

# Ego Network Structure in Online Social Networks and its Impact on Information Diffusion

Valerio Arnaboldi<sup>a,\*</sup>, Marco Conti<sup>a</sup>, Massimiliano La Gala<sup>a</sup>, Andrea Passarella<sup>a</sup>, Fabio Pezzoni<sup>a</sup>

<sup>a</sup>IIT-CNR, Via G. Moruzzi 1, 56124, Pisa, Italy

---

## Abstract

In the last few years, Online Social Networks (OSNs) attracted the interest of a large number of researchers, thanks to their central role in the society. Through the analysis of OSNs, many social phenomena have been studied, such as the viral diffusion of information amongst people. What is still unclear is the relation between micro-level structural properties of OSNs (i.e. the properties of the personal networks of the users, also known as ego networks) and the emergence of such phenomena. A better knowledge of this relation could be essential for the creation of services for the Future Internet, such as highly personalised advertisements fitted on users' needs and characteristics. In this paper, we contribute to bridge this gap by analysing the ego networks of a large sample of Facebook and Twitter users. Our results indicate that micro-level structural properties of OSNs are interestingly similar to those found in social networks formed offline. In particular, online ego networks show the same structure found offline, with social contacts arranged in layers with compatible size and composition. From the analysis of Twitter ego networks, we have been able to find a direct impact of tie strength and ego network circles on the diffusion of information in the network. Specifically, there is a high correlation between the frequency of direct contact between users and her friends in Twitter (a proxy for tie strength), and the frequency of retweets made by the users from tweets generated by their friends. We analysed the correlation for each ego network layer identified in Twitter, discovering their role in the diffusion of information.

*Keywords:* Online social networks, ego networks, tie strength, information diffusion

---

## 1. Introduction

The impressive penetration of Internet technologies and the establishment of participatory forms of content generation and exchange, such as the Web 2.0 paradigm, paved the way for the diffusion of Online Social Networks (OSNs). OSNs are nowadays a significant part of the prosumer paradigm shift in communication and data exchange, whereby users can actively create and share information with each other, rather than being passive consumers of contents as in more traditional media. In addition, OSNs are becoming one of the preferred ways to manage social relationships for an increasing number of people and their use is growing also far beyond supporting social relationships between people. Today OSNs are already successfully used, among others, for commercial recommendations, online content curation, advertising, and much more.

OSNs are significantly contributing to the so called cyber-physical world (CPW) convergence [1], which envisions a world where people actions and interactions in the

cyber (virtual) world, enabled by ICT, and in the physical world are strongly dependent upon each other and knitted into a single whole. In the CPW, actions taken in the virtual world are directly transferred to the physical world and vice versa. For example, social relationships in OSNs (i.e. in the virtual world) often depend upon those existing in the physical world and actions taken in OSNs modify the state of the physical world (e.g., mass movements or rallies that are organised and advertised exclusively over OSNs).

Characterising the properties of OSNs has been a very active research topic recently. The scale of these networks make this task challenging per se, and the diffusion of OSN services makes results obtained from this type of analysis impactful. In addition, the study of OSN structural properties is fundamental for the creation of a series of new services for the Future Internet highly customised on the user's characteristics and needs. For example, the structure of OSNs can be exploited to develop new efficient and cost-effective marketing strategies, as shown in [2]. Despite this, most of the analyses conducted so far on OSNs are focused on macro-level structural properties only (e.g. global clustering coefficient, diameter, presence of communities), whereas micro-level structures (i.e. the properties of personal social networks of the users) have not been investigated in detail. In sociology and anthropology,

---

\*Corresponding author

*Email addresses:* v.arnaboldi@iit.cnr.it (Valerio Arnaboldi), m.conti@iit.cnr.it (Marco Conti), m.lagala@iit.cnr.it (Massimiliano La Gala), a.passarella@iit.cnr.it (Andrea Passarella), f.pezzoni@iit.cnr.it (Fabio Pezzoni)

the micro-level structures of social networks formed offline (not mediated by the use of the Internet) are found to be directly related to most of the social phenomena arising in the network and, therefore, it can be reasonably expected that micro-level structures in OSNs could impact on the aforementioned OSN services.

The chief aim of this paper is twofold. On the one hand, we investigate in detail key properties of OSNs at the micro level. On the other hand, we show how these properties determine patterns of information diffusion in OSNs. In this way, we contribute to discover the relation between microscopic and macroscopic properties of OSNs, where the former are related to the social behaviour and structure of individual users, while the latter relate to social phenomena involving the network as a whole.

In order to characterise the micro-level structures of OSNs, we focused our analysis on ego networks. An ego network is defined as a portion of a social network formed of a given individual, termed *ego*, and the other persons with whom she has a social relationship, termed *alters*. Ego networks have been the subject of a very significant body of work in the sociology and anthropology literature, that has characterised some of their fundamental properties (see Section 2.3 for more details). One of the most important is the presence, in the ego network structure, of a series of concentric layers of alters with different levels of intimacy and size. The key parameter to distinguish between alters at different layers is the *tie strength* of the social relationship with the ego, which is typically approximated with the frequency of contact between them (a more precise description of these results is presented as background material in Section 2).

Section 4 of this paper reports an analysis that investigates whether similar ego network structures can also be found in OSNs. This analysis allowed us to assess the differences between the baseline results in social sciences and the properties we have observed in different OSN data sets (described in Section 3) obtained from Facebook and Twitter. Notably, we have found a layered structure in OSN ego networks similar to the one identified in offline social networks in terms of: (i) number of layers, (ii) frequency of contact of the layers, and (iii) scaling factor between the size of adjacent layers. This indicates that, as far as the structural properties of social relationships are concerned, human social behaviour seems to be unaltered by the use of OSNs. This further confirms the existence of the CPW convergence and it must be taken into account for the creation of user-centric future-Internet services.

Starting from these results, in Section 5, we report an analysis aimed at assessing the role of the ego network structure on the diffusion of information. We analysed the impact of tie strength and the presence of ego network circles on information propagation in Twitter ego networks, for which we have complete information on the creation of tweets and retweets. In accordance with the literature (see for example [3]), we found that *weak* ties, associated with lower levels of direct interactions than *strong* ties,

also transport a lower number of retweets. Despite this, the high number of weak ties in the ego networks makes the total amount of information circulating through them exceed the amount of information passing through strong ties. Then, we analysed the correlation between the frequency of interaction between users in Twitter (a proxy for their tie strength) and the frequency of retweets that flow through the social links that connect these users. The correlation has a medium/high value ( $r = 0.46$ ), but it is not sufficient to justify a model able to predict information diffusion from tie strength. Hence, we further investigated this aspect on two axes: (i) by studying the correlations within single ego network layers, and (ii) by dividing social relationships in two classes, the first related to alters who use Twitter for socialising and the second containing other types of alters, like companies, public figures, etc. The results indicate that the correlation between tie strength and information diffusion increases when we move from the outer to the inner parts of ego networks (from weak to strong ties), with values greater than 0.6 for the innermost layer. Perhaps more surprisingly, the correlations for both classes of alters are sensibly higher than those found when the two are mixed together, with the first class (i.e. people with social behaviour) showing the higher values of correlation (close to 0.8 for the innermost layer and always higher than 0.6 for the other layers).

Therefore, in summary, the results presented in the paper show a significant similarity between social network structures in online and offline environments. Not only the structure of ego networks is remarkably similar, but also these structures significantly impact on the way OSN services are used. Specifically, we have found that the patterns of information diffusion can be explained quite precisely starting from ego network structures of the individual users.

## 2. Background and Motivations

Social networks are structures composed of a set of social actors (e.g. individuals, organisations) and a set of ties (i.e. social relationships) connecting pairs of these actors. They are usually expressed in the form of graphs consisting of nodes representing social actors connected by edges, or arcs, which represent social relationships. We define *online* social networks as the social networks in which social relationships are maintained by the use of the Internet (e.g. Facebook, Twitter, e-mail exchange networks), and *offline* social networks as social networks formed outside the Internet, for example, face-to-face communication networks or phone call networks.

Both offline and online social networks have been analysed as typical complex network systems [4], i.e. with the same methodology used for other kinds of networks, such as biological and technological networks. Indeed, they have shown to present well-known properties of complex networks, such as the *small-world property* [5]. As discussed in [6], a small-world network is characterised by

short average distances between any two nodes connected via a chain of intermediate links. In addition, a small world network shows a high level of clusterisation (or network transitivity) compared to a random network, where clusterisation is the probability that two neighbours connected to a node will also be connected to each other. The small-world property directly impacts on the ability of the network to spread information quickly. Not surprisingly, another typical property found in offline and online social networks is the presence of communities [7]. These studies are normally carried out considering the unweighted network graph in which each edge (or arc) represents the mere existence of a social relationship without including information that can distinguish between different types of relationships. This is due to the fact that information about social relationships is not trivial to infer since it normally refers to qualitative aspects that are difficult to measure. Nevertheless, in particular for the analysis of microscopic social network properties, characterising (and distinguishing between) different types of social links is fundamental. In particular, *tie strength*, i.e. a quantitative measure of the importance of the social links linking two people, is a very important parameter, which has been widely investigated in the sociology and anthropology literature.

### 2.1. Measures of Tie Strength in Social Networks

In his seminal work, Mark Granovetter informally defined the strength of a social relationship as a linear combination of time, emotional intensity, intimacy, and reciprocal services [8]. Social relationships can be roughly divided into strong and weak ties, where the former denote more important relationships and the latter represent acquaintances. Besides their lower strength, weak ties are generally more numerous than strong ties. For this reason, the cumulative strength of weak ties could exceed that of strong ties and their impact on social phenomena could be substantial. Other measures of tie strength have been successively constructed and validated by Peter Marsden in [9].

Based on the results found by Marsden, several techniques to measure tie strength have been proposed also for OSNs, for example in [10, 11, 12, 13]. These studies indicate that tie strength can be effectively estimated using some measurable indicators. In particular, the frequency of contact seems to be the best among them, especially in online environments, also considering that it is easy to obtain from online communications logs. This has been confirmed in a study on Facebook [14], where the authors asked a set of Facebook users to name their closest friends in real life, and they found that contact frequency can be used to accurately discriminate closest friends from acquaintances.

### 2.2. OSN Analysis based on Tie Strength

Recently, some work has been done to characterise the differences between strong and weak ties in OSNs, and to

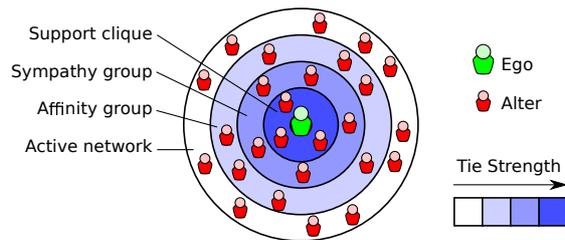


Figure 1: Ego network structure.

relate them with observable properties of the networks. Ties connecting otherwise disconnected parts of the network (also known as bridges) are associated to lower interaction levels than ties connecting clustered parts of the network [15]. This has been observed also in phone call networks [16], and is consistent with the *strength of weak ties* hypothesis of Granovetter [8] that postulates that social links connecting distant and otherwise disconnected parts of a social network must be weak ties. In [17, 18] two studies on Facebook and Google+ show that considering tie strength in the analysis of social network structures reveals properties that are not visible from the unweighted networks, and can lead to significantly different results. For example, the average distance between nodes in the Facebook network, that is less than 4 in the unweighted network, is between 5 and 10 when tie strength is considered. This is because many links in OSNs appear to be inactive, and considering them in the structural analysis of OSNs can lead to inconsistent or wrong results. For example, if we consider information diffusion, no information passes through inactive links, and they cannot be considered effective channels for information diffusion.

In addition, a relation between geographical location and tie strength has been found, with strong ties connecting most of the time people in physical proximity, and bridges connecting part of the network far from each other [15, 19]. Mobility also plays an important role in the formation of social ties and in determining their strength, since meeting other people enables social interactions [20].

