Noname manuscript No.
(will be inserted by the editor)

Discovering Homophily in Online Social Networks

Andrea De Salve · Barbara Guidi · Laura Ricci · Paolo Mori

Received: date / Accepted: date

Abstract During the last ten years, Online Social Networks (OSNs) have increased their popularity by becoming part of the real life of users. Despite their tremendous widespread, OSNs have introduced several privacy issues as a consequence of the nature of the information involved in these services. Indeed, the huge amount of private information produced by users of current OSNs expose the users to a number of risks. The analysis of the users' similarity in OSNs is attracting the attention of researchers because of its implications on privacy and social marketing. In particular, the homophily between users could be used to reveal important characteristics that users would like to keep hidden, hence violating the privacy of OSNs' users. Homophily has been well studied in existing sociology literature, however, it is not easily extensible in OSNs due to the lack of real datasets. In this paper, we propose an analysis of similarity of social profiles in terms of movie preferences. Results reveal the presence of homophily between users and its dependence from the tie strength. Moreover, we show that it is possible to profile a user (in our case by considering the age attribute) by exploiting movie preferences.

Keywords Online Social Networks, Homophily, Privacy Preserving, Viral Marketing, User Profiling

1 Introduction

Online Social Networks (OSNs) [6] are the most popular online applications that have changed the way of how people interact between them. During the years, OSNs have invaded the human life becoming the first way to share personal data. Current OSNs services are implemented by mainly exploiting two different architectural style. The most popular OSNs (such as Facebook, Twitter, or Google+) are based on a centralized architecture where the service provider stores and controls all the users' information. A huge amount of privacy problems have been arisen by centralized OSNs because data that users want to share with their friends are exploited by the service provider for different purposes, such as viral marketing, target advertising, instant advertising, recommendation, or censorship. To overcome these privacy issues, OSNs based on decentralized architecture have been proposed. Decentralized Online Social Networks (DOSNs) [10] are made up of a set of peers, such as a network of trusted servers, a P2P system or an opportunistic network, which collaborate with each other in order to provide social services.

In all the previous cases, measuring the degree of homophily between users (or similarity) plays a fundamental role because it reveals group of users exposing similar characteristics. Homophily [23] is the principle that similar users interact and share contents more often than other users. Indeed, users tend to bond more with users who have similar interests and several studies [14, 3, 13] have shown that homophily between users can impacts the predictability of the users' profiles and

Andrea De Salve - E-mail: desalve@di.unipi.it
Department of Computer Science, University of Pisa
Institute of Informatics and Telematics, National Research Council

Barbara Guidi - E-mail: guidi@di.unipi.it
Department of Computer Science, University of Pisa

Laura Ricci - E-mail: laura.ricci@unipi.it
Department of Computer Science, University of Pisa
E-mail: smith@smith.edu

Paolo Mori - E-mail: paolo.mori@iit.cnr.it
Institute of Informatics and Telematics - National Research Council

it can be successfully used for link prediction [30] and to perform product/item recommendation [7]. For instance, authors of [26] show that multiple attributes of the users' profile can be inferred from users by exploiting homophily with other users, when at least 20% of the users reveal their attribute information. In addition, the homophily is an indicator of how fast the information spreads among users and it can be exploited either to limit or to speed up the dissemination of information to the users having common interests.

In our previous study [13], we provided a preliminary analysis of the homophily in Facebook which considers the movie preferences of users. The proposed analysis uncovers interesting aspects related to the impact of the homophily on the movie preferences exposed by users and would add great value to the research on this field. Inspired by the promising results of our previous work, in this paper we investigate the homophily more in detail by focusing on a user-centric point of view. We provide a better understanding of the homophily between users and their friends by considering movie preferences and the profile information of the users. In particular, we introduce several further contributions that are highlighted in the following:

- First of all, we discuss the procedures adopted to increase the quality of data and to obtain more accurate information related to users' movie preferences.
- We exploit several distance measures in order to better quantify either the similarity between the movie preferences specified by two users or the similarity between two Facebook movie pages.
- A third contribution of this paper is the study of the macroscopic characteristics of similar groups of users in OSNs. In particular, we study the homophily of such groups by observing the implication on the age of the users' groups.
- Another novel contribution is the evaluation of the homophily from a local point of view. In particular, we investigate if friends in an ego network with higher tie strengths have more impact on the movie preferences of the ego.
- To our best knowledge, we are the first to apply the Dunbar's model for evaluating the homophily between a user and groups of distinct friends.

The rest of the paper is organized as follows. Section 2 describes the current proposals in both the DOSNs fields and the homophily in OSNs. In Section 3 we explain the impact of homophily on privacy in OSNs' services. In Section 4, we introduce the structure of our Facebook ego networks. In Section 5 we introduce the Facebook dataset used for our experiments and the preparation of such dataset. In Section 6 we describe the procedure we have exploited to discover homophily. We show in Section 7 the evaluation of the homophily and the validation of the results. In Section 8, we study the homophily by considering the age of users. Finally, we terminate in Section 9 by drawing conclusions and by discussing some future works.

2 Related work

Online Social Networks and Social Media in general have been studied in detail as concern the problem of privacy. Security and privacy issues concerning Online Social Media, both decentralized and centralized, can be summarized into three general categories:

- Privacy breaches [27]: this category contains all the attacks that try to violate the privacy exploiting the users' data available in the system;
- Impersonation attacks [22]: identity violations and creation of fake profiles used to access data which are not public accessible;
- Viral Marketing [18]: this category includes spamming attacks and phishing attacks used to disseminate unsolicited messages, or malicious software.

Viral marketing, in particular, is one of the main privacy issues. By considering both platforms, decentralized and centralized, the similarity between users, named homophily, has a strong impact on this specific privacy issue. Homophily [23] is a well-known concept in Social Networks; it means that similar individuals associate with each other more often than others. Several studies have been performed, and a detailed summary is shown in [25]. Homophily can be observed in online platforms, such as OSNs, using analytical techniques. However, there have been very few studies that involved analysis of OSNs to investigate the principle of homophily, probably for the lack of real OSNs datasets. Authors in [9] study LiveJournal and Wikipedia users' activities (such as edits) to evaluate the similarity between individuals. Simsek and Jensen [28] have proposed a technique applied in distributed systems, for navigating networks by exploiting homophily. Results show that a simple product of degree and homophily measures can be quite effective in guiding local search.

