Link Prediction in Highly Fractional Data Sets

Michael Fire, Rami Puzis, and Yuval Elovici

1 Introduction

In recent years, online social networks have grown in scale and variability and offer individuals with similar interests the possibility of exchanging ideas and networking. On the one hand, social networks create new opportunities to develop friendships, share ideas, and conduct business. On the other hand, they are also an effective media tool for plotting crime and organizing extremists groups around the world. Online social networks, such as Facebook, Google+, and Twitter are hard to track due to their massive scale and increased awareness of privacy. Criminals and terrorists strive to hide their relationships, especially those that can associate them with a executed terror act.

A large portion of recent research in social network analysis has been targeted at identifying these hidden relationships in social networks. These research endeavours are usually referred to as *link prediction* methods. These methods are able to detect existing social ties that have not been established in a particular social network [8, 10, 17, 23, 29, 33]. In the security and counter-terrorism domains link prediction can assist in identifying hidden groups of terrorists or criminals [17]. However, link prediction is also useful for civil applications, such as friend-suggestion mechanisms embedded in online social networks. For example, in bioinformatics, link prediction can be used to find interactions between proteins [3], and in e-commerce, it can help build recommendation systems [19].

M. Fire (✉) · Y. Elovici
Department of Information Systems Engineering, Telekom Innovation Laboratories,
Ben-Gurion University of the Negev, Beersheva, Israel,
e-mail: mickyfi@bgu.ac.il; elovici@bgu.ac.il

R. Puzis
University of Maryland Institute for Advanced Computer Studies (UMIACS),
College Park, MD, USA
e-mail: puzis@umd.edu

V.S. Subrahmanian (ed.), *Handbook of Computational Approaches to Counterterrorism*,
DOI 10.1007/978-1-4614-5311-6_14,

Several different methods for solving the link prediction problem have been proposed in recent years. These days, the majority of solutions are based on supervised machine learning methods, such as Bayesian probabilistic models, relational Bayesian networks, and linear algebraic methods. Further details on these approaches can be found in a thorough survey written by Hasan and Zaki [18]. We briefly describe a selected few in Sect. 2.2.

We focus on link perdition methods based on machine learning classifiers trained on a set of topological features. We use a set of well-known features, such as connectivity features, intersection and union of the friends groups, *Jaccard's-Coefficient* [34], *Preferential-Attachment* score [8, 17], and the *Friends-Measure*, introduced in [14]. The latter is a variation of the Katz measure [20] and estimates how well the friends of two users know each other.

Many link prediction methods are applied to large social networks where most of the links are assumed to be known. This is a reasonable assumption when the main objective is to predict ties that are not yet established in the social network. However, terrorist social networks mined from open sources are typically small and very partial due to the efforts of their subjects to obfuscate their activity.

In this study, we investigate the effects of dataset partiality on the effectiveness of link prediction by gradually reducing the number of visible links in the test networks. The effectiveness of classifiers was evaluated using six different types of data sets: a group of Facebook users sharing the same employer, a researchers community from Academia.edu [12], Friends and Family SMS messages social network [4], Students Cooperation Network [13], AnyBeat social network,[1] and the Profiles in Terror (PIT) dataset [31]. Information on these social networks is presented in Sect. 3. Evaluation results presented in Sect. 5 demonstrate that classification quality (in terms of Area Under the ROC Curve (AUC)) degrades with the number of visible links. Nevertheless, even a small fraction of visible links (5–20 %) helps in solving the link prediction problem with chances significantly higher than random.

The remainder of the paper is organized as follows. In Sect. 2, we review previous studies on terrorist's social networks and link prediction. In Sect. 3, we describe the datasets used in this study. Experimental setup and the features extracted from the structure of social networks are described in Sect. 4.1. We present the experimental results in Sect. 5, and conclusions in Sect. 6.

[1] http://www.anybeat.com

2 Background

2.1 Social Networks of Terrorists

Over the last two decades, social networks have been studied fairly extensively in the general context of analyzing interactions between people and determining the important structural patterns of such interactions [1]. In the previous decade, even before September 11, 2001, social network analysis was recognized as a tool for fighting the war against criminal organizations in an age where there is no well-defined enemy with a formal hierarchical organization [5]. Moreover, after the September 11, 2001 events, social network analysis became a well-known mainstream tool to help the fight against terror [26].

Several studies have analyzed terror organization social networks based on graph structural features. In the winter of 2002, Krebs [21] studied Al-Qaeda's network structural properties by collecting publicly available data on the Al-Qaeda hijackers. Rothenberg [28] conjectured on the structure of the al Qaeda network based on public media sources. After the Madrid bombing, in March 11, 2004, Rodriguez [27] used public sources to construct and study the terrorists' network. He showed that the terror organization network included mainly weak ties that are hard to detect. In 2004, Sageman [30] used various public sources, mostly records of trials, to collect and analyze 400 terrorist biographies. He discovered that 88 % of the terrorists had friendship or family bonds to the Jihad. In 2005, Basu [7] studied terrorists' organization in India. He used social network analysis, such as the betweenness measure, to identify major groups of terrorists and key players. In 2010, Wiil et al. [35] studied a recent Denmark terror plan. By using data mining techniques, they were able to construct, from public sources, the social network of David Coleman Headley, one of the terror plan conspires.

