Analysis of Users Behaviour from a Movie Preferences Perspective

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

[1] Department of Computer Science, University of Pisa, Pisa, Italy
{desalve,guidi,ricci}@di.unipi.it
[2] IIT-CNR, via G. Moruzzi 1, 56124 Pisa, Italy

Abstract. Despite their tremendous popularity, Online Social Networks (OSNs) have several issues related to the privacy of social users. These issues have motivated researchers to develop OSN services that take advantage of the decentralized platforms (such as P2P systems or opportunistic networks). Decentralized Online Social Networks (DOSNs) need specific approaches to manage the decentralization of social data. In particular, data availability is one of the main issues and current proposals exploit properties of the social relationships to manage it. At the best of our knowledge, there are no proposals which exploit similarity between users, expressed with the term homophily. Homophily has been well studied in existing sociology literature, however, it is not easily extensible in Online Social Networks due to the limitations of real datasets. In this paper, we propose a preliminary analysis of similarity of social profiles in term of movie preferences. Results reveal that user's friends are characterized by a different levels of similarity which can be exploited to propose solutions for the data availability problem.

Keywords: Decentralized Online Social Networks · Data availability Homophily

1 Introduction

Online Social Networks (OSNs) [3] are the most popular online applications that have changed the way of how people interact between them. During the years, OSNs have filled the human life and several personal data are shared on these platforms. A huge amount of privacy problems have been arisen, in particular in Facebook, which is the most well-known OSN. To overcome privacy issues, decentralized solutions have been proposed. Decentralized Online Social Networks (DOSNs) [7] 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 the social services. A DOSN is able to face the main problems of centralized solutions (i.e. scalability, privacy, etc.). However, in a decentralized environment other important issues have to be managed, such as the data availability [16]. Indeed, DOSNs are built on top of the users' devices

B. Guidi et al. (Eds.): GOODTECHS 2017, LNICST 233, pp. 226–236, 2018.
https://doi.org/10.1007/978-3-319-76111-4_23

and when a user goes offline, data which are stored on his device should be available. In this context, the knowledge of human behaviour can help to manage the data availability problem. In OSNs, users are characterized by a virtual profile which contains attributes representing their private information and interests. These information can be easily obtained and analysed to find similar behaviours or other important aspects.

As for instance, homophily [17] is the principle that a contact between similar people occurs more than among dissimilar people and it is useful to implement new replication strategies which take into account groups of users with specific interests to store data. Indeed, users tend to bond more with users who are similar and studies have shown that similarity between users is a good indicator of possible future interactions.

In addition, the similarity between users is an indicator of how fast the information spreads among users and it can be exploited either to limit or to speedup the dissemination of information to the users having common interests.

In this paper, we investigate the homophily by focusing on a user-centric point of view. We exploited the ego network model to structure the relations between the center user (named ego) and its friends (named alters) [13]. The relations between the ego and the alters are used to represent the movie preferences and the similarity between these preferences is evaluated. Furthermore, we evaluate the similarity between OSN users by exploiting the Dunbar's concept [12]. Our evaluation exploits a real Facebook dataset and all the studies concern movie genres expressed by likes to Facebook pages. Results shows that there is homophily between egos and its alters and, more in detail, the similarity is affected by the tie strength computed by considering the contact frequency between users in Facebook.

The rest of the paper is organized as follow. Section 2 describes the current proposals in both the DOSNs fields and the homophily in OSNs. In Sect. 3 we explain how use homophily in a DOSN. In Sect. 4 we introduce our Facebook dataset and in Sect. 5 we show our evaluation. Conclusions and Future Works are proposed in Sect. 6.

2 Related Work

In this section we describe current DOSN proposals and the main studies conducted in Online Social Networks to discover homophily between social users.