As concerns OSNs, in [5] authors proposed a systematic approach to study homophily concept on two OSNs, *BlogCatalog* and *Last.fm*. Another work [4] proposes an analysis of *Last.fm*. Authors show a detailed study in which they try to contrast online ties with offline links and to predict both kinds of ties automatically. Indeed, on Last.fm, users connect to online friends but also they reveal their real-life by listing a set of events they physically co-attended. Authors investigate homophily with respect to demographic, structural and taste-related attributes along online/offline ties, and if it is possible to exploit the found similarities to predict the strength of ties. An interested work, close to our analysis, is proposed by [11]. The paper proposes an analysis of ego networks in Twitter to find homophily by considering different attributes. Results show that for a few attributes, there was a consistently high homophily. In our paper, we tried to find homophily by considering preference about movies expressed by Facebook users and by clustering similar user profiles we tried to study the age attribute. On Facebook, the study proposed in [26] has demonstrated that with a little amount of information, it is possible to infer user attributes. Indeed, authors provide a study in which given data taken from Facebook, they found that users are often friends with other who share their attributes. This has concrete consequences for privacy and anonymity, as we want to prove in this paper. Finally, authors of [12] showed that homophily can be discovered by evaluating temporal information of OSNs. Indeed, the availability patterns of the egos and their alters increases when considering alters which strong ties. In addition, alters of Dunbar's circles [21] have similar temporal patterns.

3 The impact of homophily on privacy in OSNs

Nowadays, OSN analysis attracted the attention of several research communities and it is applied in different scenarios. Indeed, the users that connect to the OSNs produce a huge amount of information that can be retrieved and exploited for different purposes. One of the most interesting properties investigated by current OSN services is the concept of homophily between users. Homophily means that similar individuals associate with each other more often than others and this property is common to all OSNs because it is an intrinsic aspect of the users' life. Indeed, several studies have empirically proved the existence of homophily in real OSNs, regardless of both the type of the service and the architecture of each OSN [25, 14]. As a matter of fact, homophily has been investigated on different dimensions of users' life and the data produced by users of the OSNs can be exploited to uncover common habits (e.g., temporal patterns, interaction patterns, etc.) or similar preferences (e.g., interests, products, items, etc..) between users. For instance, the Facebook's users have a social profile which contains some information represented by pages and groups (Figure 1) which can be used to measure homophily between users. Both a like operation on a Facebook page or the join to a group gives an important information to the characterization of a user. In addition, the homophily concept has showed some relationship with trust because similar users are more likely to establish trust relations while trusted users are more similar [29]. For these reasons, homophily in OSNs can impact the privacy of the users in different ways.

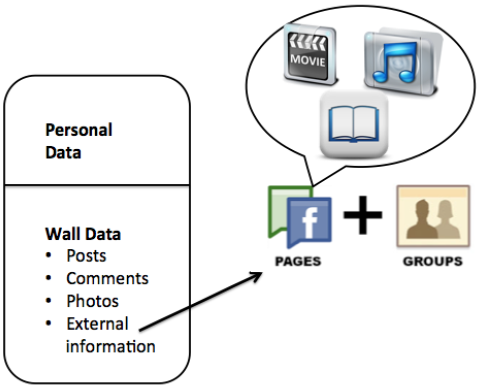


Fig. 1 User's profile overview.

For example, the homophily concept is widely exploited by instant advertising and massively target advertising to understand how the user's friends impact the predictability of his/her behaviour [14] or to perform product/item recommendation [7]. In addition, this property is used by link prediction [30] and community detection [31] algorithms in order to find hidden relationships between users of the OSNs. Recently, the homophily concept has been studied in the context of influence processes in OSNs in order to identify influential users who have maximum positive influence [32]. In all such cases, homophily is used to infer information that users want to protect from their contacts, exposing the users to a number of privacy risks, such as, improper disclosure of personal information. As a result, it is essential to know whether homophily involves other aspects of the user's life (such as movies preferences, music preferences or books preferences) because the knowledge of this information could be a threat for the privacy of the users. Understanding homophily can help in obfuscating these informations in order to prevent information leakage and avoid predictability of personal information.

4 The structure of Facebook ego networks

Ego Networks [17, 2] are a well studied model to represent Social Networks from a user's point of view. The *Ego Network* is a structure built around a user, called ego, and it is made of his/her direct friends, known as *alters*, and the existing ego-alter and alter-alter relations Fig. 2.

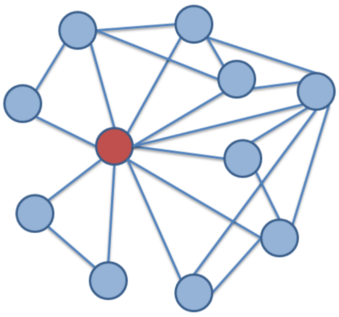


Fig. 2 Ego network of a generic node (red node).

Formally, an ego network can be modeled as a subgraph of the whole social networks. Indeed, by considering that the whole graph is composed by V vertexes, each single vertex $u \in V$ can be seen as an *ego* and $EN(u) = (V_u, E_u)$ is the ego network of u where $V_u = \{u\} \cup \{v \in V | (u, v) \in E\}$, $E_u = \{(a, b) \in E | \{a, b\} \subseteq V_u\}$ and E is the set of edges present in the original graph. $N(u) = V_u - \{u\}$ is the set of adjacent nodes of u.

