Intruder or Welcome Friend: Inferring Group Membership in Online Social Networks

Ofrit Lesser[1], Lena Tenenboim-Chekina[2], Lior Rokach[1] and Yuval Elovici[1,2]

[1]Department of Information Systems Engineering, Ben Gurion University of the Negev, Israel,
[2]Telekom Innovation Laboratories at Ben-Gurion University of the Negev

lessero@post.bgu.ac.il,{lenat,liorrk,elovici}@bgu.ac.il

Abstract. Inferring Online Social Networks (OSN) group members may help to evaluate the authenticity of an applicant asking to join a certain group, and secure vulnerable populations online, such as children. We propose machine learning based methods, which associate OSN members' affiliation with virtual groups based on personal, topological, and group affiliation features. The study applies and evaluates the methods empirically, on two social networks (Ning and TheMarker). The experimental results demonstrate that one can accurately determine the group genuine members. Our study compares personal, topological and group based classification models. The results show that topological and group affiliation attributes contribute the most to group inference accuracy. Additionally, we examine the relations among the groups and identify group clustering tendencies where some groups are more tightly connected than others.

Keywords: social networks, group prediction, machine learning

1 Introduction

Online Social Network (OSN), characteristics differ from each other, however the majority of these networks contain two main capabilities: connecting two members via a friendship connection and a group creation. These mechanisms simulate real life scenarios. While a friendship connection signifies or correlates with a tie between two individuals (e.g. relatives, friends, acquaintances, etc.), the group formation correlates with a community of multiple individuals based on similarity (i.e. residence, workplace, interests, etc.). Due to the OSN's rise and the opportunity to obtain large scale datasets, many recent studies have focused on various different aspects of the social network, such as: structure, evolution, security aspects, complex network common characteristics, and more [1] [2] [3].

The problem we contemplate is community membership inference in social networks: inferring group membership of a user based on friendship connections (i.e. topological structures), personal attributes, and affiliations with other groups. Our methodology may be used for the automatic screening of applications to join groups, protecting vulnerable internet populations (e.g. children), group recommendations and

adfa, p. 1, 2011.

more. We believe our work is the first of its kind to study this group inference problem variation, as defined here, map it to a classification problem, examine four classification models, and compare these models.

Here, we propose new group affiliation inference methods using machine learning techniques based on three sets of features: topological, groups, and personal. These methods are examined and validated on several real world datasets. Four prediction models for the OSN affiliation analysis problem were developed: (a) structural/topological features set based model. (b) group affiliation features set based model, (c) personal features set based model, and (d) a model that combines all features in the previous models (i.e., topological, group affiliation, and personal features). We evaluate our proposed models using sample datasets of two social media websites: TheMarker[1] (an online social network site in Hebrew) and Ning [2] (an online social network for Ning creators). Specifically, our novel contributions are:

- Proposal of four novel group affiliation predicting methods using machine learning techniques based on three types of features: topological, membership in other groups, and personal features.
- Introduction of several new topological measures that encompass information from the two graph structures of OSN: the social ties graph, and the group membership bipartite graph.
- We demonstrate that group clusters can be identified easily by calculating the information gain among the groups.

We describe related studies in the next section. Section 3 outlines the problem formal definition and our methods. Then, we describe the evaluation in section 4, and experiment results in section 5. We conclude and discuss future works in section 6.

2 Related Work

Identifying graph communities has been a prevalent topic in recent years. In 2002, Girvan and Newman published an innovative algorithm, which detects such communities by isolating them as separate graph components [1]. Since then, additional methods for community detection have been presented. These can be found in Fortunato’s comprehensive survey on community detection [2].

While recent studies have focused on community structure in social networks [5] [3] [2], these studies concentrated mostly on the communities’ detection on complex graphs, and examined their structure and dynamics. Traud et Al. [6][7], applied network analysis tools to study the role of university organizations and affiliations in structuring the social networks of students by examining a snapshot of the Facebook “friendships” graph at five American universities. They also compared the relative contributions of different personal characteristics to the community structure of universities.

