
Internet	SN **Hyperlinks**	75	2550	0.99	0.84
	G1 Bandwidth	75	448	0.42	0.16
	G2 Shared websites	75	2360	0.92	0.86
Terrorists networks	SN **Friends**	61	91	0.2	0.04
	G1 Financial	13	15	0.88	0.2
	G2 Places	31	82	0.61	0.18
	G3 Business	44	458	0.75	0.48
	G4 Meeting	26	63	0.41	0.2
	G5 Training	38	147	0.72	0.2
	G6 Organization	63	416	0.84	0.22
	G7 Operations	39	267	0.78	0.36
Github(directed)	SN **Followers**	595232	2551900	0.13	≈0
	G1 Commits	322461	909125	0.2	≈0
	G2 Watchers	274597	2478561	0.02	≈0
	G3 Forks	220443	673396	0.35	≈0
	G4 Pull requests	156688	379207	0.08	≈0
Brightkite	SN **Friendship**	11655	63664	0.172	≈0
	G1 Check-in	13029	1378862	0.75	0.016
Law Firm (directed)	SN **friends**	69	339	0.43	0.09
	G1 Co-work	71	726	0.41	0.15
	G2 Advice	71	717	0.42	0.14

B. Classification evaluation metrics

In this section a set of classic classification evaluation metrics is presented. These metrics are used for evaluating the classification results of the experiment. Remember that the classifier predicts for each pair of nodes (v, w) in $V(SN) \cap V(G_i)$ whether there is an edge or not. In a binary classification scheme like this, only four types of results can be obtained: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Based on these basic metrics, the following evaluation metrics result in a single number for all methods that can be more easily compared: **Precision** $\mathcal{P}$: is the ratio of TP to the number of all positive classifications. $\mathcal{P} = \frac{TP}{TP+FP}$. **Recall** $\mathcal{R}$: is also called the *true positive rate* and the *sensitivity*. $\mathcal{R} = \frac{TP}{TP+FN}$. **F-measure** $\mathcal{F}$: is the harmonic mean of precision and recall. $\mathcal{F} = \frac{2.\mathcal{P}.\mathcal{R}}{\mathcal{P}+\mathcal{R}}$.

V. Empirical results

In this section, we provide the results of the experiments where we first introduce and report a simplistic network prediction without a supervised machine learning approach and then the results using supervised machine learning.

A. Simplistic network prediction $\mathcal{SP}$

The first question to be answered is how much a single, associated interaction network can predict the social network's structure. Predicting the SN based on a single associated network G_i without applying machine learning is called a *simplistic prediction*. Simplistic Prediction $\mathcal{SP}(SN, \mathcal{G}_i)$ simply predicts that each edge in G_i also exists in the SN and that nodes not connected in G_i are not connected in SN, either. Thus, TP, TN, FP, and FN are given as follows. TP is given by the number of edges (v, w) contained in both networks. TN is given by the number of pairs of nodes not connected by an edge in both networks. FP instance means that an $e(v, w)$ does not exists in the SN but exists in $\mathcal{G}_i$. FN instance means that an $e(v, w)$ exists in the SN but exists in $\mathcal{G}_i$. The results of the simplistic prediction are shown in Table II.

The F-measure are in some cases surprisingly high. For example, the correlation between advice seeking and being friends in the law firm data set is already 0.45 and sharing a hobby is also correlated with being friends in a research group by 0.51. The F-value is very high concerning the financial ties and the social ties among the terrorist data set (0.72). But there is always the possibility that such a result is merely caused by the number of nodes and edges in the graph. For example, if G_i and SN are both complete graphs, the "prediction" is perfect by virtue of their structure. To exclude this possibility, 100 random graphs with the same number of nodes and edges as in the $\mathcal{G}_i$ [19] were built and used in the simple prediction approach for the SN. The results are also shown in Table II; in most cases, this prediction is worse by at least a factor of 10. Notable exceptions are the two interaction networks of the internet: here, the densities are overall so high, that a good prediction result is inevitable. Less pronounced but still visible is that effect in both interaction networks in the law firm data set and the business tie network with respect to the terrorist social network: all of them show a rather high density to begin with and a very low number of nodes. Here, the general structure of the two networks, the respective $\mathcal{G}_i$ and the SN, seems to dictate parts of the success of the simplistic prediction approach.

B. Predicting SN with machine learning based on $\mathcal{G}_i$

The simplistic prediction approach yielded surprisingly high congruence between interaction networks and their associated social network. However, it is clear that most networks suffer from random noise. Machine learning can help to identify those patterns in the interaction network that make a link in the interaction network likely, thereby identifying latent ties that were never observed in the interaction network

[1] Data is available here: http://snap.stanford.edu/data/loc-brightkite.html

and removing ties that just happened by chance but are not backed up by the overall structure of the interaction network. For example, a new lawyer in the law firm might seek an advice from the senior partner of the company but he never actually got the chance to meet that senior partner so far, which means that they are not friends in the friendship network. The advice seeking network will contain such a tie but the machine learning classifier might notice that most of the individuals in the same (network) position as the new lawyer do not claim to have this connection and thus the classifier gives this claimed relation a low probability to really exist. The classifier can for example learn that most edges are between people that have neighbors in common. If the new lawyer and the senior partner do not have any neighbors in common, the classifier will predict that this pair of nodes is not connected by the advice-seeking relationship, despite the claim. We thus test the following hypothesis: **H:** *The SN can be more effectively predicted using FDM$_{\mathcal{G}_i}$, i.e., single associated interaction networks provide enough structure to predict the social network structure.*

Table II shows the quality of this prediction as quantified by the evaluation metrics described earlier. Overall, the quality of the prediction is very high which confirms the hypothesis **H**. It is obvious that a prediction using the FDM$_{\mathcal{G}}$ model is more effective than the simplistic prediction $\mathcal{SP}$ performed in Section V-A: in no case, the prediction of the SN is worse than the simplistic prediction. However, the increase in quality varies strongly: The prediction of the social ties between terrorists based on their business ties does not improve by using a machine learning approach. The largest improvement is coming from the pull-request network in Github. The best prediction with the machine learning approach is achieved for the co-working relationship between lawyers: the simplistic prediction achieves an F-value of 0.54, but the machine learning approach is able to identify those edges that follow a predictable pattern and this pattern seems to be the one that (partly) determines how people choose a friend.