An interesting property of social networks concerns the discovery of a strong relation between the amount of cognitive resources and the number of social relationships an individual can actively manage. The human brain provides cognitive resources for maintaining about 150 active social relationships, on the average [15, 21]. This limit is popularly known as the Dunbar's number. Moreover, alters in the ego networks are hierarchical organized with respect to the strength of the ties. People tend to maintain social relationships at different levels of intimacy. There are a series of circles of alters arranged in a hierarchical inclusive sequence based on the level of intimacy. In detail, the innermost circle includes alters with a very strong relationship with the ego. Instead, the outermost circle includes acquaintances, with a relatively weak relationship with the ego (Figure 3). The scaling factor of the size of the circles is close to 3. To evaluate the tie strength of a relationship is used the minimum frequency of contact between the ego and the alters. This important property of the Social Network have been largely studied in Offline Social Networks, and recently discovered also in Online Social Networks [16, 1, 2]. As a consequence, decentralized Dunbar-based solutions has been proposed in DOSNs [20, 19, 8].

In this paper we evaluate the homophily between users by exploiting the ego networks structure. Furthermore, by considering that the Dunbar property is often found in Social Media, we try to find a correlation between the interests of users and the tie strength of a relationship by studying the Dunbar's circles.

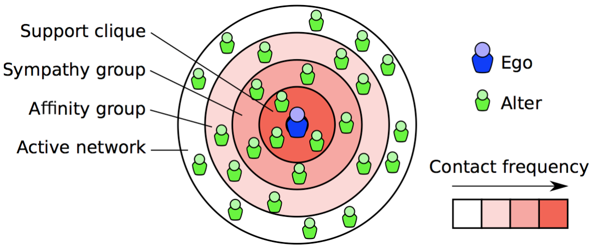


Fig. 3 Representation of the circles in the Dunbar model.

5 The Facebook Dataset

Information about movie preferences of Facebook users have been gathered by a Facebook application, called SocialCircles![1], which exploits the Facebook API to retrieve social information about registered users. The application was able to retrieve information about the Ego Network of registered users, as explained in [12]. In detail, the application retrieved:

- Topology Information. We were able to obtain friends of registered users and the friendship relations existing between them.
- Profile Information. We downloaded profile information of registered users and their friends, such as complete name, birthday, sex, current location, hometown location, works, schools, user devices, movies, music, books, interests and languages.
- Interaction Information. We collected 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 contains 337 complete Ego Networks, for a total of 144, 481 users (ego and their alters). The sample obtained from Facebook consist of 213 males and 115 females (while 9 users did not specify their gender) with age between 15 and 79, having different education, background and geographically location. We focus on the part of the profile which contains the movies that the profile owner likes. About 77% and 58% of the registered users and the registered users' friends respectively, exposes preferences about movies. The registered users have, on average, 5 favorite movies while the users' friends have about 8 favorite movies. The 90% of registered users have a fraction of friends without favorite movies that do not exceed 0.6.

Our dataset contains 69, 519 movie titles. Each title is referred to a Facebook movie page that one or more

[1] http://social.di.unipi.it

our users (registered and their friends) have added to their profiles through the *like* button. Usually, titles includes several typos and they are written in several languages. One of the main problem is that a huge amount of movie titles refers to same movies and titles are different for example, due to typos. A data preprocessing phase has been executed to obtain a refined movies dataset.

6 Discovering homophily

The aim of this paper is the evaluation of homophily in a real Facebook dataset by considering the interest of users expressed through "likes" to movie pages. Homophily between users in the same ego networks could affect the problem of privacy because of the profiling of users by exploiting profiles of friends.

The work has been logically divided into two phases: during the first one, we refined the dataset through a data preprocessing, instead during the second one, we study homophily between users.

6.1 Data Preprocessing

The first step of the data preprocessing cleans dataset by excluding all titles which are not referred to film or contain no Latin characters. The total number of film after this step was $61,918$.

The second step concerns the issue of duplicate Facebook pages which refer to the same movie. In order to discover pages which refer to the same movie we use a set of similarity metrics based on string and the public movies dataset MovieDB[2]. The similarity measures are combined to exploit the main properties of each of them. In particular, we use the following similarity metrics:

Cosine Similarity (CS). It is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. It is a common vector based similarity measure.

Levenshtein Similarity (LS). It is a metric for measuring the difference between two sequences. When we consider it as distance, it measures the minimum number of single-character edits required to change one word into the other. This metric is useful for our goal because a huge amount of titles of pages have typos and MovieDB is not able to recognize the right film associated to the page.

Smith-Waterman Similarity (SW). It is a well-known measure in the Edit-based similarity metrics. It is an algorithm which try to find the best local sequence alignment by comparing segments of all possible lengths. This metric has been chosen because it is often used to discover titles of pages whose title is partially similar. For example, if we consider the following different titles: "Harry Potter" and "Harry Potter is beautiful", this metric is able to return a high similarity between the two titles.

The similarity measures we considered take as input the field *about* containing the information (or title) about the pages.

To exploit all the properties of the chosen similarity metric, we combine the metrics as follows, in order to obtain a global similarity measure between two Facebook movie pages:

$$TitleSim = CS * LS * SW \tag{1}$$

We computed the *TitleSim* measure for each pair of pages. In this way we were able to cluster pages related to the same movie, regardless of the typos or differences in the title among pages. Cluster are computed by considering the *TitleSim* measure. In detail, for each title we find all the other titles which have a *TitleSim* higher than a threshold set to 0.3. We obtain 17185 clusters of similar movie pages. Afterwards, clusters of pages have been used to interrogate the MovieDB dataset. When a query to MovieDB produces a reply, this reply is used inside the cluster to classify the pages which does not receive a response from MovieDB. Indeed, MovieDB is not able to reply in the case of typos but the *TitleSim* measure are able to understand which pages are similar.

At the end of this phase, we obtain the final dataset used to execute our evaluation which contains 45,729 pages which have been enriched by attaching to each page the genre of the corresponding movie taken from MovieDB. The other pages (16,189) have not an associated genre, usually because MovieDB does not provide a reply or the obtained genre is empty.

6.1.1 The MovieDB dataset

The Movie Database (TMDb) is an open database of movies and television information concerning movies, television shows, production companies, and individuals in the entertainment industry. The TMDb API is a RESTful web service to obtain movie information. All content and images on the site are contributed and maintained by users.

