Computer Communications 36 (2013) 1130–1144

Contents lists available at SciVerse ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Egocentric online social networks: Analysis of key features and prediction of tie strength in Facebook

Valerio Arnaboldi [a,*], Andrea Guazzini [b], Andrea Passarella [a]

[a] Institute for Informatics and Telematics, National Research Council of Italy, Via G. Moruzzi 1, 56124 Pisa, Italy
[b] Centre for the Study of Complex Dynamics, University of Florence, Via S. Marta 3, 50139 Firenze, Italy

ARTICLE INFO

Article history:
Received 20 December 2011
Received in revised form 6 March 2013
Accepted 9 March 2013
Available online 22 March 2013

Keywords:
Online social networks
Ego networks
Tie strength
Prediction

ABSTRACT

The widespread use of online social networks, such as Facebook and Twitter, is generating a growing amount of accessible data concerning social relationships. The aim of this work is twofold. First, we present a detailed analysis of a real Facebook data set aimed at characterising the properties of human social relationships in online environments. We find that certain properties of online social networks appear to be similar to those found "offline" (i.e., on human social networks maintained without the use of social networking sites). Our experimental results indicate that on Facebook there is a limited number of social relationships an individual can actively maintain and this number is close to the well-known Dunbar's number (150) found in offline social networks. Second, we also present a number of linear models that predict tie strength (the key figure to quantitatively represent the importance of social relationships) from a reduced set of observable Facebook variables. Specifically, we are able to predict with good accuracy (i.e., higher than 80%) the strength of social ties by exploiting only four variables describing different aspects of users interaction on Facebook. We find that the recency of contact between individuals – used in other studies as the unique estimator of tie strength – has the highest relevance in the prediction of tie strength. Nevertheless, using it in combination with other observable quantities, such as indices about the social similarity between people, can lead to more accurate predictions

1. Introduction

A particularly hot research trend over the last few years has been trying to envision novel directions for the long-term evolution of the Internet, usually named Future Internet [1]. Several research visions and approaches have been proposed, ranging from novel, disruptive architectural solutions [2,3], to new paradigms incorporating mobile networks as key elements [4], to approaches where the cyber and the physical world blur in a so-called cyber-physical confluence [5]. From this perspective, online social networks (henceforth OSN) rightly deserve particular attention. Besides being a huge success in terms of concrete products (e.g., Facebook, Twitter, ...), they can be seen as a notable example of cyber-physical confluence. The growing number of new communication paradigms introduced by these services is changing the way individuals interact and link to each other, which is an example of a cyber system that can have an impact on social relationships occurring in the physical world. Moreover, OSN are fostering the availability of a huge amount of data concerning social relationships between people, that can contribute to the analysis of the social behaviour of the users.

Sociologists, anthropologists and psychologists have largely studied social relationships in humans from two different points of view. On the one hand, the analysis of personal networks starts with an individual – called ego – and studies the relationships this individual has with other people – called alters. Many researchers refer to the networks formed from the ensemble of this relationships as *personal networks* or *ego networks* [6–8]. On the other hand, social network analysis studies the relationships existing between people inside a bounded population or community (e.g., researchers community, movie directors community, ...) [9–11]. Whilst social network analysis puts more emphasis on the key features of the whole network (e.g., topology, centrality, ...), personal network analysis focuses on the relevant features of ego's social relationships. With respect to social network analysis, the analysis of ego networks does not capture relationships between alters, and is thus not able to provide a complete analysis of the social network of the users. However, ego network analysis typically studies in greater detail the properties of the individual links between ego and alters. Results from ego network analysis already highlight a

* Corresponding author. Tel.: +39 050 315 2195; fax: +39 050 315 2593.
E-mail addresses: valerio.arnaboldi@iit.cnr.it (V. Arnaboldi), andrea.guazzini@iit.cnr.it (A. Guazzini), andrea.passarella@iit.cnr.it (A. Passarella).

0140-3664/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.comcom.2013.03.003

number of key properties characterising the social behaviour of the users [7]. Offline ego networks[1] have been deeply investigated and some of the key features of these networks have been identified [12–14]. In particular, "tie strength" - the importance of the social relationship between two individuals – is found to be one of the most important features of ego networks and it is what makes social networks really "social" [6,15].

Studies on the properties of OSN are becoming increasingly popular, as there is still lack of understanding of their key features, and of their impact on social relationships between individuals. Besides being an interesting topic of research per se, a clear understanding of the properties of social relationships between users in OSN can be the basis for the design of novel OSN services, such as, for example, novel data dissemination and management schemes exploiting social relationships to optimise data replication or information diffusion more in general. Data centric solutions in the framework of the future Internet is another very hot research topic [16–19], which however, up to now, has not fully exploited the possibility of taking advantage of information concerning social relationships between users.

Although the work done so far evinced many important aspects of OSN, the relation between users' social behaviour and tie strength in virtual environments has not been fully discovered yet, and represents one of the main focuses of this paper. Specifically, we aim to make a detailed comparison between the fundamental properties of ego networks in OSN and the characteristics found in offline environments. This comparison is essential to better understand human behaviour in OSN and to effectively study social and psychological aspects of OSN. The core of our analysis is represented by the creation of a model to estimate tie strength from OSN variables. Estimating the strength of social ties is clearly important for a number of social aware services, such as data dissemination, community detection, etc. Unfortunately, direct measures of social tie strength – i.e., quantitative measures taken without explicitly asking individuals – are not possible neither in offline social network nor in OSN, as tie strength depends also on emotional factors that are not directly measurable. Nevertheless, using interaction variables – such as the frequency of contacts – has proven to be effective in estimating tie strength in offline social networks [14]. This approach has still to be fully explored in OSN.

In this work we present a detailed analysis on a real Facebook data set – that we have collected – to examine if the interaction variables that can be measured in Facebook can be used to estimate the strength of social ties between individuals. Specifically, in our data set we gathered interaction variables between a set of egos and their Facebook friends. For each friendship relationship, the ego was also asked to explicitly provide a numerical value to represent the strength of that relationship. Exploiting this information, we aim to identify the smallest possible number of variables that can be used to reliably estimate tie strength. This is a much more advanced analysis compared to other studies [20,21], that present models for tie strength prediction fed with a large number of variables to obtain the best possible fit of the data set, without focusing on prediction, and eventually leading to overfitted models.

As a preliminary step to identify the variables to be used in the model of tie strength, we perform a detailed analysis of a number of candidate variables that we collected in our data set (results are presented in Section 4). This allows us to perform an initial characterisation of ego networks, for example in terms of their size, and compare these results with similar analyses regarding offline social networks. The question whether ego network in OSN are similar to those in offline networks or not is a challenging one. On the one hand, one could note that OSN may only represent a new tool to maintain our social relationship with others and the cognitive mechanisms behind our social behaviour should remain unaltered by the adoption of this tool. On the other hand, one could also argue that OSN provide a totally new environment for social interactions, which might result in completely different structures. The results we obtain in the first part of this paper (Section 4) support the former hypothesis. Specifically, the average number of social relationships a person actively maintains in our Facebook sample (105.14) is of the same order of that found in offline ego networks (124 in [14] and 132.5 in [12]). As we show in detail in the following, also other measurable variables in our data set are compatible with analogous variables measured in offline social networks. These findings suggest that also in OSN an individual is usually tightly connected to a small set of friends and loosely connected to a larger number of people [6,14,7].

Starting from this analysis, in the second part of the paper (Section 5) we focus on how to predict tie strength through interaction variables. First, we perform a correlation analysis to find the Facebook variables strongly related to tie strength. Then we build two different families of models. The first type is based on the set of uncorrelated interaction variables that show the strongest correlation with tie strength. The second type is based on a reduced set of orthogonal factors extracted from the data set using Principal Component Analysis (PCA). Comparing both types, we show that in both cases it is possible to accurately predict tie strength with an accuracy higher than 80%. Moreover, we use PCA results to find the key factors in the set of the Facebook measurable we have collected. This analysis shows that the factors in our variables represent all the relevant dimensions of tie strength, as identified in the social networking literature, from the seminal work of Granovetter [6] to more recent analyses [22]. Specifically, to the best of our knowledge, ours is the first work that identifies a set of observable variables in Facebook that cover all the tie strength dimensions identified in [6]. This also suggests that, even if they need to be validated on a larger data set, our models remain valid in general, and their validity is not limited to our particular data set.

Our analysis, although carried out over a limited number of users, already provides interesting results. First of all, we show that it is indeed possible to predict Facebook tie strength using observable variables about users interactions. Moreover, our models show that only a small number of variables is sufficient to achieve accurate predictions. This means that our models are suitable to be used online, as they require to collect limited information about users interactions. They can be efficiently used as elements of novel social networking applications and services, that exploit predictions of the quality of social relationships between users to tune the application/service behaviour.

In addition to the results of our analysis (Sections 4 and 5) in the remainder of the paper we provide background information describing some of the key features of offline ego networks and of OSN, from the sociology and anthropology literature (Section 2). Then, in Section 3 we describe the collected data set and the data acquisition campaign we performed to gather this information. Finally, in Section 6 we discuss about how we plan to extend our analysis to a larger scale, and we draw the main conclusions of the paper.

2. Related Work

2.1. Analysis of ego networks in offline social networks

The aim of personal network analysis is to describe the properties of ego networks from the users' standpoint. It studies how the

[1] By the term "offline ego network" we mean an ego network containing social relationships formed "outside" the OSN world.

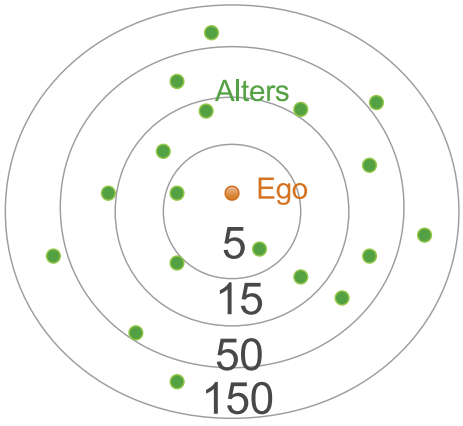

Fig. 1. Concentric hierarchical structure in offline ego networks. Each number represents the average size of the respective circle.

characteristics of ego (e.g., age, gender, personality) affect the properties of her network (e.g., network size, composition, duration and strength of the relationships, ...).

One of the most important bodies of work in this field has been provided by Dunbar, whose findings have also been confirmed by several other research groups. One of the key findings is that maintaining social relationships is costly in terms of cognitive capabilities [23]. There are constraints limiting the number of ties a person can have in an offline ego network [14,7]. This is also in line with well-known models for bounded rationality [24]. Authors of [25,14] demonstrate that these cognitive constraints impose an upper limit on the maximum number of relationships an ego can actively maintain. This limit is about 150 and is often called the *Dunbar's number*. All the relationships exceeding this number are considered inactive and represent mere acquaintances.

