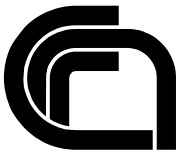

Consiglio Nazionale delle Ricerche

Analysis of Ego Network Structure in Online Social Networks

V. Arnaboldi, M. Conti, A. Passarella, F. Pezzoni

IIT TR-10/2012

Technical report

Luglio 2012

Istituto di Informatica e Telematica

Analysis of Ego Network Structure in Online Social Networks

[Technical Report]

Valerio Arnaboldi, Marco Conti, Andrea Passarella and Fabio Pezzoni
Institute of Informatics and Telematics of CNR
via G. Moruzzi, 1 - 56124 Pisa, Italy
Email: {v.arnaboldi, m.conti, a.passarella, f.pezzoni}@iit.cnr.it

***Abstract*—Results about offline social networks demonstrated that the social relationships that an individual (*ego*) maintains with other people (*alters*) can be organised into different groups according to the *ego network model*. In this model the ego can be seen as the centre of a series of layers of increasing size. Social relationships between ego and alters in layers close to ego are stronger than those belonging to more external layers. Online Social Networks are becoming a fundamental medium for humans to manage their social life, however the structure of ego networks in these virtual environments has not been investigated yet. In this work we contribute to fill this gap by analysing a large data set of Facebook relationships. We filter the data to obtain the frequency of contact of the relationships, and we check - by using different clustering techniques - whether structures similar to those found in offline social networks can be observed. The results show a strikingly similarity between the social structures in offline and Online Social Networks. In particular, the social relationships in Facebook share three of the most important features highlighted in offline ego networks: (i) they appear to be organised in four hierarchical layers; (ii) the sizes of the layers follow a scaling factor near to three; and (iii) the number of active social relationships is close to the well-known *Dunbar's number*. These results strongly suggest that, even if the ways to communicate and to maintain social relationships are changing due to the diffusion of Online Social Networks, the way people organise their social relationships seems to remain unaltered.**

I. Introduction

We are seeing a very significant process of integration between the physical world of the users of ICT technologies, and the cyber (virtual) world formed by the broad range of Internet applications. This is particularly evident in the area of social networks. Online Social Networks (OSNs) and offline social networks - which represent the social networks formed by the users due to personal interactions in the physical world - definitely influence each other. People become friends in OSNs with individuals they also know "in the real life", while OSNs can be a means of reinforcing and maintaining social relationships existing in the physical world. Facebook, Twitter and many other OSNs have introduced a set of new communication mechanisms that are becoming part of the way in which we interact socially.

Although several aspects are still under investigation, key properties regarding offline social networks have been investigated quite extensively (e.g., the difference between strong and weak ties and the importance of the latter [1], the structural properties of the network [2], just to mention a few examples). On the other hand, the analysis of the properties of OSNs is much less advanced. The interplay between social interactions in the two types of networks is only partially understood and still under investigation [3], [4]. Moreover, the structural properties of OSNs, and their differences and similarities with offline social networks are not yet fully understood.

In this paper we focus on the latter aspect, providing a characterisation of structural properties of Facebook networks, and comparing them with well known results available in the anthropology literature about offline social networks' structure [2]. Our results provide two major contributions. On the one hand, we contribute to better characterise OSNs per se. On the other hand, we compare equivalent properties on OSNs and offline social networks, thus contributing to better understanding similarities and differences of social structures in the cyber and physical worlds. Assessing such similarities can be very useful also to exploit OSNs to better understand some offline social networks' properties. For example, collecting data regarding offline social networks is a rather difficult task, which involves lengthy processes to distribute, compile and collect questionnaires. Studying similar properties on OSNs would clearly be much simpler and quicker.

In this work, we focus on characterising the properties of *ego networks* in OSNs. While a lot of work has been done to describe the global structure of OSNs [5]–[7], the study of ego networks in virtual environments has received little attention so far. Ego networks are social networks made up of an individual (called *ego*) along with all the social ties she has with other people (called *alters*). Ego networks are an important subject of investigation in anthropology, as several fundamental properties of social relationships can be characterised by studying them. In particular, it has been shown that in (offline) ego networks there are a series of "circles" of alters arranged in a hierarchical inclusive sequence based on an increasing level of intimacy [8], [9]. The innermost circle includes alters with a very strong relationship with the ego. Each subsequent circle (in hierarchy) includes all the relationships of the previous circles along with an additional set of social links with a weaker level of intimacy. The last set, included in the outermost circle only, contains simple acquaintances, with a relatively weak relationship with the ego

(we describe in greater detail this structure in Section II). The scaling factor of the size of the circles (i.e., the ratio between the sizes of two successive circles) is almost constant, and very close to three. A well-know result is that the overall size of the ego network is - on average - around 150. This is typically called the *Dunbar's number*, and identifies the maximum number of active social relationships people are able to maintain [2]. It has been shown that maintaining a social relationship *active* costs cognitive resources, and thus requires investing time in interactions, and memory to remember facts about the alter. Therefore, the Dunbar's number corresponds to cognitive limits of the human brain, i.e., it is the number that "saturates" the human cognitive capacities devoted to maintaining social relationships [8].

In this work we analyse a publicly available large-scale Facebook data set, to check whether similar clustered structures can be identified also in Facebook ego networks. It is difficult to anticipate the outcome of such an analysis. On the one hand, one could postulate that, as in the end OSNs are just one of the possible means of interactions between humans, the same structures found in offline ego networks should also be found in OSN ego networks. On the other hand, one could also postulate that the type of social relationships in the virtual and physical worlds could be different, and that therefore different structures could be found in OSN ego networks.

The Facebook data set we analyse to investigate this aspect contains information regarding social interactions between more than 3 million users. This is one of the few Facebook large-scale publicly available data sets that can be used for our purposes. As will be clear from the discussion in Section II, to characterise ego networks we don't simply need the Facebook *network* graph - i.e. the graph formed by the links between users that are friends on Facebook -, but we need the Facebook *interaction* graph - i.e., the weighted version of the network graph, where weights represent the strength of the social relationship, measured as the frequency of interaction (or contact frequency) between the two users.