### 2.3. Ego Network Model

In order to study the micro-level structural properties of social networks, researchers defined the *ego network*, as a simple social network model formed of an individual (called *ego*) and all the persons with whom the ego has a social link (*alters*). In an ego network, alters are normally arranged in a series of four or five inclusive groups (called *circles*) according to the strength of their social ties. Figuratively, an individual ego can be envisaged as sitting at the centre of the series of concentric circles [21] as depicted in Figure 1. Each of these circles has typical size and tie strength. The latter is usually estimated using the frequency of contact between the ego and the alters. Note that in the following, as typically done in the literature, a *circle* also contains all alters of the more internal circles

(with higher tie strength), while a *ring* only contains alters of a given circle that are not part of any more internal circle.

The first circle, called *support clique*, contains alters with very strong social relationships with the ego, informally identified in the literature as *best friends*. These alters are people contacted by the ego in case of a strong emotional distress or financial disasters. The size of this circle is limited, on average, to 5 members, usually contacted by the ego at least once a week. The second circle, called *sympathy group*, contains alters who can be broadly identified as *close friends*. This circle is formed of, on average, 15 members contacted by the ego at least once a month. The next circle is the *affinity group* (or *band* in the ethnographic literature), which contains about 50 alters usually representing causal friends or extended family members [22]. Although some studies tried to identify the typical frequency of contact of this circle, there are no accurate results in the literature about its properties, due to the difficulties related to the manual collection of data about the alters contained in it through interviews or surveys. The last circle in the ego network model is the *active network*, which includes all the other circles, for a total of about 150 members. This circle contains people for whom the ego actively invests a non-negligible amount of resources to maintain the related social relationships over time. People in the active network are contacted, by definition, at least once a year. The active network size coincides with the *Dunbar's number*, that identifies the average limit of the number of social relationships an individual can actively maintain due to cognitive constraints of the brain and the limited time for socialising [23]. Alters beyond the active network are considered inactive, since they are not contacted regularly by the ego. These alters are grouped in additional external circles called *mega-bands* and *large tribes*. One of the most stunning facts about ego network circular structure is that the ratio between the size of adjacent circles appears to be a constant with a value around 3, and this holds true for ego networks of users belonging to various social environments, as shown in [24]. For a complete discussion about the properties of the ego network circles we refer the reader to [25].

#### 2.4. Analyses on Ego Networks in OSNs

As far as the structure of ego networks in OSNs are concerned, and in particular the ego network model applied to online environments, still little is known. In the following, we summarise the preliminary results found on ego network properties of OSNs, which represent the starting point of this work.

The authors of [12] and [26] analysed two data sets from Facebook and Twitter respectively, founding evidences of the presence of the Dunbar's number in OSNs. Its presence indicates that, even though Facebook and Twitter allow people to have thousands of online social contacts, they only maintain a limited set of active relationships.

The authors of [27] analysed a large data set of Twitter communications discovering that the limited capacity people have for socialising bounds the amount of contacts they can actively maintain over time, as defined in the ego network model.

In a recent analysis on Twitter communication data, it has been found that the structure of Twitter ego networks is directly related to the network status of egos (defined as the ratio between followers and following), to the topic diversity of the tweets generated by egos, and to geography [28]. In particular, egos who have contacts spanning structural holes (gaps between separated groups of people) have higher network status, higher topic diversity, and more geographically sparse networks than egos with highly clusterised networks. This validates the idea that social capital is created by bridging structural holes, as proposed by sociologist Ron Burt [29]. This has been further confirmed by the work presented in [30], where the authors identify a trade-off between the diversity of information that a network can provide and the average strength of its social ties (called bandwidth). This trade-off is connected to the ability of egos to bridge structural holes. In a highly clusterised network the diversity of acquirable information is low, but bandwidth is high, with a positive impact on the quantity and quality of resources that ego can obtain from her intimate alters. On the other hand, in ego networks where egos bridge many structural holes, the diversity of information is higher, to the detriment of network bandwidth. This also confirms results in the sociology literature, which show that the higher the number of social links (and, therefore, the higher the accessible information diversity), the lower the average strength of social ties. This is, again, related to the fact that humans can only allocate a finite amount of cognitive resources for socialising, and therefore a higher number of social relationships translates into a lower average tie strength. Note that this very same phenomenon has also been observed in Twitter [26].

Although these results give a first insight on the properties of ego networks in OSNs, there is still a lack of knowledge about the micro-level structural properties of OSNs. Specifically, it is not clear if structures similar to those described by the ego network model could be found also in OSNs (e.g. the presence of ego network circles and their properties). In this paper, we aim to bridge this gap by providing a solid analysis of the ego network structures in OSNs.

This paper extends our initial analyses on Facebook and Twitter [31, 32, 33]. Specifically, in this paper we provide a more detailed analysis of ego network structures in both Online Social Networks, and we exploit it to characterise the impact of these structures on information diffusion.

#### 2.5. Analyses on Information Diffusion in OSNs

In this paper, we also investigate the role of the structural properties of OSN ego networks in the diffusion of

information, one of the most important macro-level phenomena in social networks. Specifically, we study information diffusion in Twitter ego networks and we investigate the role of tie strength and the ego network structure in the process.

Prior work on information diffusion has focused on detecting collective behaviour in the network, such as the formation of information cascades (i.e. the epidemic diffusion of information triggered by the observation and the adoption of the behaviour of others). Some works successfully predicted the outbreak of cascades in social networks [34, 35]. Other papers show that different types of users have different roles in diffusing information (e.g. normal users, opinion leaders, mass media sources) [36]. Mass media diffuse most of the information, but are focused on major topics. Opinion leaders and other users classified as “evangelists” contribute to diffuse major as well as minor topics to audiences far from the core of the network, and normal users usually have a low contribution on the diffusion process, and are more passive consumers than active producers of contents.

In social networks, and in Twitter in particular, different sources of information diffusion coexist. In fact, information may come and is propagated directly from the users within the platform following the word-of-mouth effect [37], as well as from other sources that are external to social networks (e.g. television and radio) [38]. Moreover, analyses of information cascades in Twitter confirm that various elements impact on the information diffusion process [39, 40]. Among others, the “standing” of users (e.g., their importance in their personal social network), their network centrality (according to standard complex network indices), as well as the freshness of information are directly related to the probability of propagation across nodes, and thus ultimately impact on the breadth of information cascades. In this work, we are interested in characterising in detail information diffusion through the word-of-mouth effect. In particular, we focus on the analysis of information diffusion seen from an ego network perspective, and we assess the impact of ego network circles on the process. To the best of our knowledge, this is the first detailed analysis on these particular aspects.

Before describing the analysis to characterise the structural properties of OSNs, we present a description of the Facebook and Twitter data sets that we used.

### 3. Online Communications Data Sets

To study the structural properties of OSNs and to assess their role in the diffusion of information in the network, we have analysed two data sets containing traces of communication between people in Facebook and Twitter, two amongst the most important social media nowadays (see Appendix A for a detailed description of these platforms). From the data sets, we have obtained the frequency of contact between online users, that has been used to estimate the strength of the social links (as we discuss

in Section 3.3). Hence, we have built an ego network for each user, and we have analysed their structural properties (in Section 4) and their role in the diffusion of information (in Section 5).

#### 3.1. Data Download

##### 3.1.1. Facebook

Although Facebook generates a huge amount of data regarding social communications between people, obtaining these data is not easy. In fact, publicly available data have been strongly limited by the introduction of strict privacy policies and default settings for the users after 2009. Nevertheless, before that date, most of the user profiles were public and the presence of the *network* feature, that has been removed in 2009, allowed researchers to collect large-scale data sets containing social activity between users. A network was a membership-based group of users with some properties in common (e.g. workmates, classmates or people living in the same geographical region). Each user profile was associated to a regional network based on her geographical location. By default, each user of a regional network allowed other users in the same network to access her personal information, as well as her status updates and the posts and comments that she received from her friends. Exploiting these characteristics of regional networks, some data sets have been downloaded, such as those described in [18], which have been made partly publicly available for research<sup>1</sup>. In this paper, we used the data set referred as “Regional Network A”.

The use of the regional networks feature allowed researchers to download large data sets from Facebook, however, it entails some limitations that must be taken into account for our analysis. In fact, the considered data set contains information regarding only the users and their social interactions within a regional network, excluding all the interactions and the social links that involve users external to this area. Therefore, assuming that, for each user, a part of her social relationships involve people who do not belong to the same network, this could lead to a reduction of the ego networks’ size. Moreover, we do not have specific information about the completeness of the crawling process that should have downloaded only a sample of the original regional network. For example, in [18] the same crawling agent was used for downloading several other regional networks (not publicly available) collecting, on average, 56.3% of the nodes and 43.3% of the links. We used this additional knowledge in the analysis to obtain the highest possible accuracy in the results, as explained in detail in Section 4.

##### 3.1.2. Twitter

For Twitter, we have implemented a crawling agent that is able to download user profiles and their communication data from Twitter. The agent visited the Twitter

<sup>1</sup><http://current.cs.ucsb.edu/facebook/>

Table 1: Statistics of the Facebook social graph

# Nodes	3,097,165
# Edges	23,667,394
Average degree	15.283
Average shortest path	6.181
Clustering coefficient	0.209
Assortativity	0.048

graph considering the users as nodes and following the links between them. In our study, a link between two nodes exists if at least one of the users follows the other or an interaction between them has occurred. As an indication of an interaction, we use the presence of a *mention* in a tweet (i.e. the fact that a user explicitly mentions the other in a tweet) and a *reply* (i.e. a direct response to a tweet).

The crawling agent starts from a given user profile (seed) and visits the Twitter graph following the links. For each visited node, we took advantage of the Twitter REST API to extract the user *timeline* (i.e. the list of posted tweets that can include mentions and replies), the *friends* list (i.e. the people followed by the user) and the *followers* list (i.e. the people who follow the user). Twitter REST API limits the amount of tweets that can be downloaded per user up to 3,200 tweets. This does not represent a constraint to our analysis since, as we show in the following, it is sufficient for our purposes.

The crawling agent uses 250 threads that concurrently access a single queue containing the IDs of the user profiles to download. Each thread extracts a certain number of user IDs from the queue, then it gets the related profiles and communication data from Twitter using the REST API. Finally, after extracting new user IDs from the communication data and from the friends/follower lists, the threads add them to the queue. The use of multiple threads allowed us both to speed-up the data collection and to avoid that the crawler remains trapped in visiting the neighbourhood of a node with a large number of links. The seed that we used to start the data collection is the profile of a widely know user (user id: 813286), so that her followers represent an almost random sample of the network.

The crawling agent allowed us to obtain a snowball sample of a complete portion of the Twitter network. Compared to the Facebook data set, this contains complete ego networks.

## 3.2. Data Sets Properties

### 3.2.1. Facebook

The Facebook data set that we used in this work consists of a *social graph* and four *interaction graphs*. These graphs are defined by lists of edges connecting pairs of anonymised Facebook user IDs.

Table 2: Statistics of the Facebook interaction graphs (preprocessed).

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

The social graph describes the overall structure of the downloaded network. It consists of more than 3 million nodes (Facebook users) and more than 23 million edges (social links). An edge represents the mere existence of a Facebook friendship, regardless of the quality and the quantity of the interactions between the involved users. Basic statistics<sup>2</sup> of the social graph are reported in Table 1.

The social graph can be used to study the global properties of the network, but alone it is not enough to make a detailed analysis of the structure of social ego networks in Facebook. Indeed, this analysis requires an estimation of the strength of the social relationships. To this aim, in Section 3.3, we leverage the data contained in the interaction graphs to extract the frequency of contact of the social links that can be used to estimate the tie strength.