Active networks in offline personal networks are characterised by the presence of a series of subgroupings arranged in a hierarchically inclusive sequence, depicted in Fig. 1 [26,14,8]. People socially tied to ego form a series of concentric circles around him, with a scaling factor between any two adjacent circles close to three [12]. The strength of social ties between ego and alters fades as the distance from ego increases. The first circle represents the core network, also called *support clique*. In this group reside all the individuals from whom ego seeks advice in case of severe emotional distress or financial disaster [27]. It is, on average, limited to five members. The other circles are called *sympathy group* (~15 members), *affinity group* (~50 members) and *active network* (~150 members). The frequency of contact (often replaced by the recency of contact, that is the number of days since the last contact) is used as a predictor of social tie strength to define these circles [14]. The groups are identified as the set of alters ego contacts weekly (support clique), monthly (sympathy group), and yearly (active network) [28,23]. The properties of the affinity group, both in terms of typical contact frequency and size, are not yet completely understood (see [23] for an ethnographic definition of the affinity group and of the other layers).

Ego networks are not a complete representation of social networks, as information about how ego networks are connected to each other is missing [29,30]. Nevertheless, information provided by ego network models is already sufficient to characterise many properties of social relationships (basically, all properties that depend on pairwise relationships only), such as willingness of collaboration and sharing resources. Therefore, the results presented before are a reference point for the analysis of ego networks in online environments such as OSN. In this paper we are not in the position to highlight, based on our data set, the presence of similar concentric structures in online environments. However we do find similarities between offline and online ego networks. For example, the average number of active relationships in our data set is 105.14, which is of the same order of magnitude of the active network size in offline social networks (124 in [14] and 132.5 in [12]). Moreover, interaction variables do highlight the existence of strong and weak ties also in our OSN data set. Specifically, the average number of relationships of the latter kind is of the same order of the relationships belonging to the most external circle in the offline ego network model.

2.2. Tie strength and online social networks

As mentioned in Section 1, research on OSNs is particularly active. Several specific topics are being addressed, including, amongst others, information propagation [31], scalability issues [32], social interaction patterns [33], security and privacy [34,35], and design of innovative services based on social relationships [36,37].

The area of research most closely related to this paper is about the characterisation of the strength of social ties in social networks in general, and Online Social Networks in particular. In [6] Granovetter proposes a definition of tie strength based on the combination of the amount of time, the emotional intensity, the intimacy and the reciprocal services which characterise the relationship. The author also identifies a first distinction between the different properties of strong and weak ties, with the former being useful for emotional and financial support and the latter for the acquisition of new ideas coming from groups of people socially far from ego. This distinction has been confirmed by many experiments performed on different types of social networks. In particular, weak ties appear to be crucial to maintain networks structural integrity and strong ties play an important role in the maintenance of local communities [38]. The analysis of network structure alone is not enough to fully understand the properties of social networks in terms of information diffusion, trust and so on [9]. Moreover, ties of intermediate strength seem to be the most efficient channel to spread information in the network, even if the diffusion of information also depends on many other factors [9,38]. All these results indicate the central role of tie strength in understanding social networks dynamics. Even the ego network structures described in Section 2.1 can be described through the different strength of the ties between the ego and the alters in the different circles. Thus, the study of local sources of influence on tie strength should be considered as one of the most important aspects in OSN analysis. A first detailed characterisation of tie strength in offline environment is given in [22], where tie strength is derived using an analytic model. The results of this work evinced the presence of two main dimensions of tie strength, having to do with the time spent in a relationship and the "depth" of the relationship. Moreover, the results indicate that "emotional closeness" or "intensity" of a relationship are the best indicators of tie strength and the frequency of contact only partially explains this concepts.

Although understanding tie strength is essential to study the global dynamics of a social network, many analyses conducted so far on OSN are primarily focused on general aspects of networks seen from a "planetary" perspective [11,39] and they often omit a sufficiently detailed tie strength modelling. For example, in [40] the authors analyse the entire unweighted Facebook graph trying to discover some of its intrinsic properties, such as how information could spread in it. The results presented in [40] indicate that the average distance between any two people in Facebook is 4.74 links. This means that information circulating in Facebook could reach any arbitrary users in, on average, less than five jumps. From this perspective, it seems that Facebook is reducing even further the famous *six degrees of separation*, empirically confirmed in offline social networks by the Milgram's experiment [41]. However,

as these studies do not take tie strength into consideration, the actual behaviour of social aware services might be different. Tie strength is likely to play a very significant role in determining the trust level of a social relationship, and thus the likelihood of that relationship to be used to disseminate information (at least for some class of sensible data). This means that the real number of hops separating two individuals may be higher than what predicted by the analysis of [40]. This is an open point, that the models presented in this paper can contribute to explore. Recently, an analysis of ego network structures in larger Facebook and Twitter data sets (with respect to the one used in this paper) were presented in [42,43], respectively. This analysis is complementary. The data sets used in [42,43] do not include detailed information about all types of interaction between users, which are available in the data set we use here, and the analysis is based on the frequency of interactions only. In this paper we show that, while frequency of interaction is a necessary component for predicting tie strength, additional information related to social interactions is helpful to achieve more accurate predictions.

The possibility to deduce social tie strength from OSN data was previously noted in [20]. Specifically, also [20] uses a Facebook data set and explicit evaluation of tie strength done by the users. Interaction variables are used to fit the explicit evaluation. Although a regression model of tie strength is used also in [20], the main result of that paper is fitting the explicit evaluations of tie strength, rather than predicting them (indeed, the model is not tested on a test set different from the training set, as we do in this paper). A shortcoming of the model in [20] is thus that the model can easily over fit the data, thus reducing its prediction performance. Furthermore, authors of [20] use all possible measured variables as regressors of the model, while in this paper we highlight that a small set of such variables can be used to achieve reliable predictions. This has very important implications in terms of practically applicability of the regression models. Finally, in this paper we also provide a PCA analysis on the Facebook interaction variables, in order to identify the key factors of the data set and how they can be used in the prediction models. This is not done in [20].

The authors of [44] present a study aimed at predicting tie strength from online interactions. They asked a set of participants to indicate the name of their close friends. Hence, they used the collected evaluations to train a classifier to distinguish between strong and weak ties. The classifier gives a membership probability calculated from a set of online interaction variables. This probability represents a prediction of tie strength. Since the proposed is based on evaluations of close friendships only, it is less accurate in the prediction of weak ties. This represents a strong limitation, since weak ties form about $60 - -80\%$ of an ego network [45].

This paper extends our initial work on this topic, presented in [46,47]. The present work introduces a more refined initial filtering of raw data to exclude biased data. In addition, in this work we also study the key factors emerging from the measured variables that explain tie strength, comparing them with tie strength dimensions found in offline ego networks. We also present a set of regression models able to predict tie strength from Facebook data and we present more statistics concerning Facebook ego networks properties.

3. Data set description

Before presenting our analysis, trying to clearly set the basis from which we obtained our results (as also recently advocated in [48]) we provide a description of the data set we have used.

We consider a large number of variables obtained through the use of a Facebook web application (completely described in [46]), created for this particular task. We select all the variables related to users' profiles and social relationships between individuals in Facebook. The rationale was to select a rather broad set of variables describing overall properties of ego networks and the interactions between egos and their alters, intuitively related with the properties of the tie strength, and then use statistical analysis to identify the variables that better describe and predict it. These variables are listed in Table 1. For the sake of clarity, we divide the collected variables into two distinct groups, treating them separately in the next steps of the analysis. The first category contains variables related to user's properties (e.g., age, gender, ...) usually defined as *socio-demographic* variables. The second group contains variables related to the interactions between ego and alters (e.g., number of exchanged messages, number of Facebook groups in common, number of likes made, ...), labelled as *relational* variables. In addition to the quantities concerning ego's characteristics, the first group contains also the total amount and the mean values of each *relational* variable (e.g., total and mean number of messages, posts and other quantities received or sent by ego). These variables should contribute to describe the behaviour of a user in Facebook. For example, the mean number of comments sent per alter can be seen as a descriptor of how much a person uses Facebook.

We exclude from the analysis all the user-filled fields available on Facebook profiles (e.g., political view, religion, hometown, education, ...). This choice was mainly driven by the fact that many of these fields are intentionally left blank by the users. Moreover, we think that the information contained within these variables is difficult to correctly interpret through automatic tools because people are prone to use sarcastic phrases or provoking words that cannot be easily interpreted. We argue that neither using a simple good/bad words [20] count (for *socio-demographic* variables) nor an index denoting the number of overlapping words between pairs of individuals (for *relational* variables) are enough to fully interpret people's social attitudes in OSN. Extracting information that can help to assess the type of social relationships from this fields also require to take care of cultural aspects and differences between the users, discern humorous and paradoxical expressions, and so on. While this is an exciting subject to explore, we prefer not to include it in our current analysis, and leave it for future improvements. Focusing exclusively on quantitative measurable variables, which do not require particularly refined interpretation, allows us to reduce the complexity of the analysis, and understand

Table 1
Facebook variables chosen as possible descriptors of ego networks characteristics.

Socio-demographic variables
Gender
Number of friends
Total number of status updates
Sum of each relational variable – all alters
Mean value of each relational variable – all alters
Mean value of each relational variable – active alters
Relational variables
Number of likes[a]
Number of posts[a]
Number of comments[a]
Number of private messages[a]
Number of tags on the same pictures
Number of days since first communication[b]
Number of days since last communication[b]
Number of events attended together
Number of groups in common
Number of likes on the same fan pages
Frequency of contact[a]

[a] From ego to each alter and vice versa.
[b] From ego to each alter and vice versa divided for Each type of communication (likes, posts, comments, messages).

to what extent those variables alone can be used to predict tie strength, without having to deal with possible inaccuracies and errors in interpreting the real semantic of user-filled fields. For sure, we can anticipate that including also this information – after a correct processing – will further increase the prediction accuracy of our models.