2.1 Current DOSNs

The first important proposal is Diaspora [1], which consists of a federated network of servers. On one hand, Diaspora represents the first example of a real distributed social network, on the other hand, its main drawback is the scalability, due to its architecture is not fully distributed. Other proposals exploit a fully distributed architecture, often implemented by P2P systems. In Safebook [6], data are stored in a particular social overlay named "Matryoshka", which

are concentric rings of peers around each users peer. LifeSocial [14] is a DOSN focused on the privacy issue, where user information is stored by exploiting a Distribued Hash Table (DHT) and the OSN functionalities are realized by plug-ins. PeerSon [4] is a distributed infrastructure for social networks whose focus is related to security and privacy concerns. It proposes a two-tier architecture where the first tier is a DHT and the second tier consists of the nodes representing users. All users' content is encrypted. DiDuSoNet [15] is a DOSN based on the Dunbar's concept, where social data are stored only on trusted nodes.

2.2 Homophily in OSNs

Homophily [17] means that similar individuals associate with each other more often than others. Several studies have been performed, and a detailed summary is shown in [18]. However, to the best of our knowledge, 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 [5] study the LiveJournal and Wikipedia data and used activities such as user edits to evaluate the similarity between individuals. In [2] authors proposed a systematic approach to study homophily concept on two online social media networks, *BlogCatalog* and *Last.fm*. Simsek and Jensen [19] 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. Finally, authors of [8] showed that the availability patterns of the egos and their alters increases when considering alters with a strong tie strength referred to the ego. In addition, alters of Dunbar's circles have similar temporal pattern.

3 Homophily as a Strategy for the DOSNs' Issues

DOSNs have several issues regarding the decentralization of social services. One of the main problem is the data availability, which occurs in distributed systems because data are stored among users and the online behaviour of them can affect the availability of social data. In DOSNs, replication is the most used technique to maximize data availability and it consists of storing copies of the same data on several devices and the users where data are allocated are named replica peers. The understanding of the user temporal behaviour is a crucial aspect for all those systems that rely on the users' resources for the daily operations, such as DOSNs.

Current approaches exploit properties of the social graph, such as the coverage of the social graph [9] and community structure [10,11] to face the problem of data availability. However, other important characteristics of users can be used to manage the problem of data availability. The homophily concept studied in this paper is suitable to be used for the data availability problem because we expect that users who have similar behaviour (in terms of movies preferences, music preferences, books preferences, etc.) are more interested to the same contents.

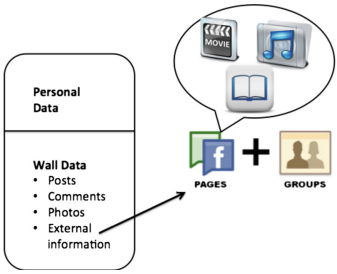


Fig. 1. User's profile overview.

As a result, the data availability strategies can be refined by allocating the data on the peers of the users having similar interests. User's profile contains several external information represented by pages and groups (Fig. 1) which can be used to measure similarity 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.

4 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 [8]. 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

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

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 registered users' friends have about 8 favorite movies. The most part of the registered users (90%) 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 of more 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 different 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.

4.1 Data Preprocessing

The first step of the data preprocessing cleans dataset by excluding all titles which both are not referred to film and titles which contains 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 Movie Database (TMDb) is an open database of movies and television information concern 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. The similarity measures used in the data preprocessing are combined to exploit the main properties of each of them. In particular, we use the following similarity metrics:

Cosine Similarity. 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. 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.

Smith-Waterman Similarity. It is a well-known measure in the Edit-based similarity metrics. It is an algorithm finding the best local sequence alignment or using the longest common subsequence.

The similarity measures we considered take as input parameters the field *about* containing the information (or title) about the pages. The measures are combined as follow, in order to obtain a global similarity measure between two Facebook movie pages:

$$TitleSim = CosineSim * LevenshteinSIM * SmithWatermanSim \quad (1)$$

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

We computed the *TitleSim* measure for each pair of pages. In this way we were able to cluster similar pages related to the same movie, regardless of the typos or differences in the title among 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.

5 Analysis of the Dataset

In this section we show the evaluation of the similarity between users as concerns the movie genres. After the preprocessing phase, explained in Sect. 4.1, we obtain a dataset of 45,792 movie pages. Each page has an associated genre which we use to categorize the behaviour of users.

