Exploiting Community Detection to Recommend Privacy Policies in Decentralized Online Social Networks

Andrea De Salve[1](✉), Barbara Guidi[2], and Andrea Michienzi[2]

[1] Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche,
Via G. Moruzzi, 1, Pisa, Italy
andrea.desalve@iit.cnr.it
[2] Department of Computer Science, University of Pisa,
Largo Bruno Pontecorvo, Pisa, Italy
{guidi,andrea.michienzi}@di.unipi.it

Abstract. The usage of Online Social Networks (OSNs) has become a daily activity for billions of people that share their contents and personal information with the other users. Regardless of the platform exploited to provide the OSNs' services, these contents' sharing could expose the OSNs' users to a number of privacy risks if proper privacy-preserving mechanisms are not provided. Indeed, users must be able to define its own privacy policies that are exploited by the OSN to regulate access to the shared contents. To reduce such users' privacy risks, we propose a Privacy Policies Recommended System (PPRS) that assists the users in defining their own privacy policies. Besides suggesting the most appropriate privacy policies to end users, the proposed system is able to exploits a certain set of properties (or attributes) of the users to define permissions on the shared contents. The evaluation results based on real OSN dataset show that our approach classifies users with a higher accuracy by recommending specific privacy policies for different communities of the users' friends.

Keywords: Decentralized online social networks
Recommendation system · Privacy · Privacy policies · Security

1 Introduction

The usage of Online Social Networks (OSNs) has become a daily activity for billions of people who share several private information on current OSNs, exposing them to privacy leaking. Indeed, current OSNs are free to use, and they make money by selling private data to advertisers. The last scandal involves Facebook and private data collected by Cambridge Analytica[1].

[1] https://www.theguardian.com/technology/2018/apr/04/facebook-cambridge-analytica-user-data-latest-more-than-thought.

G. Mencagli et al. (Eds.): Euro-Par 2018 Workshops, LNCS 11339, pp. 573–584, 2019.
https://doi.org/10.1007/978-3-030-10549-5_45

During the last decade, several solutions have been proposed to overcome the problem of current OSNs. One of the main promising is the decentralization of social services. Decentralized Online Social Networks (DOSNs) guarantee a higher level of privacy because data is distributed and users can have more control over their personal data. The access control is one of the most used technique to prevent privacy leaking in DOSNs. Indeed, the majority of existing DOSNs provide a set of privacy policies based on the knowledge derived from the relationships, content or profile information, etc...

In this paper we propose a new approach to define a Privacy Policy Recommendation System (PPRS) based on community detection in DOSNs. The motivation of this work is that users with common attributes are more likely to be friends and often form dense communities [15]. For this reason, our methodology exploits a set of attributes which describe properties of the users (such as location and school information). By considering the homophily between users [6], we compute communities based on the social graph and we exploit a decision tree learning algorithm to suggest privacy policies for such communities.

The evaluation conducted on real dataset shows that the proposed approach shall be capable of providing higher level of accuracy by correctly classifying the 80% of the users' friends in the proper community while exploiting different attributes of the users.

The paper is organized as follows. In Sect. 2 we propose an overview of the state of the art of privacy in OSNs. In Sect. 3 we introduce our approach. In Sect. 4 we describe the dataset used to evaluate our approach. Section 5 shows the evaluation of the approach, and finally, in Sect. 6 we propose our conclusion and future improvements.

2 Related Work

Nowadays, the most popular OSNs are based on centralized architectures where private data are stored in centralized storages which are under the control of the administrations. The centralized management of data exposes to several privacy risks. Indeed, malicious users, the service provider, and third-party applications can access users' private data. As explained in [11], the main attacks in OSNs are:

- Privacy breaches: attacks to strike the users' privacy. Three primary parties interact are involved: the service provider, the users, and third-party applications.
- Viral Marketing: spamming and phishing attacks which exploit information extracted from user profiles.
- Network Structural Attacks: the most famous one is the Sybil attack, in which an individual entity masquerades as multiple simultaneous identities.
- Malware Attacks: usage of OSNs to spread malicious software.