The application we used to gather the variables listed in Table 1, in addition to download Facebook data, has been augmented with an electronic survey, aimed at collecting the values of tie strength perceived by the user towards all her Facebook friends. The survey asks the user to evaluate the strength of her Facebook friendships with the following question: "How do you rate, with a value between 0 and 100, the social relationship between you and this person in Facebook?". The question is also supported by additional context information to make the users effectively express their perception of tie strength. Given the limited number of users involved in the study, it was possible to explain to them in detail the background of our study, and the purpose of the application. They were been made aware of the concept of tie strength, of different types of social relationship and the typical way used in the anthropology literature to quantify it. We used typical questions used in Dunbar's studies as examples of how they should evaluate social relationships. The qualifier "in Facebook" was also clearly explained to them. Specifically, we asked them to evaluate social relationships considering only their activity and interactions in Facebook, thus disregarding any other interactions occurring with their friends "offline". Finally, we selected a numerical range between 0 and 100. This proved to be a natural evaluation scale, once the context of the evaluation was made clear to the users. We are currently working on more automatic ways to provide similar information to the users without direct interaction, in order to collect data on a larger scale.

We use the collected evaluations of tie strength as "ground truth" to compare and calibrate our prediction models. Note that the ultimate goal of our prediction models is to avoid to ask explicit tie strength values to the users. We had to include the survey in the current data set collection, in order to calibrate the models.

We performed a data acquisition campaign over a period of three weeks, during which we selected a sample of 30 people randomly chosen within our research department. We collected all the Facebook data coming from the 30 participants, along with all the data concerning their friends, for a total of 7665 Facebook social relationships. For every participant we downloaded their Facebook data within a temporal window of three years.

From the data set of 30 participants, which is the same used in [46], we discarded two participants, because they provided only a partial evaluation of the tie strength related to their friendships (i.e., they have not completed the survey). Hence, the data set we analyse in this work is composed of 28 participants and 7103 relationships. For the figures previously analysed in [46], the refinement introduced discarding the data from the other two participants allowed us to obtain more significant and reliable results.

Although this data set needs to be enlarged in terms of number of users, it already contains a significant number of samples of social relationships to start deriving interesting results, both for ego network analysis and for tie strength prediction. In particular, this data set allows us to carry out a sensible analysis about the factors that characterise the virtual relationships, and a well grounded regression analysis to estimate the social tie strength.

For the purpose of the analysis of tie strength prediction presented in Section 5, it is worth noting that the participants with more friends could possibly have a higher impact on the determination of the coefficients of the models. To avoid this limitation, we sub-sampled 100 times our data set randomly extracting the same number of friends (100) for each participant. As far as the participants with less than 100 friends, we took all their social relationships. Hence, we come up with a set of 100 data sets that will be used in the analysis described in Section 5, where we also explain how results coming from each of the 100 data sets have been combined. Note that we did not do any sub sampling for the analysis presented in Section 4. This analysis looks, among other, to overall properties of ego networks, such as their size, and thus sub sampling would have significantly biased our results.

4. Properties of ego networks

To assess the properties of Facebook ego networks in our sample and to compare them with the characteristics found in offline ego networks, we analyse the descriptive statistics of our data. We divide this analysis in two different parts, the first related to *socio-demographic* variables and the second concerning *relational* variables.

4.1. Socio-demographic variables

As far as *socio-demographic* variables, the 28 participants within our sample are researchers, Ph.D students or master students from 24 to 48 years old ($M = 32.86$, $SD = 6.77$), 14 males and 13 females. The number of friends of each participants ranges between 86 and 1099 ($M = 253.68$, $SD = 204.14$). The distribution of this variable is shown in Fig. 2. In the figure an outlier can be clearly identified. We do not discard this outlier from the analysis, since from now on we are going to consider only the *active* part of the networks, and in such part the mentioned ego is not an outlier. According to [14,26,8,23,49] each person can maintain only a limited number of "active" social relationships, i.e., relationships for which the person invests resources (e.g., time dedicated to communicate with the alter). Other relationships, for which a person does not invest resources, are only acquaintances, and are considered "inactive". To distinguish between active and inactive social relationships in our analysis we define as "active network" the set of friends for whom the value of tie strength indicated by the user is greater than 0. This definition differs from that given by Dunbar [26], which is based on the frequency of contact (formally the set of people contacted at least once a year). We choose this definition as in our data set we consider many more variables other than frequency of contacts. Using the explicit evaluation of the users to discriminate active from inactive relationships is a way to compactly consider all variables altogether, without giving higher importance to any of them a priori. Note that this methodology was explained to the users, so that they knew that giving a score of 0 to a relationship would mean marking it as inactive.

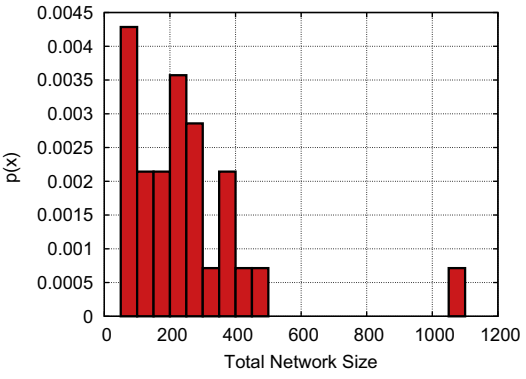

Fig. 2. Distribution of the number of Facebook friends.

Active network sizes in our sample range between 29 and 368 ($M = 105.14, SD = 85.42$). We found that people have, on average, 45.88% of their Facebook friends that can be considered active, with a 95% confidence interval equal to $(38.99\%, 54.77\%)$. The distribution of active network size of our sample, depicted in Fig. 3, is qualitatively similar to those found in other work about offline social networks [8,14]. Moreover, the mean active network size is also comparable to the same measure in offline social networks (e.g., 124 in [14] and 132.5 in [12]). This suggests that a maximum number of active relationships in the order of the Dunbar's number could hold also in OSN.

In Fig. 4 we depict the distribution of tie strength of our sample, considering active networks only (i.e., tie strength greater than zero). In the figure we show the tie strength density for each ego network divided in ten different bins of ten units of tie strength each, then averaged for all the ego networks. The shape of tie strength distribution indicates the presence of a small set of alters tightly connected to ego and a larger number of people loosely coupled with her. This is in accordance with the findings about offline ego networks [25,7,6].

4.2. Relational variables

In this section we present the descriptive statistics of the *relational* variables listed in Table 1. We also consider – for each user – the total amount and the mean values of these variables, that describe the behaviour of egos in terms of the amount of information they exchange or they have in common with others. We calculate these quantities both for all Facebook friends of a user and also for active friends only. Whilst all the other variables are self-explanatory, the concept of "like" needs to be discussed before continuing with variables description. Like-based communication relies on the "like" mechanism. Likes are a special kind of marks left on Facebook objects (e.g., pictures, comments, status updates, ...), used to give a favourable feedback towards these objects.

As indicated by the statistics reported in Table 2, people in our sample make, on average, broadly the same number of comments and likes in Facebook. This is in accordance with the results in [50] and highlights the growing importance of new kind of communications in OSN. The average number of posts sent by egos is higher than the number of likes and comments. The *number of days since first outgoing/incoming communication* give us an estimation of the mean duration of the social relationships we have considered. This duration is, on average, between 3 years and 284 days and 3 years and 175 days. This result tells us that we considered a sufficiently large temporal window. The *time since last outgoing/incoming* communication tells us how recently, on average, people have been contacted on Facebook by their friends and we will call this measure the *recency of communication*. This measure has been used in literature as an estimator of tie strength and we will see in Section 5 that this variable proves to be a good predictor of the frequency of contact between people and plays a central role in the prediction of tie strength.

The statistics presented in Table 2 indicate that all the variables representing the incoming communication received by ego from alters (i.e., the number of likes, comments, posts, ...) take values considerably higher than the variables concerning the outgoing communication made by ego to alters. This is in accordance with the findings in [49], where the authors indicate that, in Facebook, a person has a limited number of friends with whom she directly communicates and a much larger portion of people from whom she only passively receive and consume information, without reciprocating their interactions. The incoming and the outgoing communication could thus have different roles in the prediction of tie strength. We will verify this result in Section 5.

As expected, the statistics concerning the active network are always greater than the values calculated on the entire ego network (apart from the *time since last contact*). This confirms that between the ego and her active friends – defined by the tie strength – there is much more activity on Facebook than between the ego and the entire set of her friendships.

To extract additional information from the relational variables, we also analysed their complementary cumulative distribution functions (hereinafter CCDF). The typical pattern we found is that of a long tail shape. We provide one example, in Fig. 5, related to the frequency of contact. The plots for the other variables are similar, and provided for completeness in B. We have obtained the percentiles indicated in the distribution for each user, and averaged them over all users. The frequency of contact shows a distribution similar to that found in offline ego networks [14] and the shapes of these variables (see also B) is similar to the one of the tie strength in Fig. 4. This is an initial indication that this set of variables should be suitable to predict tie strength. This is validated in detail in Section 5.

Facebook variables' long tail shape indicates the presence of a large set of friends with whom egos have little communication or with whom they have little things in common (i.e., groups, pictures, ...), and a small set of alters with whom egos have a strong interaction. The "elbow" that can be noticed in the curve depicted in Fig. 5 indicates a clear distinction between these groups of "weak" and "strong" ties. Analysing the sizes of these groups, calculated for each distribution of the variables in the data set, we find that the percentage of strong ties in the ego networks of our sample is, on average, 23.53% of the total number of social relationships (i.e., 59.69 over 253.68) considering all the interactions, and 40.09% of the total number of active relationships (i.e., 42.15 over 105.14). Under the hypothesis (yet to be verified) that structures similar to those found in offline social networks are also

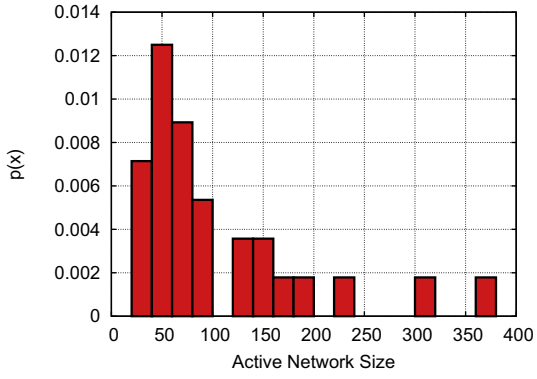

Fig. 3. Active network size distribution.

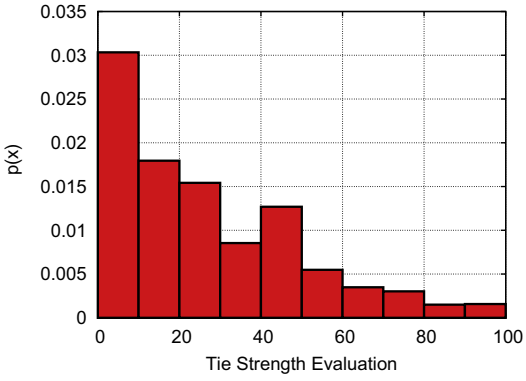

Fig. 4. Tie strength distribution.