[1] http://cafe.themarker.com
[2] http://creators.ning.com

The classical community detection problem concentrates on detecting communities within a social network based on the friendship connections between friends [1] [2]. Unlike studies where each member is associated with a unique community, our study concentrates on OSN groups where a member may belong to multiple groups or communities. Detection of the overlapping communities' problem has been addressed by Friggeri et al. [8], who introduced cohesion metrics based on network to topological features and triangles counting. Similarly, we focus on OSNs, where a member may join overlapping interest groups. OSNs include a rich set of features and information. Members' information includes social ties, group membership, and personal data. The personal and group information does not require extensive computation; therefore, our model may use this additional information in order to perform group prediction efficiently.

3 Problem Definition and Methods

We represent a social network as a graph $G = (V,E,H)$, where V is a set of n nodes (OSN members), of the same type, E is a set of edges (the friendship links), H is a set of groups that nodes can belong to, and A is a set of node attributes. The graph edge $e_{i,j} \epsilon E$ represents an undirected link between node v_i and node v_j. We describe a group as a hyper-edge $h \epsilon H$ among all the nodes that belong to that group; $h.V$ denotes the set of users who are connected through hyper-edge h. A user profile has a unique ID with which the user forms links and participates in groups. The goal is to predict for a user $v_i \in V$, whether $v_i \in h.V$, while we do not know whether v_i is a member of $h.V$ but all other group members of $h.V$ (i.e. other users who are members of the group h) are given. An alternate goal is to predict for a certain group $h \epsilon H$, which users should be included as members in $h.V$. For both cases the same method may be used, but the evaluation process is different.

We chose to use machine learning methods and develop group inference classifiers, which aim to estimate the probability that a specific user is a member in the target group. Therefore, we presented our problem as a binary classification problem where each user (OSN member), is represented as an instance that is characterized by multiple attributes (also known as features). The target class is a binary attribute indicating whether a user is a member of the group or not. For each group, a specific dataset is generated. The dataset feature attributes contain a users' information about their personal characteristic (age, gender, etc.), social ties structure (aka topological features), and affiliation with other groups. Thus, the features can be divided into three categories:

Personal Characteristics Features (PRS) - The personal information refers to the information in users' profile and usually includes demographic details such as: gender, age, residence, etc. We included each one of these information categories as a feature for our machine learning model. We assume that these personal characteristics may indicate users' groups due to homophily [9]. For example in a fan group of a kids' TV show we would expect that most of the members will be children. Therefore an applicant with an older man profile would be suspicious.

Group Affiliation Features (GRP) - This set of attributes denotes the user affiliation with all the social network groups, except for the target class group. Every instance includes a Boolean vector, where each dimension corresponds to a unique group and includes group membership information. If user v is a member in group $h_i \in H$, meaning $v \in h_i\,.V$, then the corresponding attribute is TRUE. The motivation for using other groups' affiliations is derived from the fact that similar users tend to register to the same set of groups. For example, many users that are registered to the "Data Mining" group are also registered to the "Big Data" group.

Network Topological Features (TPL) - These features are extracted from the topological structure of the graph. For each group, we extracted a set of topological features. These features assist in estimating the chances that a given user is a member in the group. For each member $v \in$V and a group $h \in H$ we calculated a set of 8 topological features as displayed below in Table 1.

Table 1. Topological Features

Attribute Name	Indication
$degree(v)$	Degree, number of immediate friends
$GRP\_F(v,h)$	Number of v's friends in group h
$GRP\_FN(v,h)$	$GRP\_F(v,h)$ normalized by total number of friends
$GRP\_CF(v,h)$	Number of v's friends connected with at least one other of v's friend in group h
$GRP\_CFN(v,h)$	$GRP\_CF(v,h)$ normalized by v's degree
$GRP\_CC(v,h)$	Number of connections v's friends in group h have among themselves
$GRP\_CCN(v,h)$	$GRP\_CC(v,h)$ normalized by number of all possible such connections
$GRP\_F\_L2(v,h)$	Number v's friends of friends in the group h sub-graph (exactly 2 hops from v)

The following definitions are the formal definitions of the topological features:

The neighbourhood $\Gamma(v)$ of v is defined as the set of v's friends, namely, vertices that are adjacent to v. The following is the formal definition of neighbourhood:

$$\Gamma(v) \coloneqq \{u | (u,v) \in E\} \tag{1}$$

The group-neighbourhood $\Gamma(v,h)$, of $v \in V$ and $h \in H$, is the set of v's friends who are also members of group h. The following is the formal definition of group-neighbourhood:

$$\Gamma(v,h) \coloneqq \left\{u \,\middle|\, u \in h \;\&\; (u,v) \in E\right\} \tag{2}$$

