UNIVERSITÀ DI PISA
Scuola di Dottorato in Ingegneria "Leonardo da Vinci"

Corso di Dottorato di Ricerca in
INGEGNERIA DELL'INFORMAZIONE

Tesi di Dottorato di Ricerca

Titolo
Analysis of the Structure of Social Networks for Information Diffusion

Autore
Massimiliano La Gala

Anno 2015

SSD ING-INF/05

UNIVERSITÀ DI PISA

Scuola di Dottorato in Ingegneria "Leonardo da Vinci"

Corso di Dottorato di Ricerca in
INGEGNERIA DELL'INFORMAZIONE

Tesi di Dottorato di Ricerca

Analysis of the Structure of Social Networks for Information Diffusion

Autore:

Massimiliano La Gala _____

Relatori:

Prof. Enzo Mingozzi _____

Dott. Marco Conti _____

Ing. Andrea Passarella _____

Anno 2015
SSD ING-INF/05

The vast proliferation of Online Social Networks (OSN) is generating many new ways to interact and create social relationships with others. In OSN, information spreads among users following existing social relationships. This spread is influenced by the local properties and structures of the social relationships at individual level. Being able to understand these properties can be fundamental for the design of new communication systems able to predict the creation and sharing of content based on social properties of the users. While substantial results have been obtained in anthropology literature describing the properties of human social networks, a clear understanding of the properties of social networks built using OSN is still to be achieved.

In this thesis, the structure of Ego networks formed online is compared with the properties of offline social relationships showing interesting similarities. These properties are exploited to provide a meaningful way to study the mechanisms controlling the formation of information diffusion chains in social networks (typically referred to as information cascades). Trough the analysis of synthetically generated diffusion cascades executed in a large Facebook communication datasets, is showed that the knowledge of tie strength of the social links is fundamental to infer which nodes will give rise to large information cascades and which links will be more used in the information diffusion process. We analysed the trade off between information spread and trustworthiness of information. Specifically, we have investigated the spread of information when only links of a certain trust value are used. Assuming, based on results from sociology, that trust can be quantised, we show that too strict limits on the minimum trust between users limit significantly information spread. In the thesis we investigate the effect of different strategies to significantly increase spread of information by minimally relaxing constraints on the minimum allowed trust level.

Contents

List of Figures

List of Tables

1

Introduction

In the past few years, Online Social Networks (OSNs) became a really popular way to communicate and exchange informations. The services provided can be really different from one OSN to another, but they usually offer easy ways to create contents that can be shared with others. Some of the most popular, allow the users to declare their social relationships (e.g. “friendships” in Facebook, “following” and “follower” in Twitter), permitting to immediately recognise the social network of the users. Since these services, record the activities of a enormous number of users, their collected data is really valuable in the analysis of the human communication behaviours and potentially can permit to obtain significant insights on how users exploit these social communication platforms. Unfortunately, OSNs typically offer limited access to these data making data collection and analysis quite difficult.

Social Networks have been extensively studied by sociologists in the “offline” world (e.g. face to face communications) using manually collected datasets that contain the interactions of relatively small amounts of individuals if compared with datasets collected from OSNs. Nevertheless they extrapolated really interesting properties on the structure of the human social networks. For example, they discovered that social networks presents “small world” properties with small average diameter [88]. The structure of social networks also shows a typical high clustering factor [92]. Moreover differentiating the social relationships in “friend” (close relationship) or in “acquaintance” (loose relationships) they have discovered interesting structural properties. In fact, the relationships that connect clusters otherwise distant in the network are usually “weak ties” (i.e. relationships maintained with an acquaintance), while the “strong ties” (i.e. relationship maintained with good friends) are usually located inside the clusters [43]. Also the kind of information that passes through these kinds of relationships are different, for example

it is more likely to acquire information about a new job trough weak ties than a strong ties. In fact, even if a friend is more motivated than an acquaintance to pass an important information like an open job position, through the latter, a user can have access to distant parts of the social network, and thus to a more diverse set of information, with respect to what it can retrieve from close friends (which are typically quite clustered) [42].

Another fundamental result obtained in the sociology literature is the analysis of the structure of the social network formed by a single person (called *ego*) with all his friends (called *alters*) in which the tie strength - a measure of the importance of the relationship - is not just binary (i.e., weak or strong) but changes as a continuous variable. The emerging structures in this social network, can be represented by 4 concentric layers - "social circles" - in which from the innermost to the outermost layer the contained relationships changes from very strong relationships to acquaintances. The interesting part is that these circles have a constant scaling factorof 3 (i.e., the number of alters in a given layer are around three times the number of alters in the layer immediately closer to the ego) and the members of a single circle share similar relationships characteristic [85].

The structure of the social networks influences the social phenomena that occurs in the network, for example an individual with many strong friendships with influential persons can affect the social network easier than a person with few, loosely connected friendships. Thus, it is interesting to analyse the impact of the structure of the ego network in a network wide phenomena like information diffusion.

In this thesis we analyse the word-of-mouth effect in Online Social Networks, i.e. the way how information spreads across users of OSN platforms, as a function of the structural properties of the social networks created by users on this platforms. This is an important phenomenon, with strong implications in marketing, which behaviour strongly depends on the structure of the social network since the established social relationships are the only channels through which the information is spread.

In the thesis we also consider information diffusion in another type of social network platforms, i.e., Distributed Online Social Networks (DOSN) [74]. DOSN are an emerging research area, mainly motivated by the fact that in conventional OSN a central operator exists, that in principle controls all the information generated by the users. DOSN aims, instead, to completely decentralise the mechanisms required to support Online Social interactions, avoiding centralised controllers. The Distributed Online Social Networks (DOSNs) are systems that provide a similar service to traditional OSNs, but without the presence of an operator that central-

ize, and thus is able to control the service. In these systems, the infrastructure necessary to deliver the services is obtained thanks to users sharing computational, storage, network resources. Specifically, in DOSN the properties of information spread are also dependent on the trust between users, because there is a lack of centralised operator entrusted by the users. In the thesis we analyse the impact of social structures on information spread also in this type of social networking environment.

1.1 Contributions

The main focus of the thesis is studying the impact of social structures formed by users in Online Social Networks on the properties of information diffusion. To this end, we have first exploited and refined some previous results on the analysis of OSN structures, and then used models of these structures to study the properties of information diffusion. The analyses are based on two datasets from Twitter and Facebook which were analysed to study the local-level properties of the structure on the social networks in online environments. The emerged characteristics are used to evaluate their impact on a network-level phenomena like the information diffusion.

1.1.1 Structure of the Ego Network in OSNs

The first contribution of this thesis is a characterization of the structure of the ego networks formed by considering separately the incoming messages and the outgoing messages. This analysis is useful to assess any difference in their in their contribution to the structural properties of the ego network. The results indicate that the two networks presents different characteristics, indeed the incoming messages network forms a bigger ego-network which only partially contains the network formed by outgoing messages. Moreover, to better approximate the tie strength, we propose an index which combine the two kind of messages and gives an higher score to relationships with reciprocated interactions. This index produces an ego network structure closer to that found in the anthropological literature. This results are presented in Section 4.2.

1.1.2 Impact of Network Properties in Information Diffusion

The second contribution of the thesis is to analyse how the local properties of the ego network of the users impact on a large scale phenomena, namely diffusion

of information (presented in Section 5). Two different analyses are presented. The first one analyses how the various social circles affect the information diffusion process, while the second analyses studies how key structural properties of nodes in the social network graph (such as their centrality and degree) impact on the diffusion of information originated by them.

Both analyses are carried out through simulations based on data crawled from real online social networks. In particular, the network graph (i.e. the set of nodes and the friendship relationships) are derived from a Facebook dataset, and the information diffusion over links depends on the estimated tie strength between users. The adopted model of information diffusion is the Independent Cascade Model (ICM).This model requires that a probability of diffusion is assigned to each link. We shows that calculating the probability of diffusion with a linear function of the frequency of contacts leads to the flooding of the network. Thus we developed two different models that solves this problem. In the first model we included the notion of information ageing with a coefficient that decrease the probability of diffusion at each steps of diffusion. The second model uses a non linear transformation validated through the comparison of the fitting parameters of the resulting distribution with the fitting parameters of real communication traces.

The first analysis studies the importance of different social circles in the information diffusion process. The first step of this analysis was to calculate the usage in the diffusion process of each social circle. The results shows that middle layers are the most active. The analysis continued in the second step in which social links were selectively cut out from the social network in order to remove single circles or groups of circles in order to assess if any circle is essential in the information diffusion process. The results shows that the removal of the social circles greatly impact on information diffusion. Noteworthy, the impact of the innermost circle, which contains only 0.3% of the links, reduced the number of exchanged messages to 17%.

The second analysis studies the impact on the characteristics of the obtained diffusion cascades of the centrality measures of the starting node. The centrality measures considered takes into account both local properties, like the number of friends or the clusterization of the ego network, and network wide properties like pagerank or eigenvector centrality. We discovered that the highest correlations is obtained with centrality measures which involve tie strength, like the eigenvector centrality. The Burt's constrain, has a medium negative correlation, which indicates that strongly connected clusters entrap the information making harder the further diffusion of the information. Interestingly, the same analysis executed in an unweighted network shows really low correlations.

1.1.3 Impact of Restrictive Privacy Settings in Information Diffusion

The third contribution of the thesis (presented in Section6) is to study the impact of trust between users on information diffusion. Specifically, based on existing results in the sociology literature, we assume that trust between users is related to the strength of the social tie between them. We analyse how information diffusion changes when only links with a minimum level of trust are considered. This is particularly relevant for Distributed Online Social Networks environments, where information flows only because users devices collaborate among them in the information diffusion process. In this analysis we selected 4 thresholds of frequency of contact, and defined 4 trust thresholds corresponding to these frequencies (that can be mapped to corresponding tie strengths). These thresholds correspond to minimum values of frequency of contact in each social circle. We analysed how information spread is affected when only links above each of these thresholds are used. As expected, this operation greatly reduced the number of nodes connected to the largest component of the social network (the giant component) especially for the more restrictive hypothesis, and also reduces the capability of the network to spread information.

To improve the capability of the network to spread informations without the use of a less restrictive threshold, we analysed the effect of the reintroduction of a single previously removed relationship per each node. With this mechanism with just a minimal reduction of the trust level between users (just one relationship does not belongs to the original minimal threshold) we obtained great improvement in both the characteristics analysed (size of the giant component, and capability to spread information).

We proposed various strategies to select the removed relationship, both deterministic and stochastic. The two deterministic strategies selects the relationship with maximum or minimum frequency of contact, while the three stochastic strategies select the links assigning a probability associated to each relationship. In two strategies, the probability was assigned proportionally/inversely proportionally to the frequency of contact, while in the third strategy the probability was uniformly distributed. We presents and compare the results of each of these strategies.

Results show that a network composed of individual willingly to communicate with highly trusted friends only has a really low capability to spread information. However, if each user accept to communicate with just one less-trusted user, the network restores much of the original capacity to spread information whatever selection strategy is used. Nevertheless, the best performing strategy is the one that selects the relationship with the maximum frequency of contact. With this

strategies, the selected relationship will not have necessarily an high interaction frequency value, because some ego network could miss some social circle.

1.2 Thesis Organisation

In Chapter 3 we present the Facebook and Twitter datasets used in this work. In Chapter 4 the datasets are examined to extrapolate the properties of the structure of the social network in these two OSNs and are compared with the findings in sociological literature. Chapter 5 analyses the impact of the properties of the ego networks in information diffusion, studying the role of the social circles and the centrality of the ego in the network. Chapter 6 presents an analysis on the capacity of a Distributed Online Social Network to diffuse information under the hypothesis that only a subset of the social relationships can be used. Chapter 7 summarises the main results of the thesis.

2

Related Works

Human sociality has always been the centre of attention of many different research fields. Indeed, in social sciences there exist a number of disciplines which aim to understand social behaviour from all its different aspects. To analyse collective social behaviour, sociologists started to use social networks as a handy model to represent groups of socially connected people. The theoretical analysis of social networks gave rise to a new research discipline called Social Network Analysis (SNA). SNA views individuals in a social network and their social relationships as vertices and edges (or arcs) of a graph, and then studies the properties of the graph to describe social phenomena in our society, in terms of the structural properties of the network. The extensive work done in SNA led to a deeper comprehension of plenty of social phenomena.

2.1 Social Network Analysis

Many important properties of social networks have been found in SNA literature. Mark Granovetter realised that a fundamental aspect of social networks is represented by the relation between micro-level interactions of social actors and macro-level patterns arising in the networks. He found that the strength of social ties, informally defined as a linear combination of time, emotional intensity, intimacy, and reciprocal services, impacts to a large extent on social networks' phenotypical properties [43]. Moreover, Social relationships can be roughly divided into strong and weak ties, where the former denote more important relationships and the latter represents acquaintances. Besides their lower strength, weak ties are generally more in number 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. Granovetter demonstrated that the analysis of tie strength is fundamental to fully assess the properties of social networks. He also introduced models of collective behaviours (e.g. the adoption of a new idea or the spread of information among social groups) based on the concept of tie strength to emulate social dynamics [44].

Another important contribution to the field has been made by Peter Marsden, who used multiple indicator techniques to construct and validate measures of tie strength [64]. Mardsen built an analytical model to explain the relation between a set of tie strength predictors (i.e. aspects of relationships that are related to, but not components of, tie strength) and tie strength indicators (emotional closeness, duration, frequency of contact, breadth of discussion topics, and confiding). The results of his analysis demonstrate that emotional closeness (or emotional intensity) is the best indicator of the strength of a social relationship. Moreover, measures of the time spent in a relationship (e.g. frequency of contact and duration) are related to the concept, even though they tend to systematically overstimate tie strength in case the involved persons are co-workers or neighbours. These results indicate that tie strength can be effectively estimated using some measurable indicators. This fact opened the door to further analyses on structural properties of social networks. In particular, evolutionary psychologists largely studied micro-level properties of social networks using measures of emotional closeness and frequency of contact to analyse social aspects of humans.

2.1.1 Micro-Level Structural Properties of Social Networks

A standard approach to the study of micro-level structural properties of social networks is the analysis of ego networks. An ego network is a simple social network model formed of an individual (called ego) and all the persons with whom the ego has a social link (alters). Ego networks are useful to study the properties of human social behaviour at a personal level, and to assess the extent to which individual characteristics of the ego affect the size and the composition of their network. For this reason they have been largely used in Sociology and Psychlogy. The most important result found on ego networks is that the cognitive constraints of human brain and the limited time that a person can use for socialising bound the number of social relationships that an ego can actively maintain in his/her network. This limit lies, on average, around 150 and is known as the *Dunbar's number* [33]. This result has been further confirmed by various experiments, and the Dunbar's number has been empirically estimated to a value equal to 132.5 [96]. The presence of the Dunbar's number in humans is in accordance with the idea of bounded rationality previously introduced by Herbert Simon [82].

An individual ego can be envisaged as sitting at the centre of a series of concentric circles of alters ordered by the strength of their social ties [77]. Each of these circles has typical size and frequency of contact between the ego and the alters contained in it. The first circle, called *support clique* contains alters with very strong social relationships with the ego, informally identified in 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 identified as *close friends*. This circle contains 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 50 alters usually representing causal friends or extended family members [78]. Although some studies tried to identify the typical frequency of contact of this circle, there are no accurate results in literature about their properties, due to the difficulties related to the manual collection of data about the alters contained in it through interviews or surveys. Indeed, people hardly remember people besides their best and close friends. The last circle in the ego network model is the *active network*, which contains all the other circles, for a total of 150 members. This circle is bounded by the limit of the Dunbar's number and 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. Alters beyond the active network are considered inactive. 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 [85].

2.1.2 Macro-Level Phenomena Observed in Social Networks

Seen from a macro-level perspective, social networks show some typical properties that have been observed in many different environments. Stanley Milgram, through his famous experiment, demonstrated the presence of the so called *small-world effect* in social networks [88]. According to this property, any two persons in the network, indirectly connected by chains of social links, have a short average distance. This is often identified as the *six degrees of separation* theory, for which everyone in a social network is six steps away. This fact directly influences the ability of the network to quickly spread information, ideas, innovations and so forth. It has been demonstrated that the diffusion of information in social networks takes place through single social links, creating the *word-of-mouth* effect. This property

has been largely used by a collection of marketing techniques whereby the presence of social links between consumers is exploited to increase sales [50].

Other distincive properties of social networks, that differentiate them from other types of networks, including technological and biological networks, are represented by the presence of a non-trivial clustering or network transitivity, that is, in other words, a high probability that two neighbours connected to a node will also be connected to each other. Moreover, social networks show positive correlations between the degrees of adjacent vertices, also called *assortativity* [69].

Based on the properties found in social networks many different models have been proposed to replicate the dynamics of several social phenomena. Similarly to what happens during a virus contagion, the diffusion of information produces a series of cascades. Hence, the traces left by the spread of information are called *information cascades*. Some models aim to reproduce information cascades relying upon the fact that nodes are "infected" by information with a probability proportional to the number of their neighbours which are already infected (see for example [37, 44]).

2.2 Online Social Networks

The power of SNA attracted plenty of disciplines, like anthropology, communication studies, biology, physics, history, political science and many others. The use of computationally intensive methods in SNA has recently originated a new category of social disciplines under the name of Computational Social Science. The advent of OSNs fostered analyses on social networks, since the abundance of online communication traces generated by social media allowed to overcome the problem of collecting large-scale social data sets that was posing strong limits to social sciences hitherto.

2.2.1 Macro-Level Properties of Online Social Networks

The availability of OSNs communication data allowed to reveal the presence of some distinctive social traits also in online environments. Specifically, the small world effect has been found in social graphs representing instant-message interactions between people [62, 31]. In [66] the authors present a detailed analysis of the macro-level structural properties of a set of different OSNs, finding results in accordance with the properties of social networks observed in offline environments.

The different roles of weak and strong ties has been confirmed in [70], where the authors analyse phone logs containing communication traces between a large number of users, revealing a relation between the frequency of contact and the presence of local structures in the network. Moreover, the authors found that social networks are robust to the removal of strong ties, but fall apart after the removal of a sufficient number of weak ties.

Although a large body of work has been done to characterise OSNs, most of the analyses have been performed on unweighted social graphs (see for example [89]), without considering the strength of social ties. This is due to the hardness of collecting information about social interactions between people in very large social networks. Nevertheless, in [93] the authors demonstrated that there is a significant difference between the properties of weighted and unweighted graphs representing the same social networks. In addition, in [40] the unweighted social graph extracted from publicly available data on Google+ has been augmented with four nodes' attributes (i.e. school, major, employer and city). The results confirm that in some cases the network of attributes shows properties significantly different from the unweighted network.

2.2.2 Measures of Tie Strength in Online Environments

In literature, several techniques to measure tie strength from OSN data have been proposed. In [36], the authors built a model to predict tie strength from OSN observable data, fitting a linear regression model with manual evaluations of tie strength collected from a small sample of users. The results indicate that the model is able to predict tie strength with sufficient accuracy. The same model has been tested on a different social medium, with consistent results across different social networks [35]. In [10] the authors found consistency between the definition of tie strength given by Granovetter in [43] and a set of factors extracted from OSN communication data used to predict reference values of tie strength manually assigned by a sample of users to their social relationships in Facebook. In [52], the authors confirmed that the frequency of contact in online interactions is a good predictor of tie strength, using explicit tie strength evaluations given by a large set of particpats. However, the work is limited to the analysis of the set of "best friends" of a sample of Facebook users and does not consider other ego network circles. A similar analysis with compatible results has been conducted in [53], where data from the "Top Friends" Facebook application is used to build a model to predict binary tie strength (i.e. strong or weak ties) using other measurable Facebook interaction variables (i.e. number of messages and pictures exchanged through Facebook posts).

