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

Evaluation of structural and temporal properties of ego networks for data availability in DOSNs

Andrea De Salve · Barbara Guidi · Laura Ricci

Received: date / Accepted: date

Abstract The large diffusion of Online Social Networks (OSNs) has influenced the way people interact with each other. OSNs present several drawbacks, one of the most important is the problem of privacy disclosures. Distributed Online Social Networks (DOSNs) have been proposed as a valid alternative solution to solve this problem. DOSNs are Online Social Networks implemented on a distributed platform, such as a P2P system or a mobile network. However, the decentralization of the control presents several challenges, one of the main ones is guaranteeing data availability without relying on a central server. To this aim, users' data allocation strategies have to be defined and this requires the knowledge of both structural and temporal characteristics of ego networks which is a difficult task due to the lack of real datasets limiting 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 the structure of communities and their evolution over the time by using a real Facebook dataset.

Keywords P2P, social networks, DOSN, data availability, temporal affinity, community detection

1 Introduction

Online Social Networks (OSNs) have hardly increased their popularity, becoming today the most used web platforms. During the years, they have created a new way of interaction/communication among people and they connect millions users that produce a significant amount of social content. The most part of current OSNs, such as Facebook, Twitter, or Google+, are based on centralized architectures which have intrinsic drawbacks, including scalability and privacy [6]. Privacy is one of the most important issues regarding the usage of centralized OSNs because contents generated by users are stored and kept available by the centralized service provides. As a result, centralized OSNs have become one of 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) [6] is an online social network implemented on a distributed platform, such as a network of trusted servers, P2P systems or a mobile network. The P2P paradigm has been largely used to implement DOSNs platforms where users' devices share the tasks needed to provide the OSN service. As a result, decentralization of the OSN allows to limit the privacy problem with respect the centralized service providers, but it introduces several new challenges such as data availability, information diffusion, or load balancing.

Besides classical P2P architectures, also the extensive use of smartphones and tablets has affected the usage of OSNs since, nowadays, most part of users connect to the OSNs by using their mobile devices. The possibility of using a wide variety of wireless radio access technologies (Bluetooth and WiFi) allows users' devices to be directly interconnected with each other via wireless links. The resulting networks established between mobile nodes (also referred as Mobile Ad hoc Network or MANET) allow their users to explicitly connect to the devices of other users, in order to share information and contents. The features of these networks make them very dynamic over time and nodes participating in the

Department of Computer Science, University of Pisa
E-mail: desalve@unipi.it, guidi@unipi.it, laura.ricci@unipi.it

system have to provide a data availability mechanism facilitating the seamless delivery of the contents.

Indeed, one of the major problems in a DOSN is to guarantee the availability of data, because of the high users' churn, i.e. the user may autonomously disconnect from the DOSN and data stored on it are no more available and accessible. As a result, the behaviour of participating users dramatically affects the data availability and it can lead to data becoming unavailable or lost. Replication is the most popular approach to manage the availability problem and it consists of replicating the user's contents on different nodes of the DOSN. Replication improves data availability, but, at the same time, it introduces both the problem of maintaining the consistency of the replicas and to define the proper number of replicas.

For these reasons, the study of good strategies for selecting nodes where replica of the contents have to be stored has recently gained momentum. The goal of these politics is to dynamically allocate data, in order to maximize the availability of the social data.

Existing DOSNs ensure data availability by considering different approaches. One solution is to encrypt social data and to store it on a Distributed Hash Table (DHT) [20] (such as [13, 1]). Another solution exploits social relations to store data on the devices of friend nodes. Current approaches consider all friends, a set of trusted friends, or random subset of the friends [14, 9, 5, 21] as possible nodes where store data. In this case, information on both the time spent online by users and the ability of the users to serve data are also considered during replicas selection.

Besides these, the study of data availability problem can benefit from information about interconnections between users of the DOSNs, i.e. communities resulting from social relations. Indeed, a very common trait of OSNs is to expose community's structures where groups of users are highly connected intra-community and weakly connected inter-community. These communities may reflect the friendship relations between users, the different environments where people met (such as workplace, school, hometown, etc.) or the different features that aggregate them (such as interests, music, religion, etc.). Having a better understanding of such community structure occurring in real OSNs can help scientists in managing the data availability problem in a distributed environment such as a DOSN. In this paper we are mainly interested in the community structures defined exclusively by the topological structure of the ego network, i.e. the friendship relations between the alters of the ego. Furthermore, it is interesting to analyse both the static structures of the communities and how they evolve when users connect/disconnect from the social network. From the perspective of data availability problem, community analysis remains unexplored in the context of DOSNs.

The main goal of this paper is to present a twofold study by using a real dataset of Facebook. Firstly, we extend the analysis proposed in our previous work [7] regarding the temporal behaviour of users by focusing on the ego networks of individual users. In particular, we evaluate the extent to which the friends of a user U have the same temporal behaviour of U, i.e. they are connected to the OSN in the same interval of time. Understanding how users navigate and interact when they connect to social networking sites creates opportunities for improvements to manage the data availability problem.

As a second step, we extend the study performed in [8] novel study concerning communities belonging to the ego networks of the users. First of all, we study static communities, i.e. communities which can be detected by considering the topological structure of an ego network. In addition, we investigate a dynamic scenario where we detect the communities arising when we consider only the users which are online at a certain time step. Next, we perform a detailed analysis aimed at studying the properties and temporal evolution of communities in each ego network, giving particular focus to a comparison of the dynamic scenario with respect to the static one. More in detail, we evaluate how community structure in ego networks evolves when the availability status of the users (online/offline) is considered. The analysis of the dynamic scenario confirms the presence of a tightly connected group of nodes in each ego network, reflecting different communities. In addition, dynamic communities become more sparse and poorly connected compared to the static scenario.