[2] https://www.themoviedb.org

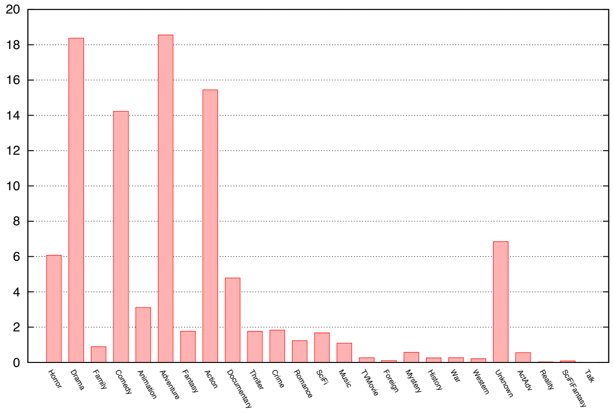


Fig. 4 Distribution of the movie genres.

7 Analysis of the dataset

In this section we show the evaluation of the similarity between users concerning the movie genres. After the preprocessing phase, explained in Section 6.1, we obtain a dataset of 45,792 movie pages. Each page has an associated genre obtained by using the TMDb API. We exploited this information to categorize the users based on their movie preferences.

7.1 Movie genres

We identify 25 genres including the genre *unknown* which contains Facebook pages which are not classified by MovieDB.

Figure 4 shows the distribution of the genres. The majority of the pages are distributed among four principal genres: *Adventure*, *Drama*, *Action*, and *Comedy*. In particular, 20% of pages are classified as *Adventure*, about 18% of pages are classified as *Drama*, and about 15% of pages are classified as *Action* and *Comedy*.

We represent the movie preferences of each user as an array of 25 items, corresponding to the distinct genres identified in the dataset. We investigate if users who like a specific genre also like other genres as well, by analyzing the correlation between genres liked by each user. Figure 5 shows the correlation matrix between movie genres liked by the users where color gradation represents the strength of the correlation (strong correlation corresponds to dark color). The matrix indicates the presence of a higher correlation between *Comedy* and *Drama*, but also *Crime* and *Thriller* with *Drama*. We have also a high correlation between *Action* and *Adventure*.

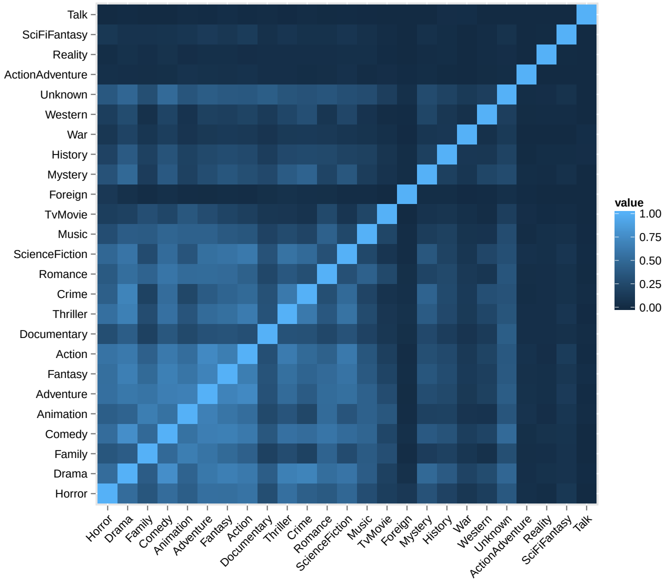


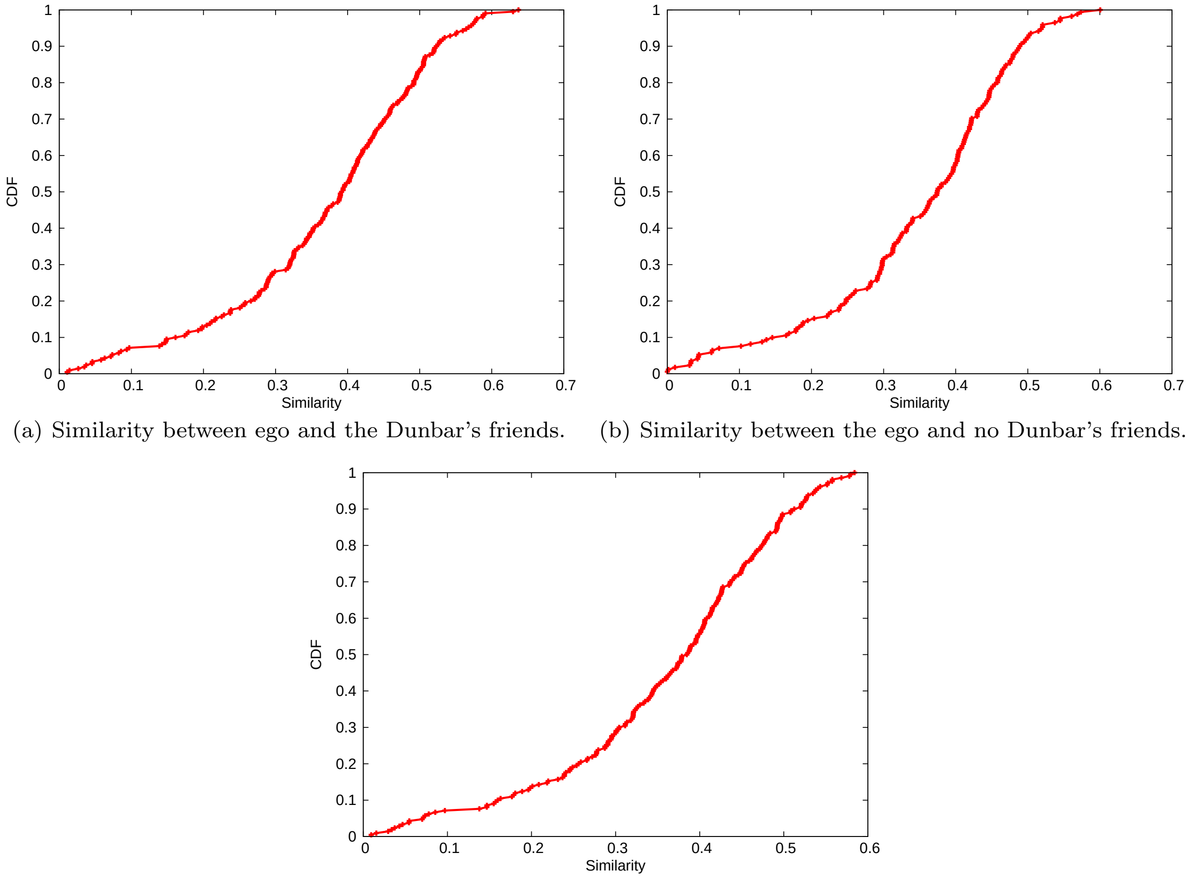
Fig. 5 Correlation Matrix between preferences on different movie genres.

