Social Network Data and Practices: the case of Friendfeed

Fabio Celli[1], F. Marta L. Di Lascio[2], matteo magnani[3], Barbara Pacelli[4], and Luca Rossi[5]

[1] Language Interaction and Computation Lab, University of Trento, fabio.celli@email.unitn.it
[2] Dept. of Statistical Science, University of Bologna, francesca.dilascio@unibo.it
[3] Dept. of Computer Science, University of Bologna, magnanim@cs.unibo.it
[4] Independent researcher, bpacelli79@gmail.com
[5] Dept. of Communication Studies, University of Urbino, luca.rossi@uniurb.it

Abstract. Due to their large worldwide adoption, Social Network Sites (SNSs) have been widely used in many global events as an important source to spread news and information. While the *searchability* and *persistence* of this information make it ideal for sociological research, a quantitative approach is still challenging because of the size and complexity of the data. In this paper we provide a first analysis of Friendfeed, a well-known and feature-rich SNS.

1 Research framework

Social Network Sites (SNSs) are undoubtedly one of the most interesting phenomena that bring together new technologies and social practices. They are going through an incredibly fast growth all over the world despite the fact many obstacles like the digital divide still exist. Despite this global success it would be hard to define a single global leader of the SNSs. Facebook, which counts more than 300 million single users mostly clustered in Europe and in the US, is surely a big player but QQ, with a high concentration of users in China, has an even larger user base. It seems that cultural diversity and local identity lead toward the choice of a specific SNS, while the shift toward the adoption of a SNS-model for online interpersonal communications seems to be global [1].

Due to this large worldwide adoption, SNSs have been widely used in many global events as an important source to spread news and information. From the terroristic attack in Mumbai in 2008 to the so-called Twitter revolution in Iran in 2009 SNSs proved several times to be a reliable way to communicate and to spread information in a quick and relatively efficient way. Within this scenario the sociological analysis of SNS based communication is still largely based on a qualitative ethnographic approach aimed at investigating living practices and uses of the SNS [2,3,4]. This approach gave us the opportunity to gain an effective insight in SNS users' lives, motivations and communicative strategies but failed in giving us a general description of how SNSs work and deal, as complex entities, with the diffusion of information.

Aim of this paper is to move a first step into a new direction of sociological SNS research coping with the topic from a multidisciplinary perspective. Within this proposed approach SNSs could be defined, at the same time, as the best and the worst place for sociological research. They can be defined as an optimal place because of the new and emerging properties that communication shows in these contexts. Information in SNSs, as boyd highlighted [5] can be defined also by *searchability* and *persistence* which are two positive characteristics for any researcher. Data can be searched and retrieved easily. At the same time the large amount of data that is published online every second can easily discourage any attempt to investigate online phenomena from a quantitative point of view. Data are out there, they can be searched and retrieved but that is still a great challenge. This paper will present some preliminary results of a larger research project that accepted this challenge and is dealing with a large quantity of SNS data in order to obtain a wider understanding of many unsolved issues in SNS research.

The #SIGSNA project, which stands for Special Interest Group on Social Network Analysis, started his research by analyzing a well known microblogging and social network service called Frienfeed (http://friendfeed.com). Friendfeed has been chosen because of several technical and sociological aspects. From a technical point of view, that will be described in the next section, Friendfeed offers a great level of access to the contents that are produced by the users. Everything that has not been marked as private is available online in RSS format. From a sociological point of view Friendfeed offers a very complex social dynamic constructed on a microblogging service (like the well-known Twitter) with the opportunity to comment the entries of other users. This very simple characteristic makes Friendfeed a microblogging platform able to host huge and complex conversations (made through comments). This is something very similar to what happens in Facebook, where conversation can arise as a sequence of comments to a specific status update.

These two socio-technical characteristics make Friendfeed a perfect starting point for the goals of the #SIGSNA project. #SIGSNA project is developed by a multidisciplinary research team that brings together social scientists, computer scientists, statistical scientists and computational linguisticians.