At the best of our knowledge, this is the first contribution of this kind in the area of DOSN definition and, more in general, this work represents the first contribution about the study of both temporal ego networks and temporal communities detected by considering the online behaviour of users. Furthermore, the analysis detects the users belonging to more than one community that may act as "hubs", able to serve accesses to the replica they store for many users of the ego network. We consider the social communities defined by the social relationships. Thanks to communities, we can be able to guarantee a high level of data availability by replicating social data on nodes belonging these communities.

The rest of the paper is organized as follow. Section 2 describes the related work. Section 3 describes the scenario we focus on and introduces the problem of data availability in DOSNs. Section 4 describes our twofold work on temporal affinity and community detection. Section 6 investigates the results of our analysis, which are obtained by exploiting a real OSN data set derived from Facebook. Finally, Section 7 draws the main conclusions.

2 Related Work

Since our study is targeted to DOSNs, we introduce the reader to the main features of the current existing DOSNs and, then, to the current techniques proposed to guarantee data availability in DOSNs.

2.1 Distributed Online Social Networks

A Distributed Online Social Network (DOSN) [6] is an Online Social Network implemented on a distributed information management platform. Current proposed approaches for implementing DOSNs can be divided into two major categories: federation of servers and OSN over P2P networks.

OSNs built on top of a federation of servers require that social network providers agree upon standards of operations in a collective fashion. An example of these kind of DOSNs is Diaspora [9].

Instead, DOSNs built on top a P2P network represent the most studied category. Authors in [6] propose a reference architecture, which consists of six layers where the lower layer is the physical communication network (the Internet or another physical communication network) and the distributed overlay which manages all the functionalities is a P2P network. Specifically, this layer provides services for looking up resources, routing, and retrieving information reliably and effectively among nodes in the overlay.

A common choice for implementing a DOSN on a P2P architecture, is to use a DHT as a support for the DOSN. LifeSocial [13] is a P2P OSN focused on the privacy issue, where user information is stored on a DHT and it presents an approach where all the functionalities of the social network are realized by plug-ins.

PeerSon [1] 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) where all users' content is stored encrypted, and the second tier consists of the nodes representing users. The idea is to use the DHT to find the necessary information for users. This approach comes without a replication scheme and stores offline messages on the DHT (OpenDHT in the prototype implementation).

Safebook [5] addresses privacy in OSNs by using a three-tier architecture where data are stored in a particular social overlay named "Matryoshka". Matryoshkas are concentric rings of peers around each user's peer that provide trusted data storage and obscure communication through indirection.

2.2 Data Availability in DOSNs

In the context of DOSNs, we consider the availability of the user's data, i.e. the content, that can be seen as the digital representation of a user, which is stored on a computing device and can be transmitted from one device to another. One of the main challenges in decentralization comes from guaranteeing availability of the data when the owner of the data is not online. When we talk about data availability, we refer to an important challenge in distributed systems. Without a central point of storage, data are distributed among all nodes in the network and, when a node leaves the network, its data should remain available in the network.

The main technique used to resolve the problem of the content availability in DOSNs is replication, which is a well-known technique in the area of distributed systems based on the storage of the same data on different storage devices.

In DOSNs, probably due to the lack of data from real social networks, only a few proposals have addressed the issue of data availability.

SuperNova [21] is an architecture for a DOSN that solves the availability issue by relying on super-peers that provide highly available storage. Other specific solutions are presented in [16], where authors propose a replication strategy based on storing the replicas of users profiles only on a set of trusted proxies.

A current trend to manage data availability is the use of social relationships to store social data. DiDuSoNet [14] is a P2P DOSN focused on the data availability problem which uses, as SafeBook, a particular overlay based on trusted connections to store data. The approach proposed by Koll et al. [15] is SOUP, a scalable, robust and secure system for storing data among heterogeneous nodes with very high availability, which considers the significantly differing characteristics of OSN users, such as device and the online time patterns.

3 Scenario and Motivations

We consider a scenario where a P2P network is used to support the DOSN and the overlay network is defined by considering the social relations existing between the users. In particular, the overlay network is modelled as an *ego network*, a well-known social network model explained in detail in [14]. This model is well suited for our purpose and it represents the local view of the social graph relative to a user, where only its friends and their interconnections are considered. To be more detailed, the ego network [11] is the network constituted by a user (*the ego*), its direct friends (*the alters*) and the social ties occurring between them.

In our previous works [3], we have studied replication strategies based on both friendships and temporal information. Taking into account these information, we are able to maximize the data availability by choosing good replica nodes.

In this work, we are interested in studying the temporal affinity between Facebook users and their friends by considering the ego network structure. Furthermore, we study how communities of nodes evolve if we consider the presence status (online/offline) of users. Indeed, the presence of densely connected groups of nodes can be exploited to increase the level of data availability. For these reasons, we first study the static ego network to evaluate some characteristics of the individuated communities. Next, we assess the characteristics of these communities over time by using the presence status of the users.

4 Ego network analysis: temporal affinities and static communities

In this section, we introduce our studies concerning the users' temporal affinity and the detection of static communities.

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 the alters in its ego network with respect to that of E.