One of the main solution to the privacy issue in OSNs has been the introduction of Decentralized Online Social Networks (DOSNs) in order to overcome the centralization of data [17]. Several works have been proposed during the last ten years which exploits P2P solutions to implement the underlying architecture [1,4,12]. An important characteristic of DOSNs is that they provide the capability to define privacy preferences on the contents produced and exchanged to define which users are allowed to see such contents [8]. Typically, privacy policies are simple statements which specify the main attributes a user can have to access contents (such as friendship type, interests, work, school,...). In detail, a DOSN proposes a privacy model [8] defined as the capability of DOSNs to provide privacy policies to specify the set of members who can access contents, and a privacy policy management which guarantees that these policies are enforced on each content by using proper security mechanisms.

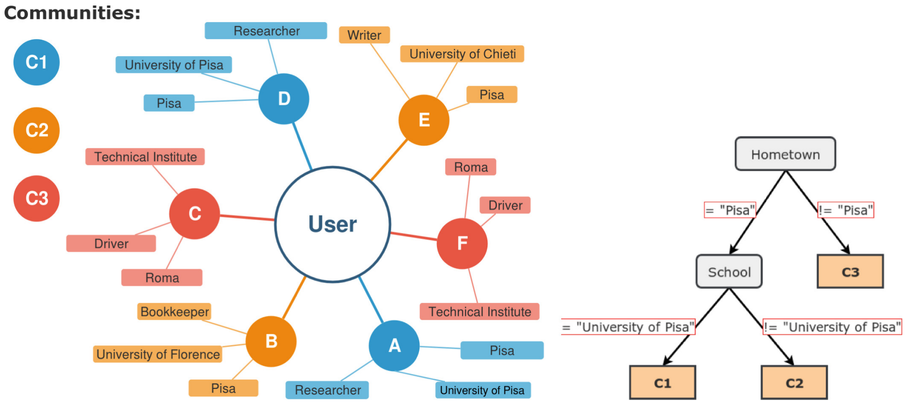
3 Our Approach

Our scenario consists of a DOSNs in which each user has information about its ego network, which includes the principal user (*ego*) together with the actors they are connected to (*alters*) and all the links among these alters [2]. Each user of the DOSNs can define privacy policies to manage the access to its content [9], and it is characterized by a set of attributes, such as information about personal profile (date of birthday, hometown, school, etc.) or information about its preferences (music, movies). Each ego node knows the friends' attributes and it can exploit them to express their privacy preferences on these friends.

3.1 Privacy Policy Recommendation

The first step of the PPRS is the application of a community detection algorithm to each ego network. The algorithm we used is DEMON [3] because it can be adapted to an ego centric approach and it is computationally not expensive, as explained in [13]. We extract the communities in the so called *ego minus ego*: a network made of the ego network of a user where we remove the ego itself and all edges connected to it. In this work, the community detection algorithm has been configured to return a set of non-overlapped communities where each alter can belong to only one community in the same ego network. The case of overlapped communities is left as future work because it demands more investigation and consideration than the case of disjointed communities. Each community has an identifier, and this identifier is used as a user's attribute, and it is inserted in the attributes list of each ego to identify how its alters are clustered. At the end of this phase, each ego node u has an array of attributes for each alter f which contains: the values of the f's attributes $p_1(f)$, $p_2(f)$, $p_3(f)$, and the communities $C = \{C_{id} | f \in C_{id}\}$ of u.