2 The Social Data Set: extraction and structure

Data has been extracted from the Friendfeed application by monitoring the public URL http://friendfeed.com/public, where the system publishes a sample of recent posts. In the following, we will indicate with *post* any text entry or comment posted by a user, with *entry* a new conversation started by a user, and with *comment* a comment to an entry.

The URL was monitored for two weeks, from September 6, 2009, 00:00 AM to September 19, 2009, 24:00 PM, at a rate of about 1 to 2 updates every second (depending on network traffic). During the monitoring phase, all the identifiers of entries appeared on the public page have been saved on our local servers,

for later retrieval. It is worth noticing that entries with many comments had a higher probability of being catched by our monitoring system. After the end of this phase, we collected all the distinct entry identifiers (9.317.499) and retrieved the corresponding XML representations using the Friendfeed API. At this point, the data has been exported to CSV files, to remove unnecessary XML formatting and for database import.

Then, at the end of the monitoring period, we have computed the network of users and followers. Starting from one random user, and retrieving all the connected graph of followers, we have extracted a related data set of more than 400.000 users, with about 15 million subscription relationships. The structure of the corresponding dataset is the following:

```
Entry(PostID, PostedBy, Timestamp, Text, Language)
Comment(PostID, EntryRef, PostedBy, Timestamp, Text, Language)
Like(User, EntryRef, Timestamp)
User(ID, Type, Name, Description)
Network(Follower, Followed)
```

All field names should be self-understandable. The only generated field is `Language`, which currently contains the most probable language identifier, e.g., `it` for Italian, etc. The final total number of posts is 10.454.195, considering both entries and comments and without private entries which could not be retrieved, for an amount of more than 2GB textual data, and 512.339 likes. In addition, this data contains a small percentage of entries and comments posted before and after the monitoring period. This happens because we have collected all the data related to entries posted during the two weeks even if their publishing date lied outside the time shift. These posts have not been used for the statistical analysis presented in the remaining of the paper.

Despite its apparent relational structure, the dataset under analysis contains a mixture of **structured**, **semi-structured** and **unstructured** data, requiring a complex data model [6]. In particular, several **data graphs** can be identified. If we consider the relationships between users, they induce a directed, labeled, weighted graph where nodes represent users, labels the kind of interaction (subscription/like/comment), weights the strength of the interaction, e.g., the number of comments, with additional text annotations. Considering different labels, we can extract sub-graphs about **active** relationships between users (comments and likes), **passive** relationships (subscriptions) and even **implicit** relationships not directly expressed in the data. For example, when Annie subscribes to John, who subscribed to Susan, Annie may see part of the content of Susan through John's feed, without having a direct subscription to her. Finally, part of the data constitutes *conversation graphs* not directly involving users, but their posting activity (entries and comments), and also in this case we can have implicit arcs[6].

[6] For instance, the `@` symbol followed by a user nickname is used inside text comments to indicate the recipient of the message.

In this last case, nodes of the graph represent short pieces of text (posts). Therefore, in addition to semi-structured data, social databases contain a large amount of **unstructured content** (text and other media) attached to different entities (users and posts).

As a consequence of the complexity of the data model, querying the Social Data Set under examination requires **recursive graph traversal** operators, **text extraction**, with Information Retrieval capabilities to evaluate the **relevance** of single text items or groups of inter-connected items, counting and other **aggregate operators**, both on nodes and on the amount and strength of arcs, and also primitive **data analysis** operators, as typical queries on Social data are often exploratory.

3 Statistical Analysis

Statistical analysis of the collected data aimed at offering a comprehensive description of the whole network and some deeper investigation of how a culturally defined part of the network (the Italian speaking sub-network) works[7]. A general overview of the Friendfeed social network is indicated in Table 1.

Table 1. Min, max, mean and standard deviation values of variables observed in the whole network

	UsrFing	UsrFed	Post	Entry	Com	Like	ComR	LikeR
min	0	0	1	0	0	0	0	0
max	3,072	412	438	316	225	2,583	422	1,235
mean	2.73	3.09	28.01	27.17	0.85	1.93	0.73	1.39
sd	21.08	17.79	46.48	44.96	4.98	22.30	5.64	15.97

A preliminary introduction to the labels is required: UsrFing is the number of users that follow the user, UsrFed is the number of users followed by the user, Com is the number of comments made by the user, ComR is the number of comments received by the user, LikeR is the number of likes[8] received by the user, Like is the number of likes made by the user, Entry is the number of entries wrote by the user. Post is a derived element and it is the number of Comments and Entries made by the user.