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

In Facebook, an interaction can occur exclusively between two users who are friends. In other words, if a link between two nodes exists in an interaction graph, an edge between the same nodes should be present in the social graph. Actually, the data set contains a few interactions between users which do not correspond to any link in the social graph. These interactions probably refer to expired relationships or to interactions made by accounts that are no longer active. To maintain consistency in the data set, we excluded these interactions from the analysis. The amount of discarded links is, on average, 6.5% of the total number of links in the data set.

In Table 2, we report some statistics regarding the different interaction graphs. Each column of the table refers to an interaction graph related to a specific temporal window. The average degree of the nodes is the average number of social links per ego that have at least one interaction

<sup>2</sup>The clustering coefficient is calculated as the average local clustering coefficient (Equation 6 in [4]).

Table 3: Twitter data set (all users) and classes statistics.

	All users	Soc. rel. users	Other users
$N$	2,463,692	1,653,436	810,256
$N_{3,200}$ (% $N_{3,200}$ )	510,119 (20.7%)	260,632 (15.8%)	249,487 (30.8%)
# Tweets	1,207	979	1,696
# Following	3,157	2,553	4,448
# Followers	7,353	2,744	17,201
% Tweets <sub>REPL</sub>	17.4%	18.4%	15.4%
% Tweets <sub>MENT</sub>	22.7%	21.6%	24.7%

in the considered temporal window. Similarly, the average edge weight represents the average number of interactions for each social link.

### 3.2.2. Twitter

We collected a data set from 2,463,692 Twitter users, whose data were downloaded between November 2012 and March 2013. In contrast to Facebook, whose users are generally people who want to socialise with others, communicating and maintaining social relationships, Twitter users are more heterogeneous. In fact, the downloaded accounts can also be related to companies, public figures, news broadcasters, bloggers and many others. We can thus classify the users in two different categories: (i) *socially relevant users*, which represent people who use Twitter for socialising, and (ii) *other users*, which use Twitter for other purposes. This classification is fundamental for our study since, in order to analyse the human social behaviour, we want to consider socially relevant users only. To automatically distinguish between the two classes of users, we built a classifier based on Support Vector Machines (SVM) that, relying on the activity logs and on the meta-data of the accounts in the data sets, distinguishes socially relevant users from other users. The accuracy of the SVM is 83%, and the false positives rate around 8%. The details of the classifier are described in Appendix B. Note that, also in Facebook, some accounts represent users that are not socially relevant (e.g. companies and public figures). Nevertheless, Facebook is more naturally used as a private communication channel, and public communications (e.g. status updates) are not considered in the data set. For this reason, and for the lack of sufficiently detailed information about the nature of Facebook users in the data set, we analysed all the Facebook accounts without splitting them into separate classes.

In the column “all users” of Table 3, we present some statistics of all the users in the data set, while in the next two columns we present the statistics of the socially relevant users and of the other users respectively. For each category, we present the number of users  $N$  and the average number of tweets, friends, and followers. Each average value is reported with its 95% confidence interval between square brackets.

We can notice that socially relevant users are the majority and their statistics indicate that they are less active than the other users. This could be explained by the fact

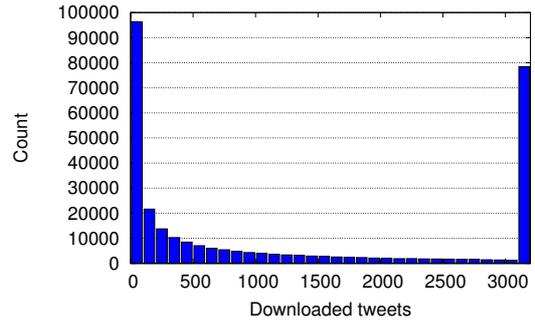


Figure 2: Downloaded tweets per user distribution.

that users in the “other users” class may be companies or other kinds of accounts managed by more than one person at the same time and aimed at advertising goods or services.

In the table, we also report, for each class of users, the average ratio of replies ( $tweets_{REPL}$ ) and mentions ( $tweets_{MENT}$ ), calculated over the total number of tweets. These values indicate that around 40% of the tweets downloaded by our crawler contain mentions or replies between people. These tweets are important for our study since they represent direct interactions, rather than broadcast communications. Moreover, socially relevant users show a slightly higher percentage of replies than other types of users (18.4% vs. 15.4%), indicating that they use more directional communications, a typical human social behaviour.

In Figure 2, we show the distribution of the number of tweets downloaded per user. We can notice the presence of a peak corresponding to the value 3,200, which is the maximum amount of tweets downloadable using the Twitter REST API. Cases where the number of tweets is lower than 3,200 correspond to users that have generated less than 3,200 tweets from the creation of their accounts. The number of users that posted a number of tweets above this threshold is indicated in the table by  $N_{3,200}$ . Note that, for socially relevant users, this is a relatively small fraction of the total number of users (15.8%). This means that our crawler was able to download the entire twitting activity for the majority of the users relevant for our study, and for those users for whom we have not obtained the entire history of outgoing communications, we still have a significant number of tweets.

In order to further investigate the behavioural differences between socially relevant users and the other users, we studied the number of replies the users send to their friends on average. In [26], a similar analysis was used to conclude that a concept similar to the Dunbar’s number (the maximum number of active social relationships an individual can actively maintain) holds also in Twitter.

Figure 3 depicts the trend of the average number of replies per friend as a function of the number of friends of the user. Differently from [26], we have divided the analysis for the two classes identified: “socially relevant

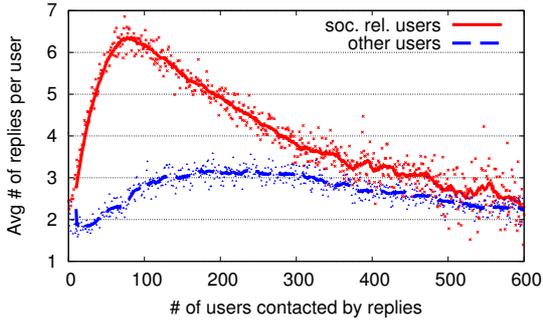


Figure 3: Points represent the average number of replies made by accounts with different number of friends; thick lines are their running averages.

users” and “other users”. The results highlight a clear distinction between the properties of the two classes.

Socially relevant users show a higher mean value of replies per friend and a maximum around 80 friends. This is an indication of the effect of the cognitive limits of human brain on the ability to maintain social relationships in OSNs. The peak of the curve identifies the threshold beyond which the effort dedicated to each social relationship decreases. This is due to the exhaustion of the available cognitive/time resources, which, therefore, have to be split over an increasing number of friends. As discussed in [26], this can be seen as an evidence of the presence of the Dunbar’s number in Twitter.

Other users show a quite different pattern, with lower average value of replies per friend without any significant discontinuities. This indicates that accounts belonging to the class “other users” are not influenced by cognitive capabilities. In fact they, are often managed by more than one person or by non-human agents.

### 3.3. Obtaining the Frequencies of Contact

#### 3.3.1. Facebook

In order to characterise tie strength in Facebook, we need to estimate the *link duration*, that is the time elapsed since the establishment of the social link. This is essential to calculate the frequency of contact between the users involved in a social link, and the latter is then used to estimate the tie strength. In the literature, the duration of a social link is commonly estimated using the time elapsed since the first interaction between the involved users [10]. Unfortunately, the data set does not provide any indication regarding the time at which the interactions occurred. To overcome this limitation, we have approximated the links duration leveraging the difference between the number of interactions made in the different temporal windows. Details on how we have estimated the link duration and the frequency of contact between users in the Facebook data set are given in Appendix C. The frequency of contact between pairs of users has been calculated as the total number of interactions occurred (obtained from the “all” interaction graph) divided by the estimated duration of

their social link. In case the users have never interacted their frequency of contact is set to zero.

#### 3.3.2. Twitter

The Twitter data set contains all the tweets sent by the users (with the limit of 3,200 tweets per user). Hence, obtaining the frequency of contact between users in Twitter is more straightforward than in Facebook. Considering socially relevant users with all their social contacts, we calculated the duration of each social link as the time elapsed between the first mention or reply exchanged between the involved users and the time of the download. Given a social link, we have thus calculated the frequency of contact for each of the two users as the number of replies sent to the other divided by the duration of the social link. In the calculation, we have used the number of replies since it is the strongest indicator of the strength of a social link in Twitter and since it has been already used in previous work [26].

## 4. Ego Networks Structure in Online Social Networks

In this section, we analyse the structures of the ego networks that can be identified in Facebook and Twitter and we compare them with the model for offline social networks presented in Section 2.3.

In order to extract the ego networks from our data sets, we have grouped the relationships of each user into different sets<sup>3</sup>. Then, to avoid including possible outliers in the analysis, we have selected only the ego networks that meet the following criteria:

1. *The account of the ego must have been created at least six months before the time of the download.* In case of the Facebook data set, the lifetime of the accounts is estimated as the time since the user made the first interaction. In case of the Twitter data set, we know the time of the account creation as it is included in the meta-data we downloaded.
2. *Ego must have made, on average, 10 or more interactions per month.* For both data sets, we can calculate the average activity as the total number of registered interactions divided by the lifetime of the account.

This selection is also motivated by the findings in other OSNs analyses (see for example [41]), in which ego networks are found to be highly unstable and with a high growing rate soon after ego joins the network, but tend to be stable after the first few months of activity. This selection allowed us to consider only users who regularly

<sup>3</sup>Since social links in the Facebook interaction graphs represent undirected edges, we have duplicated each social link in the data set in order to consider it in both the ego networks of the users connected by it.

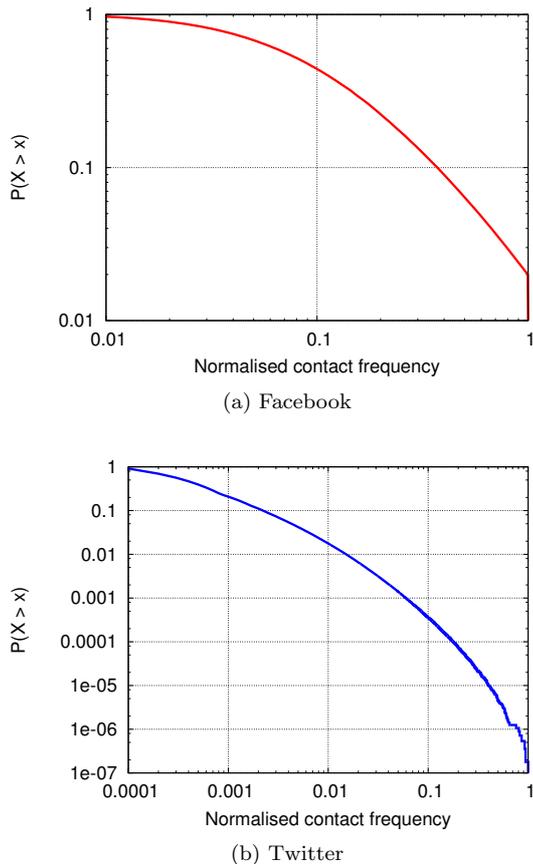


Figure 4: Aggregated CCDF of the normalised frequency of contact for all the ego networks in the data sets.

use OSNs, and filter out typical initial bursts of activities of new users. This resulted in the selection of 91,347 ego networks from the Facebook data set and 394,238 ego networks from the Twitter data set. These numbers, as we will see later, are sufficient to draw significant results about the ego network properties of OSNs. Note that the selected socially relevant users can have both socially relevant users and other users in their ego networks. In our analysis, we consider all the possible kinds of alters of socially relevant users. This is important to have a complete view of the structure of their social networks, since each ego spends cognitive efforts for communicating with all her alters, and the properties of her ego network are impacted by her cognitive and time constraints, no matter whether she spends all her time communicating with robots or with other humans.

#### 4.1. Analysis of the Aggregated Frequency Distribution

The possible presence of social structures in Facebook and Twitter may be revealed by steps in the distribution of the frequency of contact since it is the key aspect to quantify the tie strength. If the frequency of contact of an ego network gracefully degrades and does not present steps in the distribution, this suggests the absence of any structure. On the contrary, if the frequency of contact

appears clustered in different intervals, each of them may reveal the presence of a ego network layer.