2.2.3 Preliminary results on Micro-Level properties of OSNs

In [10] the authors found a first evidence of the presence of the Dunbar's number in Facebook, indicating that, even though Facebook allows people to have thousands of online social contacts, people only maintain a limited set of active relationships. This number is compatible with the results found in offline environments. As a further confirmation of this fact, authors of [39] analysed a large-scale data set of Twitter communication data, finding that the average intensity of communication of each user towards all his/her friends, as a function of the number of social contacts of the user, shows an asymptotic behaviour, ascribable to the limits imposed by the Dunbar's number. In [54], the authors demonstrated that inter-individual variability in the number of social relationships in online social networks is correlated with brain size. The authors used magnetic resonance imaging techniques to measure grey matter density of a small sample of participants, comparing the brain volume of the participants the number of their Facebook contacts.

In [79], the authors analysed online social network data of a sample of thirty participants discovering that each ego shows a typical tie strength distribution within his/her ego network. This distribution is in accordance with the ego network model. In [65], mobile-phone data extracted from the logs of a single mobile phone operator has been analysed. The results indicate that the limited capacity people have for communication limits the amount of social ties they can actively maintain.

Although these results give a first insight on the constrained nature of online social networks, revealing a similarity between online and offline human social behaviour, there is still a lack of knowledge about all the other 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.

2.2.4 Analyses on Information Diffusion in OSNs

Several models have been proposed in literature to explain the dynamics of information diffusion process in OSNs. In [2], the spread of URL links between web logs is tracked and analysed. The resulting information cascades are used to build a set of classifiers (using Support Vector Machines) able to predict the existence of links between pairs of weblogs and detect likely routes of infection. In [15], Second Life data are used to study the role of social influence (i.e. the influence that we have on our social contacts) and the diffusion of user-created content. To this aim, the authors analyse the exchange of objects in Second Life between users and they track the spread of these objects to create information cascades. The

results indicate that the spread of information is driven by social influence, and sharing among friends occurs more rapidly than sharing among strangers. This is an evidence of the word-of-mouth effect in OSNs. A similar analysis has been condicted on Facebook, where the spread of information (URLs posted on Facebook Walls) is analysed against the strength of the social ties of theusers, measured using the intensity of communication [17]. The results showed that strong ties are associated with a higher probability of diffusion compared to weak ties. Nevertheless, the total influence of weak ties in the diffusion of contents is higher due to their larger number. This confirms the ideas of Mark Granovetter [43]. Although this, the work considers strong ties all the relationships with at least an interaction since their appearance, and weak ties the social links without interactions. This assumption appears to be too simplistic to fully understand the relation between the local structural properties of OSNs and the formation of high-level social phenomena.

In [22], the author tested the ability of social networks to spread information against random networks, generating two types of artificially structured online communities and assessing the spread of adoption of health behaviour between two groups of participants who periodically received status updates about their neighbours (according to the predefined network structure). The results of this study revealed that health behaviour spread faster in networks with typical social properties (i.e. high clustering coefficient) than in random graph networks. Some of the models proposed in literature are aimed at synthetically reproducing information cascades extracted frm OSNs, like those presented in [38, 45, 61, 63]. On the other hand, some models aim to understand how variations in nodes' properties in social networks influence the spread of information [5, 76]. Other models try to discover the set of seed nodes (i.e. nodes from which the diffusion process starts) which maximises the probability of diffusion in the network [55]. This approach is of particular interest for marketing, since these models could help reducing the costs of advertisement in social networks. A different approach is to start from an unweighted social graph without knowing the probability of diffusion on the social links and, by fitting a parametrised information diffusion model with real traces of information cascades, learning these probabilities [41]. This approach could be then used to characterise the structural properties of the resulting social network graph. Although this technique is promising, we prefer to use a different approach. In fact, in our analysis we directly derive the probabilities of information diffusion on social links from the frequency of contact between pairs of users in Facebook and we study the properties of the information cascades obtained by applying a standard information diffusion model. Then, we analyse the relation between

micro-level patterns of the ego networks in the social graph and the properties of the information cascades. This allowed us to better characterise the dynamics of the information diffusion process in OSNs.

3

Datasets

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. With the information in 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.

3.1 Retrieving Datasets

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 have 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 the comments she received from her friends. Exploiting this characteristics of regional networks, some data sets have been downloaded, such as those described in [93]. The same authors

made some data sets crawled from Facebook regional networks on April 2008 publicly available for research[1]. In this paper we have used the data set referred as "Regional Network A" that has been used by other researchers for purposes different than ours [51].

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 the users within a regional network and the interactions between them only, 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 [93] 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.

3.1.2 Twitter

As far as Twitter is concerned, we have implemented a crawling agent which is able to download user profiles and their communication data from Twitter. The agent visited the Twitter 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. We use as indication of an interaction 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

[1] `http://current.cs.ucsb.edu/facebook/`

Table 3.1: 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

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 both to speed-up the data collection and to avoid the crawler to remain trapped in visiting the neighbourhood of a node with a large number of links. The seed 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.

3.2 Data Sets Properties

3.2.1 Facebook

The Facebook data set we have 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.

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[2] of the social graph are reported in Table 3.1.

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.

[2] The clustering coefficient is calculated as the average local clustering coefficient (Equation 6 in [68]).

Table 3.2: 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

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 have 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 are not connected 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 have 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 3.2 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, which have at least one interaction 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 have 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, communicat-

ing 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*, that represent the people who use Twitter for socialising, and (ii) *other users*, that use Twitter for all the other purposes. This classification is fundamental for our study since, in order to analyse the human social behaviour, we have to consider the social relevant users only. To this aim we have 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 details of the classifier are described in Appendix A. Note that also in Facebook some accounts represent users which are not socially relevant (e.g. companies and public figures). Nevertheless, Facebook is more naturally used as a private communication mean, and public communications (e.g. status updates) are not consedered 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 analyse all the Facebook accounts without splitting them into separate classes.

In the column "all users" of Table 3.3 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 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

Table 3.3: Twitter data set (all users) and classes statistics.

	All users	**Soc. rel. users**	**Other users**
N	$2.463.692$	$1,653,436$	$810,256$
$N_{3,200}$	$510,119$	$260,632$	$249,487$
$(\%\ N_{3,200})$	(20.7%)	(15.8%)	(30.8%)
$\#\ Tweets$	$1,207$	979	$1,696$
$\#\ Following$	$3,157$	$2,553$	$4,448$
$\#\ Followers$	$7,353$	$2,744$	$17,201$
$\%\ Tweets_{\mathrm{REPL}}$	17.4%	18.4%	15.4%
$\%\ Tweets_{\mathrm{MENT}}$	22.7%	21.6%	24.7%

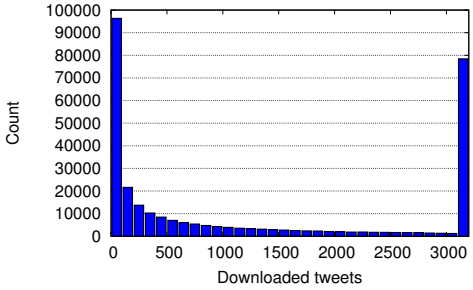


Figure 3.1: Downloaded tweets per user distribution.

the fact that users in the "other users" class could 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 ($tweets_{\mathrm{REPL}}$) and mentions ($tweets_{\mathrm{MENT}}$), 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 3.1 we show the distribution of the number of tweets downloaded per user. We can notice the presence of a peak corresponding to the value $3,200$ that 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 since their account has been created. 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%), which 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 have studied the number of replies the users send to their friends on average. In [39], a similar analysis has been used to con-

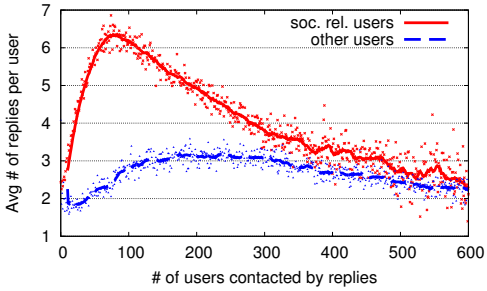


Figure 3.2: Points represent the average number of replies made by accounts with different number of friends; thick lines are their running averages.

clude 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 3.2 depicts the trend of the average number of replies per friend as a function of the number of friends of the user. Differently from [39], we have divide the analysis for the two classes identified: "socially relevant users" and "other users". The results, supported by the figure, 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 that, therefore, have to be split over an increasing number of friends. As discussed in [39], this can be seen as an evidence of the presence of the so called Dunbar's number in Twitter.

Other users show a more random pattern, with lower average value of replies per friend without any significant discontinuities. This indicates that the 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. The link

duration is needed to find the frequency of contact between the users involved in a social link that is 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 [36]. 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 B. 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 the socially relevant users only, we have 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 [39].

4

Analysis of structural properties in Online Social Networks

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, representing social relationships. We define as *online* social networks all the social networks in which social relationships are maintained by the use of the Internet (e.g. Facebook, Twitter, e-mails). On the other hand, *offline* social networks are social networks formed outside the Internet, using, for example, face-to-face communications or phone calls.

Both offline and online social networks show several distinctive properties that differentiate them from other kinds of networks, such as biological and technological networks. For example, they show the *small-world property* [88], for which any two actors in the network, indirectly connected by chains of social links, have a short average distance. As witnessed in [92], besides the short average distance, a small world network shows a high level of clusterisation (or network transitivity) compared to a random network, that 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.

The main goal of this chapter is two fold. On the one hand, we extend existing results that have analysed structural properties of Facebook and Twitter ego networks, by presenting a comparative analysis (submitted to [6]). Moreover, we also analyse the appropriateness of different types of communications (ongoing or outgoing) to extract these structural properties. In doing so, we also compare the structural properties observed for the same users in an online environment (Face-

book) and offline, highlighting correlations, similarities and differences (presented in [59]).

To make these results better understandable, we also provide at the beginning of the chapter a description of the methodology used to extract the structural properties from the Facebook and Twitter datasets, and the main results on which our original contribution is built.

4.1 Social Networks Structure

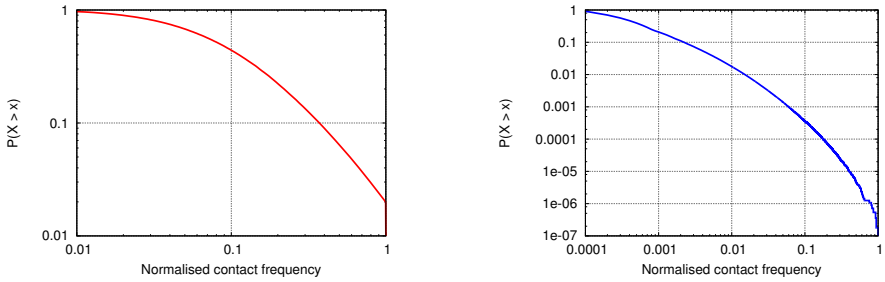
Most of the studies in the analysis of social networks focus on the presence of the small world property or other structural features, i.e. the associativity and the emergence of communities [69]. These studies are normally carried out considering the unweighted network graph in which each edge (or arc) represents the mere existence of a social relationship without including any information about it. This is due to the fact that information about social relationships is not trivial to infer since it normally refers to qualitative aspects. Nevertheless, there are several studies in sociology and anthropology that provide insights about the characterisation of the social relationships and, in particular, on the measurement of their strength.

In order to extract the ego networks from our data sets, we have grouped the relationships of each user into different sets[1]. 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 [95]), in which ego networks are found to be highly instable and with a high growing rate soon after ego joins the network, but tend to be stable after

[1] 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.



(a) Facebook (b) Twitter

Figure 4.1: Aggregated CCDF of the normalised frequency of contact for all the ego networks in the data sets.

the first few months of activity. This selection allowed us to consider only users who regularly 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.

4.1.1 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 4.1, 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's distribution had a social structure, and therefore steps in their distributions, such steps may appear at different positions from one network to another, thus resulting in a smooth aggregate CCDF (remember that also in the Dunbar's model the sizes of the layers are average value, 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. This may indicate a similarity between ego networks in offline and online social networks, as studies in socio anthropology 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). A power law shape in the CCDF is a necessary condition to have power law distributions in at least one ego network [72]. However, this is not a sufficient condition to have power law distributions in each single CCDF [71]. 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.

4.1.2 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. As shown in Figure 4.2, the CCDF distributions of individual ego networks present a series of steps that were hidden in the aggregate distribution analysed in the previous section. As previously said, the presence of these steps reveals the underlying ego network structure.

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)

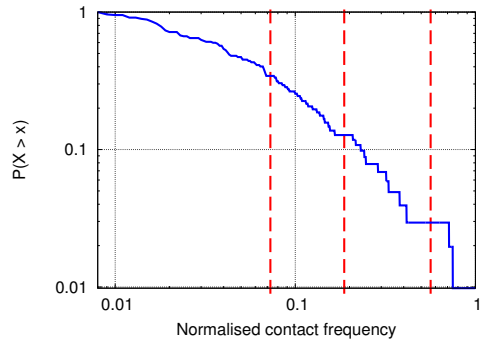


Figure 4.2: CCDF of the normalised frequency of contact of an individual Twitter ego network.

and the objects inside that cluster is minimised. In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set, that is usually considered to be noise [57]. In [8] 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 considered are substantially equivalent and that both can be used for the study of social structures in ego networks leading to the same conclusions [8].

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 [91].

In Figure 4.2 we show the result of the `Ckmeans.1d.dp` algorithm (with $k = 4$) applied to the frequencies of contact of an individual Twitter ego network. As expected, the limits between adjacent clusters (red bars in the figure) are placed by the algorithm in correspondence of the steps in the CCDF distribution.

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*, that is the proportion to which the clustering accounts for the variance of the data. In fact, a small variance in the individual clusters means that data are well described by the current configuration, 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:

$$VAR_{exp} = \frac{SS_{tot} - \sum_{j=1}^{k} SS_j}{SS_{tot}}, \tag{4.1}$$

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 the 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 4.1, the optimal solution of the 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 order to find the typical number of clusters k^*, we may calculate the optimal clustering for each k and then select the value that maximises VAR_{exp}. However, the value of VAR_{exp} increases monotonically with k, reaching its maximum when k is equal to the number of objects in the data space. Thus, there is a 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:

$$K = \underset{K}{\operatorname{argmin}}[-2L(K) + 2q(K)] \tag{4.2}$$

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 k that maximise equation 4.2. In Figure 4.3 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 around 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 Tables 4.1 and 4.2 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 95% confidence interval between square brackets.

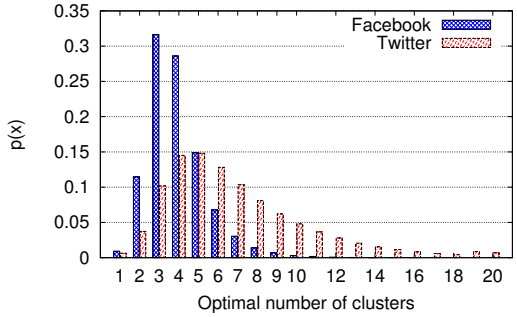


Figure 4.3: Desdity function of k^* in Facebook and Twitter ego networks.

Ego networks with only one circle tend to have similar values of contact frequency for all their links, and in many case 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. This is further confirmed by the higher activity of these ego networks compared to the immediatly next one (ego networks showing two circles). A high level of activity is another distinguishing feature of bots and spammers. 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 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. This is due to the differrences in the nature of the two platform. In fact, in Facebook users tend

Table 4.1: Optimal number of clusters (k^*) of ego networks in Facebook.

	Facebook		
k_{opt}	**# of nets**	**Net size**	**Use rate**
1	844 (0.9%)	29.68 $[\pm 1.95]$	15.30
2	$10,465$ (11.47%)	41.82 $[\pm 0.39]$	14.90
3	$28,918$ (31.66%)	39.00 $[\pm 0.27]$	18.25
4	$26,124$ (28.60%)	41.99 $[\pm 0.38]$	25.55
5	$13,584$ (14.87%)	53.89 $[\pm 0.66]$	40.92
> 5	$11,412$ (12.50%)	82.02 $[\pm 1.00]$	93.35

Table 4.2: Optimal number of clusters (k^*) of ego networks in Twitter.

	Twitter		
k_{opt}	# of nets	Net size	Use rate
1	$2,500$ (0.6%)	192.77 $[\pm 12.44]$	15.27
2	$14,683$ (3.7%)	104.93 $[\pm 2.35]$	11.70
3	$40,099$ (10.2%)	91.42 $[\pm 0.98]$	13.90
4	$57,227$ (14.5%)	89.09 $[\pm 0.73]$	18.36
5	$58,410$ (14.82%)	92.56 $[\pm 0.70]$	25.34
> 5	$221,319$ (56.1%)	100.42 $[\pm 0.31]$	81.32

to accept friendships requests only if they know the requester in person, or they recognise a real human behind her profile, whilst in Twitter the heterogeneity of profiles makes this kind of selection more difficult.

For ego networks showing more than one circle the activity of ego increases with the number of circles. Moreover, 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.

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 fount 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 Facebook. For each ego network we have 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. In fact, while clusters are disjoint groups, social circles, as depicted in Figure 6.2, 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 $C_1 \subseteq C_2 \subseteq C_3 \subseteq C_4 \subseteq C_5$.

Table 4.3: Ego network circles' properties.

		C_1	C_2	C_3	C_4	C_5
Facebook	min freq.	5.09	1.95	0.67	0.11	–
	size[a]	(1.79)	(5.83)	(17.05)	(50.46)	–
	scal. fact.	3.26	2.93	2.96	–	
Twitter	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	
Offline	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	–	

In Table 4.3 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 we have 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 as follow: *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 [96], 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 that 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. Another feature that can be used to compare the ego network circles in online and offline ego networks is the average minimum frequency of contact of the circles, since, as the scaling factor, it is not affected by possible bias derived from the sampling method.

4.1.3 Comparing Online and Offline Ego Networks

Looking at the scaling factors in Table 4.3, 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 [96]. This result is a first indication that Facebook and Twitter ego networks show a hierarchical structure remarkably similar to that found in offline environments.

Considering the average minimum frequency of contact of the circles, we can notice that there is a match between the circles of the two OSNs and those of offline social networks. Specifically, as we report in Table 4.4, 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 we used, that 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

Table 4.4: Offline/online ego networks mapping. The Facebook's size was scaled to match offline active network dimension.

		Super support clique	Support clique	Sympathy group	Affinity group	Active network
Facebook	circle	–	C_1	C_2	C_3	C_4
	min freq.	–	5.09	1.95	0.67	0.11
	size[a]	–	(4.70)	(15.31)	(44.77)	(132.50)
Twitter	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
Offline	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

their size, as reported in Table 4.4. 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 rescaling 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 rescaled 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 [93]. 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 size of C_2-C_5 in Twitter show a good match with those found offline, we can say that C_1 in Twitter, that 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 dataset of a large enough scale to reliably highlight this type of relationships [32].