As described in detail in the rest of the paper, from this data set: (i) we extract a large number of ego networks, (ii) we estimate for each network the contact frequency between the ego and her alters, and (iii) we apply different clustering techniques on the contact frequency between them. Very interestingly, we find a striking similarity between the structures in offline social networks and in OSNs. Specifically, we find that also Facebook ego networks can be seen as organised in concentric circles, and the (i) number of circles, (ii) size of the circles, and (iii) scaling factor between successive circles are remarkably similar to those found in offline social networks [10]. Note that this strongly suggests that the structural properties of social networks in the two worlds are determined by similar human cognitive processes. In particular, our work confirms the results in [11] which suggest that the Dunbar's number (and the cognitive constraints determining it) hold also in OSNs. This confirms also other results indicating that other types of communication technologies (such as cellular phones) do not significantly change the structural properties of social networks [12], [13]. However, to the best of our knowledge, this is the first time such a precise characterisation of ego networks is carried out for OSNs, one of the most important means of social communication we are using today.

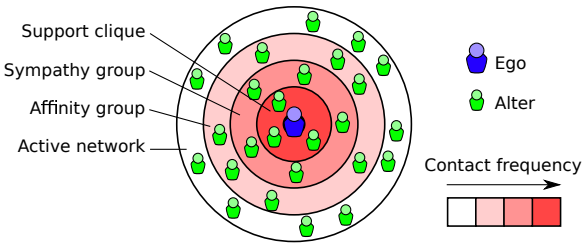

Fig. 1. Dunbar Circles

The reminder of this paper is organised as follows: in section II we give an overview of the existing work regarding offline ego networks. in Section III we describe the data set we use in the analysis. Then, in Section IV we process the data set to extract the ego networks from it. Hence, in Section V we give an overview of the techniques we use to analyse the structure of the obtained ego networks. Section VI presents the results we obtain and a discussion on the related implications. In Section VII we define a relevant subgraph of the network. Section VIII draws the main conclusions of our work.

II. Background Work on Offline Ego Networks

Studies in the anthropology literature demonstrated that the cognitive limits of the human brain constrain the number of social relationships an individual can actively maintain. Indeed, keeping a social relationship "active" requires a non negligible amount of cognitive resources, which are limited by nature. Studying the correlation between the neocortex size in primates and the dimension of their social group, anthropologists hypothesised that the average number of social ties an individual can actively maintain is approximately 150 [8], (widely known as Dunbar's number). These results have been validated in various studies on offline ego networks [9], [10].

Offline ego networks show a characteristic series of "circles" of alters arranged in a hierarchical inclusive sequence, based on an increasing level of intimacy [8], [9]. An ego can be depicted at the centre of these concentric circles (called Dunbar's circles), as shown in Fig. 1. Previous studies found that the circles of this structure have typical dimensions and characteristics [2] and that the scaling factor between their sizes is near to three [10]. The first circle (also called *support clique*) is the set of alters from whom ego seeks advice in case of severe emotional distress or financial disasters [2] and is, on average, limited to five people. The other circles are called *sympathy group* ($\sim$ 15 members), *affinity group* ($\sim$ 50 members) and *active network* ($\sim$ 150 members). The last circle delimits the boundaries dividing "active" ties, for which ego spends a non negligible amount of mental effort to maintain the relationship, and "inactive" ties, related to mere acquaintances.

Since the intimacy between people is not directly observable, the concentric structure in offline ego networks is

commonly defined using the contact frequency between ego and alters. This definition relies on the strong relation existing between the intimacy of a social link and the contact frequency between their members [9], [14]. The support clique is thus defined as the group of people ego contacts weekly, the sympathy group as the people contacted at least monthly and the active network as the people contacted at least yearly [15]. No accurate information is available in literature about the affinity group circle, neither for its typical contact frequency, nor regarding properties of the alters contained in this circle.

Apart from the anthropological studies on offline ego networks, limited research work has been done to analyse the properties of virtual ego networks. Indeed, most of the work in social network analysis focuses on the study of global properties of OSNs. Specifically, much effort has been spent to validate the famous "small world property" [5]–[7] in virtual environments. Moreover, it has been proved that in OSNs the distribution of the degree (the number of social links of an individual) typically follows a power law, with a long tailed shape [16]. Recently, in [7], [12], [13], the intensity of communication in OSNs has been used as a first attempt to discern between "active" and "inactive" virtual relationships and to validate at the same time the distinction between weak and strong ties, hypothesised in [1]. In [11], [17] authors discovered evidences of the Dunbar's number in OSNs. In addition, new models for the generation of synthetic social networks based on Dunbar's findings are beginning to spring in literature [11], [18]. Although these findings highlight some important properties of OSNs, a clear description of the structure of OSN ego networks and a detailed analysis of the differences emerging from the comparison between real and virtual ego networks are still missing. The aim of this work is to contribute to bridge the gap between offline ego networks and OSN analysis, presenting a detailed study of the structure of ego networks in Facebook.

III. Data Set Description

Public available data regarding social relationships is getting more and more difficult to be obtained from OSNs. Indeed, Facebook and other popular social networks have started to strengthen their privacy policies to limit the amount of user's sensible data that can be accessed without having the explicit consent of the user. At the same time, users have become less inclined to disclose their personal information, since they know that their data can be potentially used for commercial purposes. For this reason, people are limiting as much as possible the amount of data they put on their public profiles. Thereby, collecting large data sets of social relationships is, at present, a rather difficult task and requires a large amount of time. Each user must be contacted individually and must be prompted for special permissions before her data can be downloaded.

Before Facebook removed regional networks feature in 2009, the default privacy settings allowed people inside the same regional network to have full access to each others' personal data. At that time many crawlers were built to gather as much data as possible from the largest regional networks. Data coming from these regional networks have been widely used for social network analysis [19], [20].

TABLE I
Statistics of the Social Graph

# Nodes	$3,097,165$
# Edges	$23,667,394$
Average degree	15.283
Average clustering coefficient	0.098
Assortativity	0.048

To assess the presence in OSNs of social structures found in offline ego networks [2], we decided to analyse a large data set crawled from a Facebook regional networks on April 2008[1]. The data set has been studied in previous research work for purposes different than ours [21].

A. Data Set Features

The data set consists of a "social graph" and four "interaction graphs". These graphs are defined by lists of edges connecting pairs of anonymised Facebook user IDs.

The social graph describes the overall structure of the downloaded network. It consists of more than 3 million nodes (Facebook users) and more than 23 million edges (social links). An edge represents the mere existence of a Facebook friendship, regardless of the quality and the quantity of the interactions between the involved users. Basic statistics of the social graph are reported in Table I.

