An Analysis of Ego Network Communities and Temporal a Affinity for Online Social Networks

Andrea De Salve, Barbara Guidi(✉), and Laura Ricci

Department of Computer Science, University of Pisa, Pisa, Italy
{desalve,guidi,laura.ricci}@unipi.it

Abstract. The wide diffusion of Online Social Networks (OSNs) presents several advantages, like the definition of simple tools for information sharing and spreading. However, OSNs present also some drawbacks, one of the most important one is the problem of privacy disclosures. Distributed Online Social Networks (DOSNs), which decentralize the control of the social network, have been recently proposed to overcome these issues. The decentralization of the control has issued several challenges, one of the main ones is guaranteeing data availability without relying on a central server. To define users' data allocation strategies, the knowledge of the structure of the ego network and of the user' temporal behaviour is required. Unfortunately, the lack of real datasets limits the research in this field. The goal of this paper is the study of the behaviour of users in a real social network in order to define proper strategies to allocate the users' data on the DOSN nodes. In particular, we present an analysis of the temporal affinity and of the social communities based on a real Facebook dataset.

Keywords: P2P social networks · DOSN · Data availability · Temporal affinity · Community detection

1 Introduction

Online Social Networks (OSNs) have created a new way of interaction and communication among people. An OSN [3] is an online platform that provides services for a user to build a public profile and to establish connection between users. OSNs are almost always based on centralized structures which have intrinsic drawbacks including scalability and privacy [7]. Privacy is probably the most studied issue because social networks have become the main channel of privacy disclosure. These drawbacks have led researchers to investigate alternative solutions, such as distributed approaches. A Distributed Online Social Network (DOSN) [7] is an online social network implemented on a distributed platform, such as a network of trusted servers, P2P systems or an opportunistic network. Decentralization has several consequences, in particular in term of privacy. In

O. Gaggi et al. (Eds.): GOODTECHS 2016, LNICST 195, pp. 135–144, 2017.
DOI: 10.1007/978-3-319-61949-1_15

fact, no central entity exists that decides or changes the terms of service. However, since the social data is stored on the hosts of the users, its availability depends on the online behaviour of the users. This is a major problem in DOSNs because users' behaviour has a high fluctuation that can lead to data becoming unavailable or lost. Replication is one of the most popular approaches to manage this problem. Replica of user's profiles may be stored on nodes of the distributed systems to improve data availability, but at the same time, it introduces the problem of consistency and the problem of minimizing the number of replicas.

The study of good selection policies for the election of replica nodes with particular attention to the number of replicas has recently gained momentum. These mechanisms must allocate data on the users' nodes so to maximize the availability of the social data. Existing solutions may consider exclusively friend nodes, a set of trusted nodes among the friends, or random nodes; and existing selection policies may be based on information on both the duration and the distribution of online availability of nodes and on the ability of the replica to serve the requests for data replicated on it.

The main goal of this paper is to present a twofold study of a Facebook data set. First, we extend the analysis proposed in our previous work [8] regarding the temporal behaviour of users by considering the global ego network structure. As a second step, we introduce a first study of the communities present in the ego networks of the users. At the best of our knowledge, this is the first analysis of this kind, in this research field. The analysis may detect the nodes of the users which belong to more than one community and that may act as a "hubs", able to serve accesses to the replica they store for many nodes of the ego network.

The rest of the paper is organized as follow. Section 2 describes the related work. Section 3 introduces the problem of data availability in DOSNs. Section 4 describes our twofold work on temporal affinity and community detection. Section 5 investigates the result of our analysis. Finally, Sect. 6 draws the main conclusions.

2 Related Work

During the last years, several DOSNs have been proposed. 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 not fully distributed.

Other proposals exploit a full distributed architecture often implemented by P2P systems. LifeSocial [11] is a P2P OSN focused on the privacy issue, where user information is stored by exploiting a DHT and it presents an approach where all 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 Distributed Hash Table (DHT) and the second tier consists of the nodes representing users. The idea is to use the DHT to find the necessary

information for users connecting directly to the target nodes. This approach comes without a replication scheme and stores offline messages on the DHT (OpenDHT in the prototype implementation). All users' content is encrypted.