4.2 Incoming and Outgoing Communication

In Face-to-face communication is difficult to differentiate the impact on social relationships of incoming (e.g what we hear) and outgoing communication (e.g. what we say), while it is much easier in electronic communication (e.g. email, comments or likes in Facebook) where is clear who send the message. To execute characterise the structure of ego networks formed by incoming and outgoing communication, we analysed a small dataset crawled using a dedicated application (described in AppendixC). The obtained dataset is composed of social data regarding 27 people from our research center. For each participant, we downloaded all the social data related to the communication between them and all their friends (making sure to take care of Facebook privacy policies, and anonymising the data). Moreover, the users assigned an evaluation of the tightness of the relationship to each one of their Facebook's contacts. The evaluation consists in assigning two values in the interval $[0, 100]$, the first value evaluate the social tightness in online environments, while the second value evaluate the relationship according to offline interaction.

The following analysis, presented in [59], aims to characterize the ego network constituted by one of the two kind of communications, thus we differentiate between outgoing and incoming communication exchanged by the ego with the alters modelling each social relationship as two different directional links. Therefore we assign a weight to each link (with a similar procedure described in Section 3.3) equal to the frequency of outgoing contact as regards the links directed from ego to alters and equal to the frequency of incoming communication for the opposite direction.

From the datasets we removed the messages that Facebook implicitly strongly encourages to send (e.g., wishes for the birthday), and we consider as *active* the alters contacted at least yearly by ego. We find that in the outgoing communication the average active network size - the number of people contacted at least once in the previous year - is equal to 77.8, while the average number of relationships is 341.64. Executing the same analysis to the incoming communication we find that the average active network size is equal to 142.04. This two active network does not differs only by the size, since just 51 individual are present in both network and the 63.55% of the social links in the active network built on the incoming communication turn out to be inactive in the other active network. Considering the precedent results presented in Section 4.1.2, the active network built on the outgoing communication can better describe the properties of an ego network. This result can be explained considering that to not all the incoming message does not

correspond a expense of cognitive resources by the receiving user (e.g. it is not read by the user).

To better estimate the tie strength, we introducing an index based on the outgoing communication, with a reinforcement to all the relationships that shows reciprocity in the communication having both frequencies of contact (incoming and outgoing) greater than zero. In this way we want to capture the resources ego spends to read the messages received from alters she actively contacts. The index is defined as:

$$AdjFreq_{jk} = f_{jk} + \frac{f_{jk} * f_{kj}}{f_{jk} + f_{kj}} \tag{4.3}$$

Where f_{jk} is the frequency of outgoing communication from ego j to alter k and f_{kj} is the frequency of incoming communication from alter k to ego j. The increment given by the additional term is maximised when the incoming and the outgoing interactions are balanced. The CCDF of the frequency of outgoing communication, of the frequency of incoming communication and of the $AdjFreq$ index are depicted in Figure 4.4.

We applied the clustering methodology described in Section 4.1.2 to both outgoing frequencies and the results of the $AdjFreq$ index. We verified that the Facebook ego networks of our sample show a typical number of circles similar to that

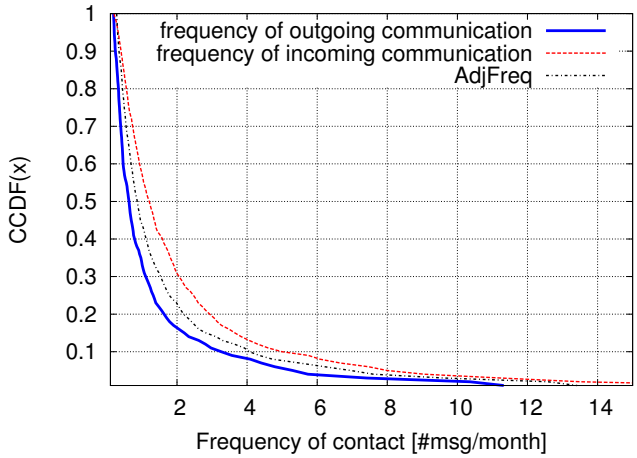


Figure 4.4: CCDF of the frequency of outgoing/incoming communication and the index AdjFreq

found in humans (i.e., equal to 4). We find that the average optimal number of clusters for outgoing communication is equal to 3.76 with a 95% confidence interval of (3.46,4.06) and a median equal to 4. As far as the $AdjFreq$ index, we obtain an optimal number of clusters equal to 3.88 with a 95% confidence interval of (3.58,4.18) and median 4. Hence, we re-apply the $k-means$ algorithm on all the ego networks fixing k to be equal to 4, using both the frequency of outgoing communication and the $AdjFreq$ index. We study the dimensions, the scaling factors and the typical contact frequency of each social circle represented by the union of each cluster found by k-means with all the other previous clusters with higher frequency of contact, comparing them with the results found in human ego networks.

Table 4.5 reports the results of the k-means analysis with k fixed to 4, indicating with "size" the dimension of the circles and "sc. f." the scaling factors between adjacent circles. The results are divided in the table into three different parts: (i) results concerning the k-means analysis applied to the frequency of outgoing communication; (ii) results of k-means applied to the $AdjFreq$ index and (iii) results in the anthropological literature. The mean value of the scaling factors is equal to 3.14 for the frequency of outgoing communication and 3.12 as regards the $AdjFreq$ index. These results are really close to the mean scaling factor found in human ego networks. The size of the circles in Facebook (for both the frequency of outgoing communication and the $AdjFreq$ index) is lower than that found in human ego networks. This could be ascribed to the fact that online ego networks represent only a partial subset of the ego networks of a person in real life and on-line ego networks are currently still in the infancy and in a growing phase (the size of an ego network also depends on how long the ego has been active in OSN). Despite this, the structure of Facebook ego networks presents the same hierarchical pattern of the structure of human ego networks. The last two rows of Table 4.5 report the projection of the results in Facebook calculated in order to make the size of the larger circle fit with its counterpart in real life, for both the frequency of outgoing communication and the $AdjFreq$ index. The results confirm the structure similarity between the social circles of the ego networks obtained by ego-net digger and those found in human real ego networks. Moreover, the $AdjFreq$ index, built as a combination of the frequency of outgoing and incoming communication, produces an ego network structure closer to that found in the anthropological literature.

The typical frequency of contact for each circle is reported in Table 4.5 as "min freq", expressed by the minimum number of posts sent from ego to alters per month within the considered circle. The resulting typical frequency of contact allows us to define the social circles in our sample data, considering the frequency

Table 4.5: Results of k-means with $k = 4$. 95% confidence intervals are reported in square brackets.

support clique	**sympathy group**	**affinity group**	**active network**

Frequency of Outgoing Communication

size	2.52 [.57]	7.84 [1.88]	23.04 [7.41]	77.8 [33.49]
sc. f.	-	3.11	2.94	3.38
min freq	10.49	3.94	1.15	.19

AdjFreq Index

size	2.56 [.63]	8.24 [2.18]	24.6 [8.28]	77.8 [33.24]
sc. f.	-	3.22	2.99	3.16
min AdjFreq	12.3	5.03	1.56	.28

Results in Human ego networks

size	4.6	14.3	42.6	132.5
sc. f.	-	3.10	2.98	3.11

Projections

outFreq	(4.29)	(13.35)	(39.24)	(132.5)
AdjFreq	(4.34)	(14.03)	(41.9)	(132.5)

of outgoing communication, the group of people contacted at least $\sim$ *three times a week* (support clique), $\sim$ *weekly* (sympathy group), $\sim$ *monthly* (affinity group) and $\sim$ *twice a year* (active network). The structure we find in Facebook ego networks is thus compatible with the one found in human ego networks.

4.3 Discussion

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 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 to characterize human social network properties. The scale at which data can be col-

lected 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. Furthermore we analyzed the different characteristics of outgoing and incoming messages in a small dataset retrieved with a dedicated application. For both kind of communication the ego network still applies but the incoming active network is bigger than the outgoing one.

5

The Role of the Ego Network Properties in Information Diffusion

Nowadays, Online Social Networks are one of the most effective channels to spread information among people. They became widely used in the last few years, due to their ability to transform users in *active* producers of contents, going sharply in opposition to more conventional communication means. The mechanisms underpinning information diffusion in social networks has recently gained attention in research community. In fact, understanding how information spreads between people could provide important insights into the dynamics of our society, revealing how the spread of ideas, innovation, influence and many other aspects take place. The advent of OSNs made available a huge amount of data regarding communications between people. The availability of these data represents a unique opportunity for the study of information diffusion.

Although some work has been done to characterise the properties of OSNs and their role in the diffusion of information, there is still a lack of understanding of some fundamental aspects controlling the process. Tie strength (i.e. the importance of social relationships) is recognised as the most important factor influencing the spread of information between pairs of individuals [43, 17]. Moreover, the concept of tie strength is strongly related to the degree of interaction between people [49, 10] and social interactions are found to be the main driver of information diffusion in social networks. This leads to the formation of the word-of-mouth effect [3], for which information travels thanks to local communications between people.

Understanding how social interactions influence information diffusion at global scale is not an easy task, since it is difficult to obtain large-scale datasets containing meaningful sample of both communication traces between people (to estimate tie strength) and information cascades. In fact, OSNs usually give limited access

to their data, especially when personal information is concerned. Moreover, tie strength is a really dynamic property and its estimation requires the collection of a large portion of the history of communication between people [7]. Obtaining information cascades is clearly not easier, since it requires the collection of a large amount of data to obtain complete trees of information paths in the network.

As discussed in Section 5.1, only a few works on information diffusion consider social networks with weighted relationships. Therefore, the relation between tie strength and the properties of information cascades is still not completely understood. The main objective of this analysis is the study of the influence of two main characteristics of the ego network on the information diffusion. The first one is the role assumed by the various social circles and the second is how the the relationships established by the ego (and thus the position of the node in the network) influences its capability to generate large information cascades.

5.1 Related Work

In literature, different approaches have been explored to characterise information diffusion in OSNs. However, none of the proposed models used to describe the generation of information cascades in social networks consider that the probability to forward information depends on tie strength and on the level of interaction between the involved users. For this reason, there is still a lack of understanding about the role of tie strength in the information diffusion process.

Research in the field started with a series of experiments aimed at collecting and studying traces of information diffusion among the population. One of the first pioneers in the field was Stanley Milgram, who showed the presence of a short average distance between randomly selected senders and a fixed target in the network, with value close to 6, confirming the the so called “small world” effect, or “six degrees of separation”. Milgram also identified the convergence of communication chains through a small set of common individuals, with a central role in the diffusion. Additional analyses on OSNs validated Milgram’s results in [31, 63, 14], showed that OSNs presents a distinctive property called “small world effect”, whereby the average distance between randomly selected nodes grows proportionally to the logarithm of the size of the network. Other analyses about the topology of OSNs found that OSNs play an important role in the information diffusion process [61]. Specifically, the clustered structure of OSNs positively impacts on the diffusion of information [22].

In [43], Mark Granovetter hypothesises that the difference in tie strength (i.e. the importance of social relationships) between different links in a social network

plays a fundamental role in the diffusion of information. Whilst strong ties are generally associated to higher level of trust and transport more information, weak ties can represent bridges connecting different communities in the network, making information travel long distances. This fact has been empirically confirmed by different studies on OSNs [70, 15]. The idea behind the role of tie strength led to the conclusion that information in social networks is moved through local links, representing personal acquaintances. For this reason social networks are said to show the so called *word-of-mouth* effect [3, 24].

Based on the properties described hitherto, a series of models for information diffusion have been created (see for example [37, 44] - the simplest and widely used models in literature). These models assume, similarly to what happens in a virus contagion, that a node is infected by information with a probability proportional to the number of its neighbours which are already infected. The produced effect is known in literature as *cascading effect*, thus the paths followed by information during the contagion are called *information cascades*.

The advent of OSNs fostered the availability of large amount of information cascades data. An interesting body of work analyses these traces to understand the mechanisms that control the spread information in the network at global level [38, 83, 67]. The relation between the exposure to information (i.e. the probability to see a message received through an OSN) and the diffusion of information have been recently explored. Results indicate that the probability to forward information is directly related to the duration of the exposure [76].

Other analyses start from information cascade data to estimate the probability of diffusion of each social link, by using different information diffusion models [41, 38, 45, 55]. Although these studies found some important properties about information diffusion in OSNs, only a few of them combine analyses about both the topology of the network and the role of tie strength in the diffusion of information (as we do here). Specifically, in [94], the authors define tie strength as the percentage of neighbours that a pair of nodes have in common. Hence, they apply a series of information diffusion models to characterise the role of tie strength in the spread of information. Results showed that “pushing” information from node to node is the best strategy to obtain higher coverage in the network. Moreover, forwarding information using the links with higher tie strength increases the success of the diffusion. Even though these results highlight some interesting properties of information diffusion, the definition of tie strength used is derived from the topology of the network and does not consider the interaction level of the nodes. Yet, tie strength is found to be strongly related to the degree of interaction between people [49, 10] and this fact must be taken into account to fully understand so-

cial dynamics of the network [43]. Only a few works analyse the influence of tie strength in the diffusion of information in OSNs [17], but they are neither exhaustive for the characterisation of information diffusion properties in OSNs nor useful for the creation of simulated environments. In fact, they are not reproducible, since the data they analyse is not publicly accessible.

5.2 The Role of Social Circles

In this section we assess how the ego networks structural properties in OSNs influence macro-level social phenomena involving the whole network and the social interactions between people. In particular, we have studied the diffusion of information in OSNs analysing the role of tie strength and social structures on the generation of information cascades. To this aim, we have used a simple model of information diffusion that allows us to study "in vitro" the generation of information cascades considering the Facebook network described in Section 3.2. Specifically, we have used the Facebook social graph (i.e. the graph describing the overall structure of the downloaded network) combined with the frequency of contact estimated from all the interactions between the users. We have used this network since it represents a significant portion (a regional network) of the entire Facebook graph. In fact, sampling a social network with geographical selection of the nodes provides a better representation of the original network than sampling nodes with a walk in the network. This work was submitted to [6].

5.2.1 Information Diffusion Model

To simulate the generation of information cascades in Facebook, we have utilised the *Independent Cascade Model* (ICM) [37] in which information spreads in the network according to a probability of diffusion defined for each link. The diffusion starts from an initial set of infected nodes which have a single chance to spread the information to each of its neighbours. Then, the process is repeated iteratively step by step until no new nodes are infected.

More formally, the diffusion process propagates the information synchronously in N discrete steps, in which each node in the network is either in one of three possible states: *not-infected*, *infected*, *contagious*. A node in the contagious state is a node that is infected and can spread information. The ICM starts with a set A_0 of contagious nodes, and all other nodes $\notin A_0$ in the state not-infected. At the i^{th} time step, with $0 < i \leq N$, all the nodes in $a \in A_{i-1}$ (contagious nodes at time step $i-1$) propagate the information to their not-infected neighbours $w \in$

W_a, with a probability $p_{a,w}$. Each not-infected node that receives the information changes its status to contagious and it is included in A_i. At the end of the step i all nodes in A_{i-1} change their status from contagious to infected (this guarantees that nodes spread the infection at most during one time step, but not continuously). The diffusion process ends at the start of the step $i = N$ when the set A_i is empty.

The ICM requires all the links in the network to be marked with their probability of information diffusion, but does not provide any constraints on how to calculate this value, so we defined a different function to obtain the probability of diffusion. Since in OSNs the spread of information occurs when two users interact, we defined the probability of diffusion of each social link as a function of its normalised frequency of contact $\varphi_{s,t}$ between a pair of users $\langle s, t \rangle$. To this aim, we have used the following transformation function to calculate the probability of diffusion:

$$p_{s,t} = max(\varphi_{s,t}, \epsilon)^{\gamma} \tag{5.1}$$

The constant ϵ is a value used to give a $p_{s,t} > 0$ to all the relationships with null normalised frequency of contact. We have introduced this constant because, even though a relationship shows no activity, its presence represents an implicit interaction between the involved users and therefore a possible (although rarely used) channel of communication, and thus we cannot assign a probability equal to zero, that indicates the total absence of communication. The value of this constant should be lower than the frequency of any other links with a non-zero frequency of contact.

The parameter γ permits to control the difference in terms of probability of diffusion between strong and weak ties, introducing a nonlinear transformation. In fact, when $\gamma = 1.0$, the probability is proportional to the frequency of contact between users. On the other hand, when it is greater than 1, the probability of the ties is penalised according to their frequency of contact. This emphasises the diffusion over strong ties.

5.2.2 Simulation Settings

We have generated a set of information cascades through simulation using the ICM model on Facebook network. We have used the frequency of contact (estimated using the methodology explained in Appendix C and normalised between 0 and 1) and we have transformed it by applying Equation 2 to obtain the probability of diffusion for the simulations. The value of ϵ has been chosen to be the highest possible value lower than all the normalised frequencies of contact in the data set. Namely, since the lowest normalised frequency of contact is equal to 0.0012 we

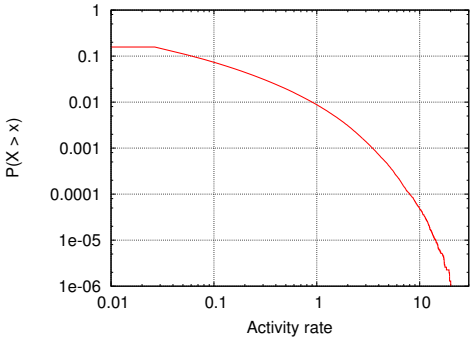


Figure 5.1: CCDF of the activity rate of the nodes.

fixed ϵ to 0.0010. This allowed us to to obtain an upper bound of usage of the social links with null frequency of contact.

As described in the previous section, the ICM model requires a set of starting nodes A_0 from which information cascades are generated. For the sake of simplicity we decided to start the simulations from a single node. Since we want to infer the average properties of information cascades in the network, we have sampled a set of $1,000$ starting points (seeds). For each seed, we performed 100 simulations, then computing its average *node coverage*, defined as the number of nodes infected. Hence, we have averaged the results obtaining the average node coverage in the network.

Anticipating the results of the simulations, we have found a positive correlation of $r = 0.73$ with a p-value of $p < 0.01$ between the *activity rate* r_s of a node s, defined as the sum of the frequencies of contact of all the relationships of the node, and the node coverage of the cascades generated by s, in simulations with $\gamma = 1.0$. This indicates that nodes with high r_s proportionally generate a much larger number of messages that are spread in the network than nodes with low r_s. Thus, to obtain significant results, we have sampled the seeds according to a uniform distribution over the activity rate of the nodes, whose distribution is depicted in Figure 5.1. In the figure we can notice that the activity rate shows a long tailed shape. This means that in our data set nodes with low activity rate are way more numerous that nodes with high activity rate. The sampling choice we have utilised avoids to select only nodes with low activity rate (that generate small cascades), as could happen with a uniform random sampling on the nodes.

With the used sampling technique, to average the results of the simulations we needed to weight the node coverage obtained from each seed by the number of

Table 5.1: Definition and statistic of rings considering our Facebook graph.