Interaction graphs describe the structure of the network during specific temporal windows, providing also the number of interaction occurred for each social link. The four temporal windows in the data set, with reference to the time of the crawl, are: *last month*, *last six months*, *last year* and *all*. The latter temporal window ("all") refers to the whole period elapsed since the establishment of each social link, thus considering all the interactions occurred between the users. In an interaction graph, an edge connects two nodes only if an interaction between two users occurred at least once in the considered temporal window. An interaction can be either a Facebook Wall post or a photo comment.

The social graph can be used to study the global properties of the network, but alone it is not enough to make a detailed analysis of the structure of social ego networks in Facebook. Indeed, this analysis requires an estimation of the intimacy between people involved in the social relationships. To this aim we leverage the data contained in the interaction graphs and we extract the contact frequency of each social link, using it as an estimate of the intimacy of the relationships.

In Facebook, an interaction can occur exclusively between two users who are friends. In other words, if a link between two nodes exists in an interaction graph, an edge between the same nodes should be present in the social graph. Actually, the data set contains a few interactions between users which are not connected in the social graph. These interactions

[1]This data set is publicly available for research at http://current.cs.ucsb.edu/facebook/, referred as "Anonymous regional network A".

TABLE II
STATISTICS OF THE INTERACTION GRAPHS (PREPROCESSED)

	Last mo.	Last 6 mo.	Last year	All
# Nodes	414,872	916,162	1,133,151	1,171,208
# Edges	671,613	2,572,520	4,275,219	4,357,660
Avg. degree	3.238	5.616	7.546	7.441
Avg. weight	1.897	2.711	3.700	3.794

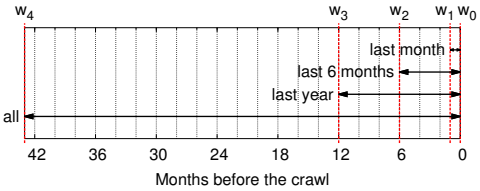

Fig. 2. Temporal windows.

probably refer to expired relationships or to interactions made by accounts that are no longer active. To maintain consistency in the data set we exclude these interactions from the analysis. The amount of discarded links is, on average, 6.5% of the total number of links in the data set.

In Table II we report some statistics regarding the different interaction graphs. Each column of the table refers to an interaction graph related to a specific temporal window. The average degree of the nodes (named "avg. degree" in the table) can be interpreted as the average number of social links per ego, which have at least one interaction in the considered temporal window. Similarly, the average link weight ("avg. weight" in the table) represents the average number of interactions for each social link. The measures reported in table are highly influenced by the presence in the data set of a large number of outliers which are identified and discarded in Section IV-D.

The data set contains only a partial view of the original Facebook regional network processed by the crawler. Since we don't have any further information regarding how the data were crawled, we assume that the crawler has picked up a random sample of the original Facebook regional network. In addition, from the statistics of other regional networks downloaded by the same crawler [19], we know that the average percentage of nodes downloaded by the crawler is 56.3% and the average percentage of downloaded links is 43.3%. Thus we might assume that the network analysed in this study represents a similar percentage of the crawled network. Starting from this assumption, in Section VI we analyse the size of the obtained ego networks and their components, by multiplying them by 2.31 (i.e., $1/0.433$). This assumption allows us to compare our results with those found in offline networks.

IV. PROCESSING DATA FOR EGO NETWORK ANALYSIS

The data set contains some relationships with no interactions associated with them. We consider these social links as "inactive". On the other hand we define as "active" all the relationships that have at least one interaction, that is to say the relationships included in the interaction graphs. The data set contains $4,357,660$ active links and $19,309,734$ inactive links. We are particularly interested in the analysis of active social relationships, since we want to assess the dimension and the structural composition of active ego networks, as defined for offline ego networks [2]. Hence, in the next part of the analysis we only consider active relationships. Since the data set contains social data over a temporal slice of several years, our definition of active relationship is compatible with that given in [15].

In order to characterise active relationships we need to estimate the *temporal span of social links* (i.e., the time elapsed since the establishment of the link), since it can be used to find the frequency of contact between the involved users. The frequency of contact is then used as an estimate of the intimacy between ego and alters, to characterise the structure of virtual ego networks. In literature, the duration of a social link is commonly estimated using the time elapsed since the first interaction between the involved users [22]. Unfortunately, the data set does not provide any indication regarding the time at which the interactions occurred. To overcome this limitation, we approximate social links duration leveraging the difference between the number of interactions made at the different temporal windows.

In Section IV-A we give some definitions we use in the following subsections. The methodologies we use to estimate the duration and the contact frequency of each social link in the data set are described in Sections IV-B and IV-C respectively. Then, in Section IV-D, we identify, from the available data, a set of ego networks that are meaningful for our study, while we discard the ego networks that we consider as outliers and hence not relevant for us.

A. Definitions

We define the temporal window "last month" as the interval of time (w_1, w_0), where $w_1 = 1$ month (before the crawl) and $w_0 = 0$ is the time of the crawl. Similarly we define the temporal windows "last six months", "last year" and "all" as the intervals (w_2, w_0), (w_3, w_0) and (w_4, w_0) respectively, where $w_2 = 6$ months, $w_3 = 12$ months and $w_4 = 43$ months. w_4 is the maximum possible duration of a social link in the data set, obtained by the difference between the time of the crawl (April 2008) and the time Facebook started (September 2004). The different temporal windows are depicted in Fig. 2.

For a social relationship r, let $n_k(r)$ with $k \in \{1, 2, 3, 4\}$ be the number of interactions occurred in the temporal window (w_k, w_0). Since all the temporal windows in the data set are nested, $n_1 \leq n_2 \leq n_3 \leq n_4$. If no interactions occurred during a temporal window (w_k, w_0), then $n_k(r) = 0$. As a consequence of our definition of active relationship, since $n_4(r)$ refers to the temporal window "all", $n_4(r) > 0$ only if r is an active relationship, otherwise, if r is inactive, $n_4(r) = 0$.

The first broad estimation we can do to discover the duration of social ties in the data set is to divide the relationships into different classes C_k, each of which indicates in which

interval of time (w_k, w_{k-1}) the relationships contained in it started (i.e., the first interaction has occurred). We can perform this classification analysing for each relationship the number of interactions in the different temporal windows. If all the temporal windows contain the same number of interactions, the relationship must be born less than one month before the time of the crawl, that is to say in the time interval (w_1, w_0). These relationships belong to the class C_1. Similarly, considering the smallest temporal window (in terms of temporal size) that contains the total number of interactions (equal to n_4), we are able to identify social links with duration between one month and six months (class C_2), six months and one year (class C_3), and greater than one year (class C_4). The classes of social relationships are summarised in Table III.