TABLE II: The results of different types of prediction.

Data Set	Interaction Network	$\mathcal{SP}_{random}$	$\mathcal{SP}_{\mathcal{G}_i}$	ψ(FDM$_{\mathcal{G}_i}$)
		$\mathcal{F}$	$\mathcal{F}$	$\mathcal{F}$
Research group	Work	0.021	0.52	0.53
	Co-author	≈0	0.472	0.72
	Lunch	0.029	0.51	0.63
	Leisure	0.03	0.46	0.67
Internet	Bandwidth	0.27	0.28	0.35
	Sharedwebsite	0.98	0.84	0.9
Terrorists	Financial	0.16	0.72	0.76
	Places	0.07	0.35	0.55
	Business	0.042	0.13	0.13
	Meeting	≈0	0.62	0.69
	Training	0.039	0.38	0.6
	Organization	0.03	0.198	0.42
	Operations	0.039	0.275	0.35
Github	Commits	≈0	0.1	0.25
	Watchers	≈0	0.1	0.16
	Forks	≈0	0.15	0.18
	Pull requests	≈0	0.02	0.13
Brightkite	Check-in	≈0	0.3	0.42
Law Firm	Co-worker	0.13	0.54	0.76
	Advice	0.1	0.45	0.63

VI. Conclusion

In this article we have shown that the tie formation process in a social network cannot only be predicted from the social network itself but that the whole structure of a social network can be satisfactorily predicted from other associated interaction networks. Such a high correlation between interaction networks and social networks does not tell us the direction of causality. However, it is clear that links in the social network are largely voluntary: nobody is forced to be the friend of the other (although some cultural pressure might apply). Some of the interaction networks are not fully controllable by the actors of the social network. For example, co-working structures are often determined by the hierarchy of the company or by sheer necessity to have people from different compartments in a project team. If such a non- or semi-controllable interaction network shows a large similarity with the associated social network structure, this indicates that a part of the social tie formation is not so much guided by an internal homophily but rather by external homophily: we have a high probability to be friends with those with whom we spend a lot of time - whether it is self-chosen or dictated by the circumstances.

References

[1] M. McPherson and et al, “Birds of a feather: Homophily in social networks,” *Annual Review of Sociology*, pp. 415–444, 2001.

[2] D. Liben-Nowell and J. Kleinberg, “The link prediction problem for social networks,” in *Proceedings of the 12th Intern. Conference on Information and Knowledge Management*. ACM, 2003, pp. 556–559.

[3] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki, “Link prediction using supervised learning,” in *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*, 2006.

[4] M. Fire and et al, “Link prediction in social networks using computationally efficient topological features,” in *Privacy, security, risk and trust (passat), third international conference on social computing (socialcom)*. IEEE, 2011, pp. 73–80.

[5] H. R. Sá and et al, “Supervised learning for link prediction in weighted networks,” in *III International Workshop on Web and Text Intelligence*, 2010.

[6] Z. Bao, Y. Zeng, and Y. Tay, “Sonlp: Social network link prediction by principal component regression,” in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2013, pp. 364–371.

[7] Z. Lu, B. Savas, W. Tang, and I. S. Dhillon, “Supervised link prediction using multiple sources,” in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 923–928.

[8] T. Zhou and et al, “Predicting missing links via local information,” *The European Physical Journal B*, vol. 71, no. 4, pp. 623–630, 2009.

[9] L. A. Adamic and E. Adar, “Friends and neighbors on the web,” *Social networks*, vol. 25, no. 3, pp. 211–230, 2003.

[10] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[11] M. E. Newman, “Clustering and preferential attachment in growing networks,” *Physical Review E*, vol. 64, no. 2, p. 025102, 2001.

[12] M. Magnani, B. Micenková, and L. Rossi, “Combinatorial analysis of multiple networks,” *The Computing Research Repository (CoRR)*, vol. abs/1303.4986, 2013.

[13] G. Barnett and et al, “Examining the international internet using multiple measures: new methods for measuring the communication base of globalized cyberspace,” *Quality & Quantity*, pp. 563–575, 2014.

[14] S. F. Everton, *The Noordin Top Terrorist Network*. Cambridge University Press, 2012. [Online]. Available: http://dx.doi.org/10.1017/CBO9781139136877.019

[15] G. Gousios, “The ghtorrent dataset and tool suite,” in *Proceedings of the 10th Working Conference on Mining Software Repositories*, ser. MSR’13, 2013, pp. 233–236.

[16] E. Cho, S. A. Myers, and J. Leskovec, “Friendship and mobility: User movement in location-based social networks,” in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’11. ACM, 2011, pp. 1082–1090.

[17] E. Lazega, *The Collegial Phenomenon: The Social Mechanisms of Cooperation among Peers in a Corporate Law Partnership*. Oxford: Oxford University Press, 2012.

[18] D. J. Watts and S. H. Strogatz, “Collective dynamics of small-worldnetworks,” *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[19] P. Erdős and A. Rényi, “On the evolution of random graphs,” *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.