Ring	**Social circles correspondence**[a]	$\%$ **of links**	**Avg** φ
R_1	support clique	0.3%	0.137
R_2	sympathy group, excluded the support clique	1.0%	0.062
R_3	affinity group, excluded the sympathy group	6.2%	0.020
R_4	active network, excluded the affinity group	10.9%	0.009
R_5	mega-band, excluded the active network	80.6%	0

nodes in the network with similar activity rate. This is based on the assumption that nodes with similar activity rate generate cascades with compatible properties. Therefore, given the seeds g_k ($k = 1, ..., 1000$) ordered by increasing values of activity rate r_{g_k}, the weight w_{g_k} to be applied to the coverage obtained with seed g_k is the number of nodes in our data set whose activity rate is within the interval around r_{g_k} defined by the intermediate points between $r_{g_{k-1}}$ and r_{g_k}, and r_{g_k} and $r_{g_{k+1}}$, respectively.

5.2.3 Social Rings and their Role in Information Diffusion

To assess the impact of the structure of ego networks in the diffusion of information we counted the number of messages that pass through specific ego network circles during the simulations. Each social link in the network has been assigned to a position in the ego network model, according to the frequency of contact between the users it connects. Remember that, by definition, the ego network model defines a hierarchical structure, and therefore outer layers include inner ones. Thus, to avoid ambiguity, as reported in Table 5.1, we have assigned each link to a *social ring*, defined as the part of a social circle that is not included in any nested one. To do so, we have used the same clustering technique described in Section 4.1.2, considering that the clusters coincide with social rings. We have assigned all the inactive relationships (i.e. with null frequency of contact) to a fifth (and external) ring, that coincide with the external part of the mega-band in the ego network taxonomy.

In Table 5.1 we report the percentage of links of the network belonging to the different rings and their average values of the normalised frequency of contact φ. As we can see, most of the social links in our network are included in the external ring R_5. This is compliant with typical models of human social networks, that clearly state that weak ties are way more numerous than strong(er) ties.

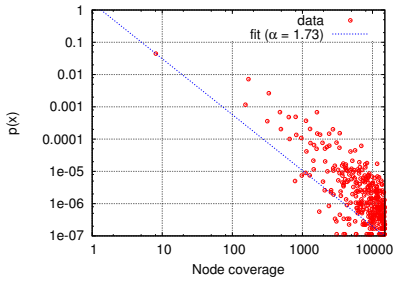
In Table 5.2 we show the usage of the rings, in percentage, for the simulations with different values of γ. Most information passes through the first four social rings (from R_1 to R_4), which correspond to the active network in the ego network model. The most used ring in diffusion process is R_3, for values of γ up to 1.25, even though it contains only 6.2% of the relationships, as reported in Table 5.1. This result is coherent with the literature, that considers the medium strength ties as the most used in the information diffusion process [31, 70]. Note that, for $\gamma = 1.5$, the diffusion takes place mostly through strong ties, thus channelling most of the messages (i.e. 85.76%) through the first three rings, that contain only 7.5% of the social links. It is also worth noting that the 5^{th} ring, the most external, is used up to only 4.55% of communications. Even though this result is strongly dependent on ϵ, for the choice we have made in the selection of the value of ϵ the results reported here can be considered an upper bound of the expected amount of communication that spreads through the 5^{th} ring.

With the results of the simulations we have also studied how the different values of the γ parameter (in the range $1 \leq \gamma \leq 2$) impact on the properties of the information cascades generated by our model.

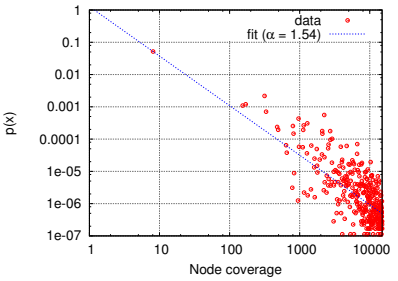
Figure 5.2 depicts the distribution of the node coverage for different values of γ, weighted as previously described. As can be seen in the figure, the node coverage shows a long-tailed behaviour for all the values of γ and decreases as γ increases. In the figure we also show the power-law functions with α that best matches the data. The values of α are compatible with the results of other analyses of information diffusion on cascades collected from real communication traces [34, 58]. This indicates that the cascades generated by our model are consistent with the real information diffusion process in OSNs. Note that in Figure 5.2 we omitted $\gamma = 2$ since it produces too small cascades.

Table 5.2: Diffusion share through links: normalized weighted share of messages through various rings.

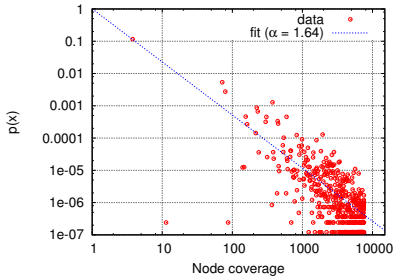
γ	**Ring share**				
	R_1	R_2	R_3	R_4	R_5
1.0	16.96%	22.60%	28.21%	27.69%	4.54%
1.125	16.97%	22.58%	28.20%	27.69%	4.55%
1.25	21.07%	25.83%	28.28%	22.74%	2.08%
1.5	29.51%	30.39%	25.86%	13.81%	0.42%



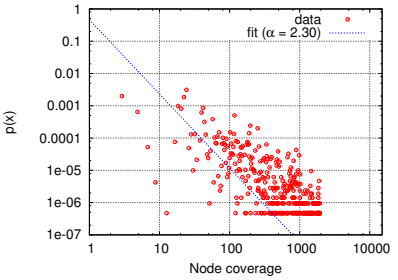
(a) $\gamma = 1.0$



(b) $\gamma = 1.125$



(c) $\gamma = 1.25$



(d) $\gamma = 1.5$

Figure 5.2: Node Coverage Density for different values of parameter γ and related power-law fit.

5.2.4 Impact of Ring's Removal on Information Diffusion

We have continued our analysis studying the effects induced by the removal of the different social rings on the diffusion of information. To this aim, we ran a set of simulations by eliminating one ring at a time from the network (and various combinations of rings) to assess to which extent they impact on the diffusion of information and to study if the diffusion can take place also without specific groups of social links. In this simulations, we have chosen the value $\gamma = 1.125$ since it produces cascades with node coverage that best matches real traces of information diffusion (see for example [34, 58]).

The results of the analysis of information diffusion with the removal of the social links, reported in in Table 5.3, show that almost all the rings are important for the diffusion of information. In fact, apart from R_5, the removal of any of the rings causes a significant drop in terms of node coverage. The low importance of R_5 could be explained by the fact that the links in this ring can be either connected

Table 5.3: Diffusion share through links in sub-network.

Removed ring(s)	**Removed links** (%)	**Coverage # nodes**	(%)	**Rings share** R_1	R_2	R_3	R_4	R_5
none	–	236.34	(100)	16.97	22.58	28.20	27.69	4.55
R_1	0.3	40.26	(17.0)	–	28.56	34.27	32.23	4.94
R_2	1.0	42.61	(18.0)	23.42	–	36.93	34.76	4.89
R_3	6.2	57.79	(24.5)	23.75	33.02	–	38.17	5.06
R_4	10.9	85.96	(36.4)	23.31	32.10	39.62	–	4.97
R_5	81.6	209.83	(88.8)	17.74	23.70	29.55	29.01	–
R_1, R_2	1.3	0.23	(0.1)	–	–	46.18	41.05	12.76
R_1, R_2, R_3	7.5	0.05	(0.0)	–	–	–	69.04	30.96
R_1, R_2, R_3, R_4	18.4	0.01	(0.0)	–	–	–	–	100
R_2, R_3, R_4, R_5	99.7	0.03	(0.0)	100	–	–	–	–
R_3, R_4, R_5	98.7	4.26	(1.8)	40.37	59.63	–	–	–
R_4, R_5	92.5	61.41	(26.0)	24.41	33.79	41.80	–	–

to peripheral nodes with very low probability of diffusion, or to already infected regions. This result does not necessarily imply that weak ties are not important. In fact, the 4^{th} ring, which contains links with very low interaction frequency, has a high relevance in the diffusion of information, even though it is used only 16.7% of the times. Moreover, its removal makes the coverage drop to 36.4% only.

It is also noteworthy that the removal of R_1, which contains only the 0.3% of the network links, reduces the node coverage to just 17.0%. According to the literature, strong ties are usually associated to the formation of clustered groups of nodes, trapping information in their cliques [43]. This fact has been also empirically confirmed in [28] on the same Facebook data set used in the present paper. Despite this, our results indicate that the role of strong ties is essential in the information diffusion process, because their removal from the network causes a drastic reduction of node coverage. This kind of relationships is essential to distribute the information within clustered regions of the network, and thus to eventually reach bridges that allow information to reach other socially distant regions.

The great impact of R_1 is influenced by the high probability of diffusion associated to its links. This high probability is supported by other work in the literature, such as [17], in which the authors empirically demonstrated that strong ties are more frequently used than other ties to diffuse information. Noticeably, simulations with only the three most internal rings (R_1, R_2, and R_3) generate cascades with

node coverage equal to 26.0% of the ones generated using all the rings, even though the links remaining after the removal of R_4 and R_5 are only 7.5% of all the links in the network. Even if we do not present it, a simulation with only the strong ties of R_1 and the weak ties of R_4 would have produced small information cascades, in fact they would be smaller than the one produced by the simulation with the removal of ring R_2.

We have not simulated the information diffusion removing rings in the middle of the ego network structure (e.g. R_2 and R_3), but we expect that the results would lead to lower node coverage than the simulations with the removal of only one of them.

The results of the simulations with the removal of social rings show that we can consider two kinds of equally important factors in the information diffusion process: (i) the intra-clique diffusion, in which the information is delivered using strong and medium ties inside closed groups of nodes and (ii) the inter-clique diffusion which uses medium and weak ties to diffuse information between different parts of the network. Both of them are essential in the information diffusion process, since the absence of either of them drastically reduces the diffusion of information in the network.

5.3 The Role of the Node Centrality

In this section we assess how the relationships established by the ego influences its capability generate large information cascades. In fact the connection with other egos and the intensity of their relation determine the position of the ego in the network, allowing him to assume a central or peripheral position in the network. In this section we characterise the position of the ego according to several centrality measurement that we use to analyse the information cascades generated by the ego trough simulations. This work was presented in [4].

5.3.1 Dataset Description

In this analysis we used the facebook dataset described in Section 3.2. Based on the frequencies of contact, we estimate, for each edge, the strength of the social tie as a numeric value in the interval between 0 (the weakest ties, non active relationships) and 1 (the strongest ties). More formally, we define the tie strength $s_{a,w}$ between two directly connected nodes a and w as the result of the linear transformation of the monthly average frequency of contact to obtain a value in the interval $[0, 1]$. Since the distribution of interaction frequencies has a characteristic long tail

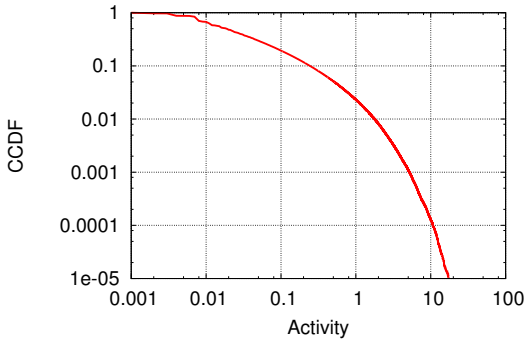


Figure 5.3: Complementary cumulative distribution function (CCDF) of the nodes' activity.

(the highest monthly average frequency of contact in the dataset is equal to 868.24 interactions per month), we assign the value of 1 to the top 500 interactions in the dataset. In other words, in our graph all edges with a number of interactions per month higher than 33 have a tie strength equal to 1. In addition, we assign to each node a a weight calculated as $\sum_{w \in W_a} s_{a,w}$ where W_a is the set of neighbours of a. We call this weight *activity*, which can be considered as a sort of weighted degree. The activity distribution in Figure 5.3 shows that it is characterised by a long tail shape.

Hereafter we use the term *weighted social graph* to identify the described social graph whose edges are labelled with the strength of the social ties. Moreover, we define the *active graph* as the sub-graph of the weighted social graph in which each node and each edge have an activity value and a tie strength greater than zero respectively. Properties of the weighted social and the active graphs are reported in Table 5.4. As we can see in the table, the active graph is considerably smaller than the social graph. In fact, both many non-active nodes and many non-active edges where removed in the active graph. The removed nodes represent both inactive users or users that have communicated with friends that do not belong to the examined regional network. It is worth noting that removed edges are not weak ties (relationship with a *low* activity) but completely inactive relationships.

Both social and active networks present the typical properties exhibited by all the social networks studied in literature [68, 69]: high level of clustering coefficient[1] and small average shortest path length (with respect to what one would expect

[1] Calculated as the average local clustering coefficient of all the nodes in the network (Eq. 6 in [68]).

on the basis of pure chance, given the observed degree distribution), thus the operation of removing inactive nodes and edges from the social network do not affect the capability of the active graph to describe a typical real life social network.

5.3.2 Experimental Environment

Using the described weighted social graph we want to generate a set of information cascades to study the relation between the properties of the network and the diffusion of information. We aim to study how information spreads in the network therefore we need a model able to create information cascades with properties similar to those find in real environments.

Information Diffusion Model

We want to simulate the word-of-mouth effect in social networks, in which a node becomes infected when reached by the information. To do so, we based the model on the *Independent Cascade Model* described in Section 5.2.1.

As we show in Section 5.3.3, the Independent Cascade Model is unable to generate realistic information cascades if we assign to each link a probability of diffusion as a linear function of interaction frequency. Thus, we modelled the *decay of interest* which indicated that the interest in information (and in propagating it) decays over time in real environments. In the model we introduce an ageing factor α, included between 0 and 1, that penalises the probability of diffusion at each step with the exception of the first one. Thus, the probability that, at step $t > 0$, a node $a \in A_{t-1}$ infects a neighbour $w \in W_a$ is equal to:

$$p_{a,w}(t) = s_{a,w} * (1 - \alpha)^{t-1} \tag{5.2}$$

where $s_{a,w}$ is a constant value (the tie strength in our case) in the interval $[0, 1]$ for each edge of the network that shows the probability of diffuse information

Table 5.4: Statistics of the graphs

	Weighted Social	Active
# Nodes	$3,097,165$	$1,171,208$
# Edges	$23,667,394$	$4,357,660$
Avg Degree	15.283	7.441
Avg Clust. Coef.	0.209	0.114
Avg Sh. Path	6.181	6.870

through the link during the first step of the diffusion. In fact, it is worth noting that if $t = 1$ (the first step) then $p_{a,w}(1) = s_{a,w}$. Moreover, if $\alpha = 0$ then the lack of the ageing factor makes the model equivalent to the Independent Cascade Model.

As we will see later, the diffusion model defined by Equation 5.2, is able to create information cascades comparable to those found in real environments.

Simulation Testbed

We tested the information diffusion model described in the previous section applying it to the weighted social graph described in Section 3.2 using the tie strength values as the parameters $s_{a,w}$.

To do so, we implemented a modular network simulator written in Java and based on `fastutil` data structure libraries[2]. We selected $1,000$ seeds from the nodes in the active graph and we ran 100 different simulations for each seed. Since the distribution of nodes' activity has a long tailed shape (as showed in Figure 5.3), we decided to adopt a sampling technique able to select the seeds in the entire spectrum of activity with the same probability. We randomly selected $1,000$ values in the range $(0, MaxActivityValue]$ and then we included in the sample the nodes with the closest value of activity. In this way, we have been able to study the correlation between the statistics of the sampled nodes and the properties of the information cascades avoiding to sample only nodes with low level of activity, that would have been selected by a random sampling of the nodes since they are the majority in the network.

In order to study the influence of the parameter α in the information diffusion model, we ran different simulations setting its value between 0.1 and 0.5, with steps of 0.1. We also executed a set of simulations with a $\alpha = 0.0$ to produce information cascades without considering the ageing factor, simulating the behaviour of the Independent Cascade Model.

Seed Nodes and Information Cascades Measures

In order to analyse how the characteristics of a seed impact on the related information cascades, we have selected a set of measures that describe some important properties of both seed nodes and information cascades.

We have considered two characteristics of the information cascades, the *node coverage*, defined as the fraction of the active nodes infected during the diffusion process, and the *cascade depth*, defined as the depth of the diffusion tree or,

[2] Available at http://fastutil.di.unimi.it/.

equivalently, as the number of the steps of the process. Since for each seed we ran 100 independent simulations, the node coverage and the cascade depth of the information cascades have been averaged over the different seeds.

For each seed we considered a set measures that describe properties of the node ranging from a local to a network perspective. We calculated each index using two different versions of the same network: the first is the original weighted network with the edges labelled with the strength of the social ties and the second is the network without the tie strength information and thus made unweighted.

The first group of measures concerns the connectivity of the nodes, taking into account the adjacent edges only. In the case of the unweighted network, we calculated the *node degree*, while for the weighted network we consider the *activity* of the node that, as introduced in Section 3.2, is the sum of the tie strengths of adjacent links. The second group of measures, the local network properties, describes how well the neighbours of the node are connected among them. For both weighted and unweighted network we calculate the *clustering coefficient*, that measures the probability of the existence of a directed connection between two neighbours. Moreover, for the weighted network only, we considered the *Burt's constraint index* [21], a structural index that shows how strongly connected between them the neighbours of the node are: a high value of this index indicates that many neighbours are connected directly among them with strong ties. The last group of measures, the network centrality properties, describes the importance of the node within the whole network. We have calculated two centrality measurement, the *eigenvector centrality* and the *PageRank*.

5.3.3 Results

In Figure 5.4(a) we show the distribution of the node coverage of the generated information cascades produced by our information diffusion model setting $\alpha = 0.0$, thus not considering any decay of the diffusion probability. The strongly-bimodal histogram suggests that each information cascade can be either very small or very large. This is due to the lack of a mechanism that simulates the ageing of the information. In fact, in this case, each node that receives the information acts as the seed of a new independent cascade process. If, during the diffusion process, the number of reached nodes becomes sufficiently large, then there is a high probability that the process speeds up increasing the number of contagious nodes at each step until all the nodes with a sufficient level of activity are reached. For example, in our testbed we found that, if the diffusion process reaches 20 nodes, there is a probability equal to 0.974 that the number of reached nodes at the end of the

process is greater than $20,000$. On the contrary, just a small number of nodes are reached in case the seed and its neighbours do not succeed in spreading the information enough during the first steps of the process.

In Figure 5.4(b) we show the node coverage histogram of the information cascades generated setting the ageing factor $\alpha = 0.1$. It's worth noting that, compared with the previous figure, the number of reached nodes is considerably reduced and that the distribution is unimodal with the peak on the head. This kind of distribution is compatible with those reported in literature for real information cascades traces [16]. On the contrary, we can reasonably consider the node coverage obtained without considering the ageing factor α as unrealistic.

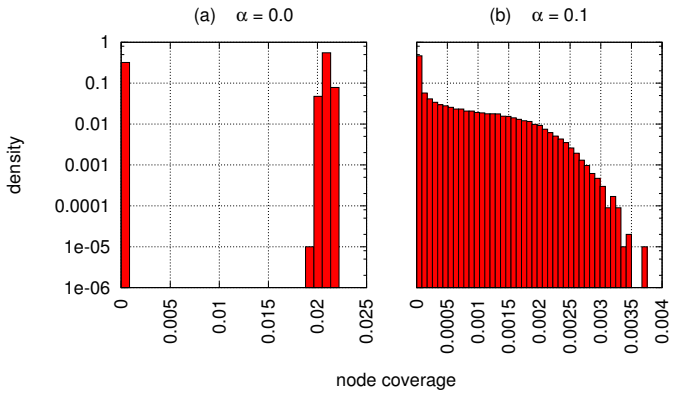


Figure 5.4: Node coverage histograms for information cascades generated using (a) $\alpha = \mathbf{0.0}$ and (b) $\alpha = \mathbf{0.1}$.