B. Estimation of the Duration of the Social Links

Although the classification given in the previous subsection is extremely useful for our analysis, the uncertainty regarding the estimation of the exact moment of the establishment of social relationships is still too high to obtain significant results from the data set. For example, the duration of a social relationship $r_3 \in C_3$ can be either a few days more than six months or a few days less than one year. To overcome this limitation, for each relationship r in the classes $C_{k\in\{2,3,4\}}$ we estimate the time of the first interaction comparing the number of interactions n_k, made within the smallest temporal window in which the first interaction occurred (w_k, w_0), with the number of interactions (n_{k-1}), made in the previous temporal window in terms of temporal size (w_{k-1}, w_0). If $n_k(r)$ is much greater than $n_{k-1}(r)$, a large number of interactions occurred within the time interval (w_k, w_{k-1}). Assuming that these interactions are distributed in time with a frequency similar to that in the window (w_{k-1}, w_0), the first occurred interaction must be near the beginning of the considered time interval. On the other hand, a little difference between $n_k(r)$ and $n_{k-1}(r)$ indicates that only few interactions occurred in the considered time interval (w_k, w_{k-1}). Thus, assuming an almost constant frequency of interactions, the first contact between the involved users must be at the end of the time interval (see Fig. 3 for a graphical representation of this concept).

In order to represent the percentage change between the number of interactions n_k and n_{k-1}, we calculate for each relationship $r \in C_k$ what we call *social interaction ratio* $h(r)$, defined as:

$$h(r) = \begin{cases} n_k(r)/n_{k-1}(r) - 1 & \text{if } r \in C_{k\in\{2,3,4\}} \\ 1 & \text{if } r \in C_1 \end{cases} . \quad (1)$$

TABLE III
CLASSES OF RELATIONSHIPS

Class	Time interval (in months)	Condition
C_1	$(w_1 = 1, w_0 = 0)$	$n_1 = n_2 = n_3 = n_4$
C_2	$(w_2 = 6, w_1 = 1)$	$n_1 < n_2 = n_3 = n_4$
C_3	$(w_3 = 12, w_2 = 6)$	$n_1 \leq n_2 < n_3 = n_4$
C_4	$(w_4 = 43, w_3 = 12)$	$n_1 \leq n_2 \leq n_3 < n_4$

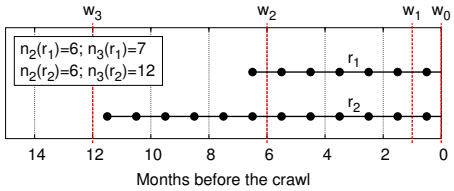

Fig. 3. Graphical representation of two different social relationships r_1, $r_2 \in C_3$. The difference between the respective values of n_2 and n_3 is small for r_1 and much larger for r_2. For this reason, fixing the frequency of contact, the estimate of the time of the first interaction of r_1 is near to w_2, while the estimate for r_2 results closer to w_3.

If $r \in C_1$ we set $h(r) = 1$ in order to be able to perform the remaining part of the processing also for these relationships. The value assigned to $h(r)$ with $r \in C_1$ is arbitrary and can be substituted by any value other than zero without affecting the final result of the data processing. Considering that $n_k(r)$ is greater than $n_{k-1}(r)$ by definition with $r \in C_{k\in 2,3,4}$, the value of $h(r)$ is always in the interval $(0, \infty)$[2].

Employing the social interaction ratio $h(r)$, we define the function $\hat{d}(r)$ that, given a social relationship $r \in C_k$, estimates the point in time at which the first interaction of r occurred, within the time interval (w_k, w_{k-1}):

$$\hat{d}(r) = w_{k-1} + (w_k - w_{k-1}) \cdot \frac{h(r)}{h(r) + a_k} \qquad r \in C_k, \quad (2)$$

where a_k is a constant, different for each class of relationship C_k.

Note that the value of $\hat{d}(r)$ is always in the interval (w_{k-1}, w_k). The greater $h(r)$ - which denotes a lot of interactions in the time window (w_k, w_{k-1}) - the more $\hat{d}(r)$ is close to w_k. The smaller $h(r)$, the more $\hat{d}(r)$ is close to w_{k-1}. Moreover, the shape of the $\hat{d}(r)$ function and the value of a_k are chosen relying on the results about the Facebook growth rate, available in [19]. Specifically, the distribution of the estimated links duration, given by the function $\hat{d}(r)$, should be as much similar as possible to the distribution of the real links duration, which can be obtained analysing the growth trend of Facebook over time. For this reason, we set the constants a_k in order to force the average link duration of each class of relationships to the value that can be obtained by observing the Facebook growth rate. In the Appendix we provide a detailed description of this step of our analysis.

C. Estimation of the Frequency of Contact

After the estimation of social links duration, we are able to calculate the frequency of contact $f(r)$ between the pair of individuals involved in each social relationship r:

$$f(r) = n_k(r)/\hat{d}(r) \qquad r \in C_k. \quad (3)$$

Previous research work demonstrated that the pairwise user interaction decays over time and it has its maximum right after

[2]In case $n_{k-1}(r) = 0$, we set $n_{k-1}(r) = 0.3$. This constant is the expected number of interactions when the number of interactions, within a temporal window, is lower than 1.

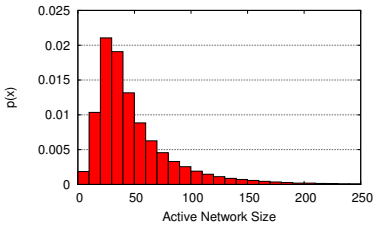

Fig. 4. Active network size distribution.

link establishment [20]. Therefore, if we assessed the intimacy level of the social relationships with their contact frequencies, this would cause an overestimation of the intimacy of the youngest relationships. In order to overcome this problem, we multiply the contact frequencies of the relationships in the classes C_1 and C_2 by the scaling factors m_1 and m_2 respectively, which correct the bias introduced by the spike of frequency close to the establishment of the link. Assuming that the relationships established more than six months before the time of the crawl are stable, we set m_1 and m_2 comparing the average contact frequency of each of the classes C_1 and C_2, with that for the classes C_3 and C_4. Obtained values of the scaling factors are: $m_1 = 0.18$, $m_2 = 0.82$. Setting $m_3 = 1$ and $m_4 = 1$, scaled frequencies of contact are defined as:

$$\hat{f}(r) = f(r) \cdot m_k \qquad r \in C_k. \tag{4}$$

To be able to extract the ego networks of the data set we group the relationships of each user into different sets R_e, where e identifies a specific ego. We duplicate each social link in the data set in order to consider it in both the ego networks of the users connected by it.