As concern the data availability problem, only few proposals have addressed this issue. Safebook [6] addresses privacy in OSNs by using a three-tiers architecture where data are stored in a particular social overlay named "Matryoshka". Matryoshkas are concentric rings of peers around each users peer that provide trusted data storage and obscure communication through indirection. SuperNova [16] is an architecture for a DOSN that solves the availability issue by relying on super-peers that provide highly available storage. DiDuSoNet [12] is a P2P DOSN focused on the data availability problem which uses, as SafeBook, a particular overlay based on trusted connections to store data. Other specific solutions are presented in [14], where authors propose a replication strategy based on storing the replicas of users profiles only on a set of trusted proxies.

3 Data Availability in DOSN

The problem of Data Availability has been mainly studied in the area of P2P networks [2,13], for instance work on P2P storage systems date back to the OceanStore [13] initiative to achieve archival storage using end-user resources.

Recently, the problem has been studied also in the context of DOSNs. Most current approaches for DOSNs rely to external storages, such as private servers or cloud. Among DOSNs, only few approaches, such as PeerSoN [4], DiDuSoNet [12], and Safebook [6] do not rely on external storage. In detail, they exploit social relations between users to decide where data has to be stored. The rationale behind this strategy is that social friends are natural candidates for replicating the data of a user, as they are interested in his/her data. In particular, the solution proposed in [12] leverages the properties related to a friendship relationship (such as trust, strength, or types of the friendship) to guide the data storage.

However, taking into account only the friendship relations and their properties is not enough to ensure to each user the availability of his/her data, at any time. Indeed, geographical proximity between users is one of the most significant factor that make the formation of a friendship (and communities) possible. As result, the majority of the user's friends live geographically close to each other and time-zone differences are negligible since users are connected during the daylight hours and disconnected during the sleeping time (i.e. at night). Consequently, taking into account the temporal behavior of users is crucial when trying to maximize data availability.

In this work we are interested in studying the temporal affinity between Facebook users and their friends by considering a well-known social network structure: the ego network [10], which is the network constituted by a user (*the ego*), his/her direct friends (*the alters*) and the social ties occurring between them.

4 An Analysis of Facebook's Ego Networks: Temporal Affinities and Communities Structure

We evaluate a real dataset gathered by *SocialCircles!*, a Facebook application deeply described in [8]. At the best of our knowledge, our dataset is the only one which contains structural and temporal information about users.

We sampled 337 registered egos and their friends every 8 min for 10 consecutive days (from Tuesday 3 June 2014 to Friday 13 June). Using this methodology we were able to access the temporal status of 308 registered users and of their friends (for a total of 95.578 users). For the purpose of clarity, we will refer to ego nodes to indicate these 308 users. We consider the availability trace of each user to determine the start of a session and its termination. More specifically, time starts at the beginning of time s_0 and it is segmented in a subsequent time periods (*time slots*).

4.1 Temporal Affinity

With the term *Temporal Affinity* we refer to the phenomena of users having the same temporal behaviour, i.e. the probability that they are online in the same interval of time. In more detail, considering the ego network of an ego E, we study how similar is the temporal behaviour of E with respect to his/her the alters. To analyse the temporal affinity, we use a specific *presence array* of 2001 entries (one for each temporal slot in our dataset) for each couple ego-alter. Each slot refers to 8 min and it contains a value: 1 if the user is online in this slot, 0 otherwise. Since our goal is to understand how the temporal behaviour of users can affect the data availability in DOSNs and considering that there are particular day periods when users tend to be offline (e.g. during the night), we propose two different metrics for the Temporal Affinity:

- *Daily Temporal Affinity (DTA)*, which exploits all the 2001 temporal slots;
- *Nighttime Temporal Affinity (NTA)*, which considers a subset of nighttime slots.