To study how the properties of the seed impact on information diffusion process considering the ageing factor $\alpha > 0$, we analysed the correlation between the measures of the seeds and of the information cascades introduced in Section 5.3.2. One of the key aspects of our analysis is to compare the correlation values obtained using weighted and unweighted social networks in order to highlight the benefits given by considering the strength of the ties. Correlation analysis results are reported in Table 5.5.

First of all we analyse the case of the unweighted social network whose correlation values are listed at the top of the table. As we can note the correlation values are low. Specifically, clustering coefficient and PageRank appear to be almost uncorrelated with both node coverage and cascade depth. On the other hand

the degree and the eigenvector centrality show medium correlation with the information cascade measures. This means that we can use these variables to predict which nodes will produce larger information cascades. It's worth noting that, despite the correlation values of the eigenvector centrality are slightly higher, the computation of the degree is significantly less expensive.

In the second part of the table we show the correlation values of the weighted social network. In general, the correlation significantly increases for all measures, showing the importance of considering the tie strength. Notice that the activity (i.e. the sum of the tie strengths of the adjacent links) is the variable with the highest values of correlation for both node coverage and cascade depth for any value of α. In order to study more in detail the role of nodes' activity in the information cascades, we report in Figure 5.5 the running average of node coverage (a) and average depth (b) of the information cascades for seeds with different activity. We notice that the nodes with higher activity produce larger information cascades both in terms of node coverage and depth. This means that if we are able to select the nodes with the highest activity in a social network and we give the information to them, the size of the generated cascades is higher than starting from nodes with lower activity. This result is intuitive and expected, since the nodes with higher activity have higher probability to infect their neighbours, thus spreading the information in a wider range. The figure also shows the effect of the parameter α of the information diffusion model. We can notice that increasing the values of α the model produces smaller information cascades. This is because the probabilities of diffusion decrease during the process more rapidly as the values of α increases. For this reason, in case of high values of α, the connectivity of the seed is fundamental to generate large information cascade. In fact, a high connectivity of the seed permits to reach a large number of nodes during the first step of the infection, when the ageing factor α has no effect. This is demonstrated by the fact that the correlation values in Table 5.5 of the connectivity measures increase with the values of α.

Interestingly, all the variants of the clustering coefficient show low values of correlation with the size of the information cascades. Considering PageRank and eigenvector centrality in case of weighted network, both measures exhibit good correlation with the information cascade size. Specifically, the node coverage has higher correlation values with the eigenvector variables, while the cascade depth better correlates with the PageRank. Moreover, the Burt's constraint index presents a medium negative correlations, which indicate that an ego network with many structural holes is more capable of diffusing the information than a node in a strongly connected ego network.

Table 5.5: Correlation analysis between nodes and cascades' properties

	Cascade Depth				
	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
Unweighted Social Graph					
Degree	0.26	0.25	0.27	0.28	0.29
Clust. Coef.	−0.05	−0.05	−0.07	−0.08	−0.11
PageRank	−0.13	−0.11	−0.10	−0.08	−0.07
Eigenv. Cent.	0.35	0.34	0.34	0.34	0.33
Weighted Social Graph					
Activity	0.72	0.75	0.77	0.78	0.79
Clust. Coef.	0.09	0.10	0.08	0.06	0.04
PageRank	0.32	0.34	0.36	0.39	0.41
Eigenv. Cent.	0.20	0.28	0.29	0.30	0.30
Burt Constr.	−0.35	−0.34	−0.36	−0.37	−0.38

5.4 Discussion

In this chapter, the role of the properties of ego network is analysed in the process of diffusion on information in OSN. In particular, we inspected the role of the social circles and the role of the ego centrality in the network. These analysis were conducted through the generation of a set of information cascades using the Independent cascade Model. We have assigned to each social link a probability of

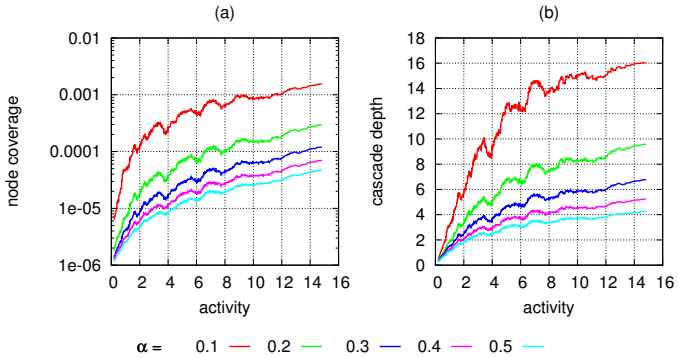


Figure 5.5: Node coverage (a) and Cascade depth (b) of the information cascades generated by seeds with different activity, considering different values for the α parameter of the model, plotted using the running average with subset size of $\mathbf{50}$ elements.

diffusion calculated as a function of the interaction frequency of respective pair of ego.

In the first analysis, we have performed an information diffusion analysis assessing the impact of the different ego network rings. We have estimate the probability of diffusion using an exponential function of the frequency of contact extracted from the dataset. The results shows that the most used social circles are the middle circles. We continued the analysis, generating another set of simulations, in which we selectively removed from the network all the links belonging to a selected social circle or a group of social circles. In the literature, social networks have been found to be more resilient to the removal of strong ties than weak ties since weak ties are often bridges representing the only connection between otherwise disconnected parts of the network. Nevertheless, our results indicate that, if we remove all the strongest ties from all the ego networks, the diffusion would be very limited. This means that strong ties are fundamental to transport information within cohesive groups of individuals because of their intrinsic high level of trust.

In the second analysis, we studied how the centrality of the node influences its capability to generate large information cascades. In the model used to calculate the probability of diffusion starting from the frequency of contact, we included the ageing of the information which decrease the probability of diffusion at each step of the simulation. The generated information cascades were analysed calculating the correlation between their properties (i.e. node coverage and cascade depth) and the characteristics of the starting node. These results indicate that the highest correlation are related to statistics of the seeds which involve tie strength, namely the activity of the seeds and the eigenvector centrality. Interestingly, the clustering coefficient shows a low correlation with the properties of the cascades. The Burt's constraint - a measure of the the number of structural holes in an ego network - has medium (and negative) correlation with the cascade depth and node coverage, indicating that more constrained ego networks limit the diffusion of information. Interestingly, having access to the unweighted graph only (without tie strength) is not sufficient to identify which seeds will be able to generate large cascades.

6

The Impact of Trust on Information Diffusion

Thanks to Online Social Networks we are witnessing a rapid shift from the physical world of face-to-face communications to the world of virtual contacts and ubiquitous services. This, and the presence of increasingly smart devices (e.g. smartphones, tablets, and smart objects) are significantly contributing to the so called Cyber-Physical World (CPW) convergence [27], for which actions in the physical world modify the state of entities in the virtual world and vice-versa. The paradigm on which OSN are built allows people to create and share content amongst themselves in the virtual world, empowering individuals and allowing them to create communities and services, which are more and more impacting on the life in our physical world. In this new scenario, the data users generate, which, for their complexity, heterogeneity and quantity, are called “big data”, represent a treasure of inestimable value for the service providers and for the final users. Big data are indeed aggregated and analysed to create novel services, such as predictions of events happening in the physical world through the analysis of communication data in OSN (e.g. car crashes detection [80], prediction of spread of diseases [81]). Although these services are important to the people, OSN service provider often centralise and limit the access to the big data their users generate. Consequently, and due to the high intrinsic value of the data, wealth is centralised as well, possibly leading to power law economies [60].

Recently, new decentralised solutions based on the paradigm of OSN, known as Distributed Online Social Networks [73] (hereinafter DOSN), have appeared on the market (like diaspora* and PeerSon [20]), accompanied by a growing research interest in the field. DOSN replicate OSN features in a decentralised way, avoiding data centralisation. Each user maintains her personal data locally or on intermediate servers, and interactions between users occur through peer-to-peer (P2P)

communications. Compared to OSN, DOSN may guarantee more transparency in data management.

This chapter presents an analysis,presented in [11] and in [12], of the capacity of a social network to spread information under the hypothesis that individuals are willingly to share contents only with trusted friends. This hypothesis is is relevant in DOSN systems, where the absence of a trusted centralised operator could induce the users to adopt more restrictive privacy policy.

6.1 Related Work

In this section we present the most relevant work in the literature concerning the context of our analysis, which includes DOSN and the analysis of information diffusion in OSN.

6.1.1 Distributed Online Social Networks

DOSN were born in recent years to address privacy concerns over OSN. Diaspora is probably the most famous DOSN nowadays. Diaspora* supports the possibility of either creating a server (called *pod*) where the user can host her personal data or using an already existing one. Social interactions are carried out through a P2P system that makes users communicate directly with each other, without passing through a single centralised server. Buchegger at al. [20] propose a similar solution, which has been also extended to be used in case of absence of stable Internet connectivity [19], a scenario particularly suited for mobile devices. Guidi et al. [46] propose a DOSN based on the automatic identification, for each user, of her ego network layers, using the contact frequency between the user and her social contacts. The differences in terms of trust between the different layers are used to automatically adjust the privacy policies towards the people in the layers. Moreover, the personal social network of each user is limited to her “active network”, and people beyond it are excluded from the main features of the system. The solutions proposed by Han and colleagues [48] and Cutillo et al. [30] further exploit trust relationships arranged in concentric layers around the users to replicate the data of the user on her friend’s devices, guaranteeing the access to her data even though her device were inaccessible due to a temporary disconnection or turnoffs.

6.1.2 Information Diffusion Analysis in OSN

In the last few years OSN has attracted a lot of interest from the Research community. One of the distinctive properties of OSN, which differentiate them from other kinds of networks, including technological and biological networks, is the presence of a non-trivial clustering coefficient (or transitivity) [69]. This indicates a high probability that two neighbours connected to a node will also be connected to each other. Moreover, social networks (including OSN) show the so called small-world property [88]. According to this property, any two persons in the network, indirectly connected by chains of social links, have a short average distance. This fact directly influences the ability of the network to quickly spread information, ideas, innovations and so forth. It has been demonstrated that the diffusion of information in social networks takes place through single social links, creating the word-of-mouth effect [37]. This property has been largely used by a collection of marketing techniques whereby the presence of social links between consumers is exploited to increase sales [50]. Recent studies on OSN quantified the capacity of the network to diffuse information and the role of different types of users in the process [16, 23]. Content locality plays an important role in OSN [24]. This means that information diffusion is often limited to the immediate neighbours (or n-hops neighbours, with small values of n) of the user who generated the content. Thus, identifying influential nodes covering the role of opinion leaders that are able to generate large diffusions, and detecting popular topics are the most important tasks for the analysis and prediction of information diffusion [47]. Several models have been proposed to simulate information diffusion in OSN [34, 25, 75, 83]. These models are generally derived from static observations of diffusion patterns in OSN. To be able to analyse the dynamic evolution of information spread, Taxidou and Fisher presented a set of methods for the analysis of real-time diffusion of information, through the online analysis of data generated by Twitter [87, 86].

6.2 Information Diffusion in DOSN

As in more traditional OSN, DOSN allow users to create and manage their *digital personal space*, where they can post and receive asynchronous messages, and insert their personal information. Moreover, DOSN support the creation of *social links* between users, giving different access policies to digital personal spaces for friends compared to strangers. DOSN also provide instant messaging functionality in the form of private communications.

Research in the field of DOSN is mainly focused on providing networking functionality even in highly dynamic scenarios with mobility of nodes and possible absence of access to network infrastructure [19]. From a technological perspective, DOSN are already based on rather solid solutions. Yet, from a higher level perspective, little is known about the capacity of DOSN to diffuse information. The ability to spread information quickly amongst people is one of the key properties of OSN, and assessing this capacity in DOSN is fundamental to understand possible limitations of the system and to design new services for distributed environments. In DOSN, the main limitation for the analysis of information diffusion is the lack of large-scale communication data sets. This is because data in DOSN are completely decentralised and content is exchanged directly between the personal devices of users. For this reason, the circulation of information cannot be traced easily. Nevertheless, DOSN clearly share similarities with OSN, for which data sets of communication data are easier to be obtained and analysed. The main difference between OSN and DOSN is that in the latter content is disseminated within the network only through chains of direct communications between users. This encourages users to have a more strict control over the data passing through their social links. In fact, too many accesses to the web page of a profile could represent a bandwidth wastage for the related user. In addition, since no central control exists over the exchanged content, accounts generating spam and other kinds of undesired content must be detected and blocked directly by the users. Therefore, it is reasonable to assume that DOSN users will be willing to help replicating and disseminating content coming primarily from a set of users they trust most. This means that, for certain applications, the *effective* social graph in DOSN may be limited to the links between users with a strong enough social relationship. Starting from this assumption, we can estimate structural properties of the social graph in DOSN by analysing OSN data, after eliminating links under a certain level of trust.

Based on these general remarks, we analyse information diffusion capacity in DOSN. To do this, we study various properties of a network graph representing a large portion of Facebook, restricting social links to trusted relationships only (we discuss how we estimate trust later on). In particular, we look at the *connectivity* and the *spreadability* properties of large components containing connected nodes in the graph after the deletion of untrusted links. These two metrics represent the size of the component and its intrinsic capacity to spread information. Each component represents an isolated portion of the original network through which information can reach all the connected nodes. Clearly, different network components (and thus different information dissemination patterns) emerge depending

on whether more or less strict trust restrictions are considered. Our analysis is aimed at discovering if these components are large enough to cover a significant portion of the users and if their structure is suitable for diffusing information. This could help to identify the impact of DOSN on the market and whether they could be considered a valid alternative to OSN or not in terms of their capacity to spread information, and, for example, for advertisement.

To perform the analysis, we firstly estimate the trust level between users in Facebook through the frequency of interactions between them. The use of the contact frequency to estimate trust between people is well backed-up by results in sociology [49, 10]. Hence, we simulate content diffusion in the Facebook network graph considering different thresholds for selecting trusted links. We assume that no central control exists on this network and we select the set of trusted contacts for its users according to their contact frequency, then studying the properties of the resulting graph. Specifically, if a social relationship is not trusted (i.e. it does not have a sufficiently high contact frequency), the respective social link is not included in the graph we use in the analysis. To assess the impact of the selection of trusted links, we define the minimum level of trust by setting a threshold on the contact frequency of the links to be included in the graph. We take values of this threshold equal to the frequencies of contact that have been used in the literature for defining different levels of social relationships [49]. In particular, we consider the well known ego network model [85], whereby social relationships of a user (ego) can be divided in concentric layers of increasing size and decreasing social intimacy (i.e. corresponding to decreasing tie strength and fewer interactions). In this way we obtain different social graphs with different minimum levels of trust, that coincide with a natural categorisation of social relationships in humans. Note that this way of estimating trust lends itself to automatic systems to decide on which social links to accept content, just by monitoring the frequency of interactions on them.

We compare the results obtained for the different thresholds to identify the values that lead to a large enough connected component, with a sufficient ability to spread information. The results of the analysis indicate that limiting content spread to social contacts that coincide with the definition of “active social contacts” of the users, which corresponds to the most external layer in the ego network model, leads to a network graph with a sufficiently large component of connected nodes, which covers more than 96% of the original Facebook network. Restricting content spread to the next layer of the ego networks, or further, makes the relative size of the biggest connected component (and therefore coverage) drop below 30%. Since the remaining components are very small compared to the largest

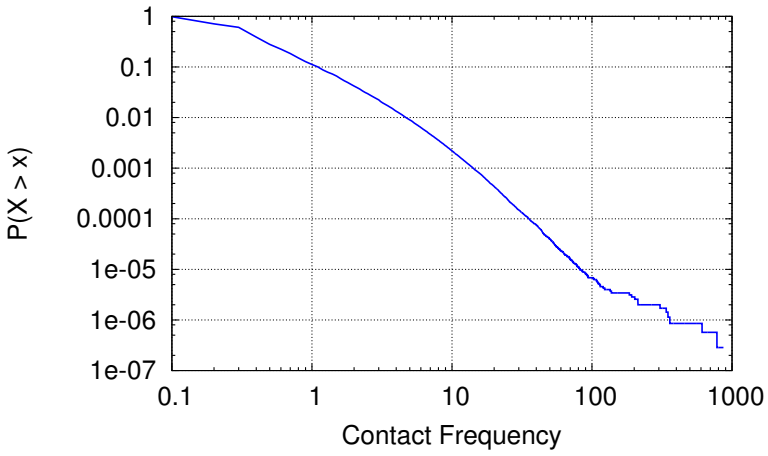


Figure 6.1: CCDF of the contact frequency for the links.

one for all the used thresholds, diffusing information in the network could be problematic when the largest component does not cover a sufficiently high number of nodes. As a possible solution to increase node coverage in case of very restrictive thresholds we investigate the effect of adding to the graph only one social contact for each user, selected with different possible strategies (e.g. select the link with highest/lowest contact frequency with the user, select a random acquaintance, etc.). The results indicate that this solution considerably increases the number of covered nodes, even in case of very strong trust. Noticeably, all the strategies, included the selection of the contact with highest contact frequency (below the minimum contact frequency imposed by the restriction), leads to very high improvement in terms of node coverage. Clearly, adding a contact to the list of trusted nodes represents a cost for the users in terms of additional unwanted content, but limiting the choice to a single node should be a reasonable solution for them since they would receive a global return in terms of quality and quantity of information circulating in the network.

6.2.1 Data Set Description

To perform our analysis, we use the same large-scale Facebook data set containing information about social interactions between users introduced in Chapter 3 in wich we applied a slightly different preprocessing. Since in our analysis we are interested in users who actively communicate with others, we select from the Face-

Figure 6.2: Ego network model.

book graph only the users with at least one active link (i.e. with contact frequency > 0) and we discard all the other users, that indeed are inactive. Moreover, we further restrict the analysis to the set of users that have communicated with other users at least 6 months before the time the data set was downloaded. This ensures that our analysis is restricted to sufficiently stable users. In fact, the contact frequency of new users in OSN is generally higher that that of older (and more stable) users [7] and could bias the analysis. The resulting graph, after this preprocess, consists of $1,083,209$ nodes and $7,709,309$ links.

More formally, we obtain a graph $G = (V, E)$ formed of a set of vertices (or nodes) V, and a set of edges (or links) E connecting pairs of nodes, each of which represents a 2-element subset of V. For convenience, we identify an edge by the nodes it connects. For example, $\{i, j\}$ represents the link between node i and j. The contact frequency of the link is defined as $f_{i,j}$. The graph in our data set is undirected. This means that $\{i, j\}$ is equal to $\{j, i\}$ and their contact frequency is also the same.

In this analysis we use the contact frequency between users in Facebook as a proxy for the level of trust between them. This is supported by results in the literature that identified a strong relation between the contact frequency and the tie strength or emotional closeness between people, both in offline and online environments [10, 49, 64]. The complementary cumulative distribution function (CCDF) of the contact frequency for the links in the graph obtained from the data set is depicted in Fig. 6.1. The figure indicates that the distribution has a power law trend, thus implying that most of the links in the network have a very low level of trust, whereas only few links have very high trust. For this reason, we expect that restricting the network to trusted links only could have a strong impact upon the structural properties of the resulting graph.