Since each ego in the data set has different Facebook usage, the calculated frequencies of contact are not directly comparable. For example, the same frequency of contact can represent, for different users, different levels of intimacy. To overcome this limitation, for each ego network in the data set, we normalise the frequency of contact related to every relationship by applying (5). Specifically, we divide it for the maximum value of frequency of contact of all the links of the ego, obtaining values between 0 and 1. This normalisation is essential to be able to compare the results of our analysis for different ego networks.

$$f_{norm}(r) = \frac{\hat{f}(r)}{\max\limits_{r^\star \in R_e} \hat{f}(r^\star)} \qquad r \in R_e. \tag{5}$$

D. Ego Networks Selection

A high number of ego networks in the data set started just before the time of the crawl while other ego networks have a very low interaction level. The analysis could be highly biased by considering these outliers. Thus, we selected a subset of the available ego networks according to the following criteria. First of all we intuitively define as "relevant" the users who joined Facebook at least six months before the time of the crawl and who have made, on average, more than 10 interactions per month. We estimate the duration of the presence of a user in Facebook as the time since she made her first interaction. The new data set obtained from the selection of relevant ego networks contains $91,347$ egos and $4,619,221$ social links[3].

The average active ego network size after the cleanup is equal to 50.6. The reader could notice a rather high discrepancy between this average active network size and those found in other studies [9], [11], [17]. The main cause of this difference is due to the fact that the data set does not contain entire ego networks, but about 43% of their size (see section III). We come back to this point in Section VIII, where we discuss in greater detail how to fairly compare the ego networks in the Facebook data set with those analysed in the anthropology literature [2], [10]. We anticipate that this comparison highlight a significant similarity between the sizes of the active ego networks in the two cases. Moreover, the active network size distribution (depicted in Fig. 4) has a similar shape to those found in other analysis both in real and virtual environments [9], [17], [18].

V. Methodology to Discover the Structure of Ego Networks

The first attempt we make in order to check whether concentric structures are present in Facebook ego networks is to observe the complementary cumulative distribution function (henceforth CCDF) of the frequency of contact calculated aggregating all the frequencies of all the ego networks. We may expect this CCDF to have some kind of irregularities (i.e., jumps) introduced by the possible presence of the clustered structure in the frequency of contact of the various ego networks. Yet, the CCDF (depicted in Fig. 5) shows a smooth trend. This is not necessarily an indication of absence of clustered structures in individual ego networks, but it could be caused by the aggregation of the different distributions of the ego networks' frequency of contact. In fact, even if the single ego networks showed the circular hierarchical structure described in Section II, the jumps between each circular cluster could appear at different positions from one ego network to another. This could mask jumps in the aggregated CCDF as we superpose the ego networks.

The CCDF of the aggregated frequency of contact shows a long tail, which can be ascribed to a power law shape. Power-law-shaped contact frequency distributions would be another indication of similarity between ego networks in real and virtual social networks, as offline ego networks are characterised by a small set of links with a very high contact frequency (corresponding to the links in the support clique). A power law shape in the aggregate CCDF is a necessary condition to have power law distributions also in each single ego networks [23]. However, this is not a sufficient condition to have power law distributions in each single CCDF [24]. Therefore, to further analyse the ego-network structure we

[3] $3,353,870$ unique social links. Some of them are counted twice because we duplicate each social link (see Section IV-C) and the ego networks of the users they connect can be both selected as relevant.

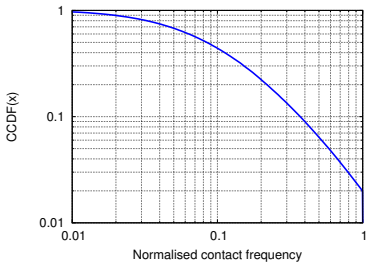

Fig. 5. Aggregated CCDF of the normalised contact frequency for all the ego networks in the data set.

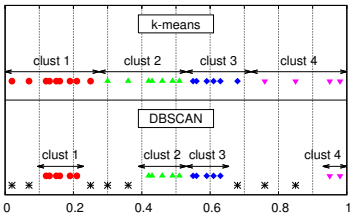

Fig. 6. Example of different results obtained applying k-means and the iterative DBSCAN over a noisy data space, using $k = 4$.

apply clustering techniques to each ego network looking for the emergence of the layered structure observed in offline social networks. Specifically, we leverage two different families of clustering techniques: *partitioning clustering* and *density-based clustering*.

Partitioning clustering algorithms start with a set of objects and divide the data space into k clusters so that the objects inside a cluster are more similar to each other than the objects in different clusters. In our analysis, similarity means closeness in contact frequency. Specifically, for each ego network, we order alters in a one dimensional space, according to the contact frequency with the ego, and we seek clusters in this one dimensional space. Density-based clustering algorithms are able to identify clusters in a space of objects with areas with different densities, see [25].

We start the structural analysis of ego networks applying a partitioning clustering technique. Specifically, we use an algorithm able to find the optimal solution of the k-means problem for the special case of one-dimensional data [26]. The k-means problem is to partition the data space into k different clusters of objects, so that the sum of squared Euclidean distance between the centre of each cluster and the objects inside that cluster is minimised. The goodness of the result of k-means algorithms is often expressed in terms of "explained variance", defined by the following formula:

$$VAR_{exp} = \frac{SS_{tot} - \sum_{j=1}^{k} SS_j}{SS_{tot}}, \tag{6}$$

where j is the j^{th} cluster, SS_j is the sum of squares within cluster j and SS_{tot} is sum of squares of the all the values in the data space. The sum of square of a vector $\mathbf{X}$ is defined by the following formula:

$$SS_{\mathbf{X}} = \sum_{i} (x_i - \mu_{\mathbf{X}})^2, \tag{7}$$

where $\mu_{\mathbf{X}}$ denotes the mean value of $\mathbf{X}$.