Attempts to reconstruct the social networks of terrorists requires a significant effort spent on mining the Web for publicly available information and free text analysis. This typically results in the ability to obtain small networks only with a high likelihood of missing information. In this study, we attempt to predict links inside social networks where a substantial amount of the network's links data were missing. A similar idea was studied by Dombroski et al. [9]. Their study examined the possibilities of using the inherent structures observed in social networks to make predictions of networks using limited and missing information.

2.2 Link Prediction

In this study, we focus on link prediction methods based on supervised machine learning algorithms. These methods were first introduced by Liben-Nowell and Kleinberg in 2003 [23], who used graph topological features in a study on five co-authorship networks, each containing several thousands of vertices. In 2006, Hasan et al. [17] increased the scale of analyzed networks to hundreds of thousands of

nodes by analyzing the DBLP and BIOBASE co-authorship networks. Supervised learning was also applied by many other researchers to solve the link prediction problems, for example in [10,22,29]. Initially, the proposed link prediction solutions were tested on bibliographic or on co-authorship data sets [10, 17, 23, 29].

In 2009, Song et al. used matrix factorization to estimate the structural similarity between profiles in online social network services, such as Facebook and MySpace [33]. In 2010, Leskovec et al. [22] studied a similar problem of predicting links' signs. Recently, Zaki and Hasan [18] published through survey on link prediction in social networks.

In 2011, the IJCNN social network challenge [24] inspired several publications on link prediction using topological network analysis. These publications proposed and evaluated different methods for predicting links in social networks. Narayanan et al. won the challenge by using a method that combined machine-learning algorithms with de-anonymization [25]. Cukierski et al. [8] took second place by extracting 94 distinct features for each one of the of several thousands of vertex pairs in the training data and analyzing it with the Random Forest algorithm. Recently, Fire et al. presented a method for predicting links inside communities; their methods used supervised learning ensemble classifiers constructed by using small training sets only which consisted of several hundreds of examples [12].

When no information on the social network besides its structure is available, it is crucial to define and calculate the features that are as informative as possible. On the one hand, large networks containing millions of vertices and links pose a scalability challenge and require the use of easy to compute topological features extracted from the neighbourhoods of the tested vertices. For example, Facebook has more than 901 million registered users and each month many new users are added [11]. On the other hand, small networks, such as the terrorist datasets available today (e.g., [21, 36]), pose a different challenge. They contain too few links and vertices to construct a large training set. Moreover, these networks are much more prone to noise than their huge counterparts because the existence or absence of every link may significantly change the values of the extracted features. In this study, we take the latter challenge to the extreme and evaluate structural link prediction methods while gradually removing random links from organizational, group, and terrorist affiliation networks.

3 Social Network Datasets

In this study, we apply link prediction classifiers to six labeled social network datasets (see Table 1), namely, Profile in Terror (PIT) [36], AnyBeat network,[2] a group of Facebook[3] users, Academia.edu [14, 15], Friends and Family study [2], and Students' Cooperation Social Network [13].

[2] http://www.anybeat.com

[3] http://www.facebook.com

Table 1 Datasets

Network	Directed	Vertices	Links	Positive examples	Visible networks	Features
PIT	No	244	840	420	200	8
AnyBeat	Yes	12,645	67,053	25,000	40	12
Co-Workers	No	165	722	361	200	8
Researchers	Yes	207	702	351	200	12
F&F	Yes	103	281	140	200	12
Students	No	185	311	155	200	8

Profile in Terror (PIT) [31] is a data set that captures intelligence information extracted from publicly available sources collected by the MIND Lab at UMD.[4] This data set contains 851 labeled relationships among terrorists and was previously used in multi-label link classification in [36]. Each relationship can have one or more labels from: colleague, congregate, contact, or family. In this study, we disregard the relationship labels and treat this data set as a flat network which contains 840 links and 244 vertices. Figure 1 depicts the graph topology.[5]

AnyBeat. "AnyBeat is an online community; a public gathering place where you can interact with people from around your neighborhood or across the world". AnyBeat is a relatively new social network in which members can log in without using their real name and members can follow any other member in the network. In this study, we evaluated our algorithm on a major part of the network's topology, which was obtained using a dedicated web crawler. The topology contained 12,645 vertices and 67,053 links (see Fig. 2). Among the networks studied in this study, AnyBeat is outstanding in its size. We included a network of more than 10,000 vertices in order to shade the results of the study and focused mainly on small partially visible social networks. As presented in Sect. 5, link prediction appears to be significantly easier for this network.

A Facebook group of co-workers (Co-Workers). Facebook is a website and social networking service that was launched in February 2004. As of March 2012, Facebook has more than 901 million registered users [11]. Facebook users may create a personal profile, add other users as friends, and interact with other members. Friendship ties in Facebook are reciprocal, therefore, we refer to the underlying group's social network as undirected. We extracted a community of co-workers who, according to their Facebook profile, worked for the same well-known high-tech company. The graph representing the co-workers' community network contains 165 vertices and 726 links and was obtained using a web crawler in the beginning of February 2012 (see Fig. 3).

[4] http://www.mindswap.org/

[5] All the social networks figures in this paper were created by Cytoscape software [32].