5.1 Movie Genres

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

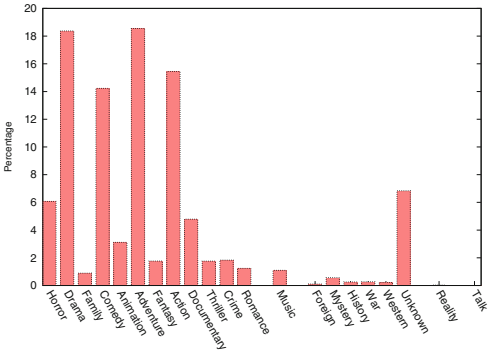


Fig. 2. Distribution of the genres.

Figure 2 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*.

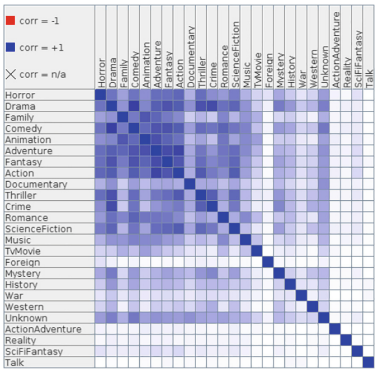


Fig. 3. Correlation matrix between preferences on different movie genres.

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 3 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*.

5.2 Evaluation of the Homophily in Ego Networks

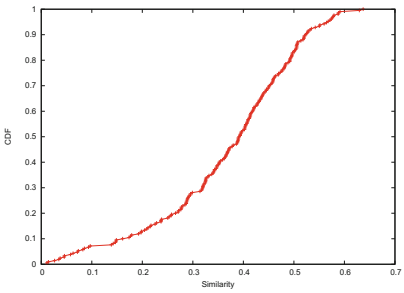
We start to study the homophily between a user with its friends by considering ego networks. We evaluate the similarity between an ego node and its alters by dividing them into three different sets:

- Dunbar's Friends. Dunbar [12] explains that human brain has cognitive limit to the number of people with whom one can maintain stable social relationships. He proposed that humans can comfortably maintain only 150 stable relationships. For these reason, relationships are classified according to the strength of the relation. 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.

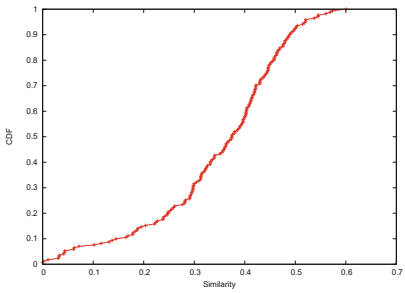
Table 1 reports some statistics about the ego networks considered in our experiments. Ego networks have on average 320 alters and the higher Standard

Table 1. Statistics of the ego networks.

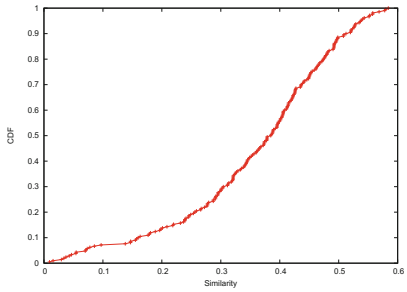
Measure	Value
Number of ego networks	229
Min ego network size	28
Max ego network size	1394
Mean ego network size	320.707
St. deviation ego net. size	227.516



(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. 4. Similarity between users by considering the movie preferences

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.

We evaluate the similarity between the movie preferences of an ego and all its alters by using the Cosine similarity. Figure 4(c) shows the Cumulative Frequency Distribution (CDF) of the cosine similarity. About the 50% of egos have a similarity with its 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 the half of its friends, by suggesting the pres-

ence 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 4(a) shows the CDF of the similarity between the egos and their Dunbar's friends while Fig. 4(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.

Consider that the half of ego nodes shows 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.

Table 2. Pearson correlation

	Dunbar's similarity	No Dunbar'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

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 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.