The explained variance is analogous to the coefficient of determination R^2 used in linear regression analysis. VAR_{exp} ranges from 0 to 1. Therefore, the goal of k-means algorithms is to assign the objects of a set of data to k clusters, so that the resulting VAR_{exp} is maximised. However, there is a inherent over fitting problem. Indeed, the maximum value of VAR_{exp} is obtained when k is equal to the number of objects in the data space. To avoid this overfitting we use an elbow method with a fixed threshold of 10% of the explained variance. This is a standard way to determine the optimal number of clusters in a data set [27]. If, after adding a new cluster, the increment in terms of VAR_{exp} is less than 0.1, we take the value of k as the optimal number of clusters. Hence, we apply this method to extract the optimal k and the cluster composition of all the ego networks in the data set.

The results obtained with k-means could be potentially affected by the presence of noisy data. We use the notion of *noise* to define points in the data space with a very low density compared to the other points around them. Noise can affect the result of k-means in two different ways: (i) the presence of noisy points between two adjacent clusters could force the algorithm to discover a single cluster instead of two (the so called "single link effect" [25]); (ii) the presence of a large number of noisy points in the data set could lead k-means to detect clusters with a size larger than it should be according to a natural and intuitive definition of clustering (see Fig. 6 for a graphical example). To verify that the noisy points in the data set do not eccessively affect k-means we compare the results of the former with the results of a density-based clustering algorithm called DBSCAN [28]. DBSCAN takes two parameters, namely ϵ and $MinPts$. If an object has more than $MinPts$ neighbours within an ϵ distance from it, it is considered a core object. A cluster is made up by a group of core objects (where two contiguous elements have a distance shorter than ϵ) and by the "border objects" of the cluster. Border objects are defined as non-core objects linked to a core object at a distance shorter than ϵ. For a more formal definition of density based clusters see [28]. Points with less than $MinPts$ neighbours within a distance equal to ϵ are considered noise by DBSCAN, and they are excluded from the clusters.

We iterate DBSCAN and we stop as we find a number of clusters equal to the number of clusters obtained by k-means. Hence, by comparing the results of k-means and DBSCAN in terms of cluster size we can verify that the former are valid and not influenced by noisy points. To allow noisy data to be identified by the iterative DBSCAN procedure we set the parameter $MinPts$ to be equal to 2. In this way isolated points are excluded from the clusters.

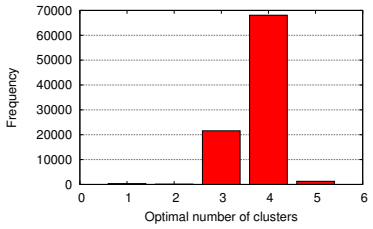

Fig. 7. Distribution of k_{opt} found using optimal 1-D k-means.

VI. The Structure of Facebook Ego Networks

Using the iterative procedure based on the k-means algorithm (see Section V) we find that the optimal number of clusters of each ego network (henceforth k_{opt}) ranges between 1 and 5. Fig. 7 depicts the number of ego networks for each k_{opt}. We find a positive correlation ($r = 0.25$, $p. < 0.01$) between k_{opt} and the active network size, as can be seen in Table IV. In the table the total contact frequency is expressed in terms of number of interactions per month made by ego to all her alters. The average value of k_{opt} is 3.76 (SD $= 0.48$) and the median is 4. The presence of a typical number of clusters near to 4 in Facebook is the first indication of similarity between the findings in offline ego networks and the ego networks in virtual environments. Since the amount of ego networks with k_{opt} equal to 1 and 2 is negligible w.r.t the total number of ego networks in the data set, we consider them as outliers and we exclude them in the subsequent part of the analysis.

We apply the iterative DBSCAN procedure (see Section V) on the ego networks with $k_{opt} = 4$. The comparison between the inclusive circles found by k-means and DBSCAN on these ego networks and those found in offline ego networks [10] are reported in Table V. For each circle we show its average size ("size" in the table) and the ratio between the latter and the average size of the previous circle in the hierarchy ("sc. f." in the table). We refer to this ratio as *scaling factor*. We find that the results of k-means and DBSCAN are compatible in terms of circles size and their respective scaling factors. This means that k-means results are not highly influenced by noisy points (see Section V). The discrepancy between the sizes of the support clique can be ascribed to the fact that DBSCAN considers isolated points as noise and, in many ego networks, the support clique could contain only one alter. The scaling factors found by k-means in Facebook (for $k_{opt} = 4$) are strikingly similar to the findings in offline ego networks (reported in Table V as "off-l"). Indeed, the average value of the scaling factors are equal to 3.12 in Facebook and 3.06 in offline ego networks. In addition, the last row of the table ("estim.") reports an estimation of the size of the circles in Facebook ego networks, obtained multiplying by 2.31 the results found by k-means - i.e., considering that the average percentage of each circle w.r.t. its real size is 43.3% (see Section III-A). Even in this case, there is a strong resemblance between the sizes of the circles in Facebook and in offline ego networks.

TABLE IV
# of Ego Networks and Average Active Network Size with 95% Confidence Intervals per each k_{opt}

k_{opt}	# of nets	Active net size [95% c. i.]	Total contact freq.
1	315	1.50 [1.23, 1.77]	15.21
2	107	3.81 [2.86, 4.75]	18.83
3	21,575	34.42 [34.09, 34.74]	26.96
4	68,079	55.23 [54.93, 55.54]	35.64
5	1,271	77.74 [74.70, 80.78]	37.87

TABLE V
Results for $k = 4$ of k-means (k-m) and DBSCAN (DB) on Ego Networks with $k_{opt} = 4$ with 95% C.I..

	support clique	sympathy group	affinity group	active network
size (k-m)	1.84 [.01]	6.36 [.03]	18.68 [.09]	55.48 [.3]
sc. f. (k-m)	-	3.45	2.94	2.97
min freq.	4.46	1.81	0.66	0.11
size (DB)	2.74 [.01]	6.85 [.04]	17.24 [.1]	49.11 [.4]
sc. f. (DB)	-	2.5	2.52	2.85
size (off-l)	4.6	14.3	42.6	132.5
sc. f. (off-l)	-	3.10	2.98	3.11
estim. (k-m)	(3.42)	(14.69)	(43.15)	(128.16)

The minumum contact frequency of the relationships within each circle ("min freq." in Table V) is expressed in number of interactions per month. Using this variable, calculated averaging the results on all the ego networks, we are able to describe the circles of the discovered structure in terms of typical frequency of contact. Our results indicate that, in Facebook, the support clique contains people contacted at least $\sim$ *weekly*, the sympathy group $\sim$ *twice a month*, the affinity group $\sim$ *eight times a year* and the active network $\sim$ *yearly*. These results indicate that also the typical frequency of contact of the Dunbar's circles in Facebook appear to be very similar to that found in offline ego networks.