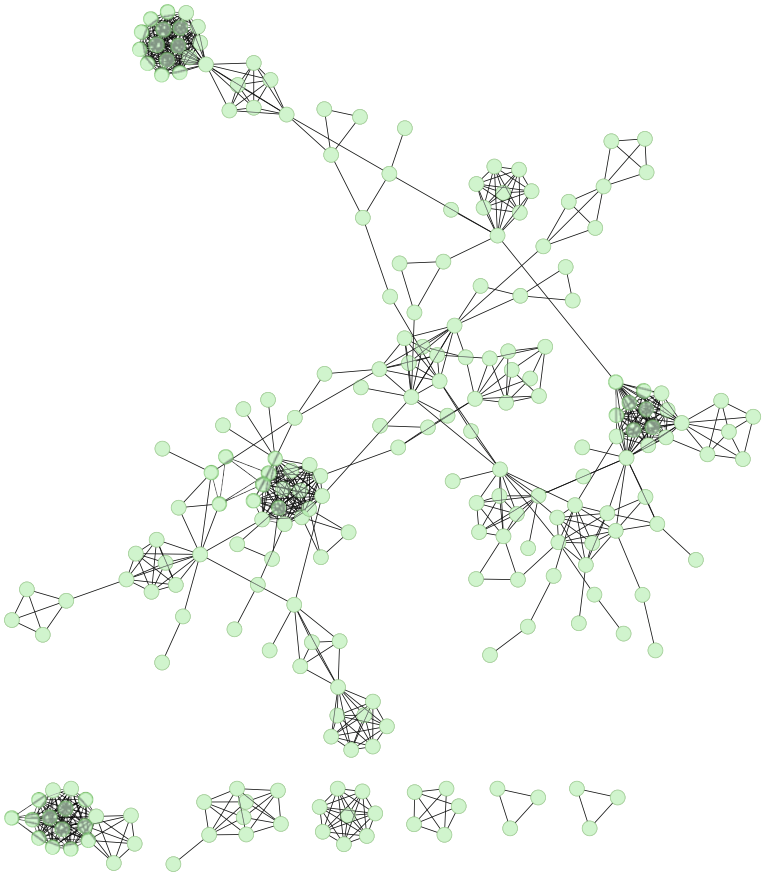

Fig. 1 Profile in terror social network

Ivy League University (Researchers). Academia.edu[6] is a platform for academics to share and follow research underway in a particular field or discipline. Members upload and share their papers with other researchers in over 100,000 fields and categories. An Academia social network member may choose to follow any of the network's members; hence, the directed nature of the links within this network. We evaluated our classifiers for a small community of researchers who, according to their Academia.edu profiles, belonged to the same Ivy League University. The researchers' community network graph contained 207 nodes and 702 links (see Fig. 4) and was obtained using a web crawler.

Friends and Family (F&F). The Friends and Family dataset contains rich data signals gathered from the smart-phones of 140 adult members of a young-family

[6] http://wwww.academia.edu

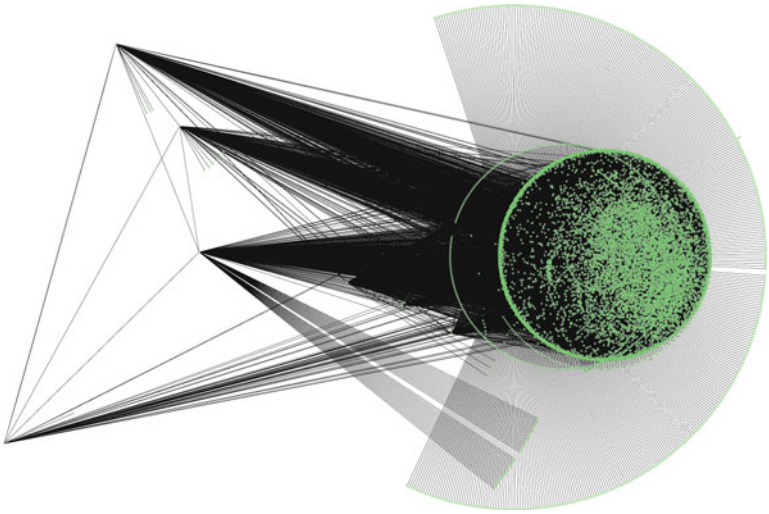

Fig. 2 AnyBeat social network

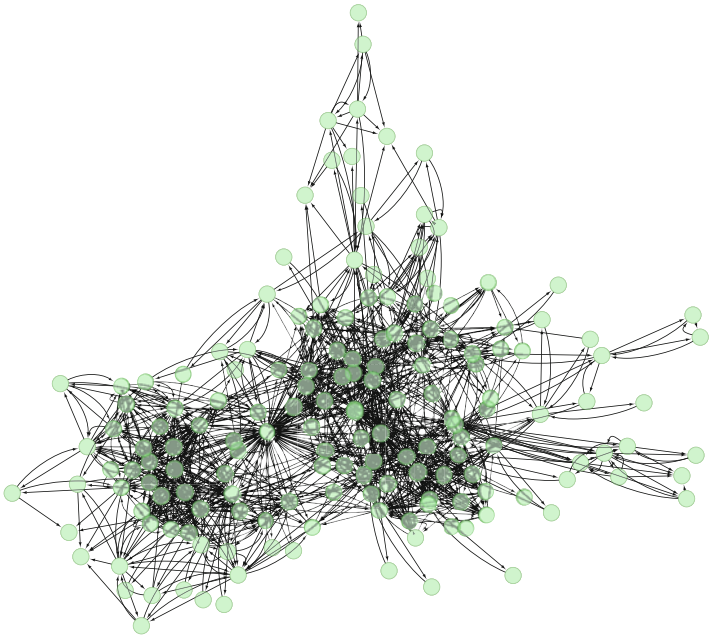

Fig. 3 Facebook coworker community social network

residential community. The data were collected over the course of 1 year [2]. We evaluated our classifiers on a social network that was constructed based on SMS messages sent and received by the members. The SMS messages social network directed graph contained 103 nodes and 281 links (see Fig. 5).