[7] Please note that in this paper we consider only users with a number of entries less or equal than 316 due to outliers. After a qualitative analysis most of the users with more than 316 entries turned out to be automated Spam bots, therefore they have not been considered as part of the social dimension of the network.

[8] A *like* is a simple way to communicate some kind of appreciation toward an entry of another user. Instead of commenting by writing something a user can simply express his level of agreement with the published sentence by hitting the *like* button below the entry. This system is not unique to Friendfeed and it can be found also in Facebook.

Table 1 shows a lively network with an average of posts equal to 28.01, which means, in the two weeks of our sampling, more than two posts (entry or comment) every day. Despite this high rate of posting, Entries are much more than comments (mean 27.17 vs 0.85) showing a large amount of users that speak alone.

Table 2. Correlation matrix of variables observed in the whole social network

	UsrFing	UsrFed	Post	Com	ComR	LikeR	Like	Entry
UsrFing	1.00	0.67	0.34	0.49	0.84	0.66	0.27	0.30
UsrFed	0.67	1.00	0.46	0.86	0.59	0.47	0.52	0.38
Post	0.34	0.46	1.00	0.35	0.28	0.24	0.22	0.99
Com	0.49	0.86	0.35	1.00	0.56	0.42	0.55	0.25
ComR	0.84	0.59	0.28	0.56	1.00	0.78	0.31	0.23
LikeR	0.66	0.47	0.24	0.42	0.78	1.00	0.41	0.20
Like	0.27	0.52	0.22	0.55	0.31	0.41	1.00	0.16
Entry	0.30	0.38	0.99	0.25	0.23	0.20	0.16	1.00

The correlation matrix represented in Table 2 can be used to have a more detailed picture of the whole network population and its habits. From the analysis of these values, it is possible to point out some not so obvious results: the correlation between entries done (Entry) and comments done (Com) is quite low as well as the correlation between comments done and comments received. These two data could suggest on one hand a distinction between the posting activity and the commenting activity and, on the other hand, the lack of reciprocity between the comments done and received.

On a more general level, we can observe in Figure 1(a) that the production of contents in the Friendfeed network has an heavy right tailed distribution that is well known in many web based services. This confirms the well-known fact that a small part of the users is very active and responsible for the largest part of the produced content. At the same time the largest part of user base contributes with very few entries. The last descriptive picture (Figure 1(b)) shows a bubble-plot of the entries related to the comments done. The size of the bubble is related to how spread is that specific cross through the network.

Figure 1(b) allows us to point out a few interesting aspects:

- The largest part of the entries does not have any comment and the largest part of the entries with no comments is made by low-activity users (with less than 29 entries in the sample)[9].
- There is a small but still significant part of users that produces only comments and no entries within the time frame of the sample.

Finally, Figure 1(c) shows the cluster of the network of *comment* relationships for a 10.000 user subset of the Italian dataset.

[9] Notice that the class intervals for the two variables compared have been chosen on the basis of the sample quantiles of their distribution function.