As regards the ego networks with k_{opt} equal to 3, it is interesting to notice that they don't have a counterpart in offline ego networks. Their size is, on average, smaller than the size of ego networks with k_{opt} equal to 4 and they show a lower rate of *Facebook usage*, defined by the total frequency of contact of each ego (see Table IV). We hypothesise that these ego networks have the same structure of the ego networks with k_{opt} equal to 4, but the results of k-means could be influenced by the presence of too few social links. To prove this fact we apply k-means on these ego networks forcing $k = 4$ and we compare the results with those found on ego networks with $k_{opt} = 4$. Table VI reports the results of this analysis. The last two rows of the table ("off-l perc." and "$k_{opt} = 4$ perc.") represent the percentage of size of the obtained circles w.r.t. the size of the respective circles found in offline ego networks and those found with k-means on the ego networks of the data set with $k_{opt} = 4$.

Ego networks with $k_{opt} = 3$ show a support clique with size near to the dimensions found in offline ego networks (81.30%) and to that found by k-means on ego networks with $k_{opt} = 4$ (86.18%). The dimensions of the other circles are noticeably lower. This result indicates that, in Facebook, users tend to

TABLE VI
RESULTS FOR $k = 4$ OF k-MEANS (k-M) ON EGO NETWORKS WITH $k_{opt} = 3$ WITH 95% C.I..

	support clique	**sympathy group**	**affinity group**	**active network**
size	1.62 [.01]	4.14 [.03]	11.9 [.1]	34.63 [.3]
sc. f.	-	2.56	2.87	2.91
min freq.	7.07	2.39	0.71	0.12
estim.	(3.74)	(9.56)	(27.49)	(80)
off-l perc.	81.30%	66.85%	64.53%	60.38%
$k_{opt} = 4$ perc.	109%	65.08%	63.71%	62.42%

have a set of core friends whom they contact frequently even if they have a lower rate of Facebook usage compared to the average. Nevertheless, the dimensions of the remaining circles are sensibly lower than the dimensions of the circles found in larger ego networks with higher Facebook usage. Still, the average scaling factor for the circles of the ego networks with $k_{opt} = 3$ - equal to 2.78 - remains close to three, as an additional proof of the similarity between virtual and real ego networks.

The typical contact frequencies of the circles of ego networks with $k_{opt} = 3$ are the following: the support clique contains people contacted at least $\sim$ *seven times a month*, the sympathy group $\sim$ *twice a month*, the affinity group $\sim$ *eight times a year* and the active network $\sim$ *yearly*.

As far as the ego networks with k_{opt} equal to 5, we add them to the ego networks with k_{opt} equal to 4 and we re-apply k-means on the resulting set, forcing $k = 4$. The results do not differ significantly (in terms of circle sizes and scaling factors) from the results found on ego networks with $k_{opt} = 4$, reported in Table V.

VII. FACEBOOK "ACTIVE" GRAPH DESCRIPTION

The Facebook network given by the data set contains a high number of low-activity users therefore, in order to analyse a network as similar as possible to an offline social network, we consider the subgraph formed by the relevant users (see Section IV-D) and the social links among them. In addiction, in order to better analyse the correlation between connected ego networks, we include in the graph the results about the ego networks' structures obtained in Section VI. We call this network *Facebook "active" graph*.

The links in the active graph are directed and connect pairs of relevant egos. Considering that for each relevant ego we can perform the analysis of its ego network structure, we can label each of its outgoing links with a layer ID. In order to maintain consistency between the structures of different ego networks, we assign the layer IDs to the links by applying k-means with $k = 4$ for each relevant ego. Note that, for 422 of 91,347 relevant egos, it is not possible to force $k = 4$ because their interaction frequencies assume less than 4 different values. These egos are thus not included in the active graph.

Statistics of the Facebook active graph are reported in Table VII.

TABLE VII
STATISTICS OF THE ACTIVE GRAPH.

# Nodes	90,925
# Edges (directed)	2,529,316
Average out-degree	27.82
Average cluster coefficient	0.109
Average shortest path	4.06
Assortativity	0.16

VIII. CONCLUSIONS

In this paper we aim to discover the presence of Dunbar's circles in OSN ego networks. With this purpose, we analyse a data set containing more than 23 million social interactions in Facebook. We extract the frequency of contact from the ego networks in the data set. Then, we apply different clustering techniques on the distributions of the frequency of contact of the different ego networks (specifically, partitioning clustering and density-based clustering). Hence, we analyse the results seeking for the presence of the possible circular structure.

We find that the properties of OSN ego networks have a strong similarity with those found in offline ego networks. Namely, the typical number of circles in the structure of virtual ego networks is, on average, equal to 4 and the average scaling factor between the concentric circles of the social structure is near to 3, as found in real environments. Moreover, the sizes of the circles, i.e. the number of social relationships of each type, is remarkably similar to those existing in offline social networks. Notably, the average size of the OSN ego networks is very close to the well known Dunbar's number, which denotes the average size of ego networks in offline social networks.

We can conclude that, even if OSNs introduce new communication paradigms and plenty of new ways to maintain social relationships with others, the structure of the personal social networks of the users maintains the same properties of ego networks formed offline.

APPENDIX
CALCULATION OF THE a_k VALUES

In order to set a_k constants properly, we leverage on the real growth trend of Facebook over time. Hence, we approximate the Facebook network's evolution reported in [19] with the piecewise function $g(t)$ defined as:

$$g(t) = \begin{cases} 8,876,376 - 720,099 \cdot t & \text{if } t < 10 \\ 3,348,056 - 167,267 \cdot t & \text{if } 10 \leq t < 18 \\ 580,070 - 13,490 \cdot t & \text{if } t \geq 18 \end{cases} \quad (8)$$

where t is the time in months before the time of the crawl. The first elbow point of the function is placed 18 months before the time of the crawl (October 2006), when Facebook opened to everyone. Before that time, the membership was restricted to university and high-school students only. The second elbow point is placed 10 months before the time of the crawl (February 2007), when Facebook starts to become popular and its growth trend shows a significant acceleration.