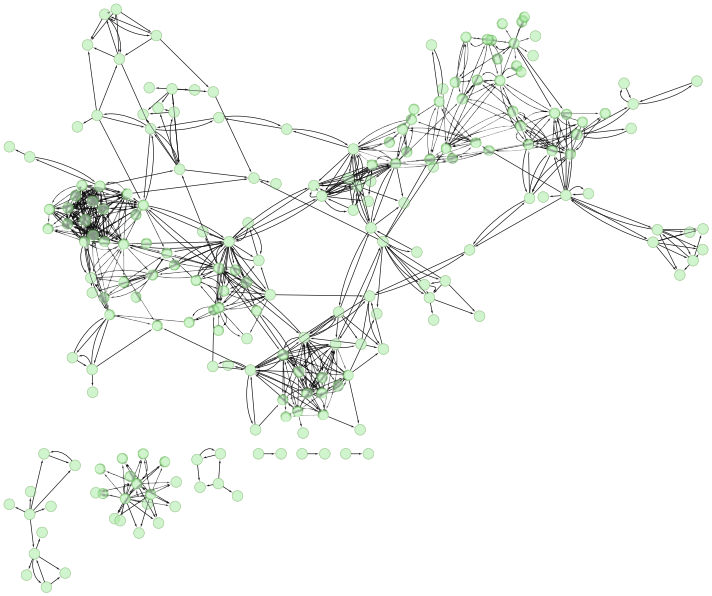

Fig. 4 Academia.edu researchers community social network

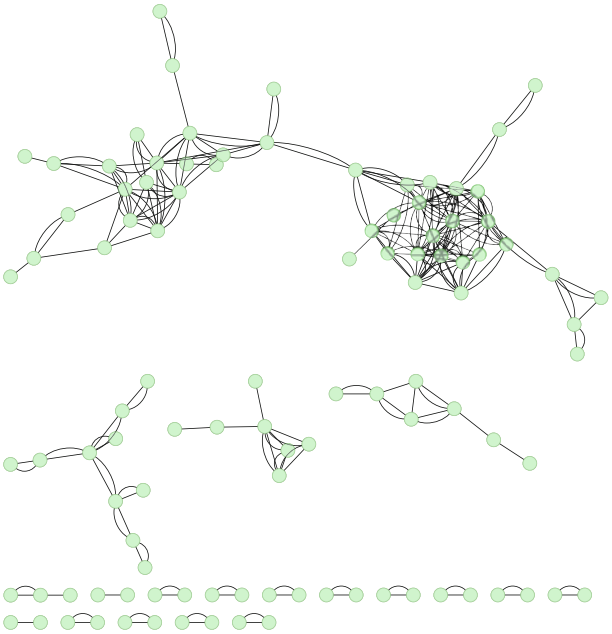

Fig. 5 Friends and family SMS messages social network

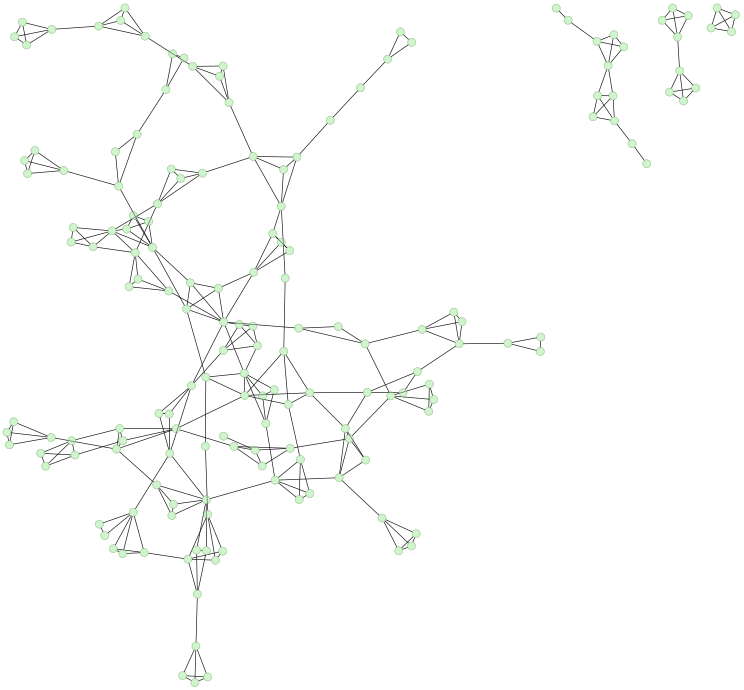

Fig. 6 Students' cooperation social network

Students' Cooperation Social Network (Students). The students' cooperation social network was constructed from data collected during a "Computer and Network Security" course; a mandatory course taught by two of this paper's authors at Ben-Gurion University [13]. The social network contains data collected from 185 participating students from two different departments. The course's social network was created by analyzing the implicit and explicit cooperation among the students while doing their homework assignments. The Students' Cooperation graph contained 185 nodes and 311 links (see Fig. 6).

4 Methods and Experiments

4.1 Experimental Setup

The goal of structural link prediction is to identify a set of hidden links within a social network's structure by analyzing the topology of the known (visible) network. As a first step in our experiments, we preprocessed each one of the social networks

described in Sect. 3. The preprocessing process is slightly different for the AnyBeat networks due to its size.