The availability of each user is represented by an availability vector of fixed size. 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 minutes and it contains the value 1 if the user is online in that time slot, 0 otherwise. Since our goal is to understand how the temporal behaviour of the users can affect the data availability in DOSNs and considering that there are particular day periods in which users tends to be offline (e.g. during the night), we propose two different metrics for the Temporal Affinity:

- *Daily Temporal Affinity (DTA)*, considers the temporal affinity during all the day. In detail, we exploit all the 2001 temporal slots;
- *Nighttime Temporal Affinity (NTA)*, considers a subset of slots, i.e. the nighttime slots.

We evaluate the temporal affinity using the cosine similarity metrics [2]. Formally, let A and B the availability vector of two users, the cosine similarity is computed as shown in (1):

$$\text{CosineSimilarity}(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||} \tag{1}$$

The resulting similarity ranges from 1 meaning perfect correlation, to 0, usually indicating no correlation between the ego and the alter.

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* as 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. Often complex networks, such as OSNs, are characterized by densely connected groups of nodes which are sparely connected to other nodes, named communities. 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 wished [23].

Numerous techniques have been developed for both efficient and effective community detection. One of the most complete survey is given by Fortunato [12], which describes all the most popular techniques for community detection. The Label Propagation algorithm (LP) [18] is considered one of the most popular technique to discover communities. For this reason, we decide to use the Label Propagation implemented by DEMON [4]. Label Propagation, according to the original work, detects communities by spreading labels through the edges of a graph and by labeling a node with the label paired with most of its neighbours. DEMON implements the classical label propagation algorithm, as explained by [18], but, unlike the original algorithm, DEMON is able to identify possible overlaps between communities. Furthermore, DEMON is suitable for a distributed implementation and it can be easily adapted to any kind of network, dense or sparse.

To evaluate the overlapping between communities, we define a new index, the *k-overlapping index (KOI)*. Taking into account the definition of an ego network as given in Section 3, the index is used to understand how many alters of a generic ego e belong to at least k communities. In detail, the index is defined as follows: considers C_S, the set of communities detected in a generic ego network EN, for each alter node $a \in EN$, we compute the number of communities in C_S a belongs to. We detect alters that appear in at least k communities and compare this value with the total number of nodes in the ego network. The formal definition of the KOI is given by (2).

$$KOI(EN, k) = \frac{|V^k|}{|V|} \tag{2}$$

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

On the basis of the KOI index, we can detect which alters belongs 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, especially if we consider mobile nodes, allocation can cause a huge amount of load on a single node characterized by a high KOI index because it could be chosen by a large set of friends to replicate data.

5 Dynamic evaluation of ego networks and communities

To ensure availability of data, user's contents must be replicated on different nodes (among those online), according to the friendships in the ego networks. However, friendship relations give few information about the quality of the replica nodes. Instead, community information provides a complete view of the structural properties of the ego networks and can be used to place a replica of the contents on the nodes located in different communities.

Our consideration about community detection, presented in the previous section, can be applied, in part or as a whole, to a dynamic scenario. An evaluation of communities in a static scenario can be a very good starting point, but it is not enough when we consider the behaviour of the users in a DOSN. Of course, the temporal affinity between users just has to consider the presence status of the users, which evolves during the time. Instead, an important characteristic that has to be considered in a distributed environment is the time evolution of the communities in term of their composition. For this reason, we have evaluated how the dynamic structure of the communities can affect data availability, by assessing how they change over time in term of number and size.

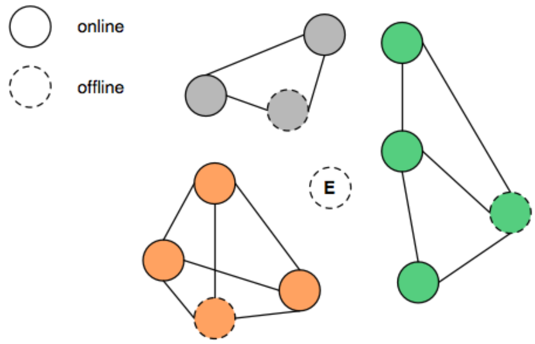


Fig. 1 Snapshot of the communities structure and availability status (online/offline) of the ego network of user E.

As explained in [17], the structure of the most part complex networks is dynamic in nature because of the addition and/or the removal of links and nodes over time. This has an impact on the real composition of communities.

Consider a dynamic scenario as a distributed system where the participating users are represented by nodes and connections between users are represented by links. The connections can be more or less stable depending on the network infrastructure used by their devices (wireless network, etc..). In this scenario the dynamic nature of the network is really hard to manage because the presence status of a user drastically changes and connections/disconnections within a small time interval could be very huge. In such a case, it is important to detect online nodes to retrieve online communities which can facilitate data availability and information diffusion. Indeed, dynamic, densely connected communities of nodes can be used to provide contents of one user to all the members of the community.

Consider the example of an ego network of a node and the communities detected in it, as shown in Figure 1, where a different colour is paired with each different community detected by the algorithm. The communities are computed by considering the topology of the ego network, i.e. nodes belong to different communities according to their social relationships with the friends of the ego. Now suppose that the ego voluntary decides to leave the network at the time t. To choose the nodes where the data of the ego have to be mapped, it is necessary to consider the snapshot (at time t) of all online nodes and dynamically compute the communities on these nodes. As shown by Figure 1, community structure may change frequently over time due to the disconnection/connection of users. Static communities can be considered as an optimal scenario which occurs only when all the nodes of the ego networks are online. In contrast, communities arising when availability status of users is considered may appear very different, because only part of their users are available.

For this reason, the evolution of the presence status of the nodes of the ego network during the offline interval of the ego node must be monitored. We study the dynamic evolution of ego networks by considering the evolution of the presence status of the nodes during the time period when an ego is offline.