To simulate the restriction of communication to a list of trusted contacts for each user in DOSN we apply a series of filters to the Facebook social graph previously described, eliminating the links with contact frequency below the chosen threshold, that defines the boundary of the trusted contact list.

6.3 Social Networks for Content Diffusion

Since in our analysis we are interested in users who actively communicate with others, we select from the Facebook graph only the users with at least one active

link (i.e., with contact frequency > 0) and we discard all the other users, that indeed are inactive. Moreover, we further restrict the analysis to the set of users that have communicated with other users at least 6 months before the time the data set was downloaded. This ensures that our analysis is restricted to sufficiently stable users. In fact, the contact frequency of new users in OSN is generally higher than that of older (and more stable) users [93] and could bias the analysis. The resulting graph, after this pre-process, consists of $1,083,209$ nodes and $7,709,309$ links.

6.3.1 Trusted Contact List Based on the Ego Network Model

The ensemble of social relationships of a person can be modelled as an ego network, that is a simple social network model which considers only an individual (called "ego") and the set of people with whom she has a social relationship (called "alters"). The main property of human ego networks is the presence of a hierarchical structure formed of layers of alters around the ego [85]. The typical structure of ego networks in human social environments shows four concentric layers containing alters at different levels of emotional closeness and with different size. Since the emotional closeness depend upon several psychological aspects of the relationship it is generally difficult to be directly measured. Nevertheless, it is strongly correlated with the contact frequency [49], and thus it is generally estimated using the latter.

In the ego network hierarchical structure, depicted in Figure 6.2, the first and innermost layer is the *support clique*, containing on average five people very close to the ego and contacted by her at least once a week. The *sympathy group* (that includes the support clique) contains fifteen members contacted at least once a month. The *affinity group* contains fifty members contacted at least $\sim$ eight times a year [8]. Lastly, the *active network* contains 150 people contacted at least once a year. The members of this last layer are people for whom the ego invests a non negligible amount of cognitive resources for the maintenance of their social relationships. Beyond the active network, alters are mere acquaintances or friends no longer contacted, and their social relationships are not actively maintained by the ego.

Alters in the same layer share similar properties in their relationships with the ego. Specifically, people in the support clique, broadly identified as "very intimate friends", are those contacted by the ego also in case of need of financial or emotional support, and thus are typically the ones the ego trusts the most. The sympathy group is composed of "close friends" to the ego, contacted less frequently than

alters in the previous layer, and generally representing the group of reliable friends on whom one can depend for a variety of exchange relationships (e.g., friendships in the social sense, protection against harassment, minimising social stress, distributed childcare) [85]. The affinity group is formed of “friends”, contacted by the ego when she cannot find enough available friends in the affinity group with specific skills she needs to solve a task or in situations requiring a group of several people. Lastly, The active network contains “casual friends”, usually contacted by the ego for a particular event or in case of need to access resources outside her social network. In fact, members of this layer are usually loosely connected with the ego and share a small number of mutual relationships, and represent bridges to reach other social groups. From this characterisation, it is clear that the trust level between the ego and her alters decays from the inner to the outer layers of the ego network. In support of this, the contact frequency between people has been found to be correlated with trust, both in anthropology [84] and social network analysis [1].

Based on the definition of the ego network layers, we can identify possible trusted contacts lists definitions that could be adopted in DOSN. This can be done by selecting social relationships belonging to a specific ego network layer, which can be identified by using the typical threshold of contact frequency of the layer. Note that the values of these thresholds have been found in several studies in offline and online social networks, indicating that they are invariant to the use of specific communication means, and they are instead determined by human cognitive and social processes [85]. Of course, contact frequency could slightly differ from the typical values we presented before since different social platforms (e.g., Facebook, Twitter, Google+) could have differences in their social interaction mechanisms, but the order of magnitude of the contact frequency for the different layers remains consistent amongst different communication media. In the case of the dataset we analysed, the thresholds which characterise the layers are very close to the typical values found in the literature [8]. Thus, we use these values to create four possible trusted contact lists for each user, and thus four different network graphs. For example, to simulate the presence of lists containing “friends” we can fix the minimum contact frequency to be considered in the analysis to eight messages a year (the affinity group). In the rest of the paper we indicate the values of the thresholds in number of messages per month, so “1/12” represents one message a year, “8/12” eight messages a year, “1” one message per month, and “4” four messages per month.

6.3.2 Network Connectivity

The *network connectivity* is the property of a social network indicating the extent to which its nodes are interconnected to each other. Usually, social networks are formed of a giant component of connected nodes covering the vast majority of the nodes, and the remaining nodes divided into a large number of small and isolated components [18]. Clearly, disconnected components do not contribute to the diffusion of information since they are not reachable from the largest part of the network, and information generated by them remains isolated. Thus, the largest the giant component the higher the potential capacity of the network to diffuse information to a large number of nodes. To quantify network connectivity, we measure the fraction of nodes in the largest component of connected nodes with respect to the total number of nodes in the network.

To further characterise the structure of the network, we define the percentage of nodes in the largest connected component of the network graph with respect to the total number of nodes in the network as "node coverage". The higher this percentage, the higher the potential of the network to make information circulating among all the nodes. Moreover, we consider the minimum number of components needed to cover a certain percentage of nodes of the network. Regarding information diffusion, the number of components needed to reach a certain node coverage represents the minimum number of message replicas to be generated. In fact, since the involved components are disconnected from each other, information cannot spread amongst them, and a replica of the message to be spread must be created and injected in each component.

Having defined the network graphs at different levels of trust using the identified thresholds on the contact frequency, we assess their potential capacity to diffuse information. Specifically, for each graph we studied its network connectivity, and the impact of the number and the size of their components considering several levels of node coverage.

6.3.3 Network Spreadability

We informally define *network spreadability* as the property of a network graph indicating its capability to diffuse information. This depends on two main factors: (i) the structure of the network, and in particular the distance in number of edges between nodes, and (ii) the contact frequency between pairs of nodes. In fact, the smaller the average number of links between nodes to be traversed by information to reach the other nodes, the higher the probability of obtaining large infections, as the propagation through each link is not guaranteed. Specifically, information

diffusion on a link depends upon the importance and the trust level of the underlying social relationship [84, 1], and, in particular, upon the contact frequency between the users [17]. For these reasons, network spreadability can be quantified by analysing the properties of the weighted shortest paths between nodes in the network graph. The shortest paths between random pairs of nodes in the graphs represent the shortest way for information to reach a destination from a random source. The analysis of the properties of these shortest paths, calculated considering the weight on each link, permits to better characterise the intrinsic ability of the graph to diffuse information. In addition, we assume that any piece of information that is exchanged over a link undergoes a loss of associated trustworthiness which is proportional to the trust between the two users. Quantitatively, we estimate the decay of trustworthiness over a link with a number between 0 and 1 (where 0 means total distrust, and 1 complete trust), and estimate the total loss of trustworthiness over a path as the product of trust decays over path links. The analysis of trustworthiness decay over shortest paths allows us to quantify the level of trustworthiness of information circulating in the graph.

Note that, whilst network connectivity indicates the maximum number of nodes which could be potentially infected by information with a certain number of message replicas, network spreadability indicates the capability of the nodes in a single component to spread information to each other. These properties describe different aspects of the network, and allows us to analyse the relation between the structure of the network weighted by the importance of the relationships between users (thus including their trust) and information diffusion.

A weighted shortest path is the sequence of links connecting a pair of nodes in a network with the lowest sum of link weights. The weight represents the cost to travel to a certain link (e.g., the distance or the cost of sending a message through the link). In our case we use the inverse of the contact frequency of the links as weight. For convenience, we also normalise the weights to be in [0,1]. The weight of social links between any pair of vertices $\{i, j\}$ in the set of vertices $V(g)$ of the network graph g is calculated as follows:

$$w_{i,j} = \frac{\min\limits_{i,j \in V(g)} f_{i,j}}{f_{i,j}} \tag{6.1}$$

where $f_{i,j}$ is the contact frequency between i and j, and thus $1/f_{i,j}$ is the weight before normalisation, while $\min\limits_{i,j \in V(g)} f_{i,j}$ is the normalisation factor.

To calculate the weighted shortest paths in the four graphs obtained by applying the thresholds, we use the Dijkstra's algorithm on the portion of the graphs

representing their giant components. Since the algorithm has a time complexity of $\mathcal{O}(V^2)$, and since some of our graphs are very large, we sample a fixed number of $10,000$ pairs of vertices for each graph. To better compare the results obtained with and without re-insertion, we sample the starting vertices of the paths within the components obtained without re-insertion, one at each value of the threshold. Then, we sample the ending vertices of the paths within the set of nodes that becomes connected to the component after re-insertion (not included in the component without re-insertion). In case of no re-insertion, both the starting and the ending vertices are chosen within the set of nodes in the components.

For each largest connected component, obtained with the different thresholds and re-insertion strategies, we calculate the set of shortest paths between the sampled pairs of starting-ending nodes. Then, we calculate four measures to quantify the content spreadability of the paths and we average them for all the obtained paths at each configuration. The measures we calculate are the following: (i) the length, in number of nodes, of the path. This represents the effective average distance in number of hops between the nodes in the component (note that we include also the starting and the ending nodes in the count); (ii) The sum of the weights of the path, which represents the cost for information to travel through a specific path. In fact, the sum of the weights on the paths indicates the total cost to pass through the path; (iii) The average contact frequency of the links of the path; and (iv) the product of the contact frequencies on the path, which can be interpreted as the effective loss in terms of trustworthiness of information travelling on the path.

6.3.4 Strategies for Link Reinsertion

As will be clear from the results in Section 6.4, for some threshold on the trust level we obtain quite small largest connected components, and a big number of extremely small additional disconnected components. To improve network connectivity we tried, as a possible alternative to lowering the trust value of the system, the re-insertion of one social contact for each user in the graph obtained at each level of trust, testing several possible re-insertion strategies. Considering that high contact frequency indicates strong relationships between users, strategies which privilege links with higher contact frequency are clearly the most appealing. For this reason, we considered the strategy “highest frequency” (indicated as “high freq” in the tables) which deterministically re-inserts the link with the highest frequency below the threshold. Note that, the reinserted link could show a contact frequency sensibly lower than the threshold value, not necessarily belonging to

the first excluded layer, since for some ego networks some of the layers could not be present. We also considered a "probabilistic" strategy ("Prob" in the tables) assigning to each excluded link a probability to be reinserted proportional to its contact frequency. Thus, compared to the previous strategy, the probabilistic one gives more chances of re-insertion to weaker links. These links are usually bridges connecting parts of the network socially far from each other, and they are shortcuts which could positively impact on network connectivity and spreadability. Note that this is compatible with the idea behind the Kleinberg's small-world model, for which nodes are densely connected to neighbour nodes, with which they share several social contacts, but they also have long-range links with other nodes socially far from them [56]. The combination of these two properties in the model leads to short average path length in the network. To assess the goodness of the strategies we introduced also a baseline "Random" strategy ("Rand" in the tables), which assigns to each link the same probability to be reinserted. For completeness, we also introduced two additional strategies which privileges the link with lowest interaction frequencies, which are the worst choices considering the trust of the relationship, but could introduce more bridges. Specifically, we considered the following strategies: "lowest frequency" ("low freq in the tables") which deterministically selects the link with the lowest contact frequency, and the "inverse probabilistic" ("Inv prob" in the tables) which assigns to each link a probability inversely proportional to the interaction frequency of the link. Obviously, many more strategies can be defined and analysed, but we though that with this five strategies we can asses whether the reinsertion of a single link leads to an improvement in terms of network connectivity and spreadability.

6.4 Results

6.4.1 Network Connectivity

The network connectivity of the graphs obtained after the pre-processing phase described in Section 6.3 is reported in Table 6.1 in the column "No insert", indicating that we have not applied any re-insertion strategy on these results. The first largest component obtained with the threshold coinciding with the contact frequency of "active social contacts" (as defined in Section 6.3.1) guarantees node coverage close to 1. This means that with this minimum level of trust the network could potentially support the diffusion of information to almost all the nodes. On the other hand, the most strict threshold ("very intimate friends") leads to a node coverage of ~ 0.03. This means that the resulting network is highly disconnected,

Table 6.1: Percentage of nodes of the original graph covered by the largest component for the different thresholds representing the minimum contact frequency on the links. Thresholds are expressed in msg/month

Threshold min. cont. freq.	**Node coverage**					
	No insert	**High freq**	**Low freq**	**Prob**	**Inv. prob**	**Rand**
$1/12$ (act. cont.)	0.966	0.994	0.994	0.994	0.994	0.994
$8/12$ (friends)	0.297	0.714	0.705	0.726	0.722	0.725
1 (close fr.)	0.191	0.642	0.634	0.661	0.657	0.661
4 (v. intimate fr.)	0.028	0.386	0.385	0.453	0.444	0.456

Figure 6.3: CCDF of the size of the components for each threshold and strategy, excluded the largest component.

and can hardly support information diffusion. The remaining thresholds represent intermediate results, with node coverage ~ 0.3 for "friends" and ~ 0.2 for "close friends". Whilst these results are clearly suboptimal with respect to the "active social contacts" threshold, they could still be exploited in circumstances not needing a complete coverage of the nodes, but requiring more trust between them.

Figure 6.3 depicts the distribution of the size of the components of connected nodes in the network excluded the largest one, in which, for all the thresholds, and especially in case of no re-insertion, these components are always very small with respect to the giant component. In fact, the distributions show power-law trends with a maximum component size of 95 nodes. This indicates that if we wanted to reach a high number of nodes in the network and the largest component was not sufficient to do so, it would be necessary to place information on a large number of additional components, without relying on automatic spread of information over trusted social links. This is further confirmed by the results in Table 6.2, 6.3, 6.4, and 6.5 under the row "No insert". The tables report the number of components which must be infected by information to reach the desired level of node coverage. Thus, only the threshold of one message per year (Table 6.2) generates a graph that requires a small amount of message replicas (just one message for 90% node coverage), whereas for the other thresholds the number of replicas needed to reach a coverage of at least 50% of the network is very high and would result in a very expensive process. Moreover, whilst it could be relatively easy to identify the largest component in the network, it is not easy to identify all the remaining components, especially in decentralised systems like DOSN. This fact could further limit the diffusion process.

To improve the network coverage, thus permitting larger information spreads expecially for disconnected graphs resulting from restrictive thresholds, we evaluate the re-insertion strategies as possible alternatives to the generalised decrease of the trust level. We apply the re-insertions on the Facebook graph at each threshold. In Table 6.1, from the third to the last column, we report the size of the largest component of connected nodes for each combination of threshold and re-insertion strategy. As can be noted, the impact of the re-insertion is substantial for thresholds $> 1/12$. The impact of link re-insertion for the threshold of $1/12$ is negligible, since most of the nodes of the original network are already present in the resulting graph. On the other hand, for the most restrictive threshold (4 messages per month) the gain due to the re-insertion is effective, bringing the node coverage to $\sim 40\%$ which indicates a giant component of about 15 times larger than the one without re-insertion.

The results of the different strategies vary significantly, with the probabilistic and the random strategies ("Prob" and "Rand" in the tables) giving the highest improvement in terms of number of nodes covered, as reported in Table 6.1. In addition, as reported in Table 6.2, 6.3, 6.4 and 6.5, these two strategies seem the most convenient (at least from the point of view of the number of connected nodes) also when all the other components, in addition to the largest one, are considered. Considering the cost for the users, the probabilistic strategy is intuitively better than the random one since guarantees that, on average, the re-inserted nodes have higher trust level than randomly selected nodes.

We also look at how the different re-insertion strategies impact on the distributions of the sizes of network components other than the largest one (see Figure 6.3). All strategies, for each threshold, produce a similar distribution of the size of the components (the largest one is not present). Nevertheless, the distributions

Table 6.2: Number of components needed to cover the specified percentage of nodes in the original network using a min. contact frequency of 1/12 msg/month (active contacts).

	Coverage						
Strategy	40%	50%	60%	70%	80%	90%	100%
No insert	1	1	1	1	1	1	$31,987$
High freq	1	1	1	1	1	1	$1,784$
Low freq	1	1	1	1	1	1	$1,784$
Prob	1	1	1	1	1	1	$1,784$
Inv prob	1	1	1	1	1	1	$1,784$
Rand	1	1	1	1	1	1	$1,784$

Table 6.3: Number of components needed to cover the specified percentage of nodes in the original network using a min. contact frequency of 8/12 msg/month (friends).

	Coverage						
Strategy	40%	50%	60%	70%	80%	90%	100%
No insert	94,218	202,539	310,860	419,181	527,502	635,823	744,143
High freq	1	1	1	1	43,045	151.366	259,686
Low freq	1	1	1	1	61,369	169,690	278,010
Prob	1	1	1	1	37,623	145,944	254,264
Inv prob	1	1	1	1	45,197	153,518	261,838
Rand	1	1	1	1	40,343	148,664	256,984

vary from the case in which no re-insertion is applied, especially for restrictive thresholds (1 message per month and 4 messages per month). This can be explained by the fact that for these thresholds the largest component is sensibly smaller than for the other thresholds, and the probability of re-inserting a node connected to this component is lower. Thus, there is the presence of a higher number of larger components disconnected from the largest one.

6.4.2 Network Spreadability

Hitherto, we analysed the ability of the network limited to a certain level of trust to maintain nodes connected to each other, as a first requirement for information diffusion. However, the mere presence of a path connecting all the nodes in one or more components of the network indicates only the possibility to reach the nodes with information, but it does not consider the effective ability of the graph to spread

Table 6.4: Number of components needed to cover the specified percentage of nodes in the original network using a min. contact frequency of 1 msg/month (close friends).

	Coverage						
Strategy	40%	50%	60%	70%	80%	90%	100%
No insert	208,769	317,090	425,411	533,732	642,053	750,374	858,694
High freq	1	1	1	8,801	87,540	195,861	304,181
Low freq	1	1	1	13,147	106,719	215,040	323,360
Prob	1	1	1	4,271	79,561	187,882	296,202
Inv prob	1	1	1	5,470	87,379	195,700	304,020
Rand	1	1	1	4,343	81,852	190,173	298,493

Table 6.5: Number of components needed to cover the specified percentage of nodes in the original network using a min. contact frequency of 4 msg/month (very intimate friends).

	Coverage						
Strategy	40%	50%	60%	70%	80%	90%	100%
No insert	391,174	499,495	607,816	716,137	824,458	932,779	1,041,099
High freq	45	2,332	12,660	47,708	144,729	253,050	361,370
Low freq	68	3,022	15,938	59,169	167,490	275,811	384,131
Prob	1	431	6,717	36,881	132,106	240,426	348,746
Inv prob	1	608	7,708	40,266	140,250	248,571	356,891
Rand	1	396	6,538	36,785	133,350	241,672	349,992

content. To delve deeper into the analysis of spreadability of the networks, we calculate the set of measures introduced in Section 6.3.3, calculating the properties of weighted shortest paths in the graphs. For this analysis we consider only the giant component of connected nodes since it is the most important part of the network for the diffusion of information. In Table 6.6 we report the average length of the shortest paths sampled in the components.

From the table we can note that comparing the basic case without re-insertion and the different re-insertion policies, there is no sensible difference in terms of average shortest path length for the first threshold (i.e., 1/12). This is in line with the results found in terms of connectivity since the largest component found at this threshold is large enough to contain most of the nodes in the original graph. For the threshold of 8/12 msg/month, the results are consistent amongst the different re-

Table 6.6: Average length (# of nodes) of the weighted shortest paths in the largest component for the different thresholds representing the minimum contact frequency (msg/month) in the network.