We will refer to the unmodified social networks as *original* in the rest of this paper. We randomly removed a portion of links from each original network to generate a *visible* network. The number of removed links ranged from 0 to 95 % in steps of 5 %. For each one of these 20 fractions we chose several random sets of removed links that spawn several different visible networks of the same size; ten random sets for the small networks and two random sets for the AnyBeat network. Overall, 200 *visible* networks were created for each one of the small *original* networks and 40 *visible* networks were created for AnyBeat. Our goal was to predict the links in the *original* network by extracting features from the *visible* network topology.

Next, in order to evaluate the machine learning classifiers, we generated training and testing data sets from each visible network with the original network treated as the ground truth. Note that due to privacy profiles and imperfect data acquirement methods, the original networks may have missing links as well. We will disregard these links and assume acquaintances between individuals that are linked together in the original networks only.

Every instance in the dataset represents a possible candidate link. The target attribute of each instance is binary, indicating the existence or absence of a link in the original social network. A set of structural features was extracted from the *visible* part of the corresponding social networks. As opposed to the 54 features discussed in [15], we have only 7 and 11 of the most efficient features for undirected and directed networks, respectively in this study. These features are briefly described in Sect. 4.2.

In order to create a balanced data set for training the classifiers, we included the same number of positive and negative examples. Positive examples are vertex pairs that are connected in the *original* network. Negative examples include the same number of vertex pairs, but this time, random pairs of vertices not connected in the *original* network. The method for selecting the negative examples in this study corresponds to the *easy* dataset, as described in [15]. Note that the features describing each positive or negative example were extracted from the structure of the *visible* networks. Table 1 summarizes the datasets used in this study.

WEKA [16], a popular suite of machine learning software written in Java and developed at the University of Waikato, New Zealand, was used as the machine learning platform for this study. We used a WEKA's J48 classifier with ten minimum objects that showed effective performance in past experiments. We evaluated our results using a ten-folds cross validation method.

Next, we describe the set of features extracted from the visible social network graphs. The features used in this study are a subset of a more extensive set of features investigated in [15].

4.2 Feature Extraction

This section describes the features extracted from the social network structure in order to build our link prediction classifiers. The extracted features are based primarily on the Friends-features subset, as suggested by Fire et al. [15].

Let $G =< V, E >$ be the graph representing the structure of a social network. Links in the graph are denoted by $e = (u, v) \in E$ where $u, v \in V$ are vertices in the graph. Our goal is to construct classifiers capable of computing the likelihood of $(u, v) \in E$ or $(u, v) \notin E$ for every two vertices $u, v \in V$. To achieve this goal, we extracted the following features for each pair, (u, v), in our datasets.

1. **Vertex degree:** Let $v \in V$ be some vertex, we can define the neighborhood of v by:
$$\Gamma(v) := \{u|(u, v) \in E \text{ or } (v, u) \in E\}$$
If G is a directed graph, we can also define the following neighborhoods:
$$\Gamma_{in}(v) := \{u|(u, v) \in E\}$$
$$\Gamma_{out}(v) := \{u|(v, u) \in E\}$$
We define the degree features for directed and undirected networks as the sizes of the respective neighborhoods:
$$degree(v) := |\Gamma(v)| \quad (1)$$
$$degree_{in}(v) := |\Gamma_{in}(v)| \quad (2)$$
$$degree_{out}(v) := |\Gamma_{out}(v)| \quad (3)$$
The degree features measure the number of friends v has. If we look at a directed graph, such as Academia.edu, the meaning of the degree feature is how many other members of the community v follows (out-degree), and how many members of the community follow v (in-degree).
2. **Common-Friends:** Let $u, v \in V$ be a pair of vertices in the network; we define the common friends of u and v to be all the vertices in the network that are friends of both u and v. Formally, common friends is the size of the intersection of the respective neighborhoods:
$$Common\text{-}Friends(u, v) := |\Gamma(v) \cap \Gamma(u)| \quad (4)$$
The Common-Friends feature was widely used in previous works for predicting links in different datasets [8, 14, 19, 23, 29, 33].
3. **Total-Friends:** Let $u, v \in V$ be a pair of vertices; we can define the number of distinct friends of u and v as the size of the union of the respective neighborhoods:
$$Total\text{-}Friends(u, v) := |\Gamma(u) \cup \Gamma(v)| \quad (5)$$

4. **Jaccard's-Coefficient:** Jaccard's-Coefficient is a well-known feature for link prediction [8, 14, 19, 23, 29, 33]. This feature, which measures the similarity among sets of nodes, is defined as the size of the intersection divided by the size of the union of the sample sets:

$$Jaccard's\text{-}Coefficient(u,v) := \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} \tag{6}$$

   In our approach, this measure indicates whether two community members have a significant number of common friends regardless of their total number of friends. A higher value of the Jaccard's-Coefficient typically indicates a stronger connection between two nodes in the network.
5. **Preferential-Attachment-Score:** The Preferential-Attachment-Score indicates the likelihood of a new link to be formed between the vertices u and v, according to the preferential attachment model [6]. It is defined as the multiplication of the number of friends of u and v.

$$Preferential\text{-}Attachment\text{-}Score(u,v) := |\Gamma(u)| \cdot |\Gamma(v)| \tag{7}$$

   The Preferential-Attachment score was used many times in past research, for examples see [8, 14, 17].
6. **Friends-Measure:** Let $u, v \in V$ be two vertices; the Friends-Measure of u and v is the extent to which their friends are interconnected. The higher the number of connections between u and v's friends, the greater the chance that u and v know each other.

$$Friends\text{-}Measure(u,v) := \sum_{x \in \Gamma(u)} \sum_{y \in \Gamma(v)} \delta(x,y) \tag{8}$$

   where $\delta(x, y)$ is defined as:

$$\delta(x,y) := \begin{cases} 1 & \text{if } x = y \text{ or } (x,y) \in E \text{ or } (y,x) \in E \\ 0 & \text{otherwise} \end{cases}$$

   The Friends-Measure was first presented by Fire et al. [14].