6 Evaluation

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

We have used a real dataset gathered by a Facebook application, named *SocialCircles!*, deeply described in [7]. 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 minutes for 10 consecutive days (from Tuesday 3 June to Friday 13 June 2014). 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 sequence of consecutive time intervals (each of 8 minutes), that we refer as *time slots* in the follow.

6.1 Temporal Affinity

The first analysis concerns the evaluation of the Daily Temporal Affinity, referenced as DTA. As explained in [7], a user can be online, offline or in an idle state.

We decide to consider the online state, and to do not distinguish the idle and offline state into the presence array so that both of them are represented by the value 0.

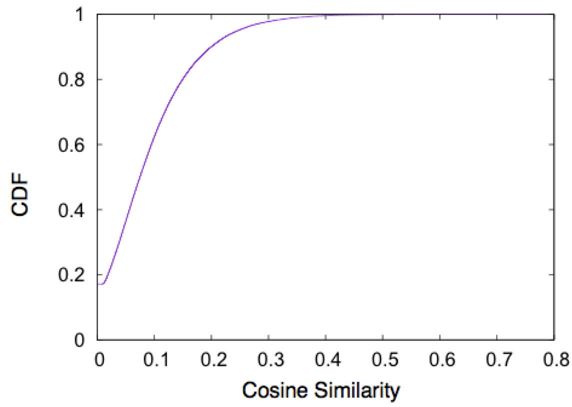


Fig. 2 Daily temporal affinity

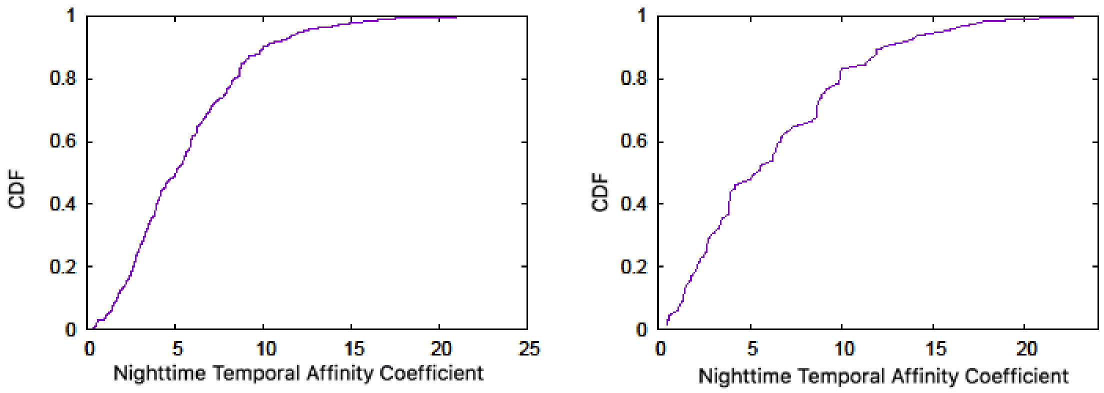
About the 80% of the couple ego-alter has a low similarity, as depicted in Figure 2. This low values can be influenced from the online behaviour of each ego. A little set of couples (less than 5%) show medium/high values of similarity. These results confirm our previous analysis on Dunbar-based ego networks [7] where users present a high similarity only with trusted nodes. In fact, extending the temporal affinity analysis to the whole ego networks, the result is similar. For the sake of clarity, Dunbar [10, 19, 22] shows that the number of friends a user can maintain stable social connections with is approximately 150 (the *Dunbar's Number*). The stability of a connection can be defined as a function of the tie strength of the relation and it quantifies the strength of the relationship between two users.

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 Section 4. For the evaluation, we consider both the online and the idle state.

Some periods of the day are critical to guarantee data availability in distributed environments, such as the night period. When an ego has a high NTA with a set of distinct alters, it means that the 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. The value k indicates the number of nights needed to consider an ego as a nighttime ego.

Figure 3 shows the CDF of the NTA of users who appear to be online for at least $k = 1$ night and $k = 4$ consecutive nights. About 90% of the nighttime users have at most 10 alters who behave in the same way, i.e. alters who appear to be online during the nigh times. In general, nighttime users and alters does not changes



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

Fig. 3 NTA with fixed indexes, by varying the number of nights k

their behaviour over time, distributions of the NTA are very similar, regardless of whether the users are online for $k = 1$ or $k = 4$ consecutive nights.

We focused on variable index that consider extended periods of the day and computed the cosine similarity between the availability vector of ego and alters who appear to be online during the night. Figure 4 show the CDF on the cosine similarity as well as the percentage of alters in each ego network. We observe that only few alters have availability pattern of the nighttime users. Analysing the NTA, we notice that less than 10% of egos can be considered nighttime. This value is drastically reduced when the constraint of the number of nights needed to be nighttime increases. 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 result tends to confirm the Dunbar's circles [10] conjecture, showing that the number of nodes which we interact with is very low if compared with the total number of alters.