Based on the group-neighbourhood definition, we define *ingroup-common-friends* of user v to be the set of v's friends who are members of group $h \in H$ and have at least another friend in this set. We denote this set of nodes as *ICF*:

$$ICF(v,h) \coloneqq \{u \,| u \in \Gamma(v,h) \;\&\; (\exists\, u' \in \Gamma(v,h) \;\&\; (u,u') \in E)\} \tag{3}$$

Based on the group-neighbourhood definition, we define *ingroup-common-connections* of user v to be all the pairs of v's friends who are members of group $h \in H$ and are also friends with each other. We denote this set of nodes as ICC:

$$ICC(v,h) := \{(u_i\ ,u_j)\ |u_i,u_j \in \Gamma(v,h)\ \&\ (u_i\ ,u_j) \in E\} \tag{4}$$

Using the above definitions, we can create the following features for vertex v:

Degree: We defined the vertex v degree as the number of vertices user v has a friendship connection with. We formally define it as:

$$degree(v) := |\Gamma(v)| \tag{5}$$

Ingroup-friends (GRP_F): We define the number of v's friends who are members in group h, as:

$$GRP\_F(v,h) := |\ \Gamma(v,h)| \tag{6}$$

Ingroup-friends-l2 (GRP_F_L2): We define the number of v's friends of friends who are all members in group h (i.e., at two hops distance from v within group h subgraph) as:

$$GRP\_F\_L2(v,h) := \left|\ \{u_j \,|\, (u_i,u_j) \in E\ \&\ u_j\ \in h,\ u_i \in \Gamma(v,h)\ \&\ u_j \notin \Gamma(v,h)\}\right| \tag{7}$$

Grp-common-friends (GRP_CF): We define the number of v's friends in group h who are connected with at least one other v's friend in group h as:

$$GRP\_CF(v,h) := |ICF(v,h)| \tag{8}$$

Grp-friends-connections (GRP_CC): We define the number of connections, which v's friends, who belong to group h, have with other v's friends in group h as:

$$GRP\_CC(v,h) := |ICC(\mathrm{v,h})| \tag{9}$$

Ingroup-friends-ratio (GRP_FN): We normalize ***GRP_F*** by the number of v's friends who are members in group h with with v's degree and define it as:

$$GRP\_FN(v,h) := \frac{|\Gamma(v,h)|}{|\Gamma(v)|} \tag{10}$$

GRP_CFN is the normalized value of ***GRP_CF***, obtained by dividing it by v's degree.

GRP_CCN is the normalized value of ***GRP_CC,*** by dividing it with the number of all possible such connections between v's friends. i.e. $(\frac{degree(v)(degree(v)-1)}{2})$.

4 Evaluation

We performed an evaluation of the proposed methodology on two OSN datasets TheMarker and Ning, and compared the four suggested group prediction models. The social networks datasets were collected using a dedicated Web crawling code. The properties of the datasets are presented in **Table 2**.

Table 2. Properties Of The Datasets.

Property	**TheMarker**	**Ning**
Number of users	87,905	11,011
Number of links	1,644,848	76,263
Number of groups	85	81
Average degree	37.4	7.4
Number of groups per user: Average (Range)	2.4(0-84)	0.4(0-53)
Group size: Average/ (Range)	2,465/ (92-8,360)	59/(1-698)
Number of personal features	28	3

Note that group affiliation inference is a highly imbalanced problem. For most of the groups there are many more non-members than members among all of the OSN users. Formally, there are many more negative links than positive links. Imbalanced datasets pose difficulties for induction algorithms as standard machine learning techniques may be “overwhelmed” by the majority class and in result ignore the minority class. For overcoming this problem, we followed the under-sampling approach in which a balanced training set is generated and used to train a classifier, which is then tested on an imbalanced test set. Two non-overlapping subsets of data, train and test, were selected from each original group data set. For the train set, half of the total of positive and the equal number of negative examples were selected, thus creating a balanced set. The rest of the positive and negative examples were used for test set (imbalanced).