7.2 Evaluation of the Homophily in Facebook

We start to study the homophily between a user and his/her friends by considering the ego networks. We evaluate the similarity between an ego node and his/her alters by considering his/her ego network and by dividing the alters into three different sets:

- Dunbar's Friends. As explained in Section 4, Dunbar explains that human brain has a cognitive limit to the number of people with whom one can maintain stable (active) social relationships. He studied that humans can comfortably maintain only 150 stable relationships. For these reasons, relationships are classified according to the strength of the relationship. In our dataset, the Dunbar friends of an ego are the friends with whom the ego has a stronger relation by taking into account the social interactions (posts, comments, etc. . .).
- No Dunbar's Friends. The Dunbar's Friends are limited to 150. The other alters contained in the ego that are not in the first 150 alters are automatically classified as no Dunbar's Friends.
- All alters. This set contains all the alters in the ego network of the users, i.e., the union of the Dunbar's Friends and the No Dunbar's Friends.

Table 1 reports some statistics about the ego networks considered in our experiments. Ego networks have on average 320 alters and the higher Standard Deviation suggests more variation in the number of alters in the ego networks. Indeed, the smallest ego network has 28 alters, instead the size of the largest one is 1,394.



(a) Similarity between ego and the Dunbar's friends.

(b) Similarity between the ego and no Dunbar's friends.

(c) Cosine similarity between the ego and all its friends.

Fig. 6 Similarity between users by considering the movie preferences

Table 1 Statistics of the ego networks.

Measure	Value
Number of Ego Networks	229
Min Ego Network Size	28
Max Ego Network Size	1394
Mean Ego Network Size	320.707
St. Deviation Ego Net. Size	227.516

We evaluate the similarity between the movie preferences of an ego and all its alters by using the Cosine similarity. Figure 6(c) shows the Cumulative Frequency Distribution (CDF) of the cosine similarity. About the 50% of egos have a similarity with their friends less than 0.4. However, more than the half of ego nodes show a high similarity (between 0.4 and 0.6). This means that the movie preference of an ego are similar to half of its friends, which suggests the presence of a sort of influence.

We investigate in more details the similarity between the ego and the different sets of alters of the ego network. Figure 6(a) shows the CDF of the similarity between the egos and their Dunbar's friends while Figure 6(b) shows the similarity between the egos and their No Dunbar's friends. The plots clearly indicate that users of the No Dunbar's friends expose lower similarity in terms of movie preferences. Indeed, about 50% of egos show a similarity lower than 0.37 with its No Dunbar's friends. Instead, the similarity between the movie preferences of the egos and those of their Dunbar's friends is slightly higher, i.e., half of the users show a similarity of about 0.4. This suggest that users that frequently interact with each other expose a higher level of similarity in terms of movie preference.

Considering that the half of ego nodes show a high similarity, we decide to evaluate if there is a relation between the cosine similarity computed by considering all the friends and the cosine similarity computed on both Dunbar or no Dunbar friends. We evaluate the Pearson correlation shown in Table 2.

We can notice that there is a positive correlation between the similarity computed by considering all the friends and the cosine similarity computed by considering only Dunbar's friends, and a positive but more scattered correlation between the similarity computed

Table 2 Pearson Correlation

	Dunbar's similarity	NoDunbar's Similarity	Alters' Similarity
Dunbar's similarity	1	0.308	0.994
No Dunbar's Similarity	0.308	1	0.287
Alters' Similarity	0.994	0.287	1

by considering all friends and the similarity computed on the set of no Dunbar's friends. In addition, Table 2 indicates that users establish friendship relations with alters having similar interests.

The obtained results show a correlation between tie strength and user's preferences. We evaluate more in detail this correlation by studying the Dunbar's circles. The circles are inclusive, as explained in [16]. However, we take into account circles as both inclusive and exclusive, built by excluding the users contained into the inner circles.

We divide alters into the four Dunbar's circles and we evaluate the cosine similarity between users inside a circle. Usually, users have low values in the preferences vector because the put a like only to a low number of pages in of each category. We consider that these low values can impact on the homophily by decreasing the level of similarity. For this reason, we evaluate the similarity between user by considering a threshold. The thresholds have been used to put a value equals to 0 in each position of the vector where the corresponding value is less than the threshold. We set the threshold to 0.05 because it results as a good trade-off by considering all the values.

In Figure 7 we present our analysis on the Dunbar's circles, both inclusive and exclusive. The analysis shows that the more intern the circle is the more the similarity increase. Indeed, users who belong to the inner circles has a higher similarity between them with respect the similarity of the users in the outer circles. This result suggests that homophily strongly affects the composition of the Dunbar's circles and it is correlated with the tie strength. In Figure 7(a), under the 40% of similarity, the behaviour is similar to the opposite due to the smaller dimension of inner circles respect to the outermost one. A more distinct behaviour of each circle is shown in Figure 7(b), where we consider only the users in the specific circle. The 60% of users has a similarity that respects the Dunbar's circles, instead the first part of the lines are overlapped due to the low number of users in the inner circles, as described before. Moreover, Figure 7 shows how the similarity is higher by considering the inclusive Dunbar's circles due to the contribution of the inner circles. From the point of view of the privacy of users, we demonstrate that there is a strong correlation between homophily and similarity of users' profile, which can be a weak point in terms of privacy. We are not able to understand if the homophily is cause or consequence of the tie strength due to the lack of information inside the dataset. However, the homophily could be a cause of the tie strength because users with a similar profile tend to interact with each other, or the homophily could be a consequence of the tie strength which means that online platforms can guide the rise of homophily by suggesting particular information.