The second step consists in the definition of decision trees which let us to classify users. An example of decision trees is shown in Fig. 1. In Fig. 1(a) we propose an example of six users with both their school attribute and the community label associated. We exploit the *School* and *Hometown* attributes to build the decision tree shown in Fig. 1(b). Finally, a privacy policy of a community



(a) Ego network

(b) Decision Tree

Fig. 1. Example of the PPRS executed on the ego network (Fig. 1(a)) of a user with 6 friends having attributes and community label associated. The results of the PPRS is the decision tree shown in Fig. 1(b)

can be provided by considering the attributes on each path from the leaves of the community to the root of the decision tree. Thereafter, the conditions on the attributes resulting from the model can be translated by the PPRS into a privacy policy format. As for instance, the most part of privacy policy languages leverage XML for defining constraints that must be satisfied by attributes of the users [9]. Finally, the PPRS is able to suggest the most suitable privacy policies to the user.

4 The Dataset

Information about attributes of Facebook users have been gathered by a Facebook application (originally described in [5]), called SocialCircles![2], which exploits the Facebook API to retrieve a set of social information about the registered users which contain information about friends of registered users and the friendship relations existing between them, profile information, and information about interactions between registered users and their friends, such as posts, comments, likes, tags and photo. Due to technical reasons (time needed to fetch all data and storage capacity), we restrict the interaction information retrieved up to 6 months prior to user application registration. The dataset we use contains 205 complete Ego Networks, for a total of 95.716 users (ego and their friends). In the following of this section we will present in more detail the collected data to understand better its nature.

[2] https://www.facebook.com/SocialCircles-244719909045196/.

4.1 Features of the Dataset

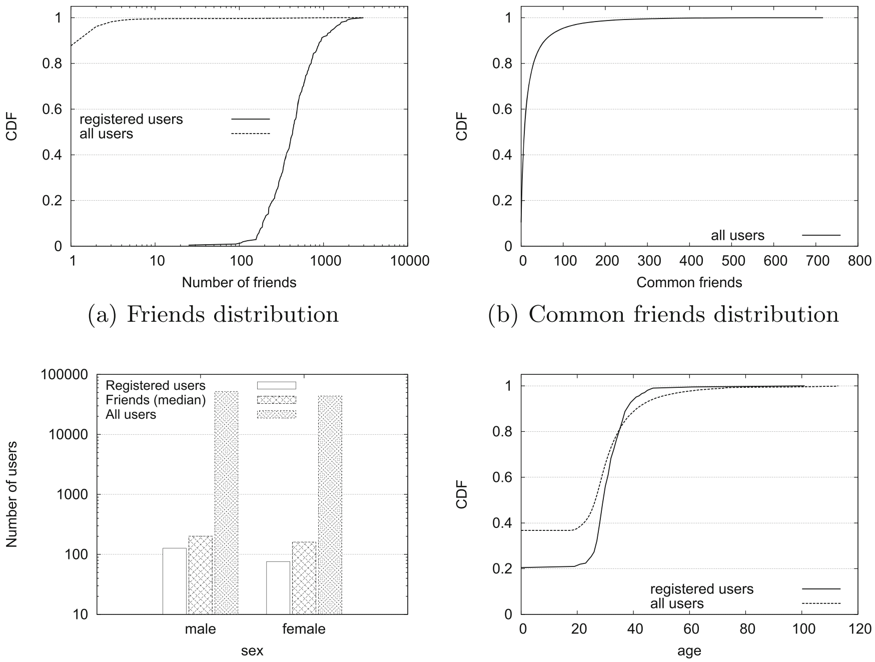
We used the information collected from the OSNs to extract different features of users that can be exploited as attributes by the PPRS. Such features (or attributes) are properties of the users that can be either obtained from the users' profile (such as, age, sex, number of common friends, etc.) or derived from the OSNs information (such as tie strength and trustiness). An overview of the statistical properties that quantitatively describes the features obtained from our collection of information are listed below.

Friends and Common friends. We first analyze the degree distribution of our dataset. The graph of Fig. 2(a) shows the Cumulative Distribution Function (CDF) of the number of friends of the registered users as well as the degree distribution of the entire sample (all users). The graph clearly indicates that 50% of the registered users have at most 432 friends in their ego networks while the most part of them (about 90%) have at most 1000 friends. In addition, the CDF of the whole set of the users in our sample points out the presence of a higher number of users (about 85%) having only one friendship relation. According to [5], this is caused by Facebook privacy setting that could not allow our application to collect enough information related to the friends of the registered users.