6 Conclusion and Future Works

In this paper, we propose a preliminary analysis of the homophily in a real Facebook dataset to face the problem of data availability in DOSNs. We study the movie preferences similarity in ego networks by exploiting the MovieDB database to retrieve information about the genre of our Facebook pages. In terms of data availability, the homophily between ego and its alters, in particular with Dunbar's friends can be exploit to predict the content requests by analysing interactions and other properties of the social graph.

Results show that there is a high homophily between an ego node with its alters, in particular with its Dunbar's friends. We plan to investigate more in detail the similarity between users by analysing the temporal behaviour of users and we want to detail how Dunbar's friends impact on the similarity by

analysing the Dunbar's circles. Moreover, we want investigate other feature of the social profile, such as music and/or book preferences to better understand the behaviour of users. Finally, we want propose a specific data availability strategy which takes into account the homophily studied in this paper.

References

1. Diaspora Website. https://diasporafoundation.org/
2. Bisgin, H., Agarwal, N., Xu, X.: Investigating homophily in online social networks. In: 2010 Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 1, pp. 533–536 (2010)
3. Boyd, D., Ellison, N.B.: Social network sites: definition, history, and scholarship. J. Comput.-Mediat. Commun. **13**(1), 210–230 (2007)
4. Buchegger, S., Schiberg, D., Vu, L.H., Datta, A.: PeerSoN: P2P social networking - early experiences and insights. In: SNS, pp. 46–52. ACM (2009)
5. Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., Suri, S.: Feedback effects between similarity and social influence in online communities. In: International Conference on Knowledge Discovery and Data Mining, KDD 2008, pp. 160–168 (2008)
6. 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)
7. Datta, A., Buchegger, S., Vu, L.H., Strufe, T., Rzadca, K.: Decentralized online social networks. In: Furht, B. (ed.) Handbook of Social Network Technologies and Applications, pp. 349–378. Springer, Boston (2010). https://doi.org/10.1007/978-1-4419-7142-5_17
8. De Salve, A., Dondio, M., Guidi, B., Ricci, L.: The impact of users availability on on-line ego networks: a facebook analysis. Comput. Commun. **73**, 211–218 (2016)
9. De Salve, A., Guidi, B., Mori, P., Ricci, L.: Distributed coverage of ego networks in F2F online social networks. In: 2016 International IEEE Conferences ATC, pp. 423–431 (2016)
10. De Salve, A., Guidi, B., Ricci, L.: An analysis of ego network communities and temporal a affinity for online social networks. In: Gaggi, O., Manzoni, P., Palazzi, C., Bujari, A., Marquez-Barja, J.M. (eds.) GOODTECHS 2016. LNICST, vol. 195, pp. 135–144. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61949-1_15
11. De Salve, A., Guidi, B., Ricci, L.: Evaluation of structural and temporal properties of ego networks for data availability in DOSNs. Mob. Netw. Appl. 1–12 (2017)
12. Dunbar, R.I.M.: The social brain hypothesis. Evol. Anthropol.: Issues, News, Rev. **6**, 178–190 (1998)
13. Everett, M.G., Borgatti, S.P.: Ego network betweenness. Soc. Netw. **27**, 31–38 (2005)
14. Graffi, K., Gross, C., Stingl, D., Hartung, D., Kovacevic, A., Steinmetz, R.: LifeSocial. KOM: a secure and P2P-based solution for online social networks. In: IEEE CCNC (2011)
15. 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**, 1177–1194 (2015)
16. Guidi, B., Conti, M., Ricci, L.: P2P architectures for distributed online social networks. In: 2013 International Conference on High Performance Computing and Simulation (HPCS), pp. 678–681. IEEE (2013)

17. Lazarsfeld, P.F., Merton, R.K.: Friendship as a social process: a substantive and methodological analysis. In: Freedom and Control in Modern Society, pp. 18–66, New York (1954)
18. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. Ann. Rev. Sociol. **27**(1), 415–444 (2001)
19. Şimşek, Ö., Jensen, D.: Navigating networks by using homophily and degree. Proc. Nat. Acad. Sci. **105**(35), 12758–12762 (2008)