	Threshold - min. contact frequency			
Strategy	**1/12 (active cont.)**	**8/12 (friends)**	**1 (close fr.)**	**4 (v. intimate fr.)**
No insert	11.67	10.81	10.51	11.07
High freq	11.72	11.75	11.95	13.74
Low freq	11.68	11.93	12.19	16.11
Prob	11.71	11.95	12.21	16.16
Inv prob	11.71	11.97	12.30	17.42
Rand	11.74	11.95	12.28	17.15

Table 6.7: Average sum of weights of the shortest paths in the largest component for the different thresholds representing the minimum contact frequency (msg/-month) in the network.

	Threshold - min. contact frequency			
Strategy	**1/12 (active cont.)**	**8/12 (friends)**	**1 (close fr.)**	**4 (v. intimate fr.)**
No insert	0.41	0.16	0.12	0.07
High freq	0.46	0.44	0.45	0.40
Low freq	0.46	0.60	0.73	1.74
Prob	0.46	0.50	0.53	0.73
Inv prob	0.46	0.55	0.64	1.42
Rand	0.46	0.53	0.58	1.08

insertion policies, and the average length sees an additional node compared to the case without re-insertion. This difference increases for the remaining thresholds, with a maximum difference of ~ 4 nodes between the best and the worst results ("high frequency" and "inverse probabilistic" respectively) for the threshold of 4 msg/month. In case of this threshold, the difference between the path length in the largest component without re-insertion and the components with re-insertion is of at least ~ 2.5 nodes.

From the data in Table 6.6 it is worth noting that the average length of weighted shortest paths is considerably higher than the same measure in unweighted social graphs, that is known to be around six [88] in social networks. This is due to the fact that the shortest weighted paths we obtained include links with very high trust belonging to the most internal ego network layers, and which, for their low travers-

Table 6.8: Average contact frequency on the weighted shortest paths in the largest component for the different thresholds representing the minimum contact frequency (msg/month) in the network.

	Threshold - min. contact frequency			
Strategy	**1/12 (active cont.)**	**8/12 (friends)**	**1 (close fr.)**	**4 (v. intimate fr.)**
No insert	0.49	0.55	0.58	0.77
High freq	0.50	0.50	0.50	0.58
Low freq	0.49	0.50	0.49	0.55
Prob	0.49	0.49	0.49	0.55
Inv prob	0.49	0.49	0.49	0.53
Rand	0.49	0.49	0.49	0.54

Table 6.9: Average product of normalised contact frequencies on the weighted shortest paths in the largest component for the different thresholds representing the minimum contact frequency (msg/month) in the network.

	Threshold - min. contact frequency			
Strategy	**1/12 (active cont.)**	**8/12 (friends)**	**1 (close fr.)**	**4 (v. intimate fr.)**
No insert	$2.721e-4$	$2.179e-3$	$4.711e-3$	$7.458e-2$
High freq	$2.317e-4$	$4.943e-4$	$5.471e-4$	$1.055e-3$
Low freq	$2.082e-4$	$4.398e-4$	$4.111e-4$	$4.607e-4$
Prob	$2.336e-4$	$4.278e-4$	$4.722e-4$	$5.753e-4$
Inv prob	$2.511e-4$	$4.368e-4$	$4.781e-4$	$4.174e-4$
Rand	$2.466e-4$	$4.261e-4$	$4.618e-4$	$4.472e-4$

ing cost, are frequently used in the paths. In unweighted networks, each link has the same cost and this effect is not present.

A similar trend can be derived from the figures in Table 6.7 and 6.8, which show the total cost of the paths and the average contact frequency of the paths, that is proportional to the average trust level of the links. For these measures the best and the worst re-insertion policies for the different thresholds appear to be the “highest frequency”, and the “lowest frequency” and “inverse probabilistic” respectively. This is an intuitive result, since the weight we used to calculate the shortest paths is directly related to the contact frequency. Note that the “highest frequency” and the “lowest frequency” policies are associated to sensibly smaller network graphs than the other policies, as reported in Table 6.1. This means that, even though the “highest frequency” policy seems to be the best choice from the point of view of network spreadability, it leads to poor node coverage.

Table 6.9 reports the product of the normalised contact frequencies on the shortest paths, that estimates the frequency at which information traverses the whole path. The measure is influenced both by the contact frequency of the links on the shortest paths and their length. As it can be noted from the results in the table, these measure is consistent with the previous results.

It is worth noting that there is a trade-off between node coverage of the component and the spreadability of its network graph. The average trust level decays when re-insertion is applied since more hops are added to the paths in the component, and the trust level of the re-inserted links is lower than the threshold of the component. The decay varies with the value of the threshold, and it could be too high in some cases. Nevertheless, the “probabilistic” strategy is the one giving a

the best node coverage, and maintaining good spreadability at the same time, and seems the most reasonable choice.

6.5 Discussion

The restriction of communications in DOSN to trusted social relationships only is essential in DOSN, since the users are willing to distribute information coming only from trusted peers to limit the resources they dedicate to communications, that are generally limited. To perform the analysis of the impact on information diffusion we analyse the topological properties of the social graph generated by DOSN with such restrictions, looking for the presence of a large component of connected nodes (at a certain threshold of trust), within which information can spread and possibly reach all its nodes. On the other hand, disconnected small components and isolated nodes represent portions of the network that are difficult to reach and that will limit the diffusion of information.

The analysis was performed on a OSN dataset, because the collection of a dataset of a DOSN is not easy for its distributed nature. This social graph was limited by selecting only links above a certain level of trust, estimated through the contact frequency between users. Hence, by applying four different thresholds corresponding to the thresholds of the social circles which are a natural classification for human social relationships, we study the connectivity of the resulting graph. The results indicate that for the threshold representing “active social contacts” for the users, the resulting graph is highly connected and contains a large component covering more than 96% of the original network. On the other hand, for more restrictive thresholds, the node coverage drops significantly.

The selected threshold on trust could be problematic due to an excessive reduction of the effective social graph. To overcome this situation, we propose a reintroduction mechanisms of one discarded social contact for each user. We investigate different strategies for selection of this social contact, which includes random methods (i.e. uniform random, proportional random, inverse proportional random) and deterministic methods (i.e. highest frequency, lowest frequency). The performed analysis allowed us to discover a series of properties of DOSN which can be useful to design new peer-to-peer services on top of DOSN communication mechanisms. Specifically, the results of our analysis highlight the following properties:

- The number of nodes connected by a DOSN could be highly influenced by the required minimum level of trust. If only very intimate friends are willing to communicate with each other, information diffusion is severely limited.
- By re-inserting in the network of trusted peers even a single social link with the highest possible trust (under the threshold) for each user is sufficient to reach a higher node coverage. Choosing the link with the highest possible trust is clearly preferable for the user than choosing a link with lower trust.
- Considering that information travels more easily through links with high trust levels, the length of average weighted shortest paths between pairs of nodes in the network is significantly higher than the length of shortest paths in unweighted "small-world" networks. In particular, the assumption that the distance between an two nodes in the network is proportional to the logarithm of the number of nodes in the network is not necessarily true when weighted links are considered. Our results indicate the presence of a backbone of links at a very high level of trust through which information can move to different parts of the network following several links at a very low cost.
- In accordance with the previous result, the best re-insertion policy is the one which selects the link with the highest contact frequency (i.e. trust) since it leads to a good increase in terms of node coverage, comparable to the other policies, with the lowest impact on the cost of the paths.
- It is noteworthy that re-insertion policies cause a limited decrease in terms of trust, and they represent a valid alternative to choosing a lower threshold in terms of trust. The re-insertion policy selecting the links to re-insert with highest contact frequency is always leading to the best performances.

7

Conclusions

The vast adoption of Online Social Networks as communication media allows the collection of vast datasets of users interactions which can be analysed to obtain useful insights about the structure of our the social relationships and the relative social phenomena. In this thesis, the properties of the structure of ego network in online environment is analysed to study the effects in network-scale phenomena like information diffusion.

Through the analysis of two large datasets, one from Facebook and one from Twitter, a comparison of the properties of the structure of the ego network in online environment was presented. The results shows that the ego network in OSNs are remarkably similar among themselves and to those found in offline social networks since the structure of offline and online ego networks are compatible. This suggests that the properties of the ego network are independent from the used media of communication, and are relative to human brain constraints.

We also analysed the differences in the ego-networks considering separately the incoming and the outgoing messages. The results shows that the two networks are fairly different, in fact just 66.52% of the nodes in the "outgoing" ego network are also present in the "incoming" one. Moreover, the later ego-network is significantly bigger than the former. This is explained considering that an incoming message does not necessarily requires the consumption of any cognitive resources of the receiver, and thus less significant than outgoing messages. Moreover we present an index to better evaluate the tie strength, which combine the incoming and the outgoing messages, to give an higher score to relationship which reciprocate the interaction. The ego network produced using this index shows better similarity to offline social network.

Building on the results about ego network structures in OSNs, we have 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. We have applied a standard information diffusion model, namely the independent cascade model, on the network graph obtained from Facebook, that is the most representative amongst the two data set we used in terms of network completeness. We assigned a probability of diffusion to each social link, estimating it from the frequency of contact extracted from the Facebook interaction graph. We also assigned labels indicating the ego network rings to which the links belong, according to a clustering analysis we performed to divide each ego network in different rings. The analysis of the information cascades produced by simulation indicates that all the ego network rings, except the most external one (containing inactive relationships), are important for the diffusion of information, since removing any of them causes a significant drop in terms of node coverage. This result, allowed us to individuate a relatively small number of relationships, the one belonging to the innermost circle, which are just the 0.3% of all the relationships of the network which the removal causes a drastic reduction in the information diffusion. In the literature, social networks have been found to be more resilient to the removal of strong ties than weak ties since weak ties are often bridges representing the only connection between otherwise disconnected parts of the network. Nevertheless, our results indicate that, if we remove all the strongest ties from all the ego networks, the diffusion would be very limited. This means that strong ties are fundamental to transport information within cohesive groups of individuals because of their intrinsic high level of trust. Without them information is not able to circulate in intra-group diffusion process. The third ring (i.e. the affinity group without the elements in the sympathy group), which contains medium strength ties, resulted to be the most used ring during the diffusion. This result confirms previous work in the field [70, 31].

Using a similar model of information diffusion, we have analysed the impact of the characteristic of a node in the information diffusion process. We have analysed the correlation between various centrality measures of the starting node of the information with the properties of the resulting information cascade. Our results indicate that the highest correlations are obtained with statistics that involve the tie strength, namely the activity of the seeds and the eigenvector centrality. The local structure of the social network influence the resulting cascades, if fact the Burt's constraint, a measure of the the number of structural holes in an ego network, has a medium negative correlation indicating that an high clustered ego networks limit the spread of information. Interestingly, executing this correlation analysis using

an unweighted graph, in which each social relationships is not differentiated by the tie strength, it is not possible to predict which nodes are more inclined to produce wider information cascades.

We have analysed the impact of trust amongst users to the information diffusion in an DOSN environment. We have considered four different scenarios in which all the users in the network limit the diffusion of content according to a threshold value. Applying these four thresholds, which were selected as the lowest interaction frequency value of each social circle, the capability of the network to spread informations was greatly reduced. To overcome this problem without lowering the threshold value, we analysed the reintroduction of a single removed social link according different strategies. All the selected strategies produces a significant improvement in both giant component size, and capability to spread information inside the connected network, but the “highest freq” strategy, which selects the relationship with the highest frequency of contact, showed the best performances.

In conclusion, in this thesis, we presented a characterisation of the properties of the ego network in online environments, which we validated with the known properties in offline social networks. Hence, using these local-level results, we analysed various aspects of the information diffusion obtaining interesting insights in this network-scale phenomenon. Our results provide significant insights on the key social reasons behind widely observed phenomena governing information diffusion in OSN. Through this, they can be exploited to predict how information spreads based on structural properties of social networks, and, conversely, how to tune social network properties (if possible) to achieve a given coverage of information spread. Finally, our results also characterise the trade off between trustworthiness of links and information spread.

All in all, therefore, our results shed light on the key social properties governing information diffusion in OSN, and our models can be used to design novel information centric services for OSN, for example by helping finding good points in the network where to inject content that needs to be spread at a certain level, or tuning trust-related parameters based on the sought level of information spread.

Appendices

A

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 Chapter 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 [29]. 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 take 10 random sub-samples of the training set, each of which contains 80% of the entries, keeping the remaining 20% for testing. Then, we apply the same methodology used to create the SVM generated from the entire training set on the 10 sub-samples. Doing so, we obtain different SVMs, trained using different sub-samples of the training set, and of which we are 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 calculate 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 consider only users falling in the "socially relevant users" class, thus it is particularly important to minimise the false positive rate[1]. Minimising the false negative

[1] 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.

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 are 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 [26].

B

Facebook Dataset

In this section we provide details about the procedure 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.

B.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. B.1.

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

The first broad estimation we can do to discover the duration of social ties in the data set is to divide the relationships into different classes C_k, each of which indicates in which interval of time (w_k, w_{k-1}) the relationships contained in it started

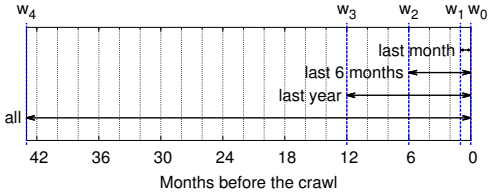


Figure B.1: Temporal windows.

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

B.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 classes $C_{k\in\{2,3,4\}}$ we estimate the time of the first interaction comparing the number of interactions

Table B.1: 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	$(w_4 = 43, w_3 = 12)$	$n_1 \leq n_2 \leq n_3 < n_4$

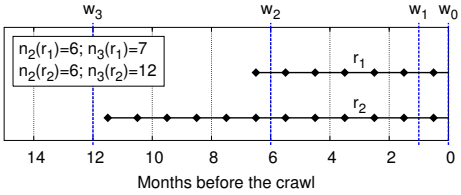


Figure B.2: Graphical representation of two social relationships with different duration.

n_k, made within the smallest temporal window in which the first interaction occurred (w_k, w_0), with the number of interactions (n_{k-1}), made in the previous temporal window in terms of temporal size (w_{k-1}, w_0). If $n_k(r)$ is much greater than $n_{k-1}(r)$, a large number of interactions occurred within the time interval (w_k, w_{k-1}). Assuming that these interactions are distributed in time with a frequency similar to that in the window (w_{k-1}, w_0), the first occurred interaction must be near the beginning of the considered time interval. On the other hand, a little difference between $n_k(r)$ and $n_{k-1}(r)$ indicates that only few interactions occurred in the considered time interval (w_k, w_{k-1}). Thus, assuming an almost constant frequency of interactions, the first contact between the involved users must be at the end of the time interval. The example in Figure B.2 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 calculate for each relationship $r \in C_k$ what we call *social interaction ratio* $h(r)$, defined as:

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

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 $(0, \infty)$[1].

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

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

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

Note that the value of $\hat{d}(r)$ is always in the interval (w_{k-1}, w_k). The greater $h(r)$ - which denotes a lot of interactions in the time window (w_k, w_{k-1}) - the more $\hat{d}(r)$ is close to w_k. The smaller $h(r)$, the more $\hat{d}(r)$ is close to w_{k-1}. Moreover, the shape of the $\hat{d}(r)$ function and the value of a_k are chosen relying on the results about the Facebook growth rate, available in [93]. 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 [9] we provide a detailed description of this step of our analysis.

B.3 Estimation of the Frequency of Contact

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

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

Previous research work demonstrated that the pairwise user interaction decays over time and it has its maximum right after link establishment [90]. Therefore, if we assessed the intimacy level of the social relationships with their contact frequencies, this would cause an overestimation of the intimacy of the youngest relationships. In order to overcome this problem, we multiply the contact frequencies of the relationships in the classes C_1 and C_2 by the scaling factors m_1 and m_2 respectively, which correct the bias introduced by the spike of frequency close

[1] 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.

to the establishment of the link. Assuming that the relationships established more than six months before the time of the crawl are stable, we set m_1 and m_2 comparing the average contact frequency of each of the classes C_1 and C_2, with that for the classes C_3 and C_4. Obtained values of the scaling factors are: $m_1 = 0.18$, $m_2 = 0.82$. Setting $m_3 = 1$ and $m_4 = 1$, scaled frequencies of contact are defined as:

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

C

Ego-Net Digger Application

Ego-Net Digger is a web-based Facebook application able to retrieve and analyse social interaction data between the user and her friends, giving as output the ego network of the user partitioned into its social circles. Ego-net digger is the result of the extensive work we performed on the basis of the know-how acquired during the creation of the prototype application described in [13]. However, ego net-digger is a much more advanced application, built to overcome some limitations of the cited prototype, summarised by the following points:

- The download of social data was only applicable to one user at a time
- The prototype did not support background data download, thus requiring the user to remain connected for the entire download process, otherwise the process would have ended before its completion
- The number of Facebook posts downloaded by the prototype (per each ego) was limited to $400/500$, far less than the complete history accessible from Facebook communication records
- The manual evaluation of Facebook tie strength was long and tedious, not giving an easy method to rank friends and to visually compare them
- People were not incentivised to use the application because of the lack of a reward for their time spent

Ego-net digger is designed to overcome these limitations, allowing a much more refined analysis of the structure of ego networks. Specifically, ego-net digger is able to retrieve the entire communication history between the user and her alters in Facebook, includings posts, comments, likes, tags, events in common, private messages, notes, photos, status updates, video, family relationship and other information related to the profile of the user and her friends. This data clearly contains

more than just the indication of contact between people, and can be leveraged to obtain a model for the estimation of tie strength much more refined than those based only on the frequency of contact between users.

We also provide a new method to collect explicit evaluations of tie strength from the users. This method is based on a simple but effective graphical interface that allows users to evaluate her friends dragging their pictures on a graphical ruler. All the pictures related to previously evaluated friends remain stick on the ruler, so that the user is able to easily compare her friends, producing a much more accurate evaluation compared to [13].

Ego-net digger introduces also the ability to obtain explicit evaluations of tie strength in the real life, in addition to those related to the Facebook world. These evaluations could be extremely useful for further characterisation of the differences between the physical and the real worlds.

In addition to the need to overcome the limitations of the prototype application [13], we developed ego-net digger with the following goals:

- support a wide range of different data analysis techniques, acting as a testbed for OSN analysis algorithms
- minimise the effort needed to update the application to cope with the OSN platform updates
- support an high number of users, without the detriment of user experience
- make the manual evaluation of tie strength as fast and simple as possible to require a limited amount of time even for users with an high number of friends

Ego-Net Digger is composed of two main modules: (i) a data fetcher component and (ii) a web based user interface for tie strength evaluation

Data fetcher module is the server side component of the application. It retrieves all the interaction data related to the user and her friends, according to her privacy settings. This data can be then elaborated using one or more data analysis algorithms. Ego-net digger is designed to be easily extended and modified to be used with other OSN different than Facebook (e.g., Twitter).

The data fetcher module performs the data retrieval and elaboration phases in background, to allow the user to disconnect from the application without blocking the data retrieval procedure. The presence of this background data download procedure also ensures a better distribution of the application workload, since the concurrent downloads can be limited or delayed to avoid network congestion.