We used the area under the ROC curve (AUC) measure, which is not influenced by the imbalance distribution of the classes [10], for evaluation of different classification models. Additionally, we used the Precision and Recall measures in order to verify the ranking performance of our algorithm. We ran the experiments with WEKA [11], a popular machine learning software suite, and used the Bagging algorithm due to its high performance and relatively low run time [4]. The Bagging algorithm was setup with its default configuration parameters and J48 (Weka's implementation of the well-known C4.5 decision tree algorithm), with the minimal number of instances per leaf set to 10 as the base learning method.

5 Experimental Results

The AUC results of the Bagging algorithm, on TheMarker and Ning networks, using various sets of attributes for the 13 selected groups are presented in Table 3. The groups were selected randomly for TheMarker network, and the largest groups were selected for the Ning. The ALL column presents the results achieved using the combination of all three subsets. For each group, the best result among all the evaluated four models is marked in bold font. The best result among the three attribute subsets (GRP, PRS and TPL) is underlined.

Table 3. AUC Results on the TheMarker (a) and Ning (b) Datasets. Baseline AUC = 0.5

Group	Attributes subset			
Size	ALL	GRP	PRS	TPL
339	**0.839**	0.822	0.694	0.747
353	**0.773**	0.759	0.574	0.733
362	0.859	**0.854**	0.570	0.778
563	0.913	**0.918**	0.622	0.807
949	**0.888**	0.871	0.627	0.695
1366	**0.882**	0.852	0.735	0.728
1671	**0.875**	0.763	0.767	0.736
1751	**0.838**	0.800	0.675	0.771
1930	**0.867**	0.838	0.710	0.753
2210	**0.823**	0.768	0.585	0.727
2248	**0.770**	0.696	0.656	0.712
3788	**0.840**	0.764	0.725	0.772
7600	**0.840**	0.767	0.656	0.689

(a)

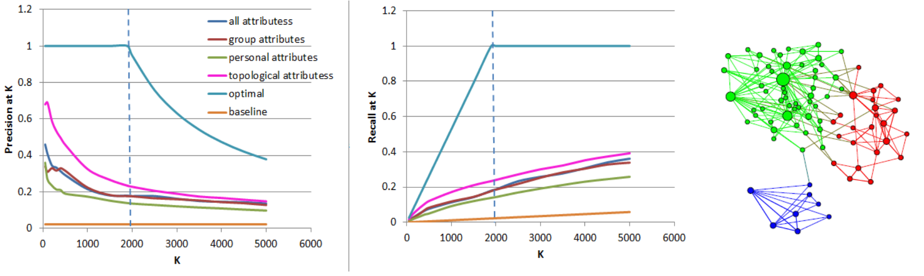
Group	Attributes subset			
Size	ALL	GRP	PRS	TPL
102	**0.909**	0.813	0.596	0.891
103	0.854	0.854	**0.895**	0.818
122	**0.831**	0.665	0.611	0.778
127	**0.901**	0.550	0.842	0.856
141	**0.879**	0.840	0.537	0.797
150	**0.866**	0.805	0.475	0.849
152	**0.851**	0.775	0.464	0.793
204	**0.934**	0.527	0.919	0.899
239	**0.872**	0.759	0.585	0.849
239	**0.910**	0.804	0.528	0.881
378	**0.883**	0.673	0.689	0.873
582	0.788	0.695	0.502	**0.789**
698	0.876	0.568	**0.886**	0.875

(b)

It can be seen that the best AUC is achieved by using all attributes for most of the groups in both networks. Among the evaluated attribute subsets, the group's affiliation provides the best prediction results for most of the groups in TheMarker network, and topological attributes perform the best for most groups of the Ning network. This suggests that the optimal model is OSN related, and depends on the social network properties. As shown in **Table 2**, TheMarker groups are larger, and the number of groups per user is higher compared to Ning (2.4 vs. 0.4 accordingly), which may explain the strength of the GRP attribute set in the case of TheMarker.

Interestingly, the prediction accuracy achieved using the ALL attribute set versus the GRP attribute set is very similar in many cases. This suggests that the GRP model (including group's affiliation attributes only), may be used in certain cases, such as in large social networks with many members and various interest groups. In these cases it allows for a better computational performance (as these attributes are easy and quick to compute), with the lowest (if any), loss in prediction accuracy. It can also be noted that in the TheMarker network, personal attributes provide AUC values only slightly above the baseline AUC (equal to 0.5). Contrarily, in the Ning network, personal attributes provide relatively high AUC values, especially in the groups which are country or language related. This strengthens our conclusion that the type of most predictive attributes depends on the investigated network itself.