5 Results

For each undirected and directed visible network, we extracted 7 and 11 features, respectively, (see Sect. 4.2) and evaluated the specified machine-learning algorithms (see Sect. 4.1) using a ten-fold cross-validation approach. The AUC results of the J48 algorithm on all the networks where no edges were removed from the networks are presented in Fig. 7.

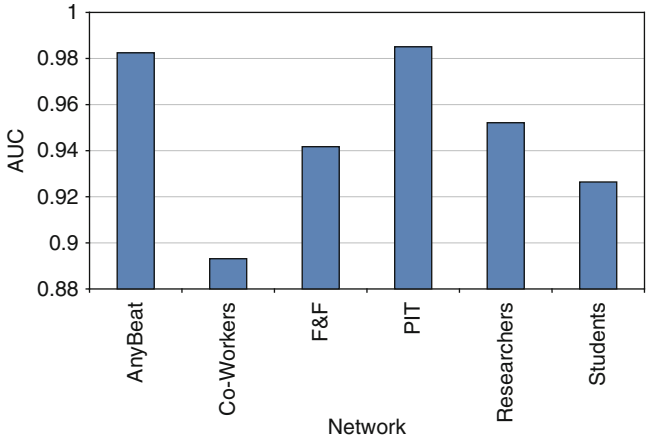

Fig. 7 AUC Results – J48 with no removed links

Table 2 Effectiveness of link prediction in fractional networks

Network	Percent of visible links				
	100 %	80 %	60 %	40 %	20 %
PIT	0.985	0.980	0.966	0.940	0.789
Co-Workers	0.893	0.858	0.820	0.791	0.738
Researchers	0.952	0.942	0.920	0.895	0.799
F&F	0.941	0.928	0.908	0.877	0.744
Students	0.926	0.877	0.841	0.762	0.629
AnyBeat	0.982	0.979	0.972	0.966	0.954

Table 2 and Fig. 8 present the AUC results of the J48 algorithm for each social network with various numbers of removed links. As expected, we noticed a degradation in the AUC values as more edges are removed from the network. The slope of the AUC as a function of the visible network size increases with the number of removed vertices. In the terrorists affiliation network (PIT), the AUC remains above 0.9, even when features in the dataset are computed according to 40 % of the original links.

Another notable observation is that the classification results (with the exception of the Students' Cooperation network) are significantly above random, even when only 5–10 % of the *original* links are *visible* with respect to the computation of structural features. The Students' Cooperation network was based on links between students in a single class taught for a short period of time. Therefore, the network did not contain as many cliques as in the other social networks. We believe that the existence of large cliques, (as in PIT network) or densely connected parts in the network, makes it easier to predict ties randomly removed from the network.

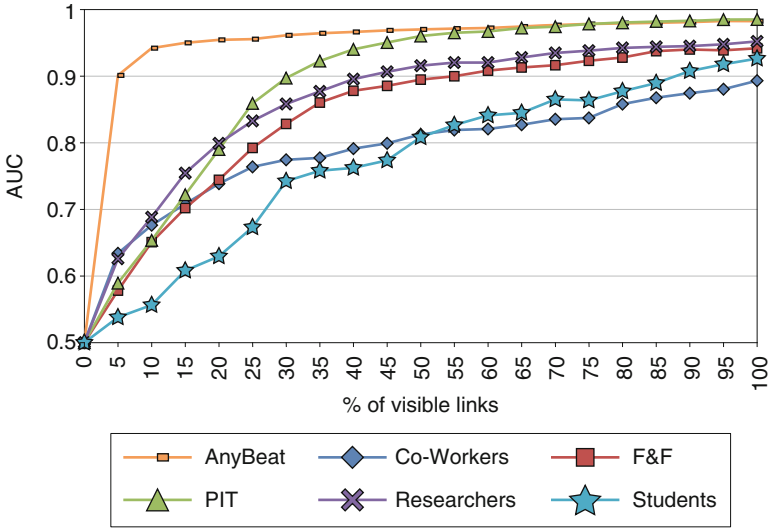

Fig. 8 AUC as a function of the fraction of visible links for various social networks

A final interesting observation is that the best AUC values were obtained for the AnyBeat network. Link prediction in this network is much easier than in the smaller networks when significant portions of the network are removed. The AUC remains above 0.9, even when there are only 5 % of visible links. Note that 5 % of visible links in AnyBeat is 3,352, which is a few orders of magnitude higher than in the other networks.

To obtain an indication of the usefulness of the various features, we analyzed their importance using Weka's Information Gain attribute selection algorithm. The the InfoGain results for different datasets are presented in Tables 3 and 4. In order to check the effect of link removal on the usefulness of various features, we computed the InfoGain values for the original networks and for networks with only 5 % of visible links. The results indicate that InfoGain of all features drops and the set of useful features changes when random links are removed from the dataset. Only the *Preferential-Attachment-Score* is useful for both collections of datasets. Nonetheless, the InfoGain values of connectivity degree features drop in Table 4 and they become more useful relative to the other features. This effect is even more apparent for the *Total-friends* feature that has the highest InfoGain value in Table 4. Yet, this value is more than two times lower than the respective value in Table 3. In contrast, we can clearly see that *Common-friends*, *Friends-Measure*, and *Jaccard's Coefficient* have high InfoGain values for networks where all links are visible and the lowest InfoGain values for networks with only 5 % of visible links.