The actual implementation of ego-net digger data fetcher module uses Facebook Query Language (FQL) to access the social data of the users. We have spent a lot of effort to make the data fetcher module easily updatable and extendible to

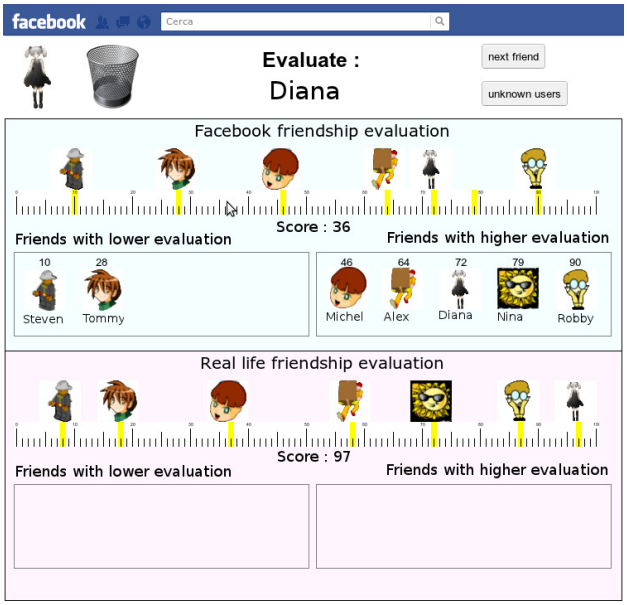


Figure C.1: Screenshot of ego-net digger tie strength evaluation module

cope with Facebook API changes, building a tool for the automatic generation of FQL queries and table relying on the database structure, described on Facebook documentation[1].

Web interface for tie strength evaluation is a module used to collect manual evaluations of tie strength from the user connected to the application. We intend to use this data only during a preliminary phase of our work, to build and tune models for automatic emotional closeness estimation. With these models we would like to study the properties of ego networks and their relation with tie strength without requiring user interaction.

As experienced in [13], for the users is not easy to assign a numeric score to their friends without a graphical comparison between them. Moreover, the duration of the ranking process highly affects the accuracy of the evaluation. Therefore is important to give an easy to use user interface for the evaluations. Using ego-net digger, a user can rank a friend clicking on a graphical “ruler”, graduated from 0 to 100. After doing so, the Facebook picture of the evaluated friend appears on the ruler at the position indicating the given score. In this way the user can

[1] https://developers.facebook.com/docs/reference/fql/

easily compare her friends and change previously assigned scores in case she wants to refine the evaluation. To facilitate the visualisation of previously evaluated friends in case they are too many to be displayed on the ruler, ego-net digger shows the pictures of five people with higher/lower tie strength considering the current position of the mouse cursor (on the ruler) in two separated panels placed beneath the ruler. Figure C.1 depicts a screenshot of the web interface for tie strength evaluation.

For each friend, the user is asked to express two different evaluations: the first one concerning the tie strength she feels with her friends in the “real life” and the second one for the tie strength in Facebook. We collect these different evaluations to analyse the differences between the users’ active networks in real and virtual environments.

From previous work [13] we know that the amount of inactive relationships which receive a tie strength evaluation equal to zero is rather substantial. To speed up the evaluation process without introduce bias in the results, we introduce a button by which the user can declare a friend as a mere acquaintance for both the physical and the cyber worlds.

Since we want to be able to check the goodness of the evaluations, collected by ego-net digger, we introduce some additional information to track the behaviour of the user. Specifically, we collect data regarding the duration of the entire evaluation process of a user and the timestamp related to each single score given. In this way we are able to study the distribution of the speed of the evaluation, performed by the user. This analysis allows us to identify if and how the evaluation changes over time and to detect not enough accurate evaluations in order to remove outliers from the collected data.

References

1. Sibel Adali, Robert Escriva, Mark K. Goldberg, Mykola Hayvanovych, Malik Magdon-Ismail, Boleslaw K. Szymanski, William a. Wallace, and Gregory Williams. Measuring behavioral trust in social networks. In *ISI '10*, pages 150–152, 2010.
2. Eytan Adar and LA Adamic. Tracking information epidemics in blogspace. In *Web Intelligence*, 2005.
3. Sinan Aral and Dylan Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 2011.
4. Valerio Arnaboldi, M. Conti, M. La Gala, A. Passarella, and F. Pezzoni. Information diffusion in osns: the impact of nodes' sociality. In *29th ACM Symposium On Applied Computing (ACM SAC 2014), Gyeongju, S. Korea, March 24 - 28.*, Gyeongju, S. Korea, 03/2014 2014.
5. Valerio Arnaboldi, Marco Conti, Massimiliano La Gala, Andrea Passarella, and Fabio Pezzoni. Information Diffusion in OSNs: the Impact of Nodes' Sociality. In *SAC '14*, pages 1–6, 2014.
6. Valerio Arnaboldi, Marco Conti, Massimiliano La Gala, Andrea Passarella, and Fabio Pezzoni. Ego-network structure in online social networks and its impact on information diffusion. *ACM Transactions on the Web (TWEB)*, 2014 Submitted to.
7. Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Robin I.M. Dunbar. Dynamics of Personal Social Relationships in Online Social Networks: a Study on Twitter. In *COSN '13*, pages 15–26, 2013.
8. Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Fabio Pezzoni. Analysis of Ego Network Structure in Online Social Networks. In *SocialCom '12*, pages 31–40, 2012.
9. Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Fabio Pezzoni. Analysis of Ego Network Structure in Online Social Networks. Technical report, 2012.
10. Valerio Arnaboldi, Andrea Guazzini, and Andrea Passarella. Egocentric Online Social Networks: Analysis of Key Features and Prediction of Tie Strength in Facebook. *Computer Communications*, 36(10-11):1130–1144, 2013.
11. Valerio Arnaboldi, Massimiliano La Gala, Andrea Passarella, and Marco Conti. The role of trusted relationships on content spread in distributed online social networks. In *Euro-Par 2014: Parallel Processing Workshops*, pages 287–298. Springer, 2014.

12. Valerio Arnaboldi, Massimiliano La Gala, Andrea Passarella, and Marco Conti. Information diffusion in distributed osn: the impact of trusted relationships. *Peer-to-Peer Networking and Applications*, 2015 Submitted to.
13. Valerio Arnaboldi, Andrea Passarella, Maurizio Tesconi, and Davide Gazzè. Towards a Characterization of Egocentric Networks in Online Social Networks. In *OTM Workshops*, volume 7046, pages 524–533, 2011.
14. Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four Degrees of Separation. *CoRR*, abs/1111.4, 2011.
15. E Bakshy, Brian Karrer, and LA Adamic. Social influence and the diffusion of user-created content. *Human Factor*, pages 325–334, 2009.
16. Eytan Bakshy, Jake M Hofman, Duncan J Watts, and Winter A Mason. Everyone's an Influencer: Quantifying Influence on Twitter. In *WSDM '11*, pages 65–74, 2011.
17. Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The Role of Social Networks in Information Diffusion. In *WWW '12*, pages 519–528, 2012.
18. Béla Bollobás. The Evolution of Random Graphs. *Transactions of the American Mathematical Society*, 286(1):257, 1984.
19. Sonja Buchegger. Delay-Tolerant Social Networking. In *Extreme Workshop on Communication*, pages 1–2, 2009.
20. Sonja Buchegger, Doris Schioberg, Le Hung Vu, and Anwitaman Datta. PeerSoN: P2P Social Networking - Early Experiences and Insights. In *SocialNets '09*, pages 46–52, 2009.
21. Ronald S Burt. *Structural Holes: The Social Structure of Competition*. 1992.
22. Damon Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–7, September 2010.
23. Meeyoung Cha and Fabrício Benevenuto. The world of connections and information flow in twitter. *Transactions on Systems, Man, and Cybernetics*, 42(4):991–998, 2012.
24. Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. *WWW*, page 721, 2009.
25. Justin Cheng, L Adamic, PA Dow, Jon Kleinberg, and Jure Leskovec. Can cascades be predicted? *WWW '14*, 2014.
26. Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is Tweeting on Twitter: Human, Bot, or Cyborg? In *ACSAC '10*, pages 21–30, 2010.
27. Marco Conti, Sajal Das, Chatschik Bisdikian, Mohan Kumar, Lionel M. Ni, Andrea Passarella, George Roussos, Gerhard Tröster, Gene Tsudik, and Franco Zambonelli. Looking Ahead in Pervasive Computing: Challenges and Opportunities in the Era of Cyber-Physical Convergence. *Pervasive and Mobile Computing*, 8(1):2–21, 2012.
28. Marco Conti, Andrea Passarella, and Fabio Pezzoni. A model to represent human social relationships in social network graphs. In *Social Informatics*, 2012.
29. Corrina Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.
30. Leucio A. Cutillo, Refik Molva, and Thorsten Strufe. Safebook: A Privacy-Preserving Online Social Network Leveraging on Real-Life Trust. *Communications Magazine, IEEE*, 47(12):94–101, 2009.
31. Peter S. Dodds, Roby Muhamad, and Duncan J. Watts. An Experimental Study of Search in Global Social Networks. *Science*, 301(5634):827–9, 2003.
32. Robin I M Dunbar. No Title. Private communication, June 2012.

33. Robin I.M. Dunbar. The Social Brain Hypothesis. *Evolutionary Anthropology*, 6(5):178–190, 1998.
34. Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. Outtweeting the Twitterers - Predicting Information Cascades in Microblogs. In *WOSN '10*, pages 3–3, 2010.
35. Eric Gilbert. Predicting Tie Strength in a New Medium. In *CSCW '12*, pages 1047–1056, 2012.
36. Eric Gilbert and Karrie Karahalios. Predicting Tie Strength with Social Media. In *CHI '09*, pages 211–220, 2009.
37. Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters*, 12(3):211–223, 2001.
38. Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring Networks of Diffusion and Influence. In *KDD '10*, pages 1019–1'28, 2010.
39. Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number. *PloS one*, 6(8):e22656, 2011.
40. Neil Z. Gong, Wenchang Xu, Ling Huang, Prateek Mittal, Emil Stefanov, Vyas Sekar, and Dawn Song. Evolution of Social-Attribute Networks: Measurements, Modeling, and Implications using Google+. In *IMC '12*, pages 131–144, 2012.
41. Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. Learning Influence Probabilities in Social Networks. In *WSDM '10*, page 241, 2010.
42. Mark Granovetter. *Getting a job: A study of contacts and careers*. University of Chicago Press, 1995.
43. Mark S. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.
44. Mark S. Granovetter. Threshold Models of Collective Behavior. *The American Journal of Sociology*, 83(6):1420–1443, 1978.
45. Daniel Gruhl, Ramanathan Guha, David Liben-Nowell, and Andrew Tomkins. Information Diffusion Through Blogspace. In *WWW '04*, pages 491–501, 2004.
46. Barbara Guidi, Marco Conti, and Laura Ricci. P2P architectures for distributed online social networks. In *International Conference on High Performance Computing & Simulation (HPCS)*, pages 678–681, 2013.
47. Adrien Guille, Hakim Hacid, Cécile Favre, and Djamel. Zighed. Information diffusion in online social networks: a survey. *ACM SIGMOD Record*, 42(2):17–28, 2013.
48. Lu Han, Badri Nath, Liviu Iftode, and S. Muthukrishnan. Social Butterfly: Social Caches for Distributed Social Networks. In *SocialCom '11*, pages 81–86, October 2011.
49. Russel A. Hill and Robin I.M. Dunbar. Social network size in humans. *Human Nature*, 14(1):53–72, 2003.
50. Shawndra Hill, Foster Provost, and Chris Volinsky. Network-Based Marketing: Identifying Likely Adopters via Consumer Networks. *Statistical Science*, 21(2):256–276, 2006.
51. Muhammad U. Ilyas, Muhammad Z. Shafiq, Alex X. Liu, and Hayder Radha. A Distributed and Privacy Preserving Algorithm for Identifying Information Hubs in Social Networks. In *INFOCOM '11*, pages 561–565, 2011.
52. Jason J. Jones, Jaime E. Settle, Robert M. Bond, Christopher J. Fariss, Cameron Marlow, and James H. Fowler. Inferring Tie Strength from Online Directed Behavior. *PLoS ONE*, 8(1):e52168, 2013.

53. Indika Kahanda and Jennifer Neville. Using Transactional Information to Predict Link Strength in Online Social Networks. In *ICWSM '09*, pages 74–81, 2009.
54. R Kanai, B Bahrami, R Roylance, and G Rees. Online social network size is reflected in human brain structure. *Biological sciences / The Royal Society*, 279(1732):1327–34, 2012.
55. David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the Spread of Influence Through a Social Network. In *KDD '03*, pages 137–146, 2003.
56. Jon Kleinberg. The Small-World Phenomenon: An Algorithmic Perspective *. In *STOC '00*, pages 163–170, 2000.
57. Hans P. Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-Based Clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240, 2011.
58. Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a Social Network or a News Media? In *WWW '10*, pages 591–600, 2010.
59. Massimiliano La Gala, Valerio Arnaboldi, Marco Conti, and Andrea Passarella. Ego-net digger: a new way to study ego networks in online social networks. In *First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research (ACM HotSocial 2012)*, 2012.
60. Jaron Lanier. *Who Owns the Future?* 2013.
61. Kristina Lerman and Rumi Ghosh. Information Contagion: An Empirical Study of the Spread of News on Digger and Twitter Social Networks. In *ICWSM '10*, pages 90–97, 2010.
62. Jure Leskovec and Eric Horvitz. Planetary-Scale Views on an Instant-Messaging Network. Technical report, 2007.
63. David Liben-Nowell and Jon Kleinberg. Tracing Information Flow on a Global Scale Using Internet Chain-Letter Data. *Academy of Sciences*, 105(12), 2008.
64. Peter V. Marsden and Karen E. Campbell. Measuring Tie Strength. *Social Forces*, 63(2):482–501, 1984.
65. Giovanna Miritello, Rubén Lara, Manuel Cebrian, and Esteban Moro. Limited communication capacity unveils strategies for human interaction. *Scientific Reports*, 3:1–7, June 2013.
66. Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07*, volume 40, page 29, 2007.
67. SA Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. *SIGKDD*, 2012.
68. Mark E.J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256, 2003.
69. Mark E.J. Newman and Juyong Park. Why Social Networks are Different From Other Types of Networks. *Physical Review E*, 68(3), 2003.
70. Jukka P. Onnela, Jari Saramäki, Jorkki Hyvönen, Gyorgy Szabó, David Lazer, Kimmo Kaski, Jànos Kertész, and Albert L. Barabási. Structure and Tie Strengths in Mobile Communication Networks. *PNAS*, 104(18):7332–7336, 2007.
71. Andrea Passarella and Marco Conti. Characterising Aggregate Inter-Contact Times in Heterogeneous Opportunistic Networks. In *Networking '11*, pages 1–12, 2011.
72. Andrea Passarella, Marco Conti, Robin I.M. Dunbar, and Chiara Boldrini. Modelling Inter-contact Times in Social Pervasive Networks. In *WSWiM '11*, pages 333–340, 2011.

73. Thomas Paul, Antonino Famulari, and Thorsten Strufe. A survey on decentralized Online Social Networks. *Computer Networks*, 75:437–452, 2014.
74. Thomas Paul, Antonino Famulari, and Thorsten Strufe. A survey on decentralized online social networks. *Computer Networks*, 75, Part A(0):437 – 452, 2014.
75. S Petrovic, Miles Osborne, and Victor Lavrenko. RT to Win! Predicting Message Propagation in Twitter. In *ICWSM*, 2011.
76. Fabio Pezzoni, Jisun An, Andrea Passarella, Jon Crowcroft, and Marco Conti. Why Do I Retweet It? An Information Propagation Model for Microblogs. In *SocInfo '13*, pages 360–369, 2013.
77. Sam G.B. Roberts. Constraints on Social Networks. In *Social Brain, Distributed Mind (Proceedings of the British Academy)*, pages 115–134. 2010.
78. Sam G.B. Roberts, Robin I.M. Dunbar, Thomas V. Pollet, and Toon Kuppens. Exploring Variation in Active Network Size: Constraints and Ego Characteristics. *Social Networks*, 31(2):138–146, 2009.
79. Jari Saramaki, Elizabeth A. Leicht, Eduardo Lopez, Sam G.B. Roberts, Felix Reed-Tsochas, and Robin I.M. Dunbar. The persistence of social signatures in human communication. *PNAS*, 111(3):942–947, 2014.
80. Axel Schulz, Petar Ristoski, and Heiko Paulheim. I see a car crash: Real-time detection of small scale incidents in microblogs. In *ESWC '13*, volume 7955, pages 22–23, 2013.
81. Charles W. Shmidt. Using social media to predict and track deseases outbreaks. *Environmental Health Perspectives*, 120(1):30–33, 2012.
82. Herbert Simon. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1):99–118, 1955.
83. Eric Sun, Itamar Rosenn, C Marlow, and TM Lento. Gesundheit! Modeling Contagion through Facebook News Feed. In *ICWSM*, number 2000, 2009.
84. Alistair Sutcliffe. Social Relationships and the Emergence of Social Networks. *Journal of Artificial Societies and Social Simulation*, 15(4), 2012.
85. Alistair Sutcliffe, Robin I.M. Dunbar, Jens Binder, and Holly Arrow. Relationships and the Social Brain: Integrating Psychological and Evolutionary Perspectives. *British Journal of Psychology*, 103(2):149–68, 2012.
86. Io Taxidou and Peter M Fischer. Online Analysis of Information Diffusion in Twitter. In *WWW Companion '14*, pages 1313–1318, 2014.
87. Io Taxidou and Peter M Fischer. RApID: A System for Real-time Analysis of Information Diffusion in Twitter. In *CIKM '14*, pages 2060–2062, 2014.
88. Jeffrey Travers and Stanley Milgram. An Experimental Study of the Small World Problem. *Sociometry*, 32(4):425, 1969.
89. Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The Anatomy of the Facebook Social Graph. *CoRR*, 2011.
90. Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. On the Evolution of User Interaction in Facebook. In *WOSN '09*, pages 37–42, 2009.
91. Haizhou Wang and Mingzhou Song. Clustering in One Dimension by Dynamic Programming. *The R Journal*, 3(2):29–33, 2011.
92. Duncan J. Watts and Steven H. Strogatz. Collective Dynamics of "Small-World" Networks. *Nature*, 393(6684):440–2, 1998.
93. Christo Wilson, Alessandra Sala, Krishna P.N. Puttaswamy, and Ben Y. Zhao. Beyond Social Graphs: User Interactions in Online Social Networks and Their Implications. *ACM Transactions on the Web*, 6(4):1–31, 2012.

94. Jichang Zhao, Junjie Wu, Xu Feng, Hui Xiong, and Ke Xu. Information propagation in online social networks: a tie-strength perspective. *Knowledge and Information Systems*, 32(3):589–608, November 2011.
95. Xiaohan Zhao, Alessandra Sala, Christo Wilson, Xiao Wang, Sabrina Gaito, Haitao Zheng, and Ben Y. Zhao. Multi-scale dynamics in a massive online social network. In *IMC '12*, pages 171–184, 2012.
96. Wei X. Zhou, Dider Sornette, Russell A. Hill, and Robin I.M. Dunbar. Discrete Hierarchical Organization of Social Group Sizes. *Biological Sciences*, 272(1561):439–44, 2005.