Table 2
Descriptive statistics of the relational variables. The first column contains variables' names. The second and the third columns contain the minimum and maximum values of the variables, considering all the possible relationships of all the users. The columns labelled "Average – All friends" show the statistics of the variables normalised dividing them by the total number of friends for each user and than averaged for all the users. The columns "Average – Active friends" present the statistics of the variables calculated on each user, then normalised dividing by the number of active friends (of the considered user) and then averaged for all the users. The last columns show the statistics related to the sum of the variables over all the relationships of each user, then averaged for all the users.

Variable	min	max	Average values – all friends				Average values – active friends				Sum – all friends			
			min	max	mean	95% c.i.	min	max	mean	95% c.i.	min	max	mean	95% c.i.
# of likes sent	0	96	.006	2.95	.82	(.52, 1.12)	.017	5.8	1.59	(1.04, 2.15)	2	968	171.25	(87.47, 255.03)
# of likes received	0	97	.015	2.86	1.03	(.71, 1.35)	.018	7.29	2.17	(1.48, 2.86)	3	2486	304.64	(119.16, 49.13)
# of posts sent	0	144	0.20	7.23	1.67	(.98, 2.35)	.28	9.87	2.73	(1.83, 3.64)	12	1626	386.86	(217.09, 556.62)
# of posts received	0	540	0	9.68	3.12	(2.08, 4.54)	0	24.3	6.75	(4.12, 9.34)	0	3795	889.46	(481.20, 1297.73)
# of comments sent	0	124	.032	2.57	.82	(.54, 1.11)	.12	6.09	1.67	(1.15, 2.19)	8	590	16.07	(102.34, 217.80)
# of received comments	0	124	.061	3.88	1.09	(.70, 1.50)	.20	12	2.37	(1.39, 3.35)	12	2168	304.86	(132.56, 477.16)
# of ego's pictures in which alters appear	0	63	0	.61	.12	(.06, .18)	0	1.86	.28	(.13, .44)	0	177	32.68	(13.28, 52.08)
# of friend's pictures in which ego appear	0	195	0	1.81	.32	(.15, .48)	0	4.86	.84	(.35, 1.33)	0	246	71.32	(42.68, 99.96)
# of pictures in which ego and alters appear together	0	87	0	1.63	.44	(.28, .61)	0	4.49	.99	(.53, 1.45)	0	564	112.68	(63.90, 161.46)
# of days since first outgoing communication	6	1761	567	1381	169	(1265, 1497)	253	1601	120	(1077, 1324)	7342	483 265	122 964	(83 676, 162 252)
# of days since first incoming communication	8	1555	888	1555	1281	(1214, 1348)	636	1555	1095	(997, 1195)	76 445	1508 461	32 433	(21 999, 43 166)
# of days since last outgoing communication	0	1241	3	365	121	(83, 158)	4	489	191	(148, 234)	159	123 163	27 864	(17 443, 38 285)
# of days since last incoming communication	0	1154	0	277	109	(80, 139)	0	464	165	(139, 242)	0	110 732	29 176	(17 642, 40 710)
# of events attended together	0	12	.015	.69	.22	(.15, .28)	0	1.15	.32	(.20, .43)	0	177	34.82	(19.77, 49.88)
# of groups in common	0	62	.07	7.87	1.70	(1.04, 2.35)	.05	12.58	2.29	(1.34, 3.23)	4	8648	692.39	(59.75, 1325.04)
# of pages in common	0	158	.32	14.67	2.32	(1.18, 3.45)	.36	14.03	2.91	(1.68, 4.14)	26	7578	706.86	(156.47, 1257.24)

present in online social networks, this result indicates that relational variables would discriminate relationships in the external part of the active network from the stronger ones, in the more internal layers of ego networks (the first three layers).

5. Models for the prediction of tie strength

In this section we present the models we use to predict tie strength, and we assess their accuracy. We first present a preliminary correlation analysis between the relational variables and the tie strength explicit values, then we present our regression models using uncorrelated variables and PCA factors, respectively. The rationale of each modelling approach is presented at the beginning of the corresponding sections. We decided to adopt a linear approach since we want to maintain our models as simple as possible. In addition, our choice was motivated by the background work in sociology, where tie strength is considered to be a mostly linear combination of social factors.

All the steps we are going to take later require all the variables to be normally distributed and standardised. Thus, we log-transform the variables having absolute value of skewness and kurtosis greater than 1 [51] and we standardise all of them.

As already pointed out in Section 3, we have divided our data set into 100 different sets, sub-sampling 100 relationships from each of the 28 egos in our data set (and including all relationships for egos with less than 100 friendships). This is to avoid that people with more Facebook friends affect the results of our analysis. Therefore, to obtain average results from the different data sets, we apply the techniques described in the following to each sub-sampled data set and then we average the obtained results. This allows us to study the average correlation of each Facebook variable with respect to tie strength and also to obtain statistically solid regression models, that can then be used to predict tie strength.

To train and validate the accuracy of the models we define a training and test set out of our data as follows. We split each of the 100 sub-sampled data sets into a training set containing 23 randomly selected ego networks and a test set with the remaining 5 ego networks. To prevent the results to be influenced by a particular combination of these sets we create five different pairs of training and test sets for each of the 100 sub-sampled data sets. Then we fit a regression model for each of the resulting 500 training sets (formed of 100 sub-sampled data set for each of the 5 different combinations of training sets) and we derive an overall model by averaging the coefficient of all the obtained 500 linear regression models. Hence, we evaluate the accuracy of the obtained model applying it on the test sets, making a comparison of the output of the model and the explicit evaluations of tie strength contained in the test sets. The accuracy results presented in the following are averaged over the 500 test sets (100 sub-samples for each of the 5 combinations of test sets).

Since our data set contains some variables that could be highly correlated between each other (e.g., the number of posts sent by ego and the number of comments received from alters and many others), linear regression could be affected by multicollinearity. Multicollinearity represents a near exact relationship between two or more variables [52], which can impact on the accuracy and correctness of the regression model. Specifically, linear regression could force the sign of the regression coefficients to be different from the sign of the correlation between the respective variables and tie strength, invalidating the correctness of the results. To avoid this problem we follow two different ways. On the one hand we calculate the correlation between all the combinations of pairs of variables and we create a regression model using uncorrelated variables only (thus excluding the sources of multicollinearity)[2]. On the other hand we use Principal Component Analysis (PCA) to extract a set of uncorrelated factors from the data set

[2] Using correlation to select the regressors in the models allows to obtain results that can be easily interpreted and reduces as much as possible the number of regressors of the models. Nevertheless, we also used stepwise regression to select the best combination of regressors in the models and the accuracy of the obtained models is of the same order of that found using correlation.

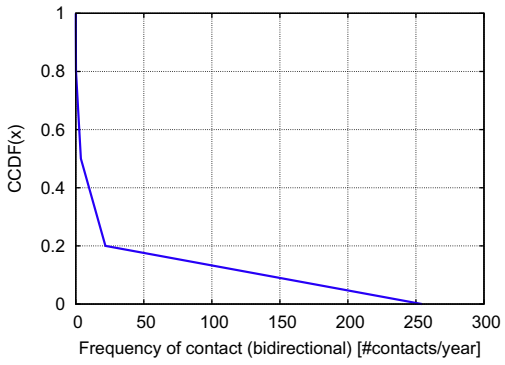

Fig. 5. CCDF of the frequency of contact (bidirectional) between ego and alters.

and we use them to create a tie strength predictive model. We present the results obtained in the two cases in Sections 5.2 and 5.3 respectively, after presenting the initial correlation analysis in Section 5.1.

5.1. Correlation between facebook variables and tie strength

We study the correlation between each variable in the data set and the evaluations of tie strength provided by the users. We use the Pearson product-moment correlation coefficient, described in greater details in A.

For *socio-demographic* variables, the correlation analysis indicates that the average tie strength of each ego network is significantly correlated with the mean bidirectional frequency of contact ($r = .474, p < .01$), the mean number of comments made by ego to her alters ($r = .418, p < .05$), the mean number of days since last communication from ego to alters ($r = -.485, p < .01$), the mean number of days since first communication from ego to her alters ($r = .376, p < .05$), the mean number of days since last communication received by ego ($r = -.473, p < .05$), the mean number of likes made by ego to her alters ($r = .476, p < .05$) and the mean number of groups in common between ego and alters ($r = .379, p < .05$). In our sample, age does not influence tie strength. This result, in contrast with [13], could be explained by the fact that we considered a quite homogeneous sample, with a narrow age difference, which could not be enough to catch the influence of age on social relationships.

As far as the *relational* variables, we calculate their correlation with tie strength for the 100 different sub-sampled data sets, averaging the obtained values for all the different data sets. The correlation values, ordered from the highest to the lowest, are reported in Table 3. In the table we omit the p-values related to the correlation, since they all satisfy $p < .01$. The variables showing the strongest correlation with tie strength are the *number of days since last communication*, the *frequency of contact* (bidirectional and related to incoming interactions) and the *number of days since first communication*. The first of these variables, representing the recency of communication, has been used in previous work as an estimator of the frequency of contact between individuals and as a tie strength estimate [14]. The correlation between Facebook variables and tie strength provides a first indication of the feasibility of the creation of a tie strength predictive model.

Table 3
Correlation between Facebook variables and tie strength.

#	Variable	r
1	Number of days since last comm.	−.56
2	Bidirectional frequency of contact	.55
3	Number of days since first comm.	.51
4	Frequency of incoming comm.	.50
5	Number of received comments	.47
6	Frequency of outgoing comm.	.44
7	Number of comments sent	.43
8	Number of received posts	.41
9	Number of received private msg	.34
10	Number of posts sent	.33
11	Number of likes sent	.32
12	Number of received likes	.29
13	Number of alters' pictures in which ego appears	.24
14	Number of fan pages in common	.20
15	Number of tags on the same objects	.20
16	Number of groups in common	.20
17	Number of ego's pictures in which alters appear	.17
18	Number of events in common	.14
19	Number of private msg sent	.11

5.2. Model with uncorrelated variables

The first family of models we create to predict tie strength and describe its composition is based on a set of uncorrelated regressors. To build these model we firstly calculate the correlation between all the possible combinations of pairs of variables and we create a set of regressors following an iterative procedure. We start with an empty set of regressors, called R_t, where t indicates the maximum value of correlation any two variables within R_t can have. Hence, we take one variable at a time from those listed in Table 3, according to a descending order – from the most correlated to tie strength to the less correlated – and we add it to R_t if all the correlation values it has with the other regressors already present in R_t are lower than t. Note that when t is equal to 1 all variables are in R_t irrespective of their mutual correlation, while t equal to 0 would result in having in R_t only the variable with the highest correlation with tie strength (i.e., only the variable that is first introduced in R_t). We iterate the procedure until all the variables are processed. Thus, R_t represents a set of uncorrelated regressors at a certain level of pairwise correlation t. For high values of t, variables in R_t are more likely to present multicollinearity, whilst this probability decreases with t. On the other hand, very low t values lead to the exclusion of most of the variables from R_t and thus to less accurate models. To find a good trade-off, we repeat the entire process described so far changing the value of the threshold t from 1 downwards, until the signs of the regressors of the models are all consistent with their value of correlation with tie strength, that indicates that multicollinearity does not exist between the variables in R_t. The corresponding R_t contains the largest possible set of variables that do not present multicollinearity. The described procedure converged at t equal to .4. Then, for each of the 500 sub-sampled training sets, we build a regression model using only the variables in R_t. The statistics of the regressors of the predictive model obtained averaging the 500 models are reported in Table 4 (this model is referred to as "model without pairwise products" in the table), using the same enumeration of Table 3. For each regressor we report the estimate of its weight, the standard error and its p-value.