The availability of each user is represented by an availability vector of fixed size. We evaluate the temporal affinity using the cosine similarity metrics.

To evaluate the NTA, we need to define when an ego can be considered active during the night (i.e. from 12:00 midnight to 06:00 a.m.) of a day i. An ego is *nighttime* during the night of the day i if and only if it has been online for at least 15 temporal slots. We define a *k-nighttime ego* an ego which has been nighttime for k nights and a *nighttime alter* as an alter which is online for the 95% in the same nighttime slots of a *k-nighttime ego*. The *nighttime affinity coefficient* is defined as the average number of nighttime alters of a *k-nighttime ego*.

4.2 Community Detection

Community structure is considered to be a significant property of social networks. Numerous techniques have been developed for both efficient and effective

community detection. A user usually has connections to several social groups like family, friends, and colleagues. Further, the number of communities an individual can belong to is actually unlimited because a person can simultaneously take part in as many groups as he/she wishes [17].

To discover communities in the ego networks, we decide to use the Label Propagation (LP) [15], implemented by DEMON [5], which defines rules to spread labels in the network so that nodes labeled with the same colour form a community. DEMON is suitable for a distributed implementation and it can be easily adapted to any kind of network, dense or sparse.

We define a new index, the *k-overlapping index, KOI,* to evaluate the overlapping of communities. The index is defined as follow: for each ego network EN including a set CS of communities, for each alter $a \in EN$, we compute the number of communities $\in$ CS, the alter a belongs to and detect alters that appear in at least k communities. We compare this value with the total number of nodes in the ego network. Formally, the index is defined in the Eq. 1.

$$KOI(EN, k) = \frac{|V^1|}{|V|} \tag{1}$$

where V is the set of alters nodes of the ego network EN and V^1 is the set of alters of EN which belong to, at least, k communities.

On the basis of the KOI index, we can detect alters which belong to a significant set of communities and are, therefore, suitable to host and spread replicated data. On the other way round, the allocation strategy must take into account that, since each user device has a limited amount of memory, allocation can causes a huge amount of load on a single node belonging to a lerge set of communities, because it could be chosen by a large set of friends to replicate data.

5 Evaluation

In this section we describe the evaluation of the temporal affinity and of the communities belonging to an ego network.

5.1 Temporal Affinity

The first analysis concerns the evaluation of the Daily Temporal Affinity, referenced as DTA. As explained in [8], a user can be online, offline or in a idle state. We decide to consider the online state, and we do not distinguish the idle and offline state into the presence array so that both of them are represented by the value 0, in the presence array.

About the 80% of the couple ego-alter has a low similarity, as depicted in Fig. 1. This low value is influenced by the online behaviour of each ego. A little set of couples (less than 5%) show medium/high values of similarity. These results confirm our previous analysis on Dunbar-based ego networks where users present

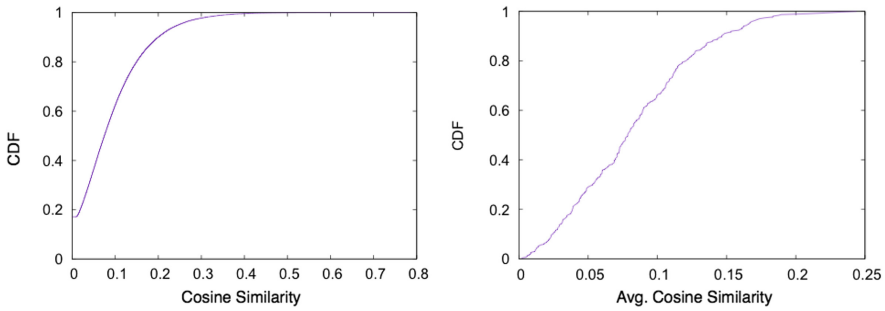
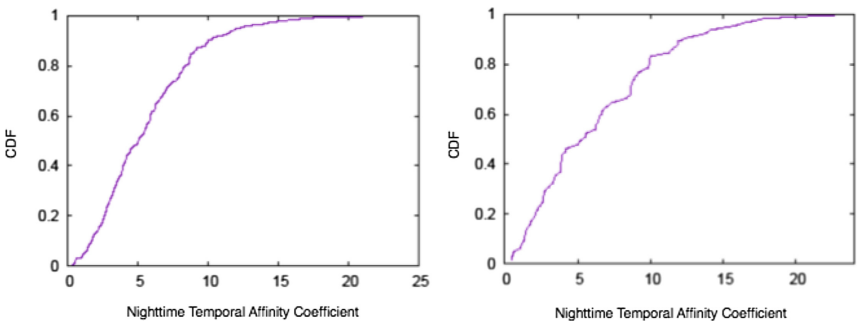