We focus on the number of common friends between users by measuring the number of mutual friends that each alter shares with the ego. As shown by Fig. 2(b), the CDF of the number of common friends has exponential shape, resulting in about 10% of the friendship relations between the egos and their alters with any friends in common while about 50% of them have less than 8 mutual friends.

Age and Gender. We investigate the gender distribution by measuring the fraction of male and female users who have registered to our application. Figure 2(c) shows the gender distribution for both the set of registered users and of all users in the dataset, as well as the median number of male and female users' friends of the registered users. We can observe that registered users registered consist of about 127 (63%) men and 76 (37%) women while the whole set of users (all users) is more balanced and it consists of about 54% men and 45% women. The typical users established, on average, 202 (55%) friendships with men and 161 (45%) friend relationships with women.

We investigated the distribution of ages in our dataset by measuring the difference between the birthday date and the current date. Figure 2(d) shows the distribution of the ages for all the users in the dataset and for the set of registered users. We observed that about 20% of the registered users and 40% of all the users did not specify their birthday date or they provided an age of 0. The median age of the registered users (30 years old) is higher than those of all users (27 years old).



(a) Friends distribution

(b) Common friends distribution

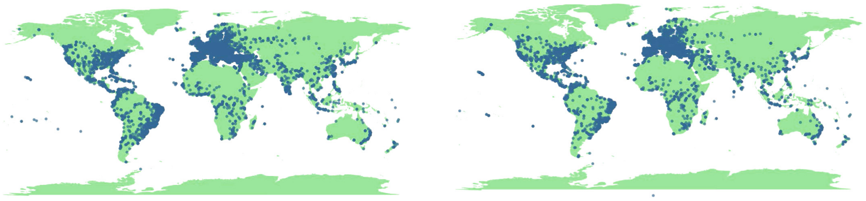
(c) Gender distribution

(d) Age distribution

Fig. 2. General description of the statistical characteristics of the Facebook dataset.

Hometown and Current Location. Geographical locations of the users are also collected and they can be exploited to measure the distance or proximity between users. The map in Fig. 3 indicates the geographical location that users have specified in their Facebook profiles. In particular, we consider the hometown (Fig. 3(a)) and the current location (Fig. 3(b)) of users because they could affect their interaction patterns. The most part of the collected users have hometown location and current location placed in Europe, where the application was initially disseminated. However, the maps indicate that our application had spread also in America and a large portion of users came from North America.

Dunbar's Circles. An interesting analysis with respect of the OSNs is the characterization of the *ego network* of the user (introduced in Sect. 3). Indeed, many studies showed that the number of active relationships that a user can establish in his ego network is limited (about 150), the so called *Dunbar number* [10]. Figure 4(a) summarizes the average number of friends in each circle as well as their 95% confidence interval in square brackets while the frequency of contacts is used to estimate the tie strength between users. In particular, the closest circle, called *support clique* (circle 0), consist of 4 [±0.17] and the typical frequency of contact between individuals is estimated to be at least once weekly. The second circle is the *sympathy group* (circle 1), being the set of individuals contacted at