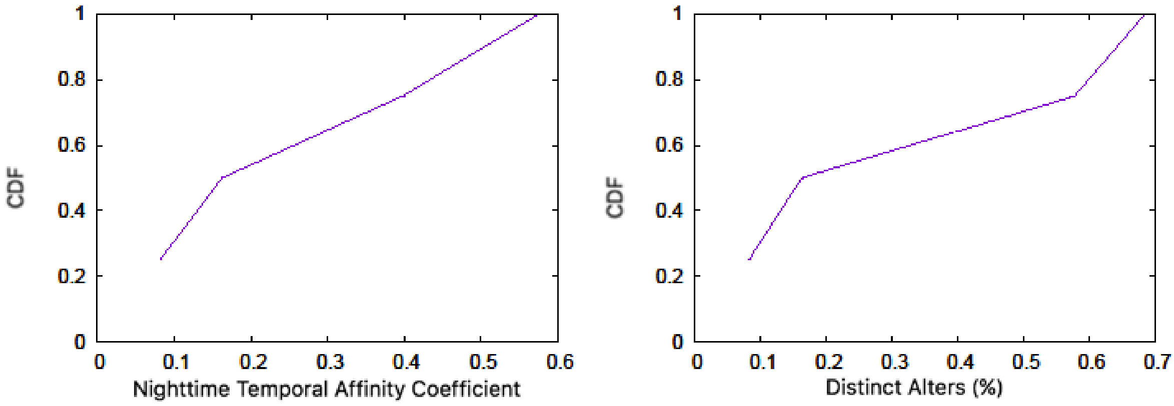
6.2 Detection of Static Communities

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).

Figure 5 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 (Figure 5(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 Section 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. 5(b), varying k from 2 to 5.

Figure 6 shows the evaluation of the overlapping index in communities. 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 (Figure 6(a) and Figure 6(d)), we can clearly notice low values of the KOI index for k equal to 5, but we note that the increase of k is not proportional to the decrease of the index.

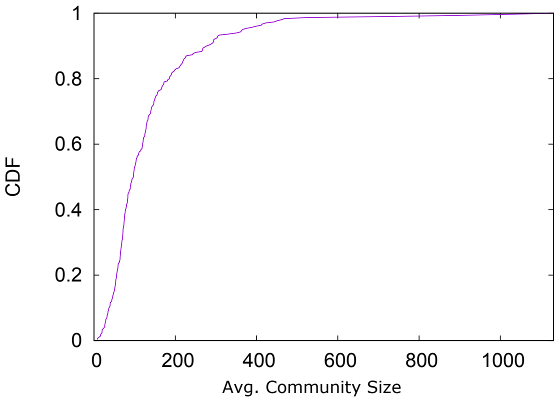
The properties we have detected for the communities highlight a guideline for the definition of a strategy for allocating the users' data replica in the ego network. A suitable node may be chosen among the set of nodes which belong to more than one community. A user paired with one of these nodes has direct friendship connections with many other nodes in 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.