Table 3 InfoGain of the extracted features for various social networks

	|Inner-subgraph(u,v)|	|nh-subgraph(u,v)|	|nh-subgraph+(u,v)|	Common-Friends(u,v)	d(u)	Density(nh-subgraph(u,v))	dout(u)	**Friends-measure(u,v)**	Jaccard's-coefficient(u,v)	Opposite-direction-friends(u,v)	Preferential-attachment-score(u,v)	scc(Inner-subgraph(u,v))	scc(nh-subgraph(u,v))	scc(nh-subgraph+(u,v))	Shortest-path(u,v)	Total-Friends(u,v)	Transitive-Friends(u,v)
Academia	0.6	0.2	0.3	0.5	0.3	0.2	0.3	**0.7**	0.5	0.3	0.4	0.2	0.2	0.2	0.4	0.3	0.4
Facebook	0.7	0.2	0.2	0.7	0.3	0.2		**0.7**	0.7		0.4	0.0	0.0	0.5	0.7	0.3	
Flickr	0.3	0.2	0.3	0.1	0.3	0.2	0.2	**0.4**	0.1	0.3	0.7	0.7	0.7	0.7	0.0	0.7	0.0
TheMarker	0.5	0.5	0.5	0.5	0.5	0.6	0.3	**0.6**	0.8	0.3	0.8	0.6	0.7	0.7	0.0	0.6	0.2
YouTube	0.5	0.4	0.5	0.3	0.5	0.4	0.5	**0.6**	0.3	0.6	0.7	0.5	0.5	0.4	0.2	0.6	0.3
Average	0.51	0.32	0.38	0.42	0.38	0.32	0.36	**0.61**	0.47	0.39	0.60	0.40	0.43	0.50	0.26	0.48	0.22

Table 4 InfoGain of the extracted features for fractional social networks with 5 % of visible links

	|Inner-subgraph(u,v)|	|nh-subgraph(u,v)|	|nh-subgraph+(u,v)|	Common-Friends(u,v)	d(u)	Density(nh-subgraph(u,v))	dout(u)	**Friends-measure(u,v)**	Jaccard's-coefficient(u,v)	Opposite-direction-friends(u,v)	Preferential-attachment-score(u,v)	scc(Inner-subgraph(u,v))	scc(nh-subgraph(u,v))	scc(nh-subgraph+(u,v))	Shortest-path(u,v)	Total-Friends(u,v)	Transitive-Friends(u,v)
Academia	0.1	0.2	0.2	0.3	0.2	0.1	0.3	**0.3**	0.3	0.3	sri	0.1		0.1	0.2	0.1	0.4
Facebook	0.2	0.2	0.2	0.4	0.3	0.2		**0.3**	0.3		0.2	0.0	0.0	0.0	0.0	0.1	
Flickr	0.5	0.5	0.5	0.5	0.5	0.6	0.3	**0.6**	0.8	0.3	0.8	0.6	0.7	0.7	0.0	0.6	0.2
TheMarker	0.4	0.4	0.4	0.3	0.4	0.4		**0.4**	0.2		0.3	0.0	0.0	0.0	0.0	0.2	
YouTube	0.2	0.3	0.3	0.5	0.3	0.2	0.3	**0.6**	0.5	0.6	0.3	0.1		0.2	0.2	0.2	0.5
Average	0.28	0.32	0.35	0.41	0.35	0.31	0.30	**0.42**	0.38	0.38	0.35	0.17	0.25	0.20	0.09	0.23	0.36

6 Conclusion

Today's terror acts are conducted by groups of individuals who link with facilitators and other groups. Together, such a network has the resources, the means, and the insights to execute these attacks [35]. It is important to establish connections between the participating parties in order to understand their group dynamics and effectively mitigate their activity [30].

In this chapter, we presented the results of a study on link prediction in small scale, highly fractional networks, such as terrorists networks obtained from an analysis of publicly available media. We employed machine learning classifiers trained on a set of features extracted from the network structure. The results indicate that, in contrast to large networks where effective link prediction can be performed even when the vast majority of links are hidden, small scale networks require at least 30–50 % visible links to get AUC values above 0.8. We also notice that networks containing large cliques, or densely connected parts, are less vulnerable to random removal of links.

In conclusion, as expected, our results show that Information Gain values of the structural features drop when most of the links are removed from the network. However, it should be noted that the decrease in the Information Gain is not uniform. Features that were the most useful for predicting links in the original network become the least useful when only 5 % of the links are visible. These features are the number of common friends, the Friends-Measure [14], and Jaccard's Coefficient [34]. All these measures refer to the connections and similarities between vertices in both neighborhoods of the two ends of the link being tested. When there are enough connections, these measures make sense, however, when only few links are visible, the most useful information is the number of acquaintances each vertex has.

Availability

An anonymous version of the Co-Workers, Researchers, Students, and the AnyBeat social network topologies crawled as a part of this study are available on our research group website http://proj.ise.bgu.ac.il/sns/.