In addition to the model using this set of regressors, we also considered other benchmark models. The first one is a very simple model used as baseline to assess the validity of the other models. It is a constant model which returns the average score of the evaluations used during the training phase for each possible input. The second model uses the set of uncorrelated regressors identified using the procedure described above and, in addition, it includes all the pairwise products between the regressors. Using pairwise products is a standard technique in regression analysis to improve the fitting introducing a set of simple non-linear terms. Moreover, we create a model using the *recency of communication* as the sole regressor, as this is the variable that correlates most with tie

Table 4
Coefficients of the regression models based on uncorrelated variables.

Variable	Estim.	Std err.	p Value
Model with one regressor			
Intercept	13.168	.376	$< .01$
1	−10.957	.375	$< .01$
Model without pairwise products			
Intercept	13.120	.357	$< .01$
1	−9.004	.379	$< .01$
11	3.798	.373	$< .01$
13	3.394	.384	$< .01$
15	.784	.388	$< .01$
Model with pairwise products			
Intercept	13.192	.376	$< .01$
1	−8.900	.380	$< .01$
11	4.317	.506	$< .01$
13	3.904	.567	$< .01$
15	.254	.419	$< .05$
$1 * 13$	−.621	.381	$< .05$
$1 * 15$	−.326	.226	$< .05$
$11 * 13$	.197	.229	$< .05$
$11 * 15$	.161	.136	$< .01$

strength (see Table 3). Lastly, we report, for completeness, the model with all the variables as regressors. This model, although suffers from multicollinearity and could be heavily overfitted, represents a reference for the other models. The coefficient estimates of the models and the respective standard errors and *p*-values are reported in Table 4. The constant model is omitted from the table since it does not have any coefficients. To not compromise readability we also omit the model with all the variables from the table. The regressors with a *p*-value greater than .05 were excluded from the models since they are not statistically significant. The standard errors indicate that the coefficient estimates are sufficiently reliable and the small *p*-values indicate their statistical significance.

For each model we compute the standard indices R^2 and the estimated standard error. Then we use each model on the test sets, computing the *rmse*. These indices allow us to understand how well the models fit the data set and their prediction accuracy on the test set. Specifically, The R^2 index indicates how much variance of the tie strength in the training set the models are able to explain. The more this measure is close to 1, the more the model is able to correctly approximate all the different values of tie strength. On the contrary, a low value of R^2 indicates that the predictions made by the model could be centred on an average value and the model is not able to capture the entire variability of the tie strength. The estimated standard error is the average value of the error made by the model while fitting the training set. The *rmse* measures the mean error made by the model during the prediction phase and is calculated comparing the output of the model and the reference values in the test set (i.e., the tie strength explicitly evaluated by the users). For a precise definition of these indices see Appendix A.

The results of the models are reported in Table 5. The first model, by definition, has a R^2 equal to 0, since it is centred on the average value of the scores in the training set. The estimated standard error and the *rmse* of the model are about 21% and should be considered as worst cases to test the effectiveness of the other models. Even though these values seem adequate for a model aimed at predicting tie strength between people, the model is not able to fit all the different values of tie strength (because of the null R^2) and fails to reproduce the typical tie strength distribution of social ego networks [45]. The model with only one regressor is able to explain 27.2% of the variance of the tie strength in the data set, according to its R^2. This represents a rather good result, considering that we are using only one variable to estimate the tie strength, that is influenced by many different sociological and psychological factors. Note that R^2 values around .3 are generally considered rather good results (e.g., [14]). Nevertheless, the remaining part of variance of the tie strength not explained by this first model is still large and this could limit the ability of the model to effectively predict all the different values of tie strength. The estimated standard error of the model is equal to 18%. This means that the model, on average, is able to fit the training set with a good accuracy. The *rmse* of the model is really close to the estimated standard error. This is a good result, since indicates that the average error made on the test set has the same magnitude of the error made on the training set. Hence, the model seems not to be affected by overfitting and remains valid also when applied to data other than that used to train it. The model with the addition of the other uncorrelated variables to the *recency of communication* shows an improvement in terms of all the presented indices. Even if the improvements in terms of estimated standard error and *rmse* are only .9% and .7% respectively, the R^2 is 7.3% higher than that of the model with one regressor. This improvement in terms of R^2 makes us prefer this model to the previous one, since it is more accurate in the fitting of all the different values of tie strength. The model with the introduction of the pairwise products of the variables does not bring a noticeable increment in terms of R^2 and *rmse* to justify its higher complexity – represented by the higher number of regressors. Lastly, the model with all the variables as regressors shows the best performances, but, as stated before, it suffers from multicollinearity.

Table 5
Statistics of the regression models based on uncorrelated variables.

Statistics of the models			
Model	R^2	stderr	rmse
Average value	0	.211	.219
One regressor	.272	.180	.184
Uncorrelated w/o pairwise prod.	.345	.171	.177
Uncorrelated with pairwise prod.	.350	.170	.177
All regressors	.454	.156	.165

Fig. 6 compares more in detail the predictions made by the models with respect to the explicit evaluations of tie strength in the training sets. Specifically, each curve in the plot shows the probability with which a given value of tie strength is predicted by the corresponding model, or appears in the explicit evaluations of tie strength. In other words, the curves show the empirical distributions of tie strength in the data set, and produced by the models. This provides a more detailed comparison with respect to the *rmse* index, which is essentially an average accuracy index.

We do not give a graphical representation of the results of the model with the addition of pairwise products of the variables since the curve is really close to that related to the model without pairwise interactions. We also omit the constant model from the figure, since its density function is a Dirac delta function centred in the average of the evaluations (i.e., 13.216). From the graphical representation in Fig. 6 we can notice that the predictions made by the models using Facebook variables as predictors have similar distributions. Both models tend to overestimate tie strength when it is close to zero, since the density of the predictions is lower than the reference at zero and higher between 5 and 30. We find that the data set is noisy in this particular region and this is likely to be the main reason for this inaccuracy. On the other hand, the models tend to overestimate tie strength for high values of tie strength. This prediction inaccuracy is likely to be due to the presence of few samples with a high value of tie strength. The both types of inaccuracy are likely to be mitigated using a larger data set. Nevertheless, Fig. 6 qualitatively confirms that the estimations made by the models are in line with the explicit values.

The results described so far indicate that the models we have obtained effectively predict tie strength using only a small set of Facebook variables (i.e., 4 in the model with all the selected uncorrelated variables). The first model, with only the *number of days since last communication* as regressor, already provides good prediction accuracy, confirming that this variable is a good predictor of tie strength. Nevertheless, using additional variables (i.e., the other regressors in the second model) provides even more accurate predictions.

5.3. Model with PCA factors

As a second approach to build a model to predict tie strength, avoiding multicollinearity at the same time, we apply PCA on the 100 sub-sampled data sets to obtain a set of orthogonal variables to be used as regressors. PCA is a standard technique that transforms a set of possibly correlated variables into a set of uncorrelated factors, obtained as linear combinations of the original variables. The results of PCA are presented here in terms of factor loadings, that is to say the weights to be given to each original variable to obtain the factors themselves. Further details on PCA technique and a detailed description of the meaning of factor loadings and of other properties of the factors are given in A. We use the obtained factors to build a regression model for each of the 500 training sets (100 sub-samples for each of the 5 combinations of training sets). Hence, we average the 500 obtained models obtaining a unique average model. Afterwards, we test the predictive power of this average model on the different test sets.

The results obtained using PCA, expressed in terms of factor loadings, are reported in Table 6. The numbering of the variables presented in the table is the same used in Table 3. We analyse only the first 5 factors since they are the only ones with eigenvalue greater than 1, as suggested in [53]. In essence, we drop all the factors with eigenvalue lower than one since they extract less than the equivalent of one original variable. Nevertheless, before putting aside the less important factors (in terms of explained variance) we study the correlation between all the factors extracted and the tie strength. We find that the first five factors, in addition to be the factors that explain the largest portion of variance of the data set (individually), are also the most correlated with tie strength. All the other factors show low values of correlation (i.e., below .1) or their p-value stays above .05. This means that they have a meaningless relation with tie strength.

Before continuing with the analysis and with the creation of a predictive model using the obtained factors, we characterise the physical meaning of each factor, based on the variables that determine it and their factor loadings (see Table 6). Doing so we can have a first broad idea on which is the nature of the principal dimensions contained in our data set and we can also sketch a preliminary comparison on the differences between these dimensions and those hypothesised by Granovetter in [6], as the most important tie strength dimensions.

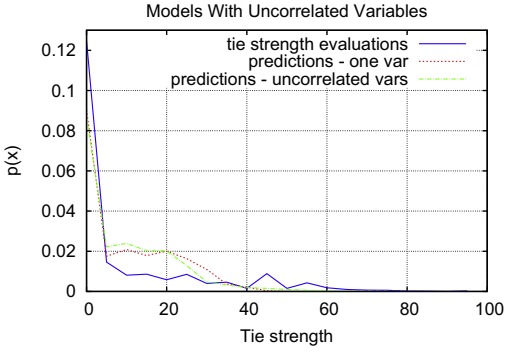

Fig. 6. Distributions of the expected values of tie strength compared to the predictions made by the models built using uncorrelated variables.

Table 6
PCA Factor Loadings. For each factor, loadings greater than .3 are in bold, to mark the variables that have a significant impact on the factor.