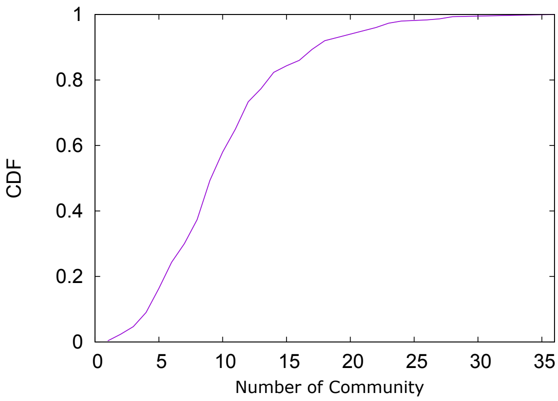
(a) NTA with k = 2

(b) Distinct alters with k = 2

Fig. 4 Variable indexes with k equals to 2



(a) Average size of communities



(b) Number of communities

Fig. 5 Community Detection: an analysis

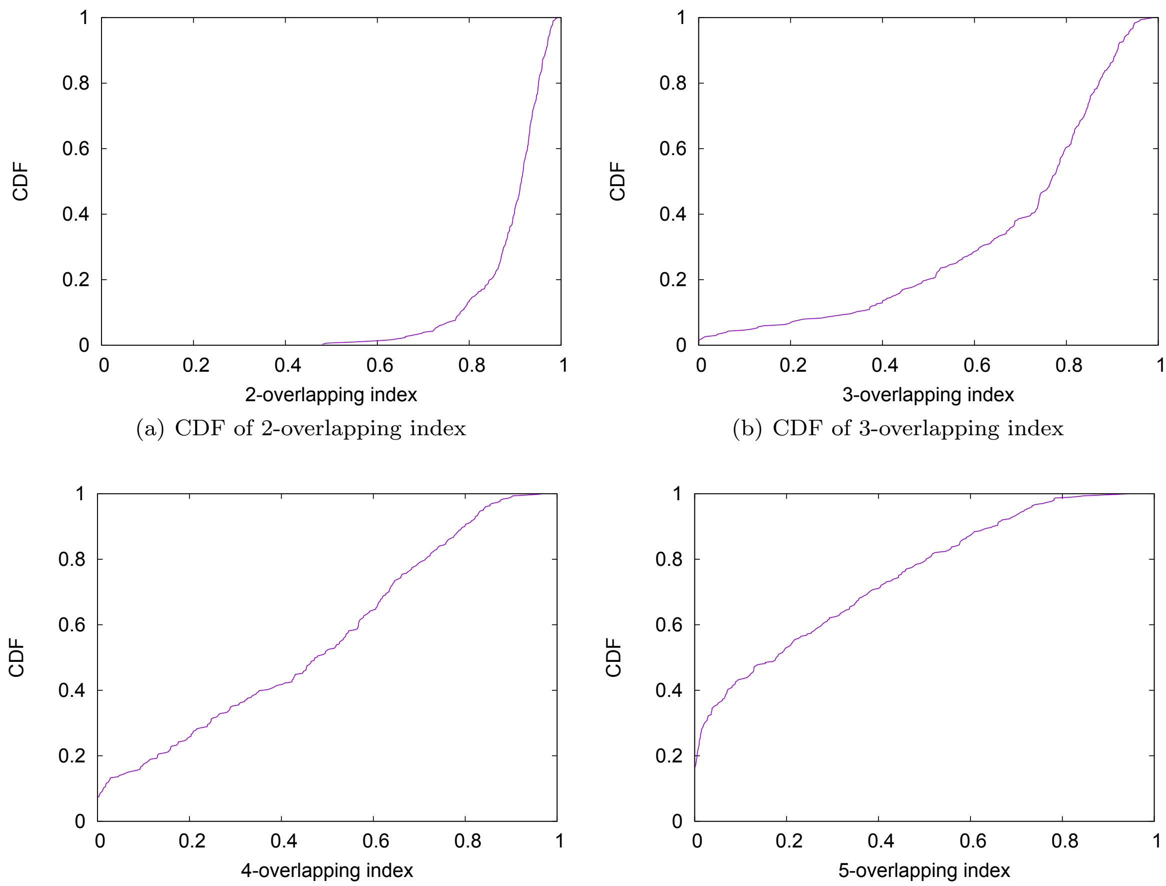
6.3 Dynamic evolution of communities

We have used the knowledge about the presence status of each ego node to build 2001 snapshots, each one related to a time slot. These snapshots are used to study the properties of the ego networks during the time and the evolution of the community structure.

Table 1 gives some details concerning the comparison between dynamic ego networks and the corresponding static view of the same ego networks. Results suggest that the size in term of both nodes and links of the dynamic snapshot of the ego network is about 10% of the static view. In general, about 8% of all nodes in the ego network are online simultaneously, which accounts for about 6% of the total friend relationships. This is motivated by considering the huge dynamism of ego networks. An important result is highlighted by the values of the degree centrality which reveals the presence of central nodes. The values are higher than those of the static ego networks because of the number of nodes are lower and nodes seems to be more central. An interesting property is disclosed by the degree centrality. In fact, the average value of the degree centrality is 0.4, which means that there are nodes that are central and can be considered as hubs inside the ego network. Finally, the average clustering coefficient od dynamic networks is 0.333.

Figures 7(a) and 7(b) show in more details the evolution of, respectively, the number of nodes and of edges during each time slot. The plots clearly indicate the presence of a temporal pattern where most of the users are connected during daylight hours. In Figure 7(c), we evaluate the *local degree* that represents the average number of common neighbours between alters and the ego. Ego networks appear to be more clustered, online alters are connected on average to at least 2 commons online alters.

Let us now introduce the *Transition Interval* as the time interval when a generic ego node e decides to go



(a) CDF of 2-overlapping index

(b) CDF of 3-overlapping index

(c) CDF of 4-overlapping index

(d) CDF of 5-overlapping index

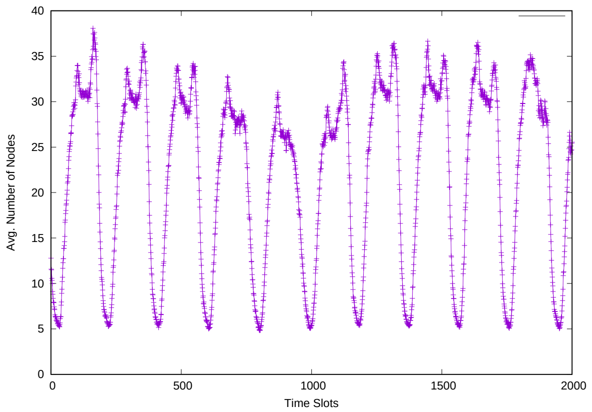
Fig. 6 Community Overlapping through the k-overlapping index

	Dynamic				Static			
	Max	Mix	Mean	StdDev	Max	Min	Mean	StdDev
Num. Nodes	379.0	1.0	38.076	31.833	2965	23	486.89	361.60
Num. Edges	14513.0	0.0	663.007	1213.0118	256968	62	10930.14	22218.14
Degree Centrality	0.9166667	0.0	0.400	0.365148	0.88	0.0003427	0.0656477	0.07920130
Clustering Coeff.	1.0	0.0	0.333	0.5773	0.88	0.42	0.63	0.078

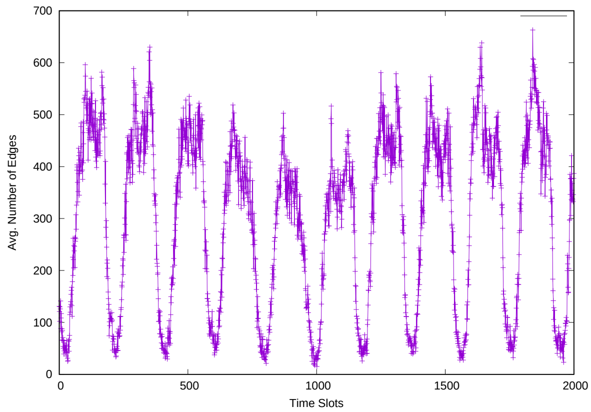
Table 1 Comparison between static and dynamic properties of ego networks.

offline and remains offline. For the sake of clarity, this is the interval starting from the last time slot when the ego node e is online to the last time slot when the ego node results offline. For example, if we consider a small presence array of 10 time slots 1110000111, the *Transition Interval* is the sequence 10000. In the following analysis, we consider the global view of the ego networks and what happens in the ego networks by focusing only on all the Transition Intervals of the ego.