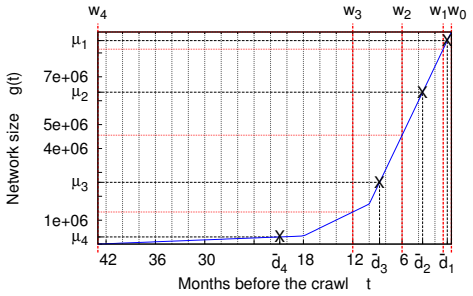

Fig. 8. The growth of Facebook over time from the time Facebook started (September 2004) to the time of the crawl (April 2008)

For each class of relationship C_k, let μ_k be the mean value of $g(t)$ with $t \in (w_k, w_{k-1})$ and let $\bar{d}_k$ be the point in time where $g(t)$ is equal to μ_k. Resulting values for $\bar{d}_k = g^{-1}(\mu_k)$ are: $\bar{d}_1 = 0.5$, $\bar{d}_2 = 3.5$, $\bar{d}_3 = 8.74$ and $\bar{d}_4 = 20.88$. The placement of these values over the Facebook growth function $g(t)$ is depicted in Fig. 8.

Reasonably assuming that the growth trend of the links is proportional to the growth trend of the nodes, we can consider $\bar{d}_k$ as the average duration of the relationships belonging to the class C_k. In order to force the means of estimated links duration to be equal to the means obtained by the Facebook growth function, we set the constants a_k to satisfy the following equation:

$$\frac{1}{|C_k|} \sum_{r \in C_k} \hat{d}(r) = \bar{d}_k \tag{9}$$

We obtain the following values of a_k: $a_1 = 1$, $a_2 = 3.18$, $a_3 = 3.69$ and $a_4 = 3.79$.

ACKNOWLEDGEMENTS

This work was partially funded by the European Commission under the SCAMPI (FP7-FIRE 258414), RECOGNITION (FP7 FET-AWARENESS 257756), and EINS (FP7-FIRE 288021) projects.

REFERENCES

[1] M. S. Granovetter, "The Strength of Weak Ties," *The American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[2] A. Sutcliffe, R. Dunbar, J. Binder, and H. Arrow, "Relationships and the social brain: Integrating psychological and evolutionary perspectives," *British Journal of Psychology*, vol. 103, no. 2, pp. 149–168, 2012.

[3] N. B. Ellison, C. Steinfield, and C. Lampe, "The Benefits of Facebook "Friends": Social Capital and College Students' Use of Online Social Network Sites," *Journal of Computer-Mediated Communication*, vol. 12, no. 4, pp. 1143–1168, 2007.

[4] M. Burke, C. Marlow, and T. Lento, "Social Network Activity and Social Well-Being," in *international conference on Human factors in computing systems*, 2010, pp. 2–5.

[5] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The Anatomy of the Facebook Social Graph," *CoRR*, vol. abs/1111.4, 2011.

[6] L. Backstrom *et al.*, "Four Degrees of Separation," *CoRR*, vol. abs/1111.4, 2011.

[7] P. S. Dodds, R. Muhamad, and D. J. Watts, "An experimental study of search in global social networks." *Science*, vol. 301, no. 5634, pp. 827–9, 2003.

[8] R. I. M. Dunbar, "The social brain hypothesis," *Evolutionary Anthropology*, vol. 6, no. 5, pp. 178–190, 1998.

[9] R. A. Hill and R. I. M. Dunbar, "Social network size in humans," *Human Nature*, vol. 14, no. 1, pp. 53–72, Mar. 2003.

[10] W.-X. Zhou, D. Sornette, R. A. Hill, and R. I. M. Dunbar, "Discrete hierarchical organization of social group sizes." in *Biological sciences*, vol. 272, no. 1561, 2005, pp. 439–44.

[11] B. Goncalves, N. Perra, and A. Vespignani, "Validation of Dunbar's number in Twitter conversations," *Networks*, vol. 2011, no. 28, p. 8, 2011.

[12] J. Onnela *et al.*, "Structure and tie strengths in mobile communication networks." in *National Academy of Sciences of the United States of America*, vol. 104, no. 18, 2007, pp. 7332–7336.

[13] J. Leskovec and E. Horvitz, "Planetary-Scale Views on an Instant-Messaging Network," Tech. Rep., 2007.

[14] P. V. Marsden and K. E. Campbell, "Measuring Tie Strength," *Social Forces*, vol. 63, no. 2, pp. 482–501, 1984.

[15] O. Curry and R. I. Dunbar, "Why birds of a feather flock together? The effects of similarity on altruism," *Social Networks*, vol. (submitted, 2011.

[16] A. Mislove *et al.*, "Measurement and analysis of online social networks," *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement IMC 07*, vol. 40, no. 6, p. 29, 2007.

[17] V. Arnaboldi, A. Passarella, M. Tesconi, and D. Gazzè, "Towards a Characterization of Egocentric Networks in Online Social Networks." in *OTM Workshops*, ser. Lecture Notes in Computer Science, vol. 7046. Springer, 2011, pp. 524–533.

[18] M. Conti, A. Passarella, and F. Pezzoni, "A Model for the Generation of Social Network Graphs," in *IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks*. IEEE, 2011, pp. 1–6.

[19] C. Wilson *et al.*, "User interactions in social networks and their implications," in *Proceedings of the 4th ACM European conference on Computer systems*, ser. EuroSys '09. New York, NY, USA: ACM, 2009, pp. 205–218.

[20] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in Facebook," in *Proceedings of the 2nd ACM workshop on Online social networks - WOSN '09*. New York, New York, USA: ACM Press, 2009, p. 37.

[21] M. U. Ilyas, M. Z. Shafiq, A. X. Liu, and H. Radha, "A distributed and privacy preserving algorithm for identifying information hubs in social networks," in *Proceedings IEEE INFOCOM*, 2011, pp. 561–565.

[22] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in *International conference on Human factors in computing systems*. New York, New York, USA: ACM Press, 2009.

[23] A. Passarella, M. Conti, R. I. M. Dunbar, and C. Boldrini, "Modelling Inter-contact Times in Social Pervasive Networks," in *ACM WSWiM*, 2011.

[24] A. Passarella and M. Conti, "Characterising aggregate inter-contact times in heterogeneous opportunistic networks ," in *IFIP Networking*, 2011, pp. 1–12.

[25] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011.

[26] H. Wang and M. Song, "Clustering in One Dimension by Dynamic Programming," *The R Journal*, vol. 3, no. 2, pp. 29–33, 2011.

[27] D. J. Ketchen and C. L. Shook, "The Application of Cluster Analysis in Strategic Management Research: an Analysis and Critique," *Strategic Management Journal*, vol. 17, no. 6, pp. 441–458, 1996.

[28] M. Ester, H. Kriegel, and J. Sander, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.