A simple initial analysis to check the presence of such steps in the distribution is considering the CCDF of the aggregate normalised frequency of interaction. In particular, we have considered the distribution obtained by taking together all the frequencies of contact of all ego networks in each data set. A normalisation of the frequencies of contact for each ego network is necessary in order to level out the differences between users in the use of the platforms. Analysing the aggregate distribution permits to focus on a single distribution, instead of analysing all individual ego networks' distributions. The obtained CCDFs, depicted in Figure 4, show a smooth trend. Clearly, this does not allow us to conclude that ego networks are clustered, but is not a sufficient condition to rule out this hypothesis. In fact, even if the individual ego network distribution had a social structure, and therefore steps in their distributions would be present, such steps may appear at different positions from one network to another, thus resulting in a smooth aggregate CCDF (remember that also in the ego network model the sizes of the layers are average values, but variations are possible at an individual ego network level).

The CCDFs show a long tail, which can be ascribed to a power law shape. A power law shape in the aggregate CCDF is a necessary condition to have power law distributions in at least one ego network [42]. However, this is not a sufficient condition to have power law distributions in each single CCDF [43]. Although formally the presence of a long tail in the CCDF is not a conclusive proof of the existence of small numbers of very active social links in the individual ego networks, this is anyway a strong indication in this direction, and a possible similarity between ego networks in offline and online social networks. Studies in the social and anthropology literature revealed that ego networks are characterised by a small set of links with very high frequencies of contact (corresponding to the links in the support clique), which appear as a heavy tail in the CCDF of individual ego network contact frequency.

#### 4.2. Revealing Ego Network Structure through Clustering

To further investigate the online ego network structures, we have applied cluster analysis on the normalised frequencies of contact of each ego network, looking for the emergence of layered structures. For each ego network, the frequencies of contact between ego and alters represent a set of values in a mono-dimensional space. Applying cluster analysis to mono-dimensional values does not require advanced clustering techniques, therefore we can consider standard widely-used methods such as *k-means clustering* and *density-based clustering* (e.g. DBSCAN algorithm). Using *k-means clustering*, given a fixed number of clusters  $k$ , the data space is partitioned so that the sum of squared euclidean distance between the centre of each cluster (centroid) and the objects inside that cluster is minimised. In density-based clustering, clusters are defined as areas of

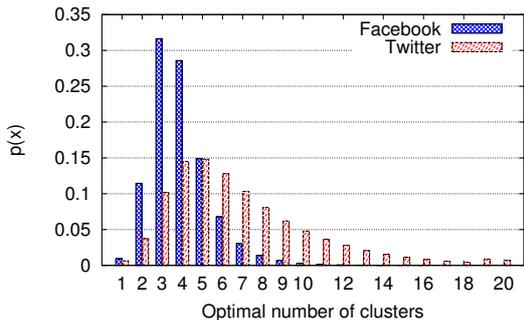


Figure 5: Density function of  $k^*$  in Facebook and Twitter ego networks.

higher density than the remainder of the data set, which is usually considered to be noise [44]. In [31], both clustering techniques have been applied on the same Facebook data set used in the present analysis. Nonetheless, results showed that the clusters identified by the two methods are substantially equivalent and that both can be used for the study of social structures in ego networks leading to the same conclusions [31].

In this work, we report the analysis using the  $k$ -means clustering since it is the simplest and the most computationally affordable method. This method is defined as an optimisation problem that is known to be NP-hard. Because of this, the common approach for  $k$ -means clustering is to search only for approximate solutions. Fortunately, in the special case of mono-dimensional space, we can use an algorithm, called `Ckmeans.1d.dp`, able to always find the optimal solution efficiently [45].

#### 4.2.1. Typical Number of Clusters

In the first step of our cluster analysis, we have sought, for each ego network, the typical number of clusters (i.e. the number  $k^*$ ) in which the frequencies of contact can be naturally partitioned. In order to do this, we have evaluated the goodness of the result of different clustering configurations. For  $k$ -means methods, this is usually expressed in terms of *explained variance*. In fact, a small variance in the individual clusters means that data are well described by the current clustering, and this is evidenced by a high value of the explained variance (up to the maximum value 1.0). Specifically, the explained variance is defined by the following formula:

$$VAR_{exp} = \frac{SS_{tot} - \sum_{j=1}^k SS_j}{SS_{tot}}, \quad (1)$$

where  $j$  is the  $j^{th}$  cluster,  $SS_j$  is the sum of squared distances within cluster  $j$  and  $SS_{tot}$  is the sum of squared distances of all the values in the data space. Given a vector  $\mathbf{X}$ , the sum of squared distances  $SS_{\mathbf{X}}$  is defined as  $SS_{\mathbf{X}} = \sum_i (x_i - \mu_{\mathbf{X}})^2$ , where  $\mu_{\mathbf{X}}$  denotes the mean value of  $\mathbf{X}$ .

Given the number of clusters  $k$ ,  $k$ -means clustering algorithms partition the space minimising the sum of squared

distance within the clusters  $\sum_{j=1}^k SS_j$ . According to Equation 1, for a given  $k$ , the solution of  $k$ -means clustering also provides the maximum value of the explained variance  $VAR_{exp}$ , since the sum of squared distances  $SS_{tot}$  is constant given the data space. In principle, the optimal number of clusters  $k^*$  would be equal to the number of objects in the data space, as the value of  $VAR_{exp}$  increases monotonically with  $k$ . Thus, there is an inherent overfitting problem. To overcome this problem and determine the typical number of clusters we used the Akaike Information Criterion (AIC), an information-theoretic measure that trades off distortion against model complexity, defined by the following equation:

$$AIC = -2L(k) + 2q(k) \quad (2)$$

where  $-L(k)$  is the negative maximum log-likelihood of the data for  $k$  clusters, and is a measure of distortion.  $q(k)$  is the number of parameters of the model with  $k$  clusters and measures complexity. The model showing the minimum value of  $AIC$  is the one with the best trade-off between distortion and complexity.

We have calculated the AIC for all the ego networks in Facebook and Twitter, by applying  $k$ -means with  $k$  from 1 to 20. For each ego network we define as  $k^*$  the value of  $k$  that minimises equation 2. In Figure 5, we report the density function of  $k^*$  for the ego networks in our data sets.

We have found that the distribution of  $k^*$  has a peak between 3 and 4 for Facebook and between 4 and 5 for Twitter. The presence of a typical number of clusters close to 4 is the first indication of similarity between the findings in offline and online ego networks.

In Table 4, we report the properties of the ego networks found with different numbers of  $k^*$ . The average network size (“net size” in the table) is reported with its 95% confidence interval between square brackets.

Ego networks with only one circle tend to have similar values of contact frequency for all their links, and in many cases the contact frequencies are exactly the same. This could be ascribed to automated forwarding of messages on all the links, associated to bots or spammers, and indicates the presence of a small set of biased ego networks in the data set. Remember that, although the classifier we used to select socially relevant users has a high accuracy, some accounts could be false positives, as probably in this case. Whilst the size of the ego networks with one circle in Facebook is relatively small, in Twitter we notice very large ego networks (i.e. with average size of 192.77 alters). This could be explained by the fact that it is more difficult for bots or spammers to create a large network of social relationships in Facebook, whereas in Twitter is easier to have a large number of followers with a significant interaction. This is due to the differences in the nature of the two platforms. In fact, in Facebook users tend to accept friendships requests only if they know the requester in person, or they recognise a real human behind her profile, whilst

Table 4: Optimal number of clusters ( $k^*$ ) of ego networks.

$k_{\text{opt}}$	Facebook		Twitter	
	# of nets	Net size	# of nets	Net size
1	844 (0.9%)	29.68 [ $\pm 1.95$ ]	2,500 (0.6%)	192.77 [ $\pm 12.44$ ]
2	10,465 (11.47%)	41.82 [ $\pm 0.39$ ]	14,683 (3.7%)	104.93 [ $\pm 2.35$ ]
3	28,918 (31.66%)	39.00 [ $\pm 0.27$ ]	40,099 (10.2%)	91.42 [ $\pm 0.98$ ]
4	26,124 (28.60%)	41.99 [ $\pm 0.38$ ]	57,227 (14.5%)	89.09 [ $\pm 0.73$ ]
5	13,584 (14.87%)	53.89 [ $\pm 0.66$ ]	58,410 (14.82%)	92.56 [ $\pm 0.70$ ]
> 5	11,412 (12.50%)	82.02 [ $\pm 1.00$ ]	221,319 (56.1%)	100.42 [ $\pm 0.31$ ]

in Twitter the heterogeneity of profiles makes this kind of selection more difficult.

The size of the ego networks seems to be almost constant between two and five circles, and it increases for networks with more than five circles.

#### 4.2.2. Ego Network Circles

According to the previous analysis, the typical number of clusters in online ego networks appears to be equal to 3 – 4 in Facebook and 4 – 5 in Twitter. Yet, to be able to compare the structure of online ego networks with that found in offline networks we have applied the algorithm `Ckmeans.1d.dp` with  $k = 4$  for Facebook and  $k = 5$  for Twitter. This choice will be more clear in the following, but we motivate it anticipating that in Twitter a new internal circles appear, that is not visible in the used Facebook data set. For each ego network, we obtained a set of clusters that we refer as  $S_1, S_2, S_3, S_4$ , and  $S_5$  (where needed), sorted by decreasing value of the centroid (i.e. the average frequency of contact of the cluster) so that  $S_1$  represents the cluster of the social links with the highest frequency of contact. The obtained clusters are not directly comparable with the circles of offline ego networks discussed in Section 2. In fact, while clusters are disjoint groups, social circles, as depicted in Figure 1, are hierarchically inclusive (i.e. the *support clique* is included in the *sympathy group* which is included in the *affinity group* which is included in the *active network*). For this reason, in order to compare social structures in online and offline ego networks, we have aggregated the clusters to form hierarchically inclusive circles. Specifically, we have defined the circles  $C_1, C_2, C_3, C_4$ , and  $C_5$  as  $C_k = \bigcup_{i=1}^k S_i$  so that  $C_1 \subseteq C_2 \subseteq C_3 \subseteq C_4 \subseteq C_5$ .

In Table 5, we compare the properties of the circles in Facebook and Twitter ego networks with those found in offline ego networks. One of the main features that we considered for the analysis is the *minimum frequency of contact*. It defines, for the alters included in each circle, the lower bound of the frequencies of contacts of their social links. In other words, this value indicates the minimum frequency of contact for an alter to be included in a given circle. In the table, we report the average value of this measure as “min freq.,” calculated for all the ego networks in terms of number of contacts per month. The minimum frequencies of contact of offline ego networks have been taken according to the definition discussed in Section 2:

*once a week* for the support clique, *once a month* for the sympathy group and *once a year* for the active network while, for the affinity group, the minimum frequency of contact has not been defined yet.

In the table, we also show the average size of the obtained circles for online ego networks while, for offline networks, we report the values presented in [24], that summarise the properties of a large number of offline social networks obtained in diverse social environments. Despite the size of the circles in Facebook and Twitter ego networks appear to be very close to each other, it is worth to remind that they should not be compared directly. In fact, as already explained in Section 3.1, the ego networks in the Facebook data set contain just a sample of the social relationships of the egos. This is because the crawling process may have not downloaded the considered regional network completely and because all the contacts external to this area have been excluded. In absence of precise information, we assume that the crawled data represent a uniform random sample of both nodes and links. On the contrary, the sizes of the circles of Twitter ego networks are more reliable, since we have at our disposal the entire outgoing communication log of each ego (given the limit of 3,200 tweets).

Rather than the size, a better feature to consider to compare the properties of online and offline ego networks is the scaling factor between the circles (“scal. fact.” in the table), defined as the ratio between the size of two hierarchically adjacent circles. This measure can provide insights about how the circles in ego network are hierarchically arranged and is not affected by a random sampling of the links. In fact, with random sampling, the size of all the circles changes proportionally without affecting the scaling factors.