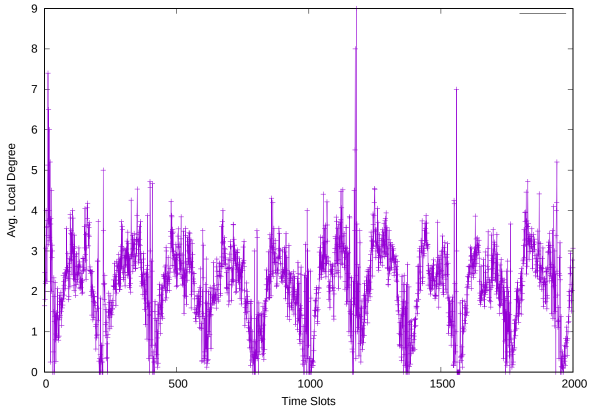
Figure 8 shows the main features of the detected communities. Figures 8(a) and 8(b) show, respectively, the number of communities and their size over time, without considering the presence status of the ego. The results clearly indicate that that the availability status of the ego (online or offline) does not strongly affect the community structure of his/her ego network. In particular, ego networks expose community structure even when the presence status of users is considered. In general, the average number of communities is quite big (400 communities distributed among the different ego networks) while their size ranges between 10 and 20 users. In particular, the number of communities arising when users is offline is quite big (on average 250 communities) and they are composed by a number of online nodes that does not exceed 20 users (see Figure 8(c) and 8(d)). In some cases (generally during the evening and in the early afternoon) communities present higher number of online members.



(a) Avg. Number of nodes over time slots



(b) Avg. Number of edges over time slots



(c) Avg. local degree evaluation

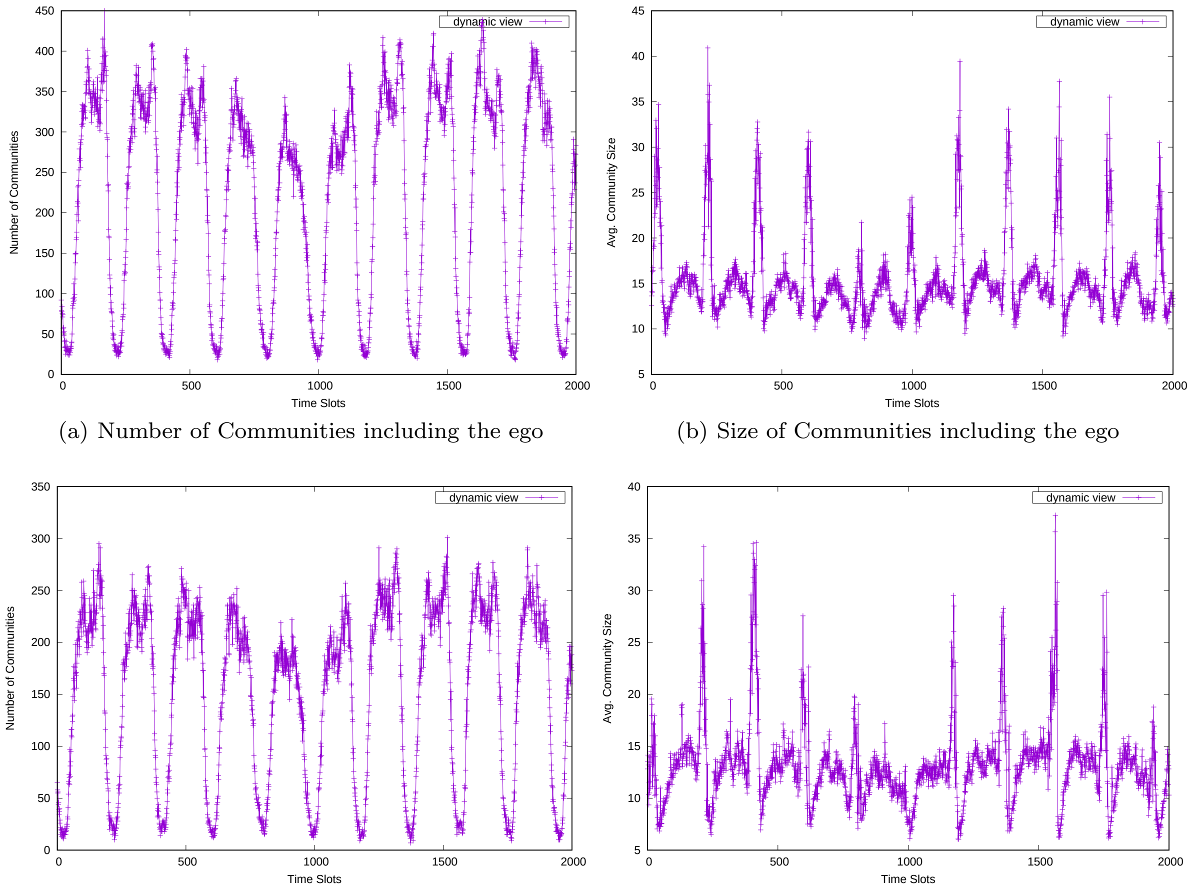
Fig. 7 Evaluation of the dynamic ego networks

	Dynamic		Static	
EN Size	Avg. Num. Comm.	Avg. Size	Avg. Num. Comm.	Avg. Size
0∼199	7	6.292697247706	4.4	44.01
200∼299	12	6.325449473	7.8	65.1
300∼399	20.7	8.0439574713	9.4	87.71
400∼499	31.13	8.92754612	11.4	103.02
500∼699	38.34	12.103263638	11.4	141.03
700∼899	87.40	13.8701502	15	175.51
900∼2999	132.76	27.80695214	16.4	342.75

Table 2 Average number and size of communities by comparing both static and dynamic ego networks

Table 2 shows some statistics of the dynamic communities as a function of the size of their ego network and compare them with the corresponding static ego network. In the static case, results suggest the presence of a significant group of communities in each ego network. We can observe that big ego networks (with a number of friends greater than 500) have a low number of communities with high size, while small ego networks (with a number of alters between 10 and 500) have a higher number of communities with a low dimension. The size of the communities is correlated to the number of alters and the nodes of each communities are highly interconnected to each other.