The Precision and Recall results at various *K* sizes for one of the groups from the TheMarker network are presented in Fig. 1(a), and compared to the optimal and baseline values of these measures. The vertical dashed line specifies the *K* equal to the actual number of positive examples in the test set (i.e. number of group members). While the absolute Precision and Recall values improvement is desired, they are much better than baseline (computed as percentage of group members out of the total amount of users in the network). These results suggest that the developed models can already be used for reducing the load and cost of group inference related tasks. These set of tasks include identifying the most suitable individuals to the group or the most suitable groups for an individual. The results on other groups in both networks follow a very similar pattern, and are thus discarded from this publication.



(a) (b)

Fig. 1. (a) Precision, Recall for different K's on the "The design works" group, TheMarker dataset; (b) TheMarker groups' clusters using information gain evaluation.

Additionally, to obtain an indication of the usefulness of the various individual features, we analysed their importance using Weka's information gain attribute selection algorithm. Generally, the results of this analysis correspond with AUC and Precision/Recall results. We also calculated the information gain between all possible group pairs in the TheMarker network and represented this information using a graph (see Fig. 1(b)). The groups in this graph are represented by nodes; directed edges (u, v) represent information gain value for group u when attempting to predict membership in target group v. For each group we chose the three edges with the highest information gain values to be included in the graph. The size of the nodes in the graph is proportional to the information gain value the group has for inferring other groups. We applied the Louvain method for community detection [12], which divided the graph into three communities. The clustered nature on this graph indicates that some groups are more tightly connected than others. Further evaluation of this clustering in OSN group graphs is one our future task plans.

6 Conclusions and Future Work

This study presents the group inference problem in OSN and proposes machine learning based methods to address it. The classification models are based on personal, topological and group affiliation features. Generally, we can see that a relatively high predictive accuracy (an AUC of about 0.8 on average, while baseline is 0.5), can be achieved using all the attributes along with the simple and quick bagging classification algorithm. The model yielding the highest accuracy varies across OSNs, and may depend on OSN properties. The precision and recall measurements also demonstrated significantly higher results compared to baseline values. Additionally, our study demonstrates that information gain values between different groups can be used for analyzing their relations and detecting group clusters. These clusters can then be used as a target class instead of individual groups.

We believe that predictive performance can be further improved using more sophisticated classification methods and by devising additional topological features. An evaluation of our approach with such methods on additional OSN datasets is one of our nearest future tasks.

References

1. Community structure in social and biological networks. M. Girvan, M.E.J. Newman. s.l.: Proc. Natl. Acad. Sci., 2002.
2. Community detection in graphs. Fortunato, S., 3-5, s.l: Physics Reports, 2010, Vol. 486.
3. Complex networks: Structure and dynamics. S. Boccalettia, V. Latora, Y. Moreno, M. Chavezf, D.-U. Hwang. 4-5, s.l. : Physics Reports, 2006, Vol. 424, pp. 175-308.
4. Link Prediction in Social Networks Using Computationally Efficient Topological Features. Fire, M., et al., et al**.** s.l. : IEEE SOCIALCOM, 2011.
5. Community structure in social and biological networks . M. Girvan, M. E. J. Newman. s.l. Proceedings of the National Academy of Sciences, 2002, Vol. 99.
6. A.L. Traud, E.D. Kelsic, P.J. Mucha, M.A. Porter. Community structure in online collegiate social networks. s.l. : eprint arXiv:0809.0690., 2008.
7. Comparing community structure to characteristics in online collegiate social networks. A.L. Traud, P.J. Mucha, M.A. Porter, E.D. Kelsic. s.l. : SIAM Review, 2011.
8. Triangles to Capture Social Cohesion. Friggeri A., Chelius G., Fleury E. passat'11, IEEE.
9. McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: Homophily in social networks. Annu Rev Sociol. 2001, Vols. 27:415–444.
10. Menon, A. and Elkan. Link prediction via matrix factorization. s.l.: Springer, 2011.
11. The weka data mining software: an update. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. s.l. : SIGKDD Explor. Newsl, 2009. 11, 10–18.
12. Fast unfolding of communities in large networks. V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre s.l.: J. Stat. Mech.: Theory and Experiment, 2008, Vol. 2008.