#### 4.3. Comparing Online and Offline Ego Networks

Looking at the scaling factors in Table 5, we can see that their values are very similar to each other and close to 3, for both Facebook and Twitter ego networks, and they are compatible with the results found offline. A scaling factor of three has been found in several offline social networks and it appears to be a fundamental property of human ego networks [24]. This result is another indication that Facebook and Twitter ego networks show a hierarchical structure remarkably similar to that found in offline environments.

Table 5: Ego network circles’ properties.

		$C_1$	$C_2$	$C_3$	$C_4$	$C_5$
<b>Facebook</b>	min freq.	5.09	1.95	0.67	0.11	–
	size <sup>4</sup>	(1.79)	(5.83)	(17.05)	(50.46)	–
	scal. fact.		3.26	2.93	2.96	–
<b>Twitter</b>	min freq.	20.55	8.91	3.98	1.36	0.18
	size	1.66	5.06	12.87	32.66	97.47
	scal. fact.		3.04	2.55	2.54	2.98
<b>Offline</b>	min freq.	4.29	1.00	–	0.08	–
	size	4.6	14.3	42.6	132.5	–
	scal. fact.		3.10	2.98	3.11	–

Table 6: Offline/online ego networks mapping.

		<b>Super support clique</b>	<b>Support clique</b>	<b>Sympathy group</b>	<b>Affinity group</b>	<b>Active network</b>
<b>Facebook</b>	circle	–	$C_1$	$C_2$	$C_3$	$C_4$
	min freq.	–	5.09	1.95	0.67	0.11
	size <sup>5</sup>	–	(4.70)	(15.31)	(44.77)	(132.50)
<b>Twitter</b>	circle	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$
	min freq.	20.55	8.91	3.98	1.36	0.18
	size	1.66	5.06	12.87	32.66	97.47
<b>Offline</b>	circle	–	$C_1$	$C_2$	$C_3$	$C_4$
	min freq.	–	4.29	1.00	–	0.08
	size	–	4.6	14.3	42.6	132.5

Considering the average minimum frequency of contact of the circles, we can note that there is a match between the circles of the two OSNs and those of offline social networks. Specifically, as we report in Table 6, we find the same magnitude in the “min freq.” values of  $C_1$  in Facebook,  $C_2$  in Twitter and  $C_1$  in offline social networks, that therefore we map to the concept of support clique. In the same way,  $C_2$  in Facebook can be matched to  $C_3$  in Twitter and  $C_2$  in offline environments (the sympathy group),  $C_3$  in Facebook matches  $C_4$  in Twitter, and we hypothesise that the two match  $C_3$  offline (affinity group).  $C_4$  in Facebook matches  $C_5$  in Twitter and  $C_4$  offline (the active network). It is worth noting that Twitter shows higher values of min. freq (nearly double) for all the circles compared to Facebook and offline ego networks. This could be ascribed to the nature of the platform, and to the measure of interaction that we used, which could be slightly different than the one used in the other environments.

Last, we have compared the ego networks according to the sizes of their layers, which is another important signature of offline ego networks. The match between  $C_2$ - $C_5$  in Twitter and  $C_1$ - $C_4$  offline is further confirmed by a strong similarity in their size, as reported in Table 6. In the case of Facebook, a direct comparison is not possible, because of the unknowns in the sampling process previously discussed. Nevertheless, we can obtain strong hints about a significant match by re-scaling the Facebook sizes, as follows. Assuming that  $C_4$  in Facebook matches  $C_4$  offline (which is suggested considering the minimum frequency and the scaling factors), we have re-scaled the size of  $C_4$

in Facebook to match the size of  $C_4$  offline (132.50). The resulting ratio has a value of 2.63 that we have applied to the other Facebook layers. Note that the value of 2.63 is compatible with the reported subsampling of other networks obtained using the same crawling agent [18]. It is interesting to note that, scaling the size of other Facebook circles ( $C_1$ ,  $C_2$  and  $C_3$ ) according to this ratio, they match very well the respective sizes of the offline layers.

Interestingly, in Twitter we have found that there is an additional circle ( $C_1$ ) with a very high minimum frequency of contact that represents a subcircle of the support clique. Since the sizes of  $C_2$ - $C_5$  in Twitter show a good match with those found offline, we can say that  $C_1$  in Twitter, which we call “super support clique”, has a typical size of 1 or 2 people. This additional circle has been already hypothesised in offline social networks, but its existence remained unconfirmed hitherto, due to absence of big enough data sets to reliably highlight this type of relationships [46].

Summarising, our results show that there is a remarkable similarity between ego networks in OSNs (both Facebook and Twitter) and offline networks, in terms of scaling factors, minimum interaction frequency and size of the layers. This suggests that the use of OSNs does not affect the structural properties of ego networks, that are instead controlled by the constrained nature of the human brain. In addition our results also highlight additional structural elements, i.e. the “super support clique” in Twitter. This is a very interesting result per se, and also shows that OSNs can be used as an extremely useful tool to collect large-scale data sets to characterise human

social network properties. The scale at which data can be collected with OSNs permits to draw statistically relevant conclusions, which is often much harder or cumbersome with more conventional data collection campaigns (such as standard questionnaires). From a more technological standpoint, our results could be useful for the creation of advanced social platforms and efficient networking solutions for the Future Internet. For example, differences in the properties of social contacts of the user, arranged into the ego network circles, could be exploited to automatically set privacy policies (e.g. giving more trust to close friends) or to facilitate the management of social relationships giving specific tools for each circle.

## 5. Analysis of Information Diffusion in Twitter Ego Networks

In this section, we assess the extent to which ego network structural properties in OSNs impact on information diffusion inside the ego networks. To do so, we analyse ego networks in Twitter, for which we have accurate information regarding the creation of tweets and retweets and we can thus understand how information is propagated by users, and we seek for the relation between direct interactions of egos with their alters and the quantity of information that these egos retweet from each social link. This kind of analysis is clearly not possible in the Facebook data set, in which no exact information is known about the data exchanged amongst users.

Since the Twitter data set contains information about complete ego networks and also about all the tweets circulating in these networks, we can perform a detailed analysis of one-hop information diffusion. Clearly, we do not have enough data for the analysis of complete information cascades in Twitter, but this would be a natural extension of our work that we are currently investigating using additional Twitter data. In addition, we did not consider the textual content of tweets, to see to what extent structural properties alone explain information diffusion patterns, without delving into analysis of the content of exchanged messages.

In this work, we analysed the same Twitter ego networks that we already described, in terms of structural properties, in Section 4. The key idea that we used to assess the impact of the structure of ego networks in the diffusion of information is to count the number of tweets originally generated by each alter that are retweeted by the considered ego, relating this measure to the tie strength of their social link, calculated - as previously done - as the frequency of direct communication between them.

All the information diffusion mechanisms induced by sources external to ego networks, such as, for example, the trending topic page of Twitter, are not considered. This allows us to study the information diffusion derived from the presence of a social relationship, eliminating possible bias derived from external sources. Of course, external

sources play an important role in the diffusion of information, but this is out of the scope of the present work and has been already characterised in Twitter [38].

Before performing the analysis, we normalised the data by the duration of each social link between users and by the differences between ego networks due to the characteristics of the egos. This ensures a homogeneous analysis, eliminating differences between ego networks due to their different duration or the different frequency of use of the users. To normalise the data, we used, as a measure of tie strength, the percentage of frequency of replies sent by ego to her alter with respect to the total frequency of replies of ego to all her alters during her entire lifespan. This measure is expressed by the following equation.

$$\overline{frep}_{e,a} = \frac{\text{link reply frequency}}{\text{ego total reply frequency}} = \frac{rep_{e,a}}{l_{e,a}} * \frac{lt_e}{reptot_e} \quad (3)$$

where  $rep_{e,a}$  is the number of replies sent from ego ( $e$ ) to alter ( $a$ ),  $l_{e,a}$  is the lifespan of the social link between  $e$  and  $a$ ,  $reptot_e$  is the total number of replies sent by  $e$  to her alters, and  $lt_e$  is the lifespan of the ego. As already done in Section 4, the frequency of contact (replies) is estimated by dividing the number of recorded replies sent from the ego to the considered alter by the time elapsed between the first mention or reply sent by ego to the alter and the time of the download (link lifespan). The lifespan of the ego is the time elapsed between the creation of her account and the time of the download. Note that  $\overline{frep}$  is different from the normalised contact frequency used in Section 4, where each contact frequency was divided by the maximum contact frequency of the respective ego network to obtain a value in  $[0,1]$ . Here the purpose of the normalisation is different and the obtained measure is more suited for capturing the differences between egos, although it does not necessarily result in a value in  $[0,1]$ .

To measure information diffusion, we used the frequency of retweets of egos generated from tweets of their alters, divided by the total frequency of retweets of the egos, as defined by the following equation.

$$\overline{fret}_{e,a} = \frac{\text{link retweet frequency}}{\text{ego total retweet frequency}} = \frac{ret_{e,a}}{lret_{e,a}} * \frac{lt_e}{rettot_e} \quad (4)$$

where  $ret_{e,a}$  is the number of retweets by ego ( $e$ ) of tweets originally generated by her alter ( $a$ ),  $lret_{e,a}$  is the maximum between  $l_{e,a}$  and the time elapsed between the first retweet done by  $e$  to a tweet originally created by  $a$  and the time of the download,  $rettot_e$  is the total number of retweets done by  $e$  of tweets originally generated by her alters, and  $lt_e$  is the lifespan of the ego.

As a first analysis of the relation between tie strength and information diffusion, we calculated the correlation between  $\overline{frep}$  and  $\overline{fret}$  for all the relationships belonging to the active network (C5) of the ego networks in our Twitter data set and we also fitted a linear function relating the

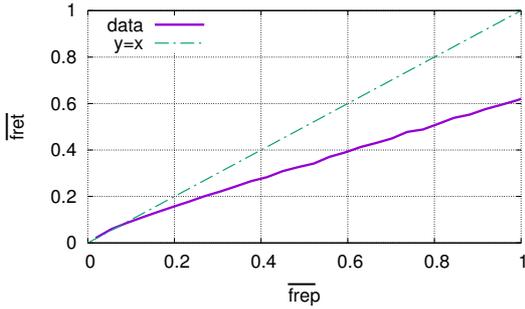


Figure 6:  $\overline{fret}$  as a function of  $\overline{frep}$ .

two measures, defined as follows.

$$\overline{fret} = \alpha + \beta * \overline{frep} \quad (5)$$

For the analysis, we have considered all the relationships of the ego networks together. Nevertheless, we also performed the same analysis by considering each ego network separately, and then averaging the results for all the egos. The results obtained with the two techniques are very similar, although for the second case the significance is sometimes not sufficient for ego networks with a small number of relationships. For this reason, in the following we report only the results obtained by taking all the ego networks together.

The correlation between  $\overline{frep}$  and  $\overline{fret}$  ( $r_{xy}$ ) and the estimated parameters  $\alpha$  and  $\beta$  are reported in the last row of Table 7, under the column “all alters”.

The medium/high value of correlation ( $r = 0.46$ ) is a first indication of a relation between tie strength and information diffusion. To further analyse the impact of ego network structure on the relation between  $\overline{frep}$  and  $\overline{fret}$ , we performed the same analysis by considering each layer of the ego networks separately. To do so, we assigned each social link in the network to a position in the ego network model, according to the contact frequency between the users it connects. Remember that, by definition, the ego network model forms a hierarchical structure, and therefore outer layers include inner ones. Thus, to avoid ambiguity, we have assigned each link to a *social ring*, defined as the part of a social circle that is not included in any nested circles. To do so, we used the same clustering technique described in Section 4.2, considering that the clusters coincide with social rings. A mapping between ego network circles and rings is provided in Table 8, where  $R_1$  represents the ego network ring containing alters with higher contact frequency, and  $R_5$  is the outermost ring.