Var	Factor				
	I	II	III	IV	V
1	−.**77**	−.29	−.07	−.17	−.12
2	.**88**	.28	.09	.17	.11
3	.**82**	.18	.10	.08	.03
4	.**90**	.08	.09	.16	.14
5	.**44**	.28	.22	.29	.30
6	.**32**	.**70**	.07	.22	.13
7	.21	.**61**	.12	.**31**	.28
8	.**78**	.02	.18	.08	.07
9	.**54**	.10	−.14	.24	.26
10	.**35**	.**51**	.10	.12	.06
11	.04	.**56**	.14	.29	.24
12	.22	.23	.24	.29	.30
13	.16	.05	.22	.30	.**34**
14	.06	.26	.15	.**35**	.**49**
15	.10	.05	.**77**	.07	.05
16	.10	.09	.**79**	.05	.03
17	.08	.19	−.03	.29	.**45**
18	.06	.05	.**46**	.13	.12
19	.01	.**37**	−.09	.20	.21

The first two factors contain all the variables related to the communication between people, like the *frequency of contact*, the *time since last/first communication*, the *number of likes/posts/msg/...* sent or received by egos. These factors are related to the time dedicated by two individuals to the social relationship that ties them together. Moreover, the factors embody the intensity of the communication related to the relationship. We can call these factors "communication factors". The first factor embodies the incoming communication and the overlap between the incoming and the outgoing communication (i.e., the incoming communication reciprocated by ego). The second factor contains the portion of outgoing communication not already contained in the first factor, that is to say the outgoing communication not reciprocated by alters. Although this requires better investigation, the fact that outgoing communication is split between the first two factors is likely to be the reason why they are uncorrelated. In general, incoming and outgoing communications are correlated, although less than expected. The correlation between the two types of communications in our data set is .33 ($p. < .01$). This fact is induced by the nature of Facebook, that allows people to consume information received by other users, but does not require that these people directly communicate with those specific users, as previously described by the authors of [49].

The third factor is a combination of the *number of groups* and *events* in common and the *number of tags on the same objects*. This factor represents how similar two Facebook profiles are and we call it "social similarity factor". The last two factors share broadly the same variables and we hypothesise that they could be related to the intimacy and the emotional intensity of a relationship, since they also contain the number of pictures in which two users appear together, which we hypothesise as an indicator of the emotional affinity of individuals.

Although it is necessary to extend this analysis to a larger number of users, these results suggest that our data set contains broadly the same dimensions hypothesised by Granovetter, even though the presence of variables indicating the intimacy and the reciprocal services in our data still remain as an hypothesis.

Using the factor scores obtained using the PCA we create three different regression models for tie strength prediction. The first

model uses only the first PCA factor as regressor (the factor with the strongest correlation with tie strength). Then we create a second model using all the five PCA factors and a third model with all the factors and their pairwise products. We report in Table 7 the coefficient estimates of the models along with their respective standard error and their p-values. The regressors with a p-value higher than .05 have been excluded from the models. These statistics indicate that all the regressors reported in the table are significant and their estimates are sufficiently accurate.

The R^2, the estimated standard error and the *rmse* of the models are reported in Table 8. The first model has a noticeably lower value of R^2 compared to the other models and the error it makes during prediction is higher (almost 20%). The second model, instead, shows a good R^2, with a sensible improvement compared to the previous one. Also the *rmse* and the average standard error indicate better performances, not far from the reference model built using all the possible regressors reported in Table 5. The third model introduces an additional improvement in terms of R^2, but its augmented complexity is not supported by a noticeable increment in terms of prediction accuracy. In fact the *rmse* is equal to that of the model without pairwise interactions and the presence of additional regressors could introduce overfitting. Hence, the second model turns out to be the best one, since it is simpler than the third one - maintaining a similar R^2 at the same time - and has a far better R^2 compared to the first model.

Fig. 7 shows a graphical comparison between the distributions of the tie strength explicit evaluations in the test sets and the distributions of the tie strength predicted by the first and the second models, built using the PCA factors as regressors. The predictions made by the first model – with only the first factor as regressor – are not enough accurate, especially between 0 and 30. The model with the five PCA factors shows a good accuracy in the prediction, since the distribution of its output indicatively follows the distribution of the tie strength in the test set, even if the major part of the error is still concentrated between 0 and 25. The same hypothesis on the nature of the error made by the models, already highlighted for the models with uncorrelated variables in Section 5.2, holds also for Fig. 7.

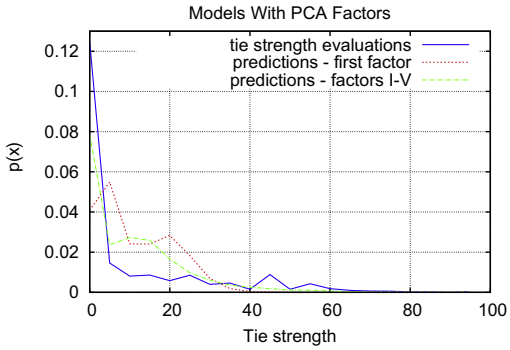

Fig. 7. Distributions of the expected values of tie strength compared to the predictions made by the models built using PCA factors.

Table 7
Statistics of the regression models with PCA factors.

Regressor	estim.	std err.	p value
Model without pairwise products			
Intercept	13.342	.361	$< .01$
Factor I	10.565	.374	$< .01$
Model without pairwise products			
Intercept	13.143	.341	$< .01$
Factor I	9.224	.340	$< .01$
Factor II	5.444	.338	$< .01$
Factor III	4.418	.338	$< .01$
Factor IV	4.669	.339	$< .05$
Factor V	4.080	.339	$< .05$
Model with pairwise products			
Intercept	13.143	.336	$< .01$
Factor I	9.028	.367	$< .01$
Factor II	6.086	.453	$< .01$
Factor III	4.201	.355	$< .01$
Factor IV	5.167	.516	$< .05$
Factor V	4.718	.492	$< .05$
I∗III	.913	.335	$< .05$
II∗III	.530	.254	$< .05$
II∗IV	−.372	.231	$< .01$
II∗V	−.471	.204	$< .05$
IV∗V	−.242	.208	$< .05$

Table 8
Statistics of the regression models with PCA factors.

Model	R^2	residual std err	rmse
First PCA factor	.193	.189	.198
PCA factors I-V	.404	.163	.171
PCA factors I-V + pairwise prod.	.423	.161	.171

5.4. Comparison between the different models

The models described so far have approximately the same predictive power in terms of *rmse* ($M = .180, SD = .010$). This represents a good result, since all the models we have built are able to predict tie strength with an accuracy greater than 80%. The little difference in terms of *rmse* between the different models seems to indicate the model with only the *recency of contact* as regressor as the best choice, since it is the simplest one. Nevertheless, the noticeable difference in terms of R^2 makes us prefer the models built using all the five PCA factors. In fact, the higher value of R^2 (not far from the R^2 of the reference model built with all the variables) makes this model able to better approximate all the possible values of tie strength and assure that the model does not produce always the same average score, but it effectively follows real tie strength distribution. For this reason the model with the five PCA factors is also the most general, and its validity is not limited to our particular data set. A drawback of this model is that it needs all relational variables. In cases where this is not feasible, the model using the four uncorrelated variables is a very good trade off. It is not much more complex than the one using only one regressor (*recency of contact*), and is able to provide higher R^2 and lower *rmse*, although it does not reach the performance of the model with all PCA factors.

It is noteworthy that the most important variable for tie strength prediction remains the *time since last contact* in all the models. Our results also confirm that this variable is a good estimator of the frequency of contact, since it has a very high correlation with the *bidirectional frequency of contact* ($r = -.86, p < .01$). Moreover, it also represents a large portion of the first PCA factor. The *time since last contact* is also really simple to be obtained from Facebook and the model that uses only this variable as predictor requires only a small amount of information. In fact, it is sufficient to download only the last communication record and not the whole history of interactions between the users to obtain the time at which the last contact between two online users occurred.

6. Conclusions and future work

In this paper we have presented a detailed analysis of Facebook ego networks. This analysis has a double aim. On the one hand it provides a fine-grained characterisation of OSN ego networks properties and it studies the relation between these properties

and the tie strength. On the other hand, the analysis is aimed at building a set of models able to predict tie strength from OSN data. To perform this analysis, we downloaded a set of observations of Facebook variables concerning social relationships through a data acquisition campaign. Then we collected the tie strength estimation related to the involved relationships, asking the participants of our experiment to explicitly evaluate their Facebook friendships through an electronic survey. From this data, we find that the properties of Facebook ego networks are compatible with the findings regarding offline ego networks. In particular, the number of active relationships an individual can maintain – found in offline social networks (the well-known "Dunbar number") is comparable with the one we have found in Facebook. Facebook users in our sample have, on average, a maximum number of active friends equal to 105.14. This value falls inside the boundaries hypothesised for offline ego networks [25] and is similar to other active network mean sizes found in human social networks [12,14]. Moreover, the distribution of active network sizes of our data set is similar to those found in offline ego networks [8,14].

To study the composition of tie strength and to predict it using a set of Facebook variables we create a series of regression models dividing the analysis in a first phase dedicated to the training of these models and a second phase in which we test them on another portion of the data set, different from that used for training. To build the predictive models we take two different approaches. On the one hand we select a group of variables not correlated between each other – discarding variables that can lead to multicollinearity – but having a sufficiently high correlation with tie strength, and we create a first regression model. On the other hand we use PCA to extract the principal factors of the data set – that are orthogonal and thus uncorrelated by definition – and we use them to create a second type of predictive model. The PCA also allows us to compare the dimensions of tie strength hypothesised in the seminal work by Granovetter [6] and in [28] with the factors derived from the data set. The results of this analysis suggest that the variables we have collected represent all the dimensions of tie strength as hypothesised in [6].

The regression models perform quite well. They show R^2 indices comparable to other models in literature, namely to that presented in [14], regarding "offline" social networks, and that in [20], as far as online environments. Moreover, the validity of our models is also tested on a test set containing data different from that used to train them. They show good results in terms of prediction accuracy. On average, they achieve accuracy greater than 80%. The best one among them is the one using all the PCA factors as regressors. The main drawback of this model is that it needs to collect all the variables present in our data set, which may pose concerns in terms of practical applicability.

It is noteworthy that the most important regressors of the model, in terms of prediction power, are those implying the recency of communication and the frequency of contact between people involved in a social relationship, already used as tie strength predictors in other studies regarding "offline" social networks analysis [14]. Models using only this variable as predictor (which is similar to what has been typically done in the anthropology literature [14]) perform quite well, although do not match models based on the PCA factors. They are appealing, as they can be implemented by monitoring only one variable, which is a very low cost. An interesting trade off between one-regressor models and full-PCA models is achieved by a model that uses only four uncorrelated variables, and provides better fitting and prediction performance with respect to the model with one regressor. The 4-variable model keeps the complexity at a reasonable level, providing good (although sub-optimal) performance in terms of fitting and prediction accuracy.