Fig. 1. The cosine similarity and the average value for the *daily temporal affinity*

a high similarity only with trusted nodes. In fact, extending the temporal affinity analysis to the ego networks, the result is similar.

The second step concerns the analysis of the ego networks to find the Nighttime Temporal Affinity, indicated with NTA through the *Temporal Affinity index* introduced in Sect. 4. For the evaluation, we consider both the online and the idle state, and we vary the parameter k which defines the k-nighttime ego. Notice that when an ego present a high NTA with a set of distinct alters means that tie strength between the ego and these alters is strong.

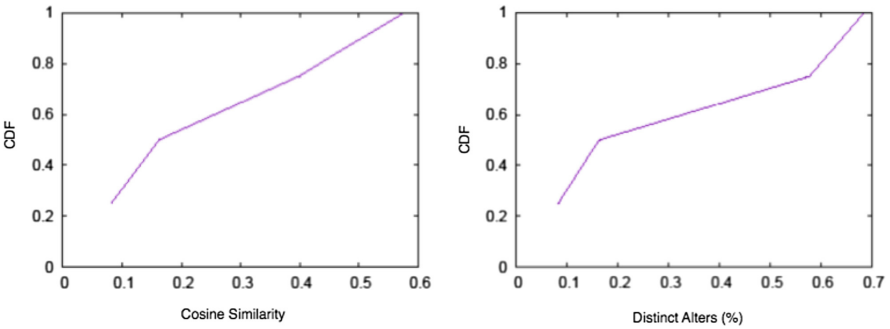
For the evaluation, we use two indexes: *fixed indexes*, when the night is divided into 38 slots from about the 03.00 A.M. to the 08.00 A.M. and *variable indexes*, when the night is defined by the slots containing less than 10000 users.



(a) NTA with k = 1 (b) NTA with k = 4

Fig. 2. NTA with fixed indexes, by varying the number of nights k

Analyzing the NTA we notice that less than 10% of egos can be considered nighttime. This value is drastically reduced when the constraints of the number of nights needed to be nighttime increase.



(a) NTA with k = 2 (b) Distinct alters with k = 2

Fig. 3. Variable indexes with k equals to 2

For example, if we consider an ego network of 400 nodes, the number of alters which contribute to the NTA is less than 40 considering the state online and idle and this number decreases when we consider only the online state (less than 10). This results tends to confirm the Dunbar's result about social circles [9] showing that the number of nodes which we interact with is very low if compared with the total number of alters.

5.2 Community Detection

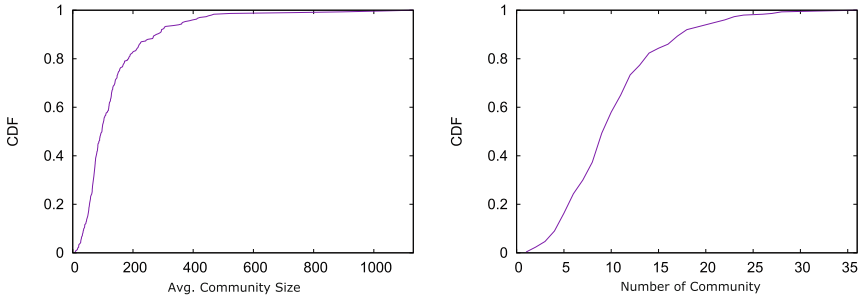
Our goal is to identify, in each ego network, a group of users which are densely connected among them. We have considered the ego network EN of each of the 337 registered users and we have executed DEMON on EN to compute the number of communities in EN. As expected, the dataset has revealed the typical structure of the social network, including groups of nodes which are strongly interconnected between them (high cluster coefficient).