The results of the analysis for the different rings are reported in Table 7 in the rows related to  $R_1$ - $R_5$ . The correlation between  $\overline{frep}$  and  $\overline{fret}$  ranges between 0.61 for the innermost ego network ring and 0.22 for the outermost one (note that we are referring here to the “all alters” column only). Compared to the average correlation calculated on all the relationships in the ego networks (equal

to 0.46), these values denote that the correlation is higher for more internal rings, and decreases as we move from inner to outer rings. This indicates that in the outer part of the ego networks the two measures are less dependent, and this could be explained by the fact that in this part of the network alters are more heterogeneous. The lower correlation might also be due to the fact that for information coming from outer rings, the content of information is more important than the strength of the ties (also because in general tie strength is quite low on outer layers), and therefore the retweeting behaviour is less correlated with social interactions.

The value of  $\beta$  increases from inner to outer rings. In the inner rings, a direct contact is related to less than one retweet ( $\beta < 1$ ), whilst in the outer circles a direct contact is related to a higher number of retweets ( $\beta > 1$ ). This is visible in Figure 6, which depicts  $\overline{fret}$  as a function of  $\overline{frep}$ .

A possible explanation of the lower values of  $\beta$  in the innermost layers could be that the relative gain in terms of information diffusion due to an increment in terms of tie strength may saturate after a certain level of strength (i.e., there is a sort of marginal utility law governing the dependency between tie strength and information diffusion). From the literature, we know that information coming from strong ties tends to remain trapped into highly clustered parts of the network formed of nodes socially close to egos. Therefore, the information circulating between strong ties tend to be not very diverse, and thus an increment in terms of tie strength could not be accompanied by an equal increment of information diffusion. On the other hand, for lower values of tie strength, the information coming from alters is generally much more diverse, and this could explain a lower dependency upon tie strength, and thus higher values of  $\beta$ . In other words, the diversity of information coming from weak ties increases its probability of being retweeted, with respect to the sole effect of the strength of the social tie over which information arrives to the ego. Another possible explanation for this trend is that egos may choose their strong ties primarily on the basis of their emotional closeness, while weak ties could be picked primarily because of the information they allow the ego to access, which is therefore proportionally more likely to be retweeted.

As described in Section 3, users in Twitter can be divided in two distinct groups: socially relevant users and other users. We have already seen the differences in terms of social behaviour of the users in these groups, with the former containing people who use Twitter for socialising and maintaining relationships with others, and the latter containing users with a less “human” social behaviour, e.g. companies, public figures, bots, and others. In the analysis

<sup>4</sup>Facebook circles’ size are affected by the data set subsampling discussed in Section 3.1.

<sup>5</sup>Scaled size to match offline active network dimension.

<sup>6</sup>Social circles are defined in Section 2.3.

Table 7: Information diffusion properties of ego network rings in Twitter, where  $x$  and  $y$  are  $\overline{frep}$  and  $\overline{fret}$ .

Ring	all alters			soc. rel alters			other alters		
	$r_{xy}$	$\hat{\beta}$	$\hat{\alpha}$	$r_{xy}$	$\hat{\beta}$	$\hat{\alpha}$	$r_{xy}$	$\hat{\beta}$	$\hat{\alpha}$
$R_1$	0.61	0.49	0.03	0.80	0.74	0.03	0.74	0.58	-0.01
$R_2$	0.52	0.62	0.01	0.76	0.76	0.02	0.71	0.59	0.02
$R_3$	0.44	0.74	0.00	0.72	0.80	0.03	0.67	0.64	0.02
$R_4$	0.34	0.97	0.00	0.66	0.85	0.06	0.65	0.72	0.02
$R_5$	0.22	1.58	0.00	0.61	0.99	0.09	0.65	0.93	0.03
<b>Whole net (C5)</b>	0.46	0.57	0.02	0.68	0.83	0.09	0.65	0.78	0.03

Table 8: Ego network rings.

Ring	Social circles correspondence <sup>6</sup>
$R_1$	super support clique
$R_2$	support clique, excluded the super support clique
$R_3$	sympathy group, excluded the support clique
$R_4$	affinity group, excluded the sympathy group
$R_5$	active network, excluded the affinity group

performed in Section 4, we studied the structure of the ego networks of socially relevant users, considering all their alters, which could belong to both classes. Going deeply into detail into the relation between information diffusion and the structural properties of ego networks, we need to analyse the behavioural differences of socially relevant users towards their different classes of alters. For this reason, the analysis introduced in this section is performed not only on all alters without distinction, but also considering socially relevant alters and other alters separately. Unfortunately, we have complete information about the nature of an alter only in case her profile has been downloaded by our crawler, so, for each ego network, we only have a fraction of alters that we can classify, which is, on average,  $\sim 30\%$  of the ego network. Nevertheless, the sample of classifiable relationships for each ego network can be considered a random sample of the relationships of egos. Therefore, the results presented for the different classes are estimated by using the set of classified alters. Moreover, as we consider all the social relationships of each circle for all the ego networks in the data set, the number of relationships is sufficient to obtain significant results.

In the Twitter ego networks that we have crawled we find that there are, on average, 27.8% socially relevant alters, and 72.2% other alters. Note that, as already shown in Section 3, the majority of egos in the initial data set are socially relevant users. This means that, also according to the statistics presented in Section 3, socially relevant users have less connections than other users, and this results in a higher proportion of connections towards “other alters” than to socially relevant alters, also for socially relevant egos.

As shown in Table 7, the correlations between  $\overline{frep}$  and  $\overline{fret}$  considering all the social relationships in the ego networks (C5) for the two classes of alters are respectively 0.68 for socially relevant alters and 0.65 for other alters. In addition, Table 7 reports the statistics regarding the relation between  $\overline{frep}$  and  $\overline{fret}$  for the different categories

of alters divided into the different social rings.

When considering socially relevant alters and other alters separately, the correlation is significantly higher in both cases than the case where alters are taken altogether. This indicates that there are two separate processes underpinning the relation between tie strength and information diffusion for the two classes, and, when the processes are mixed together, this difference is less visible. The different values of  $\alpha$  and  $\beta$  for the two classes support this hypothesis. The higher and more homogeneous values of  $\beta$  for socially relevant users indicates that these alters are treated in a more homogeneous way by egos across the different rings. The increasing value of  $\beta$  for both classes moving from internal to external rings, in addition, confirms the same phenomena already discussed when analysing all alters at the same time.

Figure 7 depicts the average number of retweets per link, for the different rings. Inner circles show a higher number of retweets per link, in accordance with the values of correlation and the estimated values of the regressors of equation 5. It is worth noting that this value, when multiplied by the average number of alters in each ring (calculated from the size of the ego network circles obtained in Section 4), indicates that, cumulatively, the quantity of information diffused through the outer layers is higher than that in the inner layers. This is visible in Figure 8 and it is in accordance with the idea of “the strength of weak ties” [8], as also empirically found in Facebook [3]. Interestingly, the first three rings show approximately the same amount of diffusion, whilst the outermost rings,  $R_4$  and  $R_5$ , bring significantly higher levels of diffusion. When we divide socially relevant alters and other types of alters, we find that the amount of information diffused in the first four rings coming from the former class is higher than that coming from the latter. In the outermost ring, we found the opposite behaviour, with socially relevant alters providing less information than other types of alters. This is perhaps not too surprising, as we expect to find alters that are not socially relevant prevalently in the outermost ring, where the level of emotional closeness or intimacy with the alters is lower than more internal rings.

Finally, to understand also the relation between the properties of the individual egos and their information diffusion patterns, we analysed the relation between the activity of the egos, defined as the sum of contact frequencies of the links of each ego network (see Section 4), and the properties of tweets originally generated by egos and

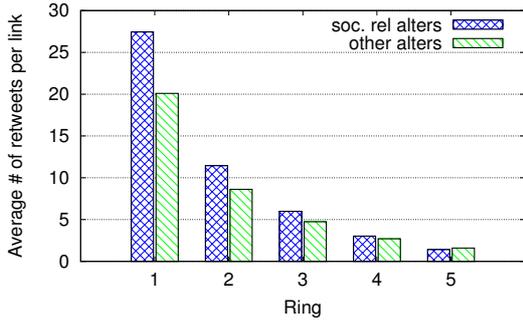


Figure 7: Average number of retweets per link divided by rings.

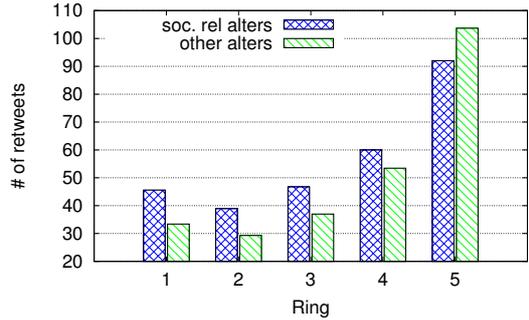


Figure 8: Average number of retweets per ego network, divided by ring.

of her retweets. The correlation between the activity of the egos and the number of tweets they generate is 0.38. By applying a logarithmic transformation to both the activity and the number of tweets generated by the egos we obtain a higher correlation (0.51), indicating a non-linear relation between the two measures. The correlation between the activity and the number of retweets generated by egos (after the logarithmic transformation) is 0.38. We also analysed the relation between activity of the egos and the average popularity of their tweets, calculated as the number of retweets they received. Activity on Twitter appears to be uncorrelated to popularity, showing a correlation value of 0.004.

## 6. Conclusion

In this paper, we presented an analysis aimed at characterising the micro-level properties of OSNs and to understand how these properties influence the formation of macro-level social phenomena, specifically the diffusion of information in the network following the word-of-mouth effect.

As far as the structure of OSNs is concerned, we have found that online ego networks show properties that are remarkably similar to those found in offline social networks. Specifically, we have analysed two data sets containing interaction data collected from Facebook and Twitter, that have been processed to obtain the online ego networks of a large number of users. The results of our analysis indicate that the structures of offline and online ego networks are compatible. In fact, we have found that the typical number of social circles in online ego networks is equal to 4 and the scaling factor between hierarchically adjacent circles is very close to 3. Moreover, the characteristic frequency of communication inside the circles is comparable with that measured offline, and that the sizes of the circles are very similar. These results are in line with the fundamental properties of human social networks found offline.

Looking in detail at the properties of the circles obtained from Facebook and Twitter, we matched them with those defined in sociology and anthropology. The results indicate that the four circles in Facebook are directly mapped with their offline equivalents. The higher richness of the

Twitter data set allowed us to discover an additional circle nested in the support clique, that we called “super support clique”. This circle is characterised by a very high frequency of contact (17.28 interaction per month) and small size, with one or two members on average. Alters inside this circle could be a partner and/or a best friend of the ego. For a long time this result has been hypothesised in sociology and psychology, but, for the lack of data, the presence of this additional circle has never been experimentally demonstrated until now.

Building on the results on ego network structures in OSNs, we performed an information diffusion analysis assessing the impact of the different ego network rings (i.e. portion of each circle not containing the other nested circles) on the process. Specifically, we performed a correlation analysis on Twitter data to assess the relation between direct contact frequency (of Twitter replies) and the frequency of retweets passing through social links. The results indicate that the two measures are highly correlated, with links in the internal ego network layers showing the highest correlation. As a further refinement of the analysis, we classified the alters of each ego network into “socially relevant users” and “other users”, and we calculated the correlations for these classes separately. Interestingly, the correlations of both classes are higher when taken separately rather than analysing them together. This could indicate the presence of two separate processes governing the diffusion of information for the two classes of alters. The different values of angular coefficients of the function explaining the relation between the measures of tie strength and information diffusion for the two classes, estimated through linear regression, support the presence of these two separate processes. The correlations found for socially relevant alters are high (higher than 0.8 for the innermost layers), indicating that the diffusion of information can be accurately explained as a function of tie strength.