Table 3 Age Clusters with the corresponding interval of age.

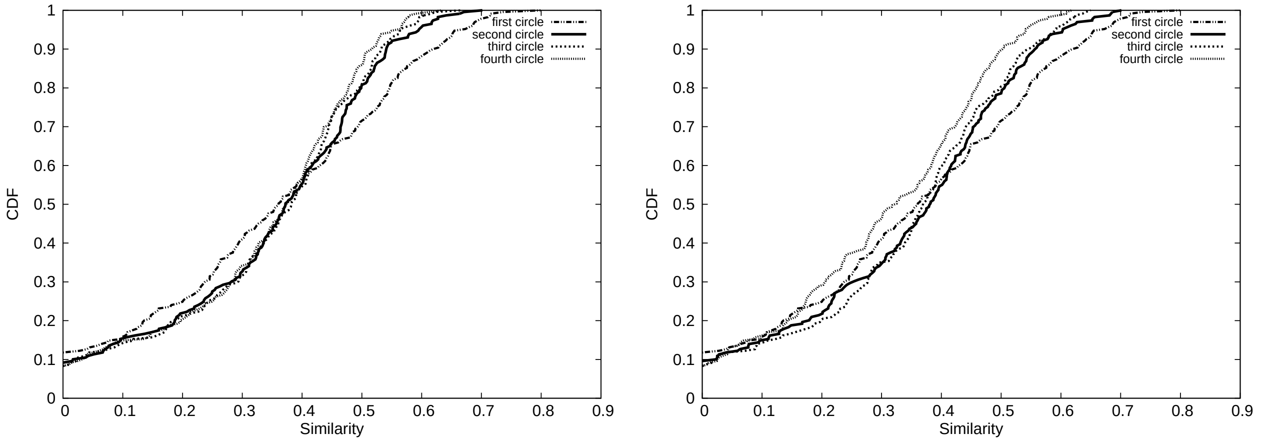
Cluster	Low Value	High Value
Cluster 0	28	35
Cluster 1	80	109
Cluster 2	50	79
Cluster 3	13	21
Cluster 4	36	49
Cluster 5	22	27

8 Homophily and the age attribute

One of the main problems concerning the privacy issue is that homophily can reveal important attributes of the user's profile. In our case, we exploit the film pages of Facebook and we compute cluster of users according to their age attribute. We evaluate homophily inside each cluster to understand if it is possible to retrieve information about users by exploiting homophily. This study has been executed outside the ego networks by considering the whole information of users. We use the k-means algorithm [24] to obtain the cluster of users. K-means executes an iterative procedure which classifies a given set of data through a certain number of clusters fixed a priori. The main idea is to define K centroids, one for each cluster. To understand the right value of K, we execute the k-means repeatedly by changing the value of K, until the average square error seems to be stable. We compute the *Mean Square Error* (MSE) which permits us to estimate the right K value, and we choose K equals to 6 (Figure 8).

Table 3 shows the obtained clusters and the minimum and maximum age of the users of each cluster.

At the end of the first phase in which we compute the 6 clusters, we evaluate the cosine similarity within each cluster. Table 4 shows the results obtained.



(a) Inclusive Dunbar's circle with a threshold of 0.05 (CDF) (b) Exclusive Dunbar's circle with a threshold of 0.05 (CDF)

Fig. 7 Similarity by exploiting the Dunbar's circles (CDF)