References

1. Aggarwal C (2011) Social network data analytics. Springer, New York
2. Aharony N, Pan W, Ip C, Khayal I, Pentland A (2011) Social fMRI: investigating and shaping social mechanisms in the real world. Pervasive and Mobile Computing 7(6):643–659
3. Airoldi E, Blei D, Fienberg S, Xing E, Jaakkola T (2006) Mixed membership stochastic block models for relational data with application to protein-protein interactions. In: Proceedings of the international biometrics society annual meeting. ENAR, Tampa, FL, USA

4. Altshuler Y, Fire M, Elovici Y, Pentland A (2012) How many makes a crowd? On the evolution of learning as a factor of community coverage, Social Computing, Behavioral – Cultural Modeling and Prediction, Lecture Notes in Computer Science, (7227), Springer, Berlin/Heidelberg, pp 43–52
5. Arquilla J, Ronfeldt D (2001) Networks and netwars: the future of terror, crime, and militancy. 1382. Rand Corp
6. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512
7. Basu A (2005) Social network analysis of terrorist organizations in india. In: North American Association for Computational Social and Organizational Science (NAACSOS) conference. CASOS, Notre Dame, Indiana, USA, pp 26–28
8. Cukierski WJ, Hamner B, Yang B (2011) Graph-based features for supervised link prediction. International joint conference on neural networks. IEEE, San Jose, California
9. Dombroski M, Fischbeck P, Carley K (2003) Estimating the shape of covert networks. In: Proceedings of the 8th international command and control research and technology symposium
10. Doppa JR, Yu J, Tadepalli P, Getoor L (2009) Chance-constrained programs for link prediction. In Proceedings of workshop on analyzing networks and learning with graphs at NIPS conference
11. Facebook-Newsroom. http://www.facebook.com
12. Fire M, Katz G, Rokach L, Elovici Y (2012) Links reconstruction attack using link prediction, Security and Privacy in Social Networks , pp 181–196, Springer, Berlin/Heidelberg
13. Fire M, Katz G, Elovici Y, Shapria B, Rokach L (2012) Predicting student exam's scores by analyzing social network data, AMT 2012, LNCS 7669, pp. 584–595. Springer, Heidelberg
14. Fire M, Tenenboim L, Lesser O, Puzis R, Rokach L, Elovici Y (2011) Link prediction in social networks using computationally efficient topological features. In: Privacy, security, risk and trust (PASSAT), 2011 IEEE third international conference on and 2011 IEEE third international confernece on social computing (SocialCom). IEEE, Washington, DC, USA pp 73–80
15. Fire M, Tenenboim L, Puzis R, Lesser O, Rokach L, Elovici Y. Computationally efficient link prediction in variety of social networks, (working paper)
16. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: an update. SIGKDD Explor Newsl 11:10–18. doi:http://doi.acm.org/10.1145/1656274.1656278
17. Hasan MA, Chaoji V, Salem S, Zaki M (2006) Link prediction using supervised learning. SDM workshop of link analysis, counterterrorism and security. SIAM, Lake Buena Vista, Florida
18. Hasan MA, Zaki MJ (2011) Social network data analytics. Springer, New York
19. Huang Z, Li X, Chen H (2005) Link prediction approach to collaborative filtering. Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries. ACM, Denver, CO, USA
20. Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1):39–43. http://ideas.repec.org/a/spr/psycho/v18y1953i1p39-43.html
21. Krebs V (2001) Mapping networks of terrorist cells. Connections 24(3):43–52
22. 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. ACM, New York, NY, USA pp 641–650
23. Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. J Am soc Inf Sci Technol 58(7):1019–1031
24. Nachbar D (2010) IJCNN social network challenge. http://www.kaggle.com/c/socialNetwork/Data
25. Narayanan A, Shi E, Rubinstein B (2011) Link prediction by de-anonymization: How we won the kaggle social network challenge. In: The 2011 international joint conference on neural networks (IJCNN). IEEE, Washington, DC, USA pp 1825–1834
26. Ressler S (2006) Social network analysis as an approach to combat terrorism: Past, present, and future research. Homel Secur Aff 2(2):1–10
27. Rodriquez J (2005) The march 11th terrorist network: in its weakness lies its strength, VIII Congreso Espaol de Sociologa, Alicante, Spain

28. Rothenberg R (2001) From whole cloth: making up the terrorist network. Connections 24(3):36–42
29. Sa HR, Prudencio RBC (2010) Supervised learning for link prediction in weighted networks. III international workshop on web and text intelligence. São Bernardo do Campo, Brazil
30. Sageman M (2004) Understanding terror networks. University of Pennsylvania Pr
31. Sen P, Namata GM, Bilgic M, Getoor L, Gallagher B, Eliassi-Rad T (2008) Collective classification in network data. AI Mag 29(3):93–106
32. Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, Amin, N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504
33. Song HH, Cho TW, Dave V, Zhang Y, Qiu L (2009) Scalable proximity estimation and link prediction in online social networks. In: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference, IMC '09. ACM, New York, pp 322–335. doi:http://doi.acm.org/10.1145/1644893.1644932
34. Tan PN, Steinbach M, Kumar V (2005) Introduction to Data Mining. Addison Wesley, Boston, Massachusetts, USA
35. Wiil U, Memon N, Karampelas P (2010) Detecting new trends in terrorist networks. In: 2010 international conference on advances in social networks analysis and mining (ASONAM). IEEE, Washington, DC, USA pp 435–440
36. Zhao B, Sen P, Getoor L (2006) Event classification and relationship labeling in affiliation networks. In: Proceedings of the workshop on statistical network analysis (SNA) at the 23rd international conference on machine learning (ICML). Pittsburgh, Pennsylvania