A possible practical application of our results may be to use them to improve existing information diffusion models. The knowledge about the role of ego network rings in the diffusion process may lead to more representative synthetic diffusion traces than traditional models. Differences

in the structural properties of ego networks (e.g. size, number of layers, tie strength distribution) could also be useful for identifying influential information spreaders in the network. Another possible application field for the results of our analysis is related to distributed online social networks, an alternative to OSNs based on peer-to-peer communications. DOSN users would probably like to replicate their data on nodes that they trust and help disseminate content coming primarily from these nodes [47, 48]. From the analysis presented in this paper, we know that there is a significant influence of tie strength on information propagation, and DOSN system could exploit knowledge about tie strength between users to estimate the level of information diffusion, and replicate it accordingly (an initial effort in this sense is presented in [49]).

## APPENDIX

### Appendix A. Facebook and Twitter

In this section, we present a brief discussion about the main features of Facebook and Twitter with particular regard to the mechanisms they provide to the users to communicate with each other.

#### Appendix A.1. Facebook

Facebook is the most used online social networking service in the world, with roughly 1.26 billion users as of 2013. Facebook was founded in 2004 and is open to everyone over 13 years old. Facebook provides several features to the users. First, each user has a *profile* which reports her personal information and it is accessible by other users according to their permissions and the privacy settings of the user. Connected to her profile, the user has a special message board called *wall*, which reports all the asynchronous messages made by the user (*status updates*) or messages received from other users (*posts*). Posts (that include status updates) can contain multimedia information such as pictures, URLs and videos. Users can *comment* posts to create discussions around them. Comments have the same format as posts. To be able to access the personal page of other users, a user must obtain their *friendship*. A friendship is a bi-directional relation between two users. Once a friendship is established, the involved users can communicate with each other and view their personal information - depending on their privacy settings. The users can visualise the activity of their *friends* by using a special page called *news feed*.

#### Appendix A.2. Twitter

Twitter is an online social networking and microblogging service founded in 2006, with more than 500 million registered users as of 2012<sup>7</sup>. In Twitter, users can post short public messages (with at most 140 characters) called

*tweets*. All the users' tweets are accessible by other users, unless the users' profiles are private or the access is restricted by other specific settings. Users can also automatically receive notifications of new tweets created by other users by "following" them (i.e. creating a subscription to their notifications). People following a specific user are called her *followers*, whilst the set of people followed by the user are her *friends*.

Tweets can be enriched with multimedia content (i.e. URLs, videos, pictures) and by using special text characters to insert additional information. Specifically, a tweet can reference one or more users with a special mark called *mention*. Users mentioned in a tweet automatically receive a notification, even though they are not followers of the tweet's author. Users can also *reply* to tweets. In this case, a tweet is generated with an implicit mention to the author of the replied tweet. This implies that replies represent directional communications. Replies often require additional effort in terms of cognitive resources compared to other tweets since they presuppose that the user creating the reply has read the tweet she is replying. Twitter has also a private messaging system, however, since private messages are not publicly accessible, we did not collect them in our data set.

In addition to mentions and replies, Twitter provides a series of mechanisms for broadcast communication that represent the most popular features of the platform. First, all the tweets are automatically sent towards all the followers of their authors. Moreover, tweets can also be *retweeted*. A user can make a retweet to forward a tweet it to all her followers. Each tweet can be assigned to a topic through the use of a special character called hashtag (i.e. "#") placed before the text indicating the topic. Hashtags are used by Twitter to classify the tweets and to obtain *trending topics*.

### Appendix B. Classifier for the selection of socially relevant users in Twitter

To build the supervised learning classifier used to select socially relevant users from Twitter data set (see Section 3 for more details), we manually classified a sample of 500 accounts, randomly drawn from the data set, and we used this classifications to train a Support Vector Machine [50]. This SVM uses a set of 115 variables: 15 of them related to the user's profile (e.g., number of tweets, number of following and followers, account lifespan) and 100 obtained from her timeline (e.g., percentage of mentions, replies and retweets, average tweets length, number of tweets made using external applications).

To test the generality of the SVM (i.e., the ability to categorise correctly new examples that differ from those used for training), we took 10 random sub-samples of the training set, each of which contains 80% of the entries, keeping the remaining 20% for testing. Then, we applied the same methodology used to create the SVM generated

<sup>7</sup>According to Twitter CEO Dick Costolo in October 2012.

from the entire training set on the 10 sub-samples. Doing so, we obtained different SVMs, trained using different sub-samples of the training set, and of which we were able to assess the accuracy. The average accuracy of these SVMs can be seen as an estimate of the accuracy of the SVM derived from the complete training set. Specifically, we calculated the *accuracy* index, defined as the rate of correct classifications, and the *false positives rate*, where false positives are accounts wrongly assigned to the “socially relevant user” class. In our analysis, we considered only users falling in the “socially relevant users” class, thus it is particularly important to minimise the false positive rate<sup>8</sup>. Minimising the false negative rate is also important but less critical, as false negatives result in a reduction of the number of users on which we base our analysis.

The average accuracy of our classification system is equal to 0.813 [ $\pm 0.024$ ] and the average false positives rate is 0.083 [ $\pm 0.012$ ] (values between brackets are 95% confidence interval). These results indicate that we were able to identify socially relevant people in Twitter with sufficient accuracy, even if people have different behaviours and characteristics (e.g., different culture, religion, age). Moreover, the false positive rate is quite low (below 10%). The results are of the same magnitude as those found in a similar classification performed in Twitter [51].

## Appendix C. Frequency of contact estimation in Facebook

In this section, we provide details about the procedure that we used to estimate the frequency of contact between users in the Facebook data set described in Section 3. As described in the text, the data set is divided into snapshots representing four temporal windows containing the number of interactions occurred between the users during the considered time period.

### Appendix C.1. Definitions

We define the temporal window “last month” as the interval of time  $(w_1, w_0)$ , where  $w_1 = 1$  month (before the crawl) and  $w_0 = 0$  is the time of the crawl. Similarly, we define the temporal windows “last six months”, “last year” and “all” as the intervals  $(w_2, w_0)$ ,  $(w_3, w_0)$  and  $(w_4, w_0)$  respectively, where  $w_2 = 6$  months,  $w_3 = 12$  months and  $w_4 = 43$  months.  $w_4$  is the maximum possible duration of a social link in the data set, obtained by the difference between the time of the crawl (April 2008) and the time Facebook started (September 2004). The different temporal windows are depicted in Fig. C.9.

For a social relationship  $r$ , let  $n_k(r)$  with  $k \in \{1, 2, 3, 4\}$  be the number of interactions occurred in the temporal

<sup>8</sup>False negatives are “socially relevant users” with behaviour similar to the subjects in the “other users” class. For this reason we consider them as outliers, since our analysis is focused on Twitter average users.

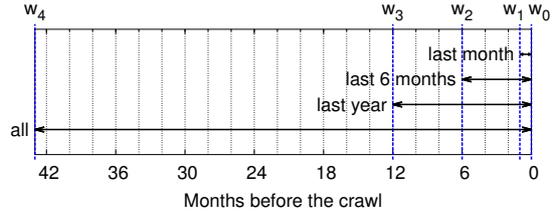


Figure C.9: Temporal windows.

window  $(w_k, w_0)$ . Since all the temporal windows in the data set are nested,  $n_1 \leq n_2 \leq n_3 \leq n_4$ . If no interactions occurred during a temporal window  $(w_k, w_0)$ , then  $n_k(r) = 0$ . As a consequence of our definition of active relationship, since  $n_4(r)$  refers to the temporal window “all”,  $n_4(r) > 0$  only if  $r$  is an active relationship, otherwise, if  $r$  is inactive,  $n_4(r) = 0$ .

The first broad estimation that we can do to discover the duration of social ties in the data set is to divide the relationships into different classes  $C_k$ , each of which indicates in which interval of time  $(w_k, w_{k-1})$  the relationships contained in it has started (i.e. the first interaction has occurred). We can perform this classification by analysing, for each relationship, the number of interactions in the different temporal windows. If all the temporal windows contain the same number of interactions, the relationship must be born less than one month before the time of the crawl, that is to say in the time interval  $(w_1, w_0)$ . These relationships belong to the class  $C_1$ . Similarly, considering the smallest temporal window (in terms of temporal size) that contains the total number of interactions (equal to  $n_4$ ), we were able to identify social links with duration between one month and six months (class  $C_2$ ), six months and one year (class  $C_3$ ), and greater than one year (class  $C_4$ ). The classes of social relationships are summarised in Table C.9.

### Appendix C.2. Estimation of the Duration of the Social Links

Although the classification given in the previous subsection is extremely useful for our analysis, the uncertainty regarding the estimation of the exact moment of the establishment of social relationships is still too high to obtain significant results from the data set. For example, the duration of a social relationship  $r_3 \in C_3$  can be either a few days more than six months or a few days less than one year. To overcome this limitation, for each relationship  $r$  in the

Table C.9: Facebook classes of relationships.

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

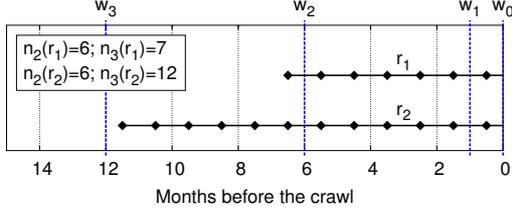


Figure C.10: Graphical representation of two social relationships with different duration.

classes  $C_{k \in \{2,3,4\}}$ , we estimate the time of the first interaction comparing the number of interactions  $n_k$ , made within the smallest temporal window in which the first interaction occurred  $(w_k, w_0)$ , with the number of interactions  $(n_{k-1})$ , made in the previous temporal window in terms of temporal size  $(w_{k-1}, w_0)$ . If  $n_k(r)$  is much greater than  $n_{k-1}(r)$ , a large number of interactions occurred within the time interval  $(w_k, w_{k-1})$ . Assuming that these interactions are distributed in time with a frequency similar to that in the window  $(w_{k-1}, w_0)$ , the first occurred interaction must be near the beginning of the considered time interval. On the other hand, a little difference between  $n_k(r)$  and  $n_{k-1}(r)$  indicates that only few interactions occurred in the considered time interval  $(w_k, w_{k-1})$ . Thus, assuming an almost constant frequency of interaction, the first contact between the involved users must be at the end of the time interval. The example in Figure C.10 is a graphical representation of this concept.

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

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

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

If  $r \in C_1$  we set  $h(r) = 1$  in order to be able to perform the remaining part of the processing also for these relationships. The value assigned to  $h(r)$  with  $r \in C_1$  is arbitrary and can be substituted by any value other than zero without affecting the final result of the data processing. Considering that  $n_k(r)$  is greater than  $n_{k-1}(r)$  by definition with  $r \in C_{k \in \{2,3,4\}}$ , the value of  $h(r)$  is always in the interval  $(0, \infty)$ <sup>9</sup>.

Employing the social interaction ratio  $h(r)$ , we define the function  $\hat{d}(r)$  which, given a social relationship  $r \in C_k$ ,

<sup>9</sup>In case  $n_{k-1}(r) = 0$ , we set  $n_{k-1}(r) = 0.3$ . This constant is the expected number of interactions when the number of interactions, within a temporal window, is lower than 1.

estimates the point in time at which the first interaction of  $r$  occurred, within the time interval  $(w_k, w_{k-1})$ :

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

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

Note that the value of  $\hat{d}(r)$  is always in the interval  $(w_{k-1}, w_k)$ . The greater  $h(r)$  - which denotes a lot of interactions in the time window  $(w_k, w_{k-1})$  - the closer  $\hat{d}(r)$  is to  $w_k$ . The smaller  $h(r)$ , the closer  $\hat{d}(r)$  is to  $w_{k-1}$ . Moreover, the shape of  $\hat{d}(r)$  and the value of  $a_k$  are chosen relying on the results about the Facebook growth rate, available in [18]. Specifically, the distribution of the estimated links duration, given by the function  $\hat{d}(r)$ , should be as much similar as possible to the distribution of the real links duration, which can be obtained analysing the growth trend of Facebook over time. For this reason, we set the constants  $a_k$  in order to force the average link duration of each class of relationships to the value that can be obtained by observing the Facebook growth rate. In [52] we provide a detailed description of this step of our analysis.