(a) Hometown location

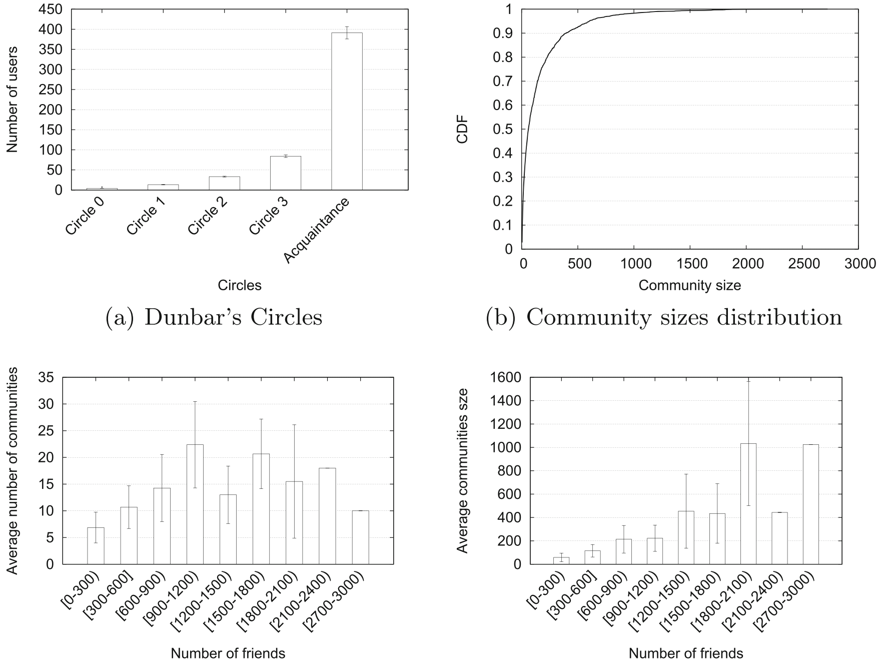
(b) Current location

Fig. 3. Geographic map representing the hometown and current location of the users.

least monthly and it consists of 13 [±0.48] users. The next circle is the *affinity group* (circle 2) which consists of 33 [±1.21] users, and finally the outermost circle is the *active network* (circle 3), which consists of 84 [±3.2] users. The total number of users that belong to the Dunbar's circles is equals to 134 [±5] while any other contact is considered to be a simple acquaintance. In particular, acquaintances are friends of an ego e that are occasionally contacted by e and the total number of acquaintances of an ego is equals to 391 [±15]. The procedure performed to obtain the Dunbar's circles are described in more detail in [5].

Community Structure. A significant property of OSNs is the community structure, i.e., densely connected groups of users which are sparely connected to other users. An important step for the PPRS is to identify, for each ego network, to which community a specific user belongs to. For this reason, we consider the ego networks of registered users and we utilize the community detection algorithm exploited in [6,13] to compute both the number and the structure of the communities. The total number of communities discovered in this step is equal to 2237 and in Fig. 4(b) we show the CDF of the community size. The average size of communities is about 60 contacts while about 80% of communities has less than 250 nodes. Figure 4(c) shows the average number of communities discovered in ego networks having different number of alters. The plot indicates that the number of communities defined by users is weakly correlated with the number of users' friends: as long as the number of users' friends increases the number of communities of such users remains bounded to 30 while the average and median number of communities for each ego network is about 14.

We focused on the size of the communities discovered in each ego network and we showed in Fig. 4(d) the average size of the communities defined by users having different number of friends. We observe that the size of the communities of users strongly depends on the number of friendships established by such users, resulting in left skewed distribution with average and median community size equal to 444 and 176, respectively.



(a) Dunbar's Circles

(b) Community sizes distribution

(c) Ego network: number of communities

(d) Ego network: size of the communities

Fig. 4. Analysis of the Dunbar-based and static communities of the ego network

5 Experimental Methodology and Results

We focused on the evaluation of the classification task, i.e., on the capability of the PPRS to correctly identify the communities of users in each ego network by exploiting the users' attributes derived from the users' profiles.

5.1 Training and Classification Algorithm

Since our dataset is a collection of registered users, we considered the ego network of each registered user individually for the classification task. Given an ego network of the registered user u, we perform a transformation phase on the original dataset in order to construct an input dataset which contains a record R for each friend in the ego network. The record R consists of the set of attributes that a registered user u wants to exploit to define privacy policies on his ego network. In particular, the attributes are derived from the dataset's features (described in Sect. 4.1) and they include, for each friend f of u: *a)* the sex of f, *b)* the age of f, *c)* the number of common friends between u and f, *d)* the distance (in meters) between the hometown location of u and f, *e)* the distance (in meters) between the current location of u and f, and *f)* the Dunbar's circle to which f belongs, in the ego network of u. In addition to these attributes, a target attribute is created for each friend f of the registered user u to indicate the community to which the user f belongs, in the ego network of u, and it is used as (discrete, unordered) class label by the classification task. For this reason, we create a class $label(c)$ for each community c in the ego network of the registered user u and we set the target attribute of each friend f in the community c to the class $label(c)$. The goal of the classification task is to create a model for each registered user u that can be used to classify each friend f of u in the proper community c, based on a number of attributes related to f:

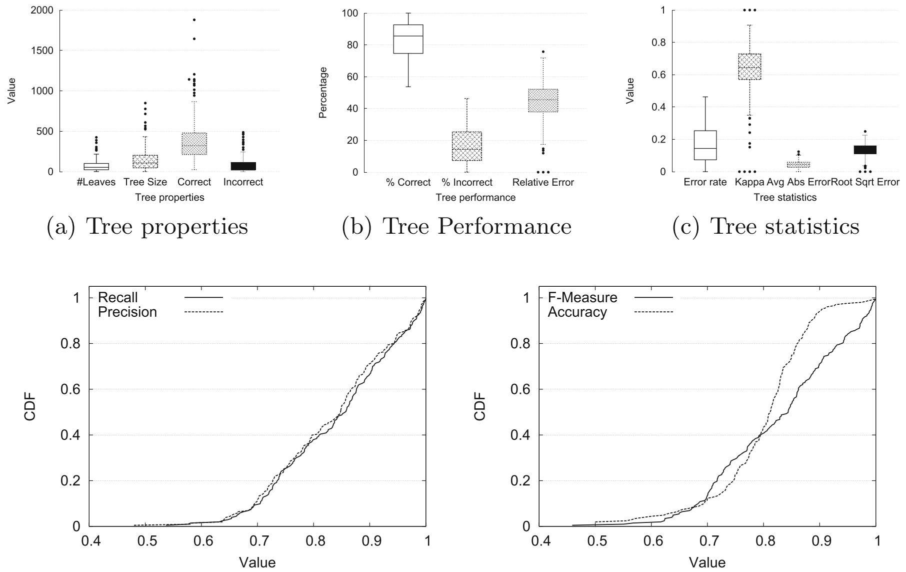
$$[gender, year, comm, distH, distC, dunbarCircle] \Rightarrow label(c) \tag{1}$$

The model of each registered user u is created by exploiting the C4.5 Decision Tree Learner [16], that is one of the most used methods for classification decision tree. The algorithm builds a hierarchical decision tree which is used for classifying the class label of a user. The attributes of f are used by the tree to routes f towards a leaf node which contains a class label of the community in the ego network of u. The conditions on internal nodes of the tree are generated by splitting the domain of attributes in two partitions (i.e., using a binary split) and the Gini index is used as a quality measure to calculate such splitting point.

5.2 Results Validation

In order to evaluate the models resulting from the supervised learning algorithm we collected several performance measures for each decision tree of a registered user u. The box plot of Fig. 5(a) shows the minimum, lower quartile, median, upper quartile, and maximum, of different tree properties. In particular, we consider the number of leaves in the tree (*#Leaves*), the size of the tree model (*Tree size*), and both the number of friends correctly and incorrectly classified in their community (*Correct* and *Incorrect*, respectively). The decision trees of the registered users have an average number of leaves equals to 76 [±0.31], which account for an average size of 151 [±0.63]. The size of the tree clearly depends on the number of leaves but, in general, the 80% of the trees have less than 120 leaves. The attributes selected from the users allow to correctly classify about 396 [±1.3] friends in the proper community while the average number of incorrectly classified friends is equals to 80 [±0.4]. Indeed, the amount of friends correctly classified by the model is high: 80% of the registered users have correctly classified at most 535 friends in their ego network while the number of friends incorrectly classified by the model is less than 130.