In conclusion, our findings indicate that the characteristics of OSN ego networks are not so different from those found in offline ego networks, both in terms of their structure and tie strength composition. This means that, even if OSN like Facebook and Twitter give us many new and different ways to communicate, our social behaviour and our capacity to maintain social relationships with others seem to remain unaltered. These results clearly need further investigation on a much larger data set, more representative of the entire Facebook population, but they still represent a first interesting indication of the similarity existing between offline and online social networks. The tie strength models presented in this paper clearly indicate that a lot of work still need to be done in OSN analysis to fully understand the global "social" properties of the networks. Our models are still preliminary and their performance must be improved, especially in terms of predictive power. Although this, our work demonstrates the feasibility of the creation of a general model for tie strength prediction that could represent the basis for more advanced studies in OSN analysis.

Acknowledgements

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

Appendix A. Methodology to create and evaluate tie strength prediction models

A.1. Correlation analysis

To perform the correlation analysis, we used the sample correlation coefficient r that, given two random variables X and Y, is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y}, \tag{A.1}$$

where $\bar{x}$ and $\bar{y}$ are the sample mean, and S_x and S_y are the sample standard deviations, of X and Y, respectively. r is an estimator of the Pearson product-moment correlation coefficient (also known as Pearson's r), defined as:

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y}, \tag{A.2}$$

where $COV(X,Y)$ is the covariance and σ_X and σ_Y are the standard deviations. r-values are typically presented together with a corresponding value of a parameter p, which describe the significance of the r-value. Small p-values indicate that r is reliable (see [54] for a precise definition of p).

A.2. Multiple linear regression

Regression analysis implies that the relation between a dependent variable Y and one or more independent variables X – also called regressors – can be described with the following linear model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon, \tag{A.3}$$

that is, for a given vector $(x_1, x_2, \ldots, x_n)$, a corresponding observation y consists of the value calculated as a linear combination of the different x_i (where i is between 1 and n) plus an error term ϵ. We say that a model is linear if it is linear in its parameters [55]. The values of the β coefficients are unknown, but they can be estimated from the observed data using, for example, the method of least squares. On the contrary, ϵ changes for each observation and it cannot be

estimated. Thus the model that predicts Y from the regressors X_i is the following:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_n X_n. \quad (A.4)$$

The "hats" above the β signify that those parameters are being estimated. The hat above Y means that the dependent variable is being predicted [52].

We feed the models with the records of our training sets, where each record is formed of a pair $(\hat{X}_{ij}, Y_{ij})$, where $\hat{X}$ is a vector of Facebook variables related to the social relationship between two subjects i (ego) and j (alter) and Y_{ij} is a variable representing the evaluation of tie strength given by i with respect to j.

We fit a model for each of the 5 training sets of the 100 bootstrapped data sets, obtaining a total of 500 different models. Hence, we assess how well the models fit the training sets, calculating the R^2 index for each of them. The R^2 index is a standard measure of the worth of a regression equation and is defined as follows:

$$R^2 = \frac{\sum(\hat{Y} - \overline{Y})^2}{\sum(Y - \overline{Y})^2}, \quad (A.5)$$

where $\hat{Y}$ is the output of the model, Y is the observed variable used to estimate the parameters and $\overline{Y}$ is the mean of the latter. R^2 represents the total amount of variance explained by the model.

To show the predictive power of the models, fitted on the training sets, we apply them on the records contained in the test sets (represented by tuples of the form $(x_1, x_2, \ldots, x_n)$ where each x is an occurrence of a Facebook variable X. Hence we calculate how accurate is the result of each model compared to the expected value in the data set (i.e., the tie strength evaluations in the test sets). To do so, we calculate the root mean squared error (henceforth *rmse*) of each model on its relative test set. The *rmse* indicates the magnitude of the error made by a model while predicting tie strength and is defined as follows:

$$rmse = \sqrt{\frac{\sum(\hat{Y} - Y)^2}{n}}, \quad (A.6)$$

where $\hat{Y}$ is the tie strength predicted by the models, Y is the tie strength in the test set – which we want to predict – and n is the number of relationships in the test set.

A.3. Principal component analysis

PCA is a technique aimed reducing the dimensionality of a data set without eliminating relevant information. Intuitively, PCA groups the original variables in a – typically small – set of groups. Inside each group, variables are strongly correlated with each other. Each group is represented by a unique variable (called principal component), computed as a linear combination of the variables of the group. Principal components of different groups are uncorrelated. The operational details of PCA can be found in [56,51]. The key idea of PCA is therefore to identify a – possibly small – number of orthogonal dimensions (the principal components) of the data set, replacing groups of correlated variables with a unique variable, which is representative of the dimension. PCA is able, given a data set, to highlight the key dimensions represented in the data set. In addition, components of a PCA are automatically ordered. The first component is the one that is most relevant in explaining the overall variance of the data set.

The results of PCA are usually discussed in terms of component loadings and component scores. The former are the weights by which each original variable is multiplied in the linear combination to determine the principal factors. The latter are the values of the original variables multiplied by their respective weight in the linear combination. PCA is useful to identify the principal factors of the data set and therefore we use it to check whether our data set includes all the dimensions of tie strength described by Granovetter in [20]. However, it could be inappropriate to use these factors for prediction, since the factors which explain only few variance might have high predictive relevance. Thus, we calculate the correlation between the PCA factors and the tie strength and we select the factors showing a high correlation index before creating the predictive model.

Appendix B. Additional statistics of the data set

In this Appendix we present the additional descriptive statistics of the *relational* variables of the data set. Specifically, Figs. B.8, B.9, B.10, B.11, B.12, B.13, B.14, B.15, B.16, B.17 depict the CCDF of all the relational variables not already shown in Section 4. The figures related to the communication between ego and alters report both

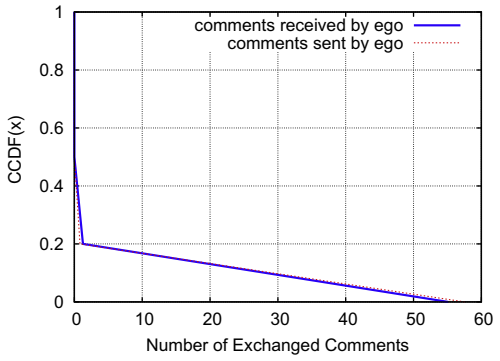

Fig. B.8. Number of Comments.

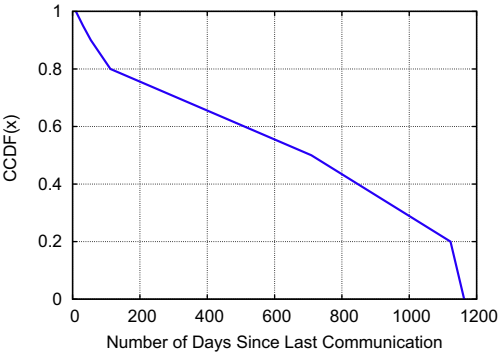

Fig. B.9. Number of days since last communication.

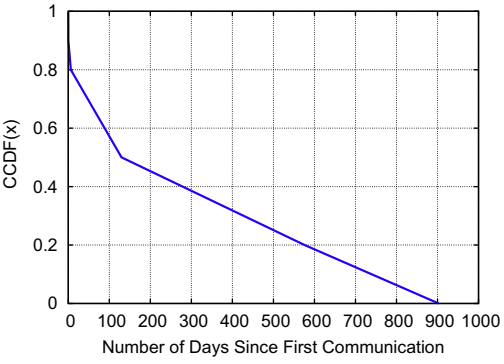

Fig. B.10. Number of days since first communication.

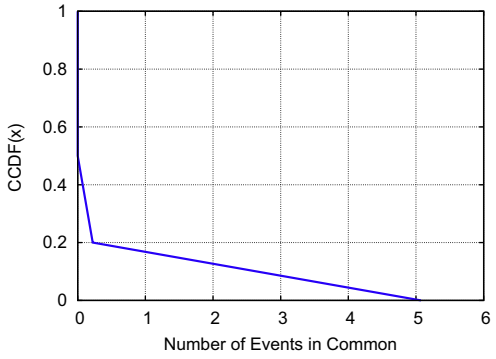

Fig. B.11. Number of events in common.

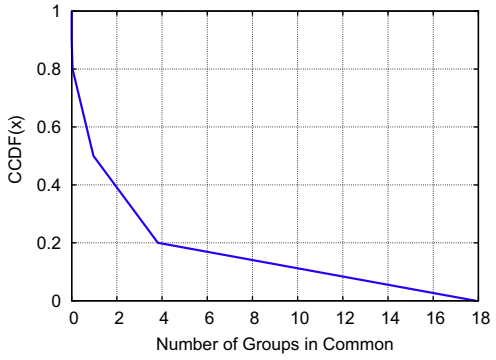

Fig. B.12. Number of groups in common.

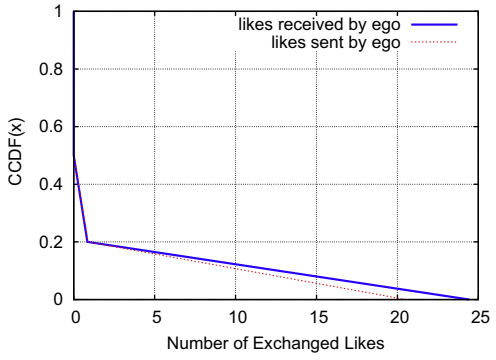

Fig. B.13. Number of likes.

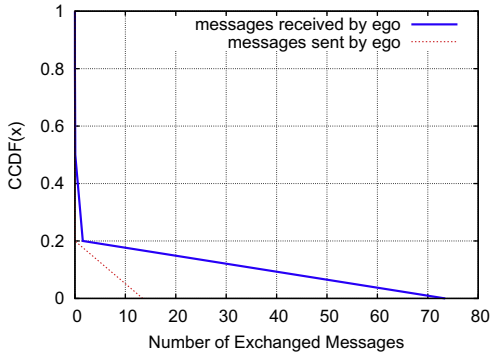

Fig. B.14. Number of messages.

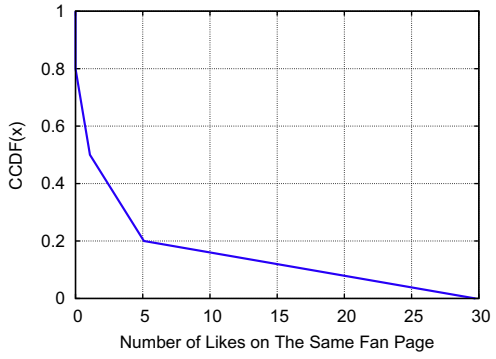

Fig. B.15. Number of fan page in common.