### Appendix C.3. Estimation of the Frequency of Contact

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

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

Previous research work demonstrated that the pairwise user interaction decays over time and it has its maximum right after link establishment [53]. Therefore, if we assessed the intimacy level of the social relationships with their contact frequencies, this would cause an overestimation of the intimacy of the youngest relationships. In order to overcome this problem, we multiplied the contact frequencies of the relationships in the classes  $C_1$  and  $C_2$  by the scaling factors  $m_1$  and  $m_2$  respectively, which correct the bias introduced by the spike of frequency close to the establishment of the link. Assuming that the relationships established more than six months before the time of the crawl are stable, we set  $m_1$  and  $m_2$  comparing the average contact frequency of each of the classes  $C_1$  and  $C_2$ , with that for the classes  $C_3$  and  $C_4$ . The obtained values of the scaling factors are:  $m_1 = 0.18$ ,  $m_2 = 0.82$ . Setting  $m_3 = 1$  and  $m_4 = 1$ , the scaled frequencies of contact are defined as:

$$\hat{f}(r) = f(r) \cdot m_k \quad r \in C_k. \quad (\text{C.4})$$

- [1] M. Conti, S. Das, C. Bisdikian, M. Kumar, L. M. Ni, A. Passarella, G. Roussos, G. Tröster, G. Tsudik, F. Zambonelli, Looking Ahead in Pervasive Computing: Challenges and Opportunities in the Era of Cyber-Physical Convergence, *Pervasive and Mobile Computing* 8 (1) (2012) 2–21. doi:<http://dx.doi.org/10.1016/j.pmcj.2011.10.001>.

- [2] P. Domingos, M. Richardson, Mining the Network Value of Customers, in: KDD '01, 2001, pp. 57–66.
- [3] E. Bakshy, I. Rosenn, C. Marlow, L. Adamic, The Role of Social Networks in Information Diffusion, in: WWW '12, 2012, pp. 519–528.
- [4] M. E. Newman, The Structure and Function of Complex Networks, *SIAM Review* 45 (2) (2003) 167–256. doi:10.1137/S003614450342480.
- [5] J. Travers, S. Milgram, An Experimental Study of the Small World Problem, *Sociometry* 32 (4) (1969) 425. doi:10.2307/2786545.
- [6] D. J. Watts, S. H. Strogatz, Collective Dynamics of "Small-World" Networks, *Nature* 393 (6684) (1998) 440–2. doi:10.1038/30918.
- [7] M. E. Newman, J. Park, Why Social Networks are Different From Other Types of Networks, *Physical Review E* 68 (3). doi:10.1103/PhysRevE.68.036122.
- [8] M. S. Granovetter, The Strength of Weak Ties, *The American Journal of Sociology* 78 (6) (1973) 1360–1380. doi:10.2307/2776392.
- [9] P. V. Marsden, K. E. Campbell, Measuring Tie Strength, *Social Forces* 63 (2) (1984) 482–501. doi:10.2307/2579058.
- [10] E. Gilbert, K. Karahalios, Predicting Tie Strength with Social Media, in: CHI '09, 2009, pp. 211–220.
- [11] E. Gilbert, Predicting Tie Strength in a New Medium, in: CSCW '12, 2012, pp. 1047–1056.
- [12] V. Arnaboldi, A. Guazzini, A. Passarella, Egocentric Online Social Networks: Analysis of Key Features and Prediction of Tie Strength in Facebook, *Computer Communications* 36 (10–11) (2013) 1130–1144. doi:10.1016/j.comcom.2013.03.003.
- [13] I. Kahanda, J. Neville, Using Transactional Information to Predict Link Strength in Online Social Networks, in: ICWSM '09, 2009, pp. 74–81.
- [14] 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 (1) (2013) e52168. doi:10.1371/journal.pone.0052168.
- [15] Y. Volkovich, S. Scellato, D. Laniado, C. Mascolo, A. Kaltenbrunner, The Length of Bridge Ties: Structural and Geographic Properties of Online Social Interactions, in: ICWSM '12, 2012.
- [16] J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A. L. Barabási, Structure and Tie Strengths in Mobile Communication Networks, *PNAS* 104 (18) (2007) 7332–7336. doi:10.1073/pnas.0610245104.
- [17] N. Z. Gong, W. Xu, L. Huang, P. Mittal, E. Stefanov, V. Sekar, D. Song, Evolution of Social-Attribute Networks: Measurements, Modeling, and Implications using Google+, in: IMC '12, 2012, pp. 131–144.
- [18] C. Wilson, A. Sala, K. P. Puttaswamy, B. Y. Zhao, Beyond Social Graphs: User Interactions in Online Social Networks and Their Implications, *ACM Transactions on the Web* 6 (4) (2012) 1–31. doi:10.1145/2382616.2382620.
- [19] J. Leskovec, E. Horvitz, Planetary-Scale Views on an Instant-Messaging Network, Tech. rep. (2007). arXiv:arXiv:0803.0939v1.
- [20] P. A. Grabowicz, J. J. Ramasco, B. Gonçalves, V. M. Eguíluz, Entangling mobility and interactions in social media., *PloS one* 9 (3) (2014) e92196. doi:10.1371/journal.pone.0092196.
- [21] S. G. Roberts, Constraints on Social Networks, in: *Social Brain, Distributed Mind (Proceedings of the British Academy)*, 2010, pp. 115–134. doi:10.5871/bacad/9780197264522.001.0001.
- [22] S. G. Roberts, R. I. Dunbar, T. V. Pollet, T. Kuppens, Exploring Variation in Active Network Size: Constraints and Ego Characteristics, *Social Networks* 31 (2) (2009) 138–146. doi:10.1016/j.socnet.2008.12.002.
- [23] R. I. Dunbar, The Social Brain Hypothesis, *Evolutionary Anthropology* 6 (5) (1998) 178–190. doi:10.1002/(SICI)1520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-8.
- [24] W. X. Zhou, D. Sornette, R. A. Hill, R. I. Dunbar, Discrete Hierarchical Organization of Social Group Sizes, *Biological Sciences* 272 (1561) (2005) 439–44. doi:10.1098/rspb.2004.2970.
- [25] A. Sutcliffe, R. I. Dunbar, J. Binder, H. Arrow, Relationships and the Social Brain: Integrating Psychological and Evolutionary Perspectives, *British Journal of Psychology* 103 (2) (2012) 149–68. doi:10.1111/j.2044-8295.2011.02061.x.
- [26] B. Gonçalves, N. Perra, A. Vespignani, Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number, *PloS one* 6 (8) (2011) e22656. doi:10.1371/journal.pone.0022656.
- [27] V. Arnaboldi, M. Conti, A. Passarella, R. I. Dunbar, Dynamics of Personal Social Relationships in Online Social Networks: a Study on Twitter, in: COSN '13, 2013, pp. 15–26.
- [28] D. Quercia, L. Capra, J. Crowcroft, The social world of Twitter: Topics, geography, and emotions, *ICWSM '12*.
- [29] R. S. Burt, *Structural Holes versus Network Closure as Social Capital*, 2001.
- [30] S. Aral, M. V. Alstytne, The Diversity-Bandwidth Tradeoff.
- [31] V. Arnaboldi, M. Conti, A. Passarella, F. Pezzoni, Analysis of Ego Network Structure in Online Social Networks, in: *Social-Com '12*, 2012, pp. 31–40.
- [32] V. Arnaboldi, M. Conti, A. Passarella, F. Pezzoni, Ego Networks in Twitter: an Experimental Analysis, in: *NetSciCom '13*, 2013, pp. 3459–3464.
- [33] R. I. Dunbar, V. Arnaboldi, M. Conti, A. Passarella, The structure of online social networks mirrors those in the offline world, *Social Networks* 43 (2015) 39–47. doi:10.1016/j.socnet.2015.04.005.
- [34] P. Cui, S. Jin, L. Yu, F. Wang, W. Zhu, S. Yang, Cascading outbreak prediction in networks: a data-driven approach, in: *KDD '13*, 2013.
- [35] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance, Cost-effective outbreak detection in networks, in: *KDD '07*, ACM Press, New York, New York, USA, 2007, p. 420. doi:10.1145/1281192.1281239.
- [36] M. Cha, F. Benevenuto, The world of connections and information flow in twitter, *Transactions on Systems, Man, and Cybernetics* 42 (4) (2012) 991–998.
- [37] J. Goldenberg, B. Libai, E. Muller, Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth, *Marketing Letters* 12 (3) (2001) 211–223. doi:10.1023/a:1011122126881.
- [38] S. Myers, C. Zhu, J. Leskovec, Information diffusion and external influence in networks, in: *SIGKDD '12*, 2012. arXiv:arXiv:1206.1331v1.
- [39] V. Arnaboldi, M. Conti, M. La Gala, A. Passarella, F. Pezzoni, Information Diffusion in OSNs: the Impact of Nodes' Sociality, in: *SAC '14*, 2014, pp. 1–6.
- [40] F. Pezzoni, J. An, A. Passarella, J. Crowcroft, M. Conti, Why Do I Retweet It? An Information Propagation Model for Microblogs, in: *SocInfo '13*, 2013, pp. 360–369.
- [41] X. Zhao, A. Sala, C. Wilson, X. Wang, S. Gaito, H. Zheng, B. Y. Zhao, Multi-scale dynamics in a massive online social network, in: *IMC '12*, 2012, pp. 171–184.
- [42] A. Passarella, M. Conti, R. I. Dunbar, C. Boldrini, Modelling Inter-contact Times in Social Pervasive Networks, in: *WSWiM '11*, 2011, pp. 333–340.
- [43] A. Passarella, M. Conti, Characterising Aggregate Inter-Contact Times in Heterogeneous Opportunistic Networks, in: *Networking '11*, 2011, pp. 1–12.
- [44] H. P. Kriegel, P. Kröger, J. Sander, A. Zimek, Density-Based Clustering, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1 (3) (2011) 231–240.
- [45] H. Wang, M. Song, Clustering in One Dimension by Dynamic Programming, *The R Journal* 3 (2) (2011) 29–33.
- [46] R. I. M. Dunbar, No Title, Private communication (Jun. 2012).
- [47] V. Arnaboldi, M. L. La Gala, A. Passarella, M. Conti, The Role of Trusted Relationships on Content Spread in Distributed Online Social Networks, in: *LSDVE '14*, 2014, pp. 287–298.
- [48] V. Arnaboldi, M. La Gala, A. Passarella, M. Conti, Information Diffusion in Distributed OSN: the Impact of Trusted Relationships, Peer-to-Peer Networking and Applications (accepted for publication).

- [49] M. Conti, A. D. Salve, B. Guidi, L. Ricci, Epidemic Diffusion of Social Updates in Dunbar-Based DOSN, in: EuroPar '14, 2014, pp. 311–322.
- [50] C. Cortes, V. Vapnik, Support-Vector Networks, Machine Learning 20 (3) (1995) 273–297. doi:10.1023/A:1022627411411.
- [51] Z. Chu, S. Gianvecchio, H. Wang, S. Jajodia, Who is Tweeting on Twitter: Human, Bot, or Cyborg?, in: ACSAC '10, 2010, pp. 21–30.
- [52] V. Arnaboldi, M. Conti, A. Passarella, F. Pezzoni, Analysis of Ego Network Structure in Online Social Networks, Tech. rep. (2012).
- [53] B. Viswanath, A. Mislove, M. Cha, K. P. Gummadi, On the Evolution of User Interaction in Facebook, in: WOSN, 2009, pp. 37–42.