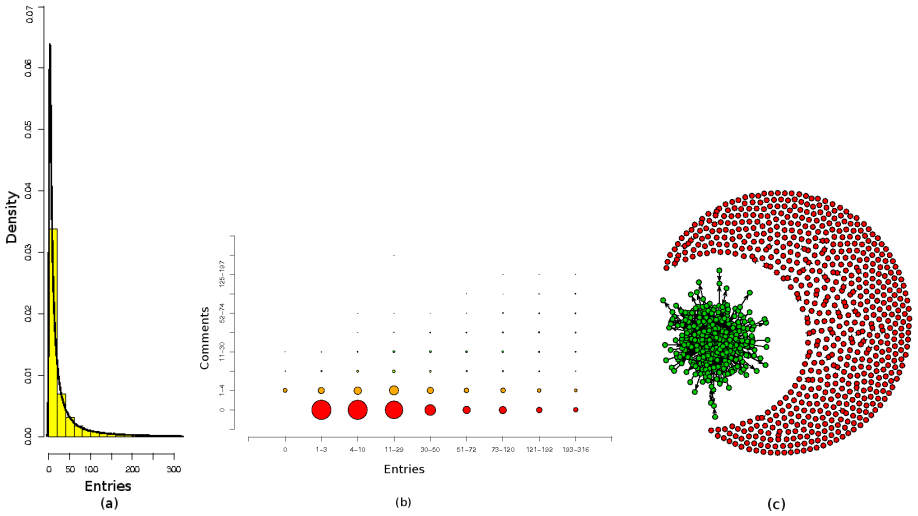

Fig. 1. (a) Density histogram with kernel density estimates of entries in the whole network, (b) relationship between number of posted entries and number of comments done, and (c) graphical representation of comments between Italian users

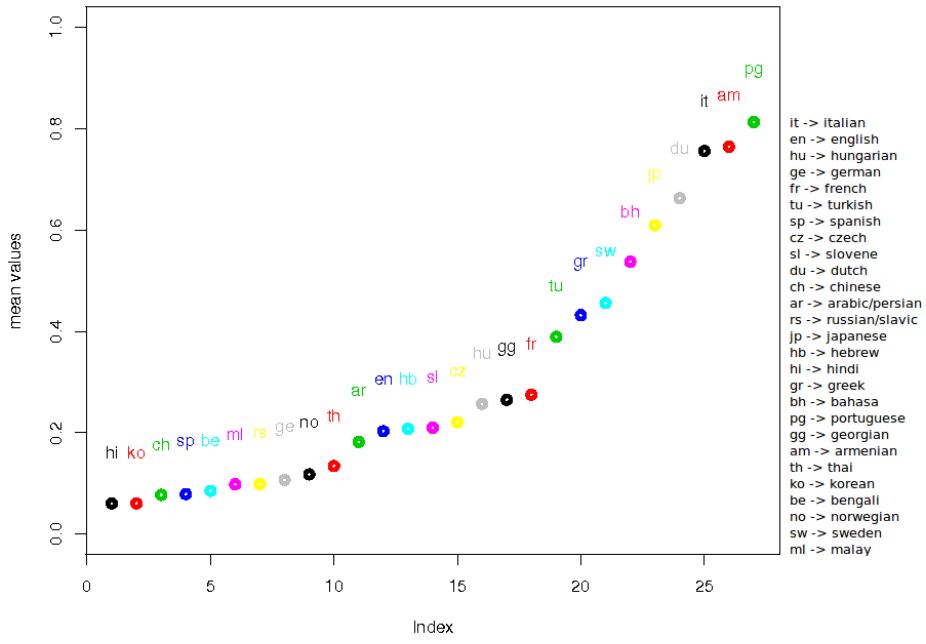

Fig. 2. Mean values of Number of Posts in each language divided by the total Number of Posts of each user with an entry in that language

As stated before, a specific goal of the #SIGSNA project is to start investigating SNSs by comparing how they are used in different cultural contexts. A preliminary analysis has been done using the language as a way to identify specific cultural contexts. The language identification has been obtained using specifically developed software called SLIde (Simple Language Identifier in Perl) [7], which enabled the creation of language annotations without requiring inefficient connections to Web-based services. Language identification appears to be a suboptimal strategy to identify the cultural context of the users. Many users can use several languages to address to different audiences. This will be surely true especially for languages, like English, widely spoken worldwide.

In order to be able to describe the limitations of the chosen method we have calculated the language fidelity level of every language in the sample. The language fidelity level (Figure 2) shows the average level of posting in different languages for every user that posted, at least once, in a specific language. This level, obtained by calculating the mean values of the number of posts in each language divided by the total number of posts of each user with an entry in that language, allows us to establish the level of average fidelity toward a language. As shown in Figure 2 several languages show a high level of fidelity (close to mean value 1.0) that means that users that write in that language usually keep writing in the same language all the time.