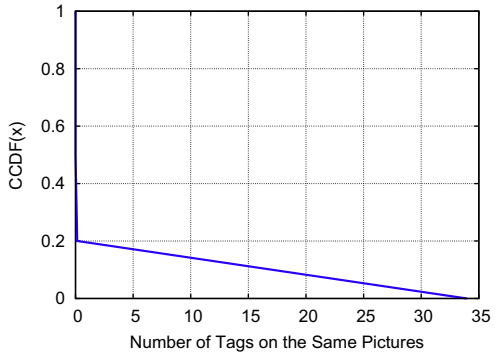

Fig. B.16. Number of tags on the same pictures.

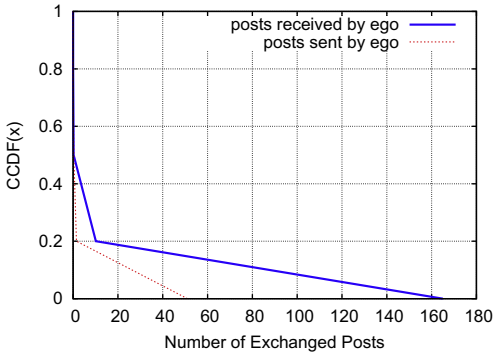

Fig. B.17. Number of posts.

the interactions made by ego to alters and the communication made by alters to ego.

References

[1] M. Conti, S. Chong, S. Fdida, W. Jia, H. Karl, Y.-D. Lin, P. Mähönen, M. Maier, R. Molva, S. Uhlig, M. Zukerman, Research challenges towards the future internet, Computer Communications 34 (2011) 2115–2134.
[2] S. Paul, J. Pan, R. Jain, Architectures for the future networks and the next generation internet: a survey, Computer Communications 34 (2011) 2–42.
[3] D. Trossen, J. Riihijarvi, P. Nikander, P. Jokela, J. KjÃllman, J. Rajahalme, Designing, implementing and evaluating a new internetworking architecture, Computer Communications 35 (2012) 2069–2081.
[4] M. Conti, Special section on mobile opportunistic networking, Pervasive and Mobile Computing 7 (2011) 159.
[5] M. Conti, S.K. Das, C. Bisdikian, M. Kumar, L.M. Ni, A. Passarella, G. Roussos, G. Troester, G. Tsudik, F. Zambonelli, Looking ahead in pervasive computing: challenges and opportunities in the era of cyberâ, physical convergence, Pervasive and Mobile Computing 8 (2012) 2–21.

[6] M.S. Granovetter, The strength of weak ties, The American Journal of Sociology 78 (1973) 1360–1380.
[7] S.G. Roberts, Constraints on Social Networks, in: Social Brain, Distributed Mind, Proceedings of the British Academy, 2010, pp. 115–134.
[8] S.G. Roberts, R.I. Dunbar, T.V. Pollet, T. Kuppens, Exploring variation in active network size: constraints and ego characteristics, Social Networks 31 (2009) 138–146.
[9] P.S. Dodds, R. Muhamad, D.J. Watts, An experimental study of search in global social networks, Science 301 (2003) 827–829.
[10] M.E.J. Newman, D.J. Watts, S.H. Strogatz, Random graph models of social networks, Proceedings of the National Academy of Sciences of the United States of America 99 (Suppl. 1) (2002) 2566–2572.
[11] J. Leskovec, E. Horvitz, Planetary-scale views on an instant-messaging network, Technical Report, 2007.
[12] W.-X. Zhou, D. Sornette, R.A. Hill, R.I.M. Dunbar, Discrete hierarchical organization of social group sizes, in: Biological Sciences, vol. 272, 2005, pp. 439–44.
[13] R.I.M. Dunbar, S. Roberts, Communication in social networks: effects of kinship, Network Size and Emotional Closeness, Personal Relationships 18 (2011) 439–452.
[14] R.A. Hill, R.I.M. Dunbar, Social network size in humans, Human Nature 14 (2003) 53–72.
[15] P.V. Marsden, K.E. Campbell, Measuring tie strength, Social Forces 63 (1984) 482–501.
[16] A. Passarella, A survey on content-centric technologies for the current internet: cdn and p2p solutions, Computer Communications 35 (2012) 1–32.
[17] S. Tang, E. Jaho, I. Stavrakakis, I. Koukoutsidis, P.V. Mieghem, Modeling gossip-based content dissemination and search in distributed networking, Computer Communications 34 (2011) 765–779.
[18] E. Jaho, I. Koukoutsidis, I. Stavrakakis, I. Jaho, Cooperative content replication in networks with autonomous nodes, Computer Communications 35 (2012) 637–647.
[19] P. Van Mieghem, The viral conductance of a network, Computer Communications 35 (2012) 1494–1506.
[20] E. Gilbert, K. Karahalios, Predicting tie strength with social media, in: International Conference on Human Factors in Computing Systems, ACM Press, New York, USA, 2009.
[21] R. Xiang, J. Neville, M. Rogati, Modeling relationship strength in online social networks, in: Proceedings of the 19th International Conference on World Wide Web, 2010, pp. 1–8.
[22] P.V. Marsden, Core discussion networks of americans, American Sociological Review 52 (1987) 122–131.
[23] A. Sutcliffe, R. Dunbar, J. Binder, H. Arrow, Relationships and the social brain: integrating psychological and evolutionary perspectives, British Journal of Psychology 103 (2012) 149–168.
[24] H. Simon, A behavioral model of rational choice, The Quarterly Journal of Economics 69 (2007) 99–118.
[25] R.I.M. Dunbar, The social brain hypothesis, Evolutionary Anthropology 6 (1998) 178–190.
[26] R.I.M. Dunbar, The social brain hypothesis and its implications for social evolution, Annals of Human Biology 36 (1998) 562–572.
[27] R.I.M. Dunbar, M. Spoors, Social networks, Support Cliques and Kinship, Human Nature 6 (1995) 273–290.
[28] O. Curry, R. Dunbar, Why birds of a feather flock together: the effects of similarity on altruism, Personal and Social Relationships (2011), submitted for publication.
[29] M. Conti, A. Passarella, F. Pezzoni, From ego network to social network models, in: Proceedings of the Third ACM International Workshop on Mobile Opportunistic Networks, MobiOpp '12, ACM, New York, NY, USA, 2012, pp. 91–92.
[30] M. Conti, A. Passarella, F. Pezzoni, A model to represent human social relationships in social network graphs, in: Fourth International Conference on Social Informatics, (SocInfo 2012), 2012.
[31] C. Doerr, N. Blenn, S. Tang, P. Van Mieghem, Are friends overrated? a study for the social news aggregator digg.com, Computer Communications 35 (2012) 796–809.
[32] S. Ghosh, A. Srivastava, N. Ganguly, Effects of a soft cut-off on node-degree in the twitter social network, Computer Communications 35 (2012) 784–795.
[33] N. Aharony, W. Pan, C. Ip, I. Khayal, A. Pentland, Social fmri: investigating and shaping social mechanisms in the real world, Pervasive Mobile Computing 7 (2011) 643–659.
[34] M. Önen, T. Strufe, Special section on security and social networking, Computer Communications 35 (2012) 47.
[35] L.M. Aiello, G. Ruffo, Lotusnet: tunable privacy for distributed online social network services, Computer Communications 35 (2012) 75–88.
[36] A. Passarella, R.I. Dunbar, M. Conti, F. Pezzoni, Ego network models for future internet social networking environments, Computer Communications 35 (2012) 2201–2217.
[37] I. Parris, G. Bigwood, T. Henderson, Privacy-enhanced social network routing in opportunistic networks, in: Eighth IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010, pp. 624–629.
[38] J. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A. Barabási, Structure and tie strengths in mobile communication networks, in: National Academy of Sciences of the United States of America, vol. 104, 2007, pp. 7332–7336.
[39] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, S. Vigna, Four degrees of separation, CoRR abs/1111.4, 2011.
[40] J. Ugander, B. Karrer, L. Backstrom, C. Marlow, The anatomy of the facebook social graph, CoRR abs/1111.4 (2011).
[41] J. Travers, S. Milgram, An Experimental Study of the Small World Problem, Sociometry 32 (1969) 425.
[42] V. Arnaboldi, M. Conti, A. Passarella, F. Pezzoni, Analysis of ego network structure in online social networks, in: ASE-IEEE International Conference on Social Computing (SocialCom), 2012.
[43] V. Arnaboldi, M. Conti, A. Passarella, F. Pezzoni, Ego networks in twitter: an experimental analysis, in: The Fifth IEEE International Workshop on Network Science for Communication Networks (NetSciCom 2013), 2013.
[44] J.J. Jones, J.E. Settle, R.M. Bond, C.J. Fariss, C. Marlow, J.H. Fowler, Inferring tie strength from online directed behavior, PLoS One 8 (2013) e52168.
[45] J. Saramäki, E.A. Leicht, E. Lopez, S.G.B. Roberts, F. Reed-Tsochas, R.I.M. Dunbar, The persistence of social signatures in human communication, 2012, pp. 1–16. arXiv:1204.5602.
[46] V. Arnaboldi, A. Passarella, M. Tesconi, D. Gazzè, Towards a characterization of egocentric networks in online social networks, in: OTM Workshops, vol. 7046, 2011, pp. 524–533.
[47] M. La Gala, V. Arnaboldi, M. Conti, A. Passarella, Ego-net digger: a new way to study ego networks in online social networks, in: Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research, HotSocial '12, ACM, New York, NY, USA, 2012, pp. 9–16.
[48] B. Krishnamurthy, W. Willinger, P. Gill, M. Arlitt, A socratic method for validation of measurement-based networking research, Computer Communications 34 (2011) 43–53.
[49] C. Marlow, Maintained relationships on facebook, http://www.overstated.net/2009/03/09/maintained-relationships-on-facebook, 2009
[50] K.N. Hampton, L.S. Goulet, L. Rainie, P. Kristen, Social networking sites and our lives, Technical Report, Pew Internet & American Life Project, 2011.
[51] A.L. Comrey, H.B. Lee, A First Course in Factor Analysis, vol. 2, Lawrence Earlbaum Associates, 1992.
[52] T.P. Ryan, Modern Regression Methods, vol. 39, Wiley, 1997.
[53] H. Kaiser, The varimax criterion for analytic rotation in factor analysis, Psychometrika 23 (1958) 187–200.
[54] R.M. Sirkin, Statistics for the Social Sciences, third ed., SAGE Publications, 2005.
[55] N.R. Draper, H. Smith, Applied regression analysis, in: Wiley Series in Probability and Statistics, third ed., Wiley-Interscience, 1998.
[56] I.T. Jolliffe, Principal Component Analysis, Second Edition, Technometrics 30 (2002) 487.