The scenario is completely different when we consider the dynamic case and many important characteristics are visible by comparing the two scenarios. In the dynamic case, the larger is the ego network size, the larger is the average number of communities. This means that large ego networks tend to be more fragmented if we consider the time evolution. This depends also by the fact that big ego networks tend to have big communities as shown in the static scenario. Due to the users' churn, these communities tend to become frag-



(a) Number of Communities including the ego

(b) Size of Communities including the ego

(c) Number of Communities by considering Transition Intervals

(d) Size of Communities by considering Transition Intervals

Fig. 8 Evaluation of the number and size of communities discovered in each ego networks: regardless of the online status of ego (Fig. 8(a) and 8(b)) and when ego is offline (Fig. 8(c) and 8(d)).

mented over the time slots. Another important characteristic is highlighted by the average size of the communities which increases as a function of the ego network size. However, the gap is reduced and the value of the difference between small ego networks and big ones is about 4.

7 Conclusion and Future Works

This paper presents an analysis of the temporal behaviour of users and of the communities of a real OSN to understand how these issues can affect the data availability in DOSN.

We noticed that not all egos have a similar behaviour and that only a subset of nodes in their ego networks are good candidates to store replicas of profiles, by confirming the results obtained in [7]. The analysis reported interesting results when the temporal affinity is evaluated during the night. Furthermore, we investigated the static 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 on the ego network size and that there is a high level of overlapping of these communities.

The analysis was extended with the evaluation of the evolution of community structures in ego networks. In particular, we investigated the behaviour of the communities over time, when the ego is offline. We detected that ego networks expose community structures, even if online/offline status of users is considered. More in detail, our work showed there is an evolution of the dynamic community structure in ego networks, which became smaller and poorly connected, with a low clustering coefficient compared to the static communities. However, analysis of dynamic ego networks confirmed the presence of tightly connected groups of nodes, reflecting different aspects of the ego life.

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, from a more systematic study of the communities for a better characterization of their features, also considering a larger dataset, to the definition of a good strategy to select the nodes to store data replica.

Acknowledgement

The authors gratefully acknowledge Andrea Michienzi for his useful support.

References

1. Buchegger S, Schiberg D, Vu LH, Datta A (2009) PeerSoN: P2P social networking - early experiences and insights. In: SNS, ACM, pp 46–52
2. Choi SS, Cha SH, Tappert CC (2010) A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics 8(1):43–48
3. 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
4. Coscia M, Rossetti G, Giannotti F, Pedreschi D (2012) DEMON: A local-first discovery method for overlapping communities. In: Proceedings of the 18th ACM SIGKDD, KDD '12
5. Cutillo LA, Molva R, Strufe T (2009) Safebook: a Privacy Preserving Online Social Network Leveraging on Real-Life Trust. Communication Magazine, IEEE 47
6. 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
7. 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
8. De Salve A, Guidi B, Ricci L (2016) An analysis of ego network communities and temporal affinity for online social networks. In: Proceedings of the 2nd EAI International Conference on Smart Objects and Technologies for Social Good
9. Diaspora (2010) Website. `https://diasporafoundation.org/`
10. Dunbar RIM (1998) The Social Brain Hypothesis. Evolutionary Anthropology: Issues, News, and Reviews 6
11. Everett MG, Borgatti SP (2005) Ego network betweenness. Social Networks 27:31–38
12. Fortunato S (2010) Community detection in graphs. Physics reports 486(3):75–174
13. Graffi K, Gross C, Stingl D, Hartung D, Kovacevic A, Steinmetz R (2011) LifeSocial.KOM: a Secure and P2P-Based Solution for Online Social Networks. In: IEEE CCNC
14. 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
15. Koll D, Li J, Fu X (2013) With a little help from my friends: replica placement in decentralized online social networks. Technique Report: TR-IFI-TB-2013-01, University of Goettingen, Germany
16. Narendula R, Papaioannou TG, Aberer K (2012) A decentralized online social network with efficient user-driven replication. In: Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom), IEEE, pp 166–175
17. Oliveira M, Guerreiro A, Gama J (2014) Dynamic communities in evolving customer networks: an analysis using landmark and sliding windows. Social Network Analysis and Mining 4(1):1–19
18. Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76
19. Roberts SGB, Dunbar RIM, Pollet TV, Kuppens T (2009) Exploring Variation in Active Network Size: Constraints and Ego Characteristics. Social Networks 31
20. Rowstron AIT, Druschel P (2001) Pastry: Scalable, Decentralized Object Location, and Routing for Large-Scale Peer-to-Peer Systems. In: Proceedings of the International Conference on Distributed Systems Platforms
21. Sharma R, Datta A (2011) SuperNova: Super-Peers Based Architecture for Decentralized Online Social Networks. CoRR abs/1105.0074
22. Sutcliffe A, Dunbar R, Binder J, Arrow H (2012) Relationships and the Social Brain: Integrating Psychological and Evolutionary Perspectives. British Journal of Psychology 103
23. 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)