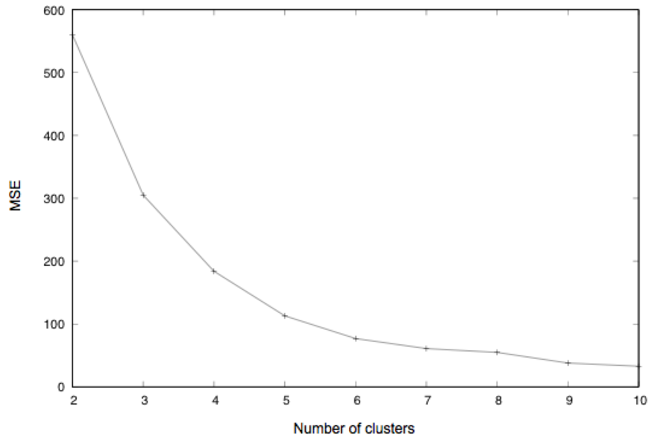


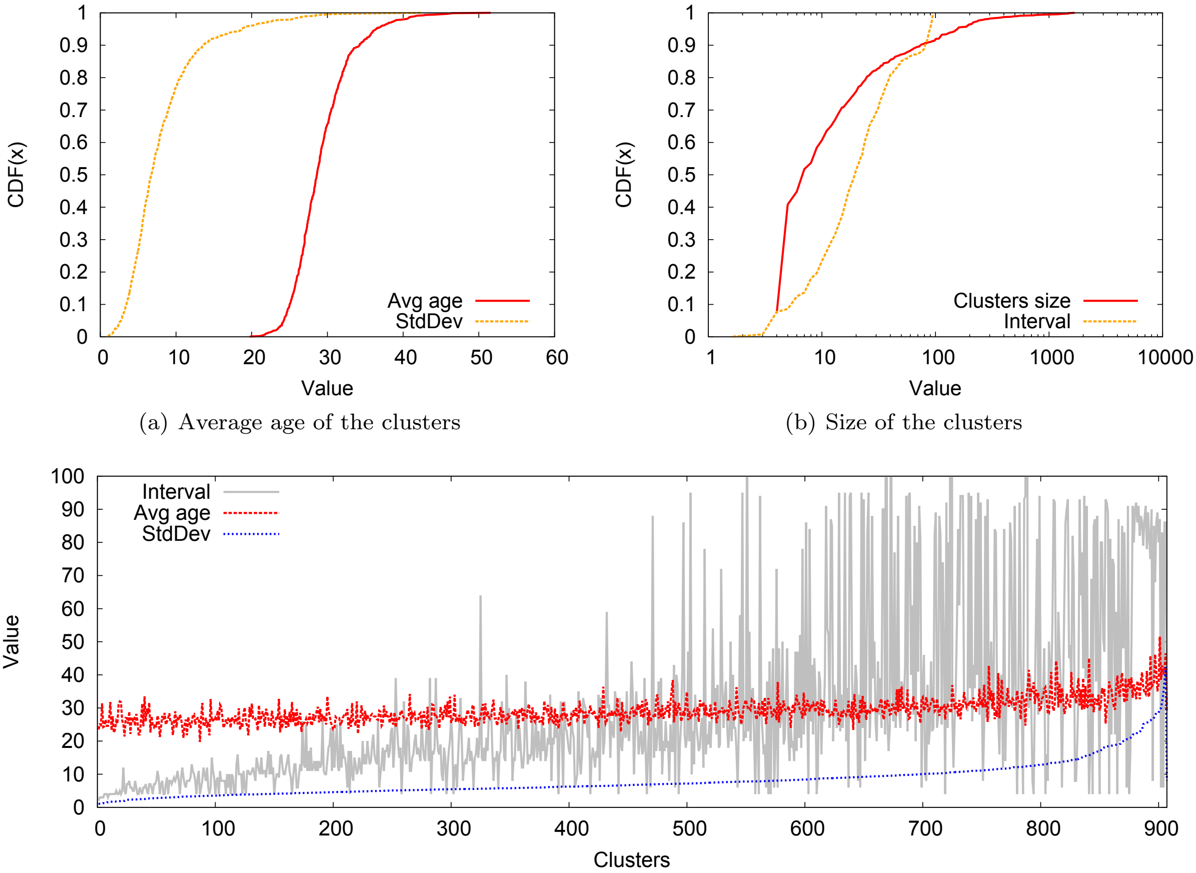
Fig. 8 Computation of the MSE.

Table 4 Cosine Similarity for each cluster

	Min	Max	Avg	St.Dev
Cluster0	0	1	0.26455564	0.3074781
Cluster1	0	1	0.2344554	0.30599057
Cluster2	0	1	0.13961	0.26568424
Cluster3	0	1	0.345162788	0.30753427
Cluster4	0	1	0.20616305	0.292033067
Cluster5	0	1	0.332689	0.31323947

Results show the lack of similarity between users in the same clusters when we consider the whole dataset (global view) instead of the ego networks (local view). We have low values for each cluster and this means that users having the same age do not necessarily share similar movie preferences. Instead, users having similar age expose heterogeneous behaviors in term of movie preferences. In this case, we can say that it is a good point in terms of privacy, because we are not able to retrieve information about other users by considering the age of users. Cluster 3 and 5 have the higher similarity value and the intervals of these two clusters are small with respect to the others. Probably, the obtained clusters are too big and they cover a set of age that are not similar.

The previous analysis aimed at understanding if users with similar age expose the same movie preferences. As a further analysis concerning the movie preferences, we investigated in more detail if users with the same movie interests have also similar ages. In contrast to the previous analysis, which clusters users with respect their ages, in this experiment we organize users in different clusters based on their movie preferences. We removed from this set, the clusters having less than two users and the clusters of users who have not specified any movie preference. In particular, we obtained 906 clusters which correspond to distinct movie preferences where each cluster contains homogeneous users who specified the same movie preferences. Figure 9(b) shows the CDF of the number of users in each cluster (*Cluster size*) while the associated 95% confidence interval (C.I.) is reported in square brackets in the discussion below. The average number of users in each cluster is about 52 [95% C.I. ±15.42]. However, we observed that the distribution of clusters' size is heavily left skewed because the 50% of the clusters have less than 9 users. Indeed, more than 75% of the users belong to clusters which have less than 1000 members. In particular, the first 3 larger clusters contain about 20% of the users of our dataset and they represent groups of users who liked Drama movies (6.8%), Comedy movies (7.3%), and both of them (6.6%). We showed in Figure 9(b) the difference between the age of the younger and the oldest user in each cluster (Interval). The plot clearly indicates that 50% of the clusters consist of users



(a) Average age of the clusters

(b) Size of the clusters

(c) Characteristichs of the clusters

Fig. 9 Comparison between groups of users with similar movie preferences and the ages of these users.

whose age difference does not exceed 20 years.
The Figure 9(a) shows the cumulative distribution of both the average age of cluster users (Avg age) and the standard deviation (StdDev). The distribution is almost normal because the mean and median values are almost coincident (28 with 95% C.I. ±0.26). The average standard deviation value of 6 [95% C.I. ±0.34] suggests strong similarity (in terms of age) between users of the same cluster. To better show the characteristics of each cluster, we plot in Figure 9(c) the average number of user (Avg age), the standard deviation (StdDev), and the distance between the younger and the oldest user (Interval) for each cluster. The graph clearly shows that a large portion of clusters (66%) has low variation of the ages (that is less than 10). This low standard deviation indicates that users with similar movie preferences tend to be close in age. Furthermore, a higher variation of the users' ages (i.e., more than 20) is exposed by less than 0.60% of the clusters.

9 Conclusion and Future Works

In this paper, we propose an analysis of homophily by exploiting real data obtained from Facebook users. This paper is an extended version of our previous work [13] in which we propose a preliminary analysis of homophily in Facebook to manage the problem of the data availability in a DOSNs. In this paper, we analyse homophily more in details and from a general point of view without considering the nature of the platform (decentralized or centralized). We studied the movie preferences exposed by users by exploiting the MovieDB database to retrieve accurate information about the genre of our Facebook pages. Indeed, our analysis has been executed on a real Facebook dataset, from which we have extrapolated more than 45.000 movie pages of Facebook that represent the preferences of about 300 Facebook users. Results show that egos are similar to alters, in particular they are similar to friends with whom they have a strong relation tie. Indeed, by considering the Dunbar's circles, users with a high tie strength are more similar, and this similarity decreases moving from the inner circle to the outermost one. This homophily can be easily used for spamming or viral marketing in general. Another important point we address in this paper is the user profiling. To understand if it is possible to profile a user, we study how we can infer information about age from the movie preferences and viceversa by considering all the Facebook users of our dataset. Results show that by clustering users with respect to the age attribute we are not able to find similar movie preferences, instead by clustering with respect to the movie preferences, we found that 50% of the clusters consist of users whose age difference does not exceed 20 years. This means that it is possible to obtain information about users, such as age in our case, by exploiting information about user profile, like the movie preferences.