As shown by Fig. 5(b), the average fraction of friends correctly identified in each ego network amounts to 85.6% with average relative absolute error equals to 44 [±0.05] while only 14.3% of the friends cannot be classified by exploiting the selected attributes. In particular, the most part of the registered users (about 80%) have classified an average fraction of friends which ranges between 80% and 95%. Figure 5(c) shows the average error rate is quite low (about 0.14) while the median Kappa index, which measures the degree of accuracy and reliability of the classification task, is equal to 0.64. As explained also in [7],



(a) Tree properties (b) Tree Performance (c) Tree statistics

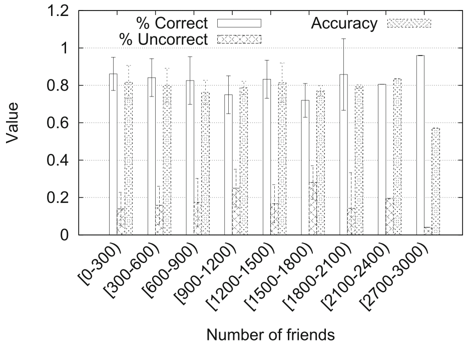
(d) Precision (e) Accuracy

Fig. 5. Analysis of the performance of the classification task

depending on the value of kappa, the index can be interpreted as [14]: *(i)* no agreement if $k \in [0, 0.2]$, *(ii)* fair agreement if $k \in [0.21, 0.4]$, *(iii)* moderate agreement if $k \in [0.41, 0.6]$, *(iv)* substantial agreement if $k \in [0.61, 0.8]$, and (v) perfect agreement if $k \in [0.81, 1]$. Instead, the error of the predicted probability distribution (Avg Abs Error) and the root mean square error (Root Sqrt Error) are quite low and their media value does not exceed 0.13.

We investigated in more detail the ability of the predictors to correctly derive the community to which users belong to. Figure 5(d) shows the CDF of the recall and precision on the resulting models. As we expected, the recall of the classifier in predicting the community of the users is very high (about 0.84) and the precision of the most part of the users (80%) is higher that than 0.75. In addition, the predictors show to have similar precision, indicating that the most part of users are correctly classified by the models by exploiting the values of the users' attributes.

The last step in our analysis consists in evaluating the accuracy of the results, indicating the ability of the model to classify friends that belong to different communities while the *F-measure* summarizes the performance of each predictor. We showed in Fig. 5(e) the CDF of both the accuracy and the F-measure of the models. The average accuracy achieved by the model is 0.83 [±0.0004] and it is to the median accuracy (about 0.80), suggesting that about 50% of the models expose an accuracy higher than 0.80. Figure 6(a) shows the average performance achieved by models which are build on users having different number of friends, while Fig. 6(b) shows the attributes' importance which is measured as



(a) Accuracy of the prediction

Attributes	Rank
Distance current location	0,205
Age	0,200
Sex	0,191
Common friends	0,165
Distance hometown location	0,146
Dunbar Circles	0,093

(b) Attribute importance

Fig. 6. Analysis of the accuracy of the model for ego network having different size.

the information gain provided by an attribute in order to identify the class label of a community. The graph clearly indicates that the most part of the models have either accuracy or fraction of correctly classified friends higher than 0.8 while some users (those in the ranges [900–1200] and [1500,1800]) expose a lower accuracy because the attributes specified by friends cannot be used to correctly classify them in the proper community.

6 Conclusion

In this paper, we focused on the privacy issues related to the contents sharing in Distributed Online Social Networks (DOSNs) by proposing a Privacy Policy Recommendation System (PPRS) that suggests to users the most appropriate privacy policy that expressing the groups of friends who can read the content they share. In particular, the privacy policy specify the authorized users in terms of a set of features encoded by attributes. Those attributes model different properties of users (such as gender, age, common relationship, preferences, location, etc.) and they are exploited by the PPRS to predict the privacy preference a user would give to their friends. We investigated the capability of the PPRS to select the attributes of the privacy policies by considering a real dataset obtained from Facebook, and the communities of friends arising from the friendships defined by each user. The experimental results performed on six attributes reveal that the PPRS is able to suggest privacy policies which correctly grant access to about 80% of the members of the communities, achieving higher level of accuracy.