Table 1. Analysis of the communities.

Ego network size	Avg. number of communities	Avg. community size
0–199	43	44.01
200–299	53	65.1
300–399	51	87.71
400–499	49	103.02
500–699	48	141.03
700–899	29	175.51
900–2999	28	342.75

Table 1 shows the characteristics of the detected communities with respect to the size of the corresponding ego network. We show the size of an ego network

and both the average number and the average size of its communities. We can observe a low number of communities with a high dimension in big ego networks, while a high set of communities with a low dimension, in small ego networks.



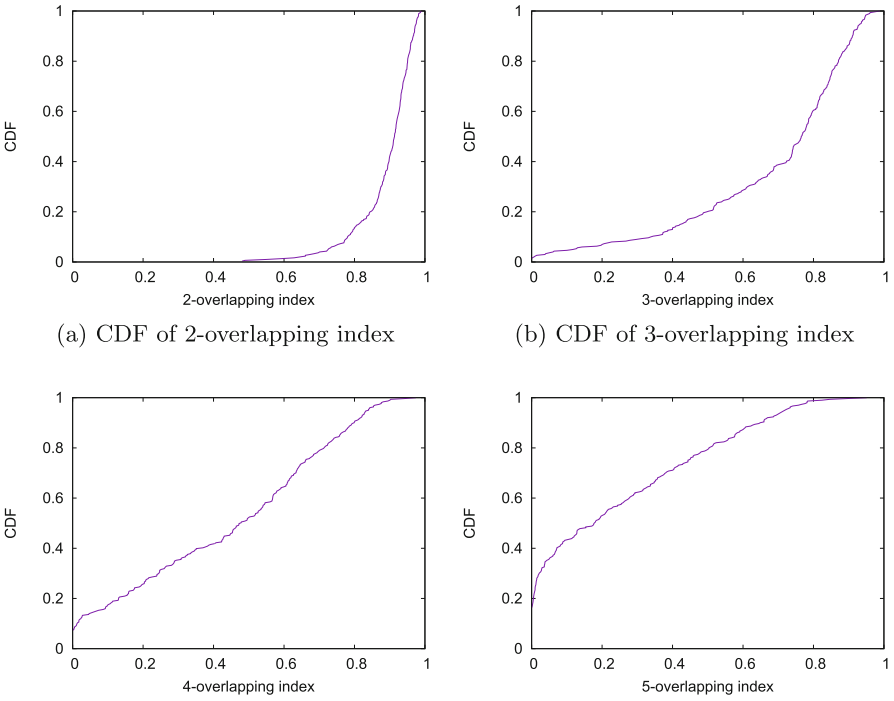
(a) CDF of the Average Community Size (b) CDF of the number of Community

Fig. 4. Community detection: an analysis

Figure 4 depicts the CDF of the community size and of the number of communities. About the 80% of communities has less than 250 nodes. The ego networks show a complex structure and about 80% of them has a number of communities less than 15 (Fig. 4(b)). This information is important when we consider the goal of our analysis. In fact, it permits us to estimate the number of replicas which can be allocated for the ego's data. Another interesting analysis evaluates the overlapping of communities. We exploit the *k-overlapping index* defined in Sect. 4 to evaluate, for each ego, the number of communities an alter belongs to. To choose the value k, we consider the average number of communities an ego node could have, as shown in Fig. 4(b), varying k from 2 to 5.

Figure 5 shows the communities overlapping evaluation. Communities present a considerable overlap, also when we consider increasing values of k. This means that ego networks have a set of nodes which belong to more than 2 communities, these nodes represent a bridge between different communities and are central in the ego network. In detail, when we consider the two bounds of k (Fig. 5(a) and (d)), we can clearly notice that low values of the KOI index for k equals to 5, but the increase of k is not proportional to the decrease of the index.