We plan to extend our work by investigating more in detail the similarity between users and by analysing their temporal behaviour. Moreover, we want to investigate other features of the social profile, such as music and/or book preferences to better understand the behaviour of users. Finally, we plan to extend the analysis by considering other clustering algorithms which provide communities of different types.

References

1. Arnaboldi V, Conti M, Passarella A, Pezzoni F (2012) Analysis of ego network structure in online social networks. In: Privacy, security, risk and trust (PASSAT), 2012 international conference on and 2012 international confernece on social computing (SocialCom), IEEE, pp 31–40
2. Arnaboldi V, Conti M, Passarella A, Pezzoni F (2013) Ego networks in twitter: an experimental analysis. In: Computer Communications Workshops (INFOCOM WKSHPS), 2013 IEEE Conference on, IEEE, pp 229–234
3. Balduzzi M, Platzer C, Holz T, Kirda E, Balzarotti D, Kruegel C (2010) Abusing social networks for automated user profiling. In: Recent Advances in Intrusion Detection, pp 422–441
4. Bischoff K (2012) We love rock 'n' roll: Analyzing and predicting friendship links in last.fm. In: Proceedings of the 4th Annual ACM Web Science Conference, WebSci '12, pp 47–56
5. Bisgin H, Agarwal N, Xu X (2010) Investigating homophily in online social networks. In: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010, vol 1, pp 533–536
6. Boyd D, Ellison NB (2007) Social network sites: Definition, history, and scholarship. J Computer-Mediated Communication 13(1):210–230
7. Carullo G, Castiglione A, De Santis A, Palmieri F (2015) A triadic closure and homophily-based recommendation system for online social networks. World Wide Web 18(6):1579–1601
8. Conti M, De Salve A, Guidi B, Pitto F, Ricci L (2014) Trusted Dynamic Storage for Dunbar-Based P2P Online Social Networks. In: On the Move to Meaningful Internet Systems: OTM 2014 Conferences, pp 400–417
9. Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S (2008) Feedback effects between similarity and social influence in online communities. In: International Conference on Knowledge Discovery and Data Mining, KDD '08, pp 160–168
10. Datta A, Buchegger S, Vu LH, Strufe T, Rzadca K (2010) Decentralized online social networks. In: Handbook of Social Network Technologies and Applications, pp 349–378
11. De Choudhury M (2011) Tie formation on twitter: Homophily and structure of egocentric networks. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, IEEE, pp 465–470
12. De Salve A, Dondio M, Guidi B, Ricci L (2016) The impact of users availability on on-line ego networks: a facebook analysis. Computer Communications 73:211–218
13. De Salve A, Guidi B, Ricci L (2017) Analysis of users behaviour from a movie preferences perspective. In: 3rd EAI International Conference on Smart Objects and Technologies for Social Good
14. De Salve A, Mori P, Ricci L (2017) Evaluating the impact of friends in predicting users availability in online social networks. In: International Workshop on Personal Analytics and Privacy, Springer, pp 51–63
15. Dunbar R (1998) The social brain hypothesis. brain 9(10):178–190
16. Dunbar RI, Arnaboldi V, Conti M, Passarella A (2015) The structure of online social networks mirrors those in the offline world. Social Networks 43:39–47
17. Everett M, Borgatti SP (2005) Ego network betweenness. Social networks 27(1):31–38
18. Gao H, Hu J, Huang T, Wang J, Chen Y (2011) Security issues in online social networks. IEEE Internet Computing 15(4):56–63
19. Guidi B, Conti M, Ricci L (2013) P2p architectures for distributed online social networks. In: High Performance Computing and Simulation (HPCS), 2013 International Conference on, IEEE, pp 678–681

20. Guidi B, Amft T, De Salve A, Graffi K, Ricci L (2015) Didusonet: A p2p architecture for distributed dunbar-based social networks. Peer-to-Peer Networking and Applications pp 1–18
21. Hill RA, Dunbar RI (2003) Social network size in humans. Human nature 14(1):53–72
22. Kontaxis G, Polakis I, Ioannidis S, Markatos EP (2011) Detecting social network profile cloning. In: Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on, pp 295–300
23. Lazarsfeld PF, Merton RK (1954) Friendship as a social process: A substantive and methodological analysis. In: Freedom and Control in Modern Society, New York, pp 18–66
24. MacQueen J, et al (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol 1, pp 281–297
25. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. Annual review of sociology 27(1):415–444
26. Mislove A, Viswanath B, Gummadi KP, Druschel P (2010) 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
27. Ruan X, Yue C, Wang H (2013) Unveiling Privacy Setting Breaches in Online Social Networks, Springer International Publishing, pp 323–341
28. Şimşek Ö, Jensen D (2008) Navigating networks by using homophily and degree. Proceedings of the National Academy of Sciences 105(35):12,758–12,762
29. Tang J, Gao H, Hu X, Liu H (2013) Exploiting homophily effect for trust prediction. In: Proceedings of the sixth ACM international conference on Web search and data mining, ACM, pp 53–62
30. Wang D, Pedreschi D, Song C, Giannotti F, Barabasi AL (2011) Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 1100–1108
31. Xie J, Kelley S, Szymanski BK (2013) Overlapping community detection in networks: The state-of-the-art and comparative study. ACM Comput Surv 45(4):43:1–43:35
32. Zhang C, Lu T, Chen S, Zhang C (2017) Integrating ego, homophily, and structural factors to measure user influence in online community. IEEE Transactions on Professional Communication 60(3):292–305