We plan to enhance the proposed system by introducing proper mechanisms that adjust the conditions generated on each attribute by the PPRS in order to refine the set of authorized members. A further extension is also the investigation of the effect that different configuration parameters have on the performance of the classification algorithm.

References

1. Bielenberg, A., Helm, L., Gentilucci, A., Stefanescu, D., Zhang, H.: The growth of diaspora - a decentralized online social network in the wild. In: INFOCOM Workshops, pp. 13–18. IEEE (2012)
2. Conti, M., De Salve, A., Guidi, B., Pitto, F., Ricci, L.: Trusted dynamic storage for dunbar-based P2P online social networks. In: Meersman, R., et al. (eds.) OTM 2014. LNCS, vol. 8841, pp. 400–417. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-45563-0_23
3. Coscia, M., Rossetti, G., Giannotti, F., Pedreschi, D.: Demon: a local-first discovery method for overlapping communities. In: Proceedings of the 18th ACM KDD, pp. 615–623 (2012)
4. Cutillo, L.A., Molva, R., Strufe, T.: Safebook: a privacy-preserving online social network leveraging on real-life trust. IEEE Commun. Mag. **47**(12), 94–101 (2009)
5. De Salve, A., Dondio, M., Guidi, B., Ricci, L.: The impact of user's availability on on-line ego networks: a facebook analysis. Comput. Commun. **73**, 211–218 (2016)
6. De Salve, A., Guidi, B., Ricci, L.: Evaluation of structural and temporal properties of ego networks for data availability in DOSNs. Mob. Netw. Appl. **23**(1), 155–166 (2018)
7. De Salve, A., Mori, P., Ricci, L.: Evaluating the impact of friends in predicting user's availability in online social networks. In: Guidotti, R., Monreale, A., Pedreschi, D., Abiteboul, S. (eds.) PAP 2017. LNCS, vol. 10708, pp. 51–63. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71970-2_6
8. De Salve, A., Mori, P., Ricci, L.: A survey on privacy in decentralized online social networks. Comput. Sci. Rev. **27**, 154–176 (2018)
9. De Salve, A., Mori, P., Ricci, L., Al-Aaridhi, R., Graffi, K.: Privacy-preserving data allocation in decentralized online social networks. In: Jelasity, M., Kalyvianaki, E. (eds.) DAIS 2016. LNCS, vol. 9687, pp. 47–60. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39577-7_4
10. Dunbar, R.: The social brain hypothesis. Brain **9**(10), 178–190 (1998)
11. Gao, H., Hu, J., Huang, T., Wang, J., Chen, Y.: Security issues in online social networks. IEEE Internet Comput. **15**(4), 56–63 (2011)
12. Guidi, B., Amft, T., De Salve, A., Graffi, K., Ricci, L.: DiDuSoNet: A P2P architecture for distributed dunbar-based social networks. Peer-to-Peer Netw. Appl. **9**(6), 1177–1194 (2016)
13. Guidi, B., Michienzi, A., Rossetti, G.: Dynamic community analysis in decentralized online social networks. In: Heras, D.B., Bougé, L. (eds.) Euro-Par 2017. LNCS, vol. 10659, pp. 517–528. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75178-8_42
14. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)
15. Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: inferring user profiles in online social networks. In: Proceedings of the third ACM International Conference on Web search and Data mining, pp. 251–260 (2010)
16. Quinlan, J.R.: C4. 5: Programs for Machine Learning. Elsevier, Amsterdam (2014)
17. Yeung, C.-m.A., Liccardi, I., Lu, K., Seneviratne, O., Berners-Lee, T.: Decentralization: the future of online social networking. In: W3C Workshop on the Future of Social Networking Position Papers, vol. 2, pp. 2–7 (2009)