The properties of detected communities define for us a guideline for the definition of a strategy for allocating the users' data replica in the ego network. A suitable node to store replicas of the user's data may be chosen among the set of nodes which belong to more than one community. A user paired with one of these nodes has direct friendship connections with many other nodes of the ego network and may guarantee a good coverage of the ego network. On the other hand, a clever allocation has to take into account also the problem of load balancing, because nodes belonging to a set of communities could be affected by a huge load.



(a) CDF of 2-overlapping index

(b) CDF of 3-overlapping index

(c) CDF of 4-overlapping index

(d) CDF of 5-overlapping index

Fig. 5. Communit overlapping through the k-overlapping index

6 Conclusion and Future Works

This paper has analysed the temporal behaviour of users and the communities of a real OSN to understand how these issues can affect the data availability in DOSN. We have noticed that egos have not a similar behaviour by considering their ego network and only a subset of nodes are good candidate to store replicas of profiles, by confirming the results obtained in [8] The analysis has reported interesting results when the temporal affinity has been evaluated during the night. Furthermore, we have investigated the communities in the ego networks and shown that they are heterogeneous in term of structure and size. We found that ego networks are characterized by a low number of communities, which does not depend from the ego network size and that there is an high level of overlapping of these communities. These results can be used to support a proper data allocation strategy, for example by using the number of overlapping communities as parameter of the selection strategy. We plan to extend our work in several directions. First, we will extend our study to the investigation of dynamic communities, i.e. communities detected by considering each time slot separately. We plan also to define a good strategy to select the nodes to store data replica.

Acknowledgement. This work has been funded by the project *Big Data, Social Mining and Risk Management* (PRA_2016_15), University of Pisa.

References

1. Diaspora Website. https://diasporafoundation.org/
2. Bhagwan, R., Savage, S., Voelker, G.M.: Understanding availability. In: Kaashoek, M.F., Stoica, I. (eds.) IPTPS 2003. LNCS, vol. 2735, pp. 256–267. Springer, Heidelberg (2003). doi:10.1007/978-3-540-45172-3_24
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. Coscia, M., Rossetti, G., Giannotti, F., Pedreschi, D.: DEMON: a local-first discovery method for overlapping communities. In: Proceedings of 18th ACM SIGKDD, KDD 2012 (2012)
6. Cutillo, L.A., Molva, R., Strufe, T.: Safebook: a privacy preserving online social network leveraging on real-life trust. Commun. Mag. IEEE **47**, 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, Heidelberg (2010)
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. Dunbar, R.I.M.: The social brain hypothesis. Evol. Anthropol.: Issues, News Rev. **6**, 178–190 (1998)
10. Everett, M.G., Borgatti, S.P.: Ego network betweenness. Soc. Netw. **27**, 31–38 (2005)
11. 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)
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**, 1–18 (2015)
13. Kubiatowicz, J., Bindel, D., Chen, Y., Czerwinski, S., Eaton, P., Geels, D., Gummadi, R., Rhea, S., Weatherspoon, H., Weimer, W., et al.: Oceanstore: an architecture for global-scale persistent storage. ACM SIGPLAN Not. **35**(11), 190–201 (2000)
14. Narendula, R., Papaioannou, T.G., Aberer, K.: A decentralized online social network with efficient user-driven replication. In: 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT), and 2012 International Confernece on Social Computing (SocialCom), pp. 166–175. IEEE (2012)
15. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E **76**, 036106 (2007)
16. Sharma, R., Datta, A.: SuperNova: super-peers based architecture for decentralized online social networks. CoRR abs/1105.0074 (2011)
17. Xie, J., Kelley, S., Szymanski, B.K.: Overlapping community detection in networks: the state-of-the-art and comparative study. ACM Comput. Surv. **45**(4), 43 (2013)