The language fidelity index has a double value for the #SIGSNA project: on one hand it allows us to identify the languages that can be used as good indicators of specific cultural contexts and, on the other hand, it suggests the existence of specific nation-wide sub-networks that are loosely connected with the larger Friendfeed network.

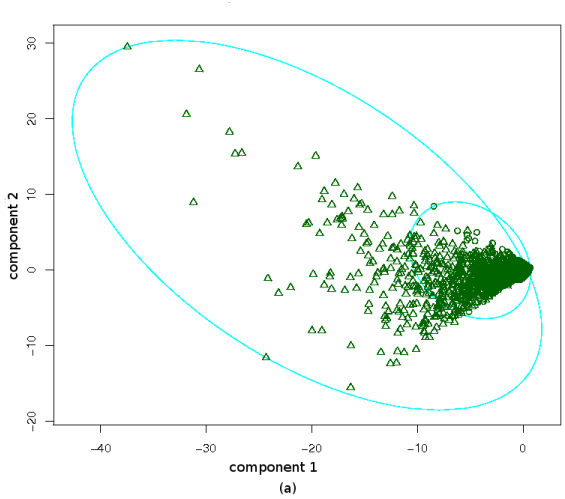

Fig. 3. Cluster plot of users with Italian posts

The Italian sub-network has been further investigated with a cluster analysis of the users. Cluster analysis has been performed by using the *clara* algorithm: a partitioning method for finding clusters [8] into very large datasets. The number of clusters k has been chosen on the basis of the overall silhouette width [8] by varying k from 2 to 100. The average silhouette width is useful for evaluating the goodness of both the obtained clustering and the selected number of clusters, and it turned out that for our dataset the best number of clusters was two.

The cluster average widths are 0.85 and 0.13 for the two founded clusters, respectively whereas the average width of the clustering is 0.59 indicating that the founded clustering is quite appropriate. By Table 3 we note that the first cluster looks like the sample of users weakly active in the network whereas the second one looks like the sample of most active users. Notice that the biggest cluster is the first one. The graphical representation illustrated in Figure 3 is based on the work of Rousseeuw (1990).

Table 3. Mean and standard deviation values of the variables observed into clusters 1 and 2

	UsrFing	UsrFed	Com	ComR	LikeR	Like	Entry
mean 1	2.30	2.90	1.28	0.96	1.06	1.73	22.41
sd 1	9.50	9.22	4.28	4.05	5.73	8.51	18.96
mean 2	29.41	29.69	6.88	7.00	12.41	12.36	149.59
sd 2	93.35	63.84	17.97	22.90	52.09	45.37	65.57

Cluster analysis of Italian users gave us the opportunity to point out the existence of two different groups of users within the Friendfeed social network. A larger loosely connected and weakly active group coexists with a smaller heavy engaged group. This suggests, as we are going to discuss further in the sociological analysis, a wide range of uses of the Friendfeed social network.

4 Sociological Analysis

The statistical analysis of the Friendfeed social network depicts an interesting scenario. The descriptive analysis suggests a lively social network with a high level of production of content. Even if the overall level is quite high there are large differences between users' level of participation in the process. This suggests a highly personal use of Friendfeed according to many individual needs. Heavy users can post new updates continuously while light users could post a message once in a while.

An interesting aspect is the use, shown by the descriptive analysis, of Friendfeed only as a conversational space. Several users, in fact, didn't use Friendfeed to actually post something about themselves but just to comment someone else's posts. This suggests that Friendfeed can be used in a different way from how Twitter is used — Twitter is probably the most famous microblogging site. While

in Twitter every conversation is a connected sequence of micro-posts, in Friendfeed conversations may take place in a dedicated comment space. This makes Friendfeed conversations much more similar to what happens on Facebook than to what takes place on Twitter.

Another aspect that makes the conversational practices of Friendfeed similar to Facebook conversations is the identity of the audience. In Friendfeed, as well as in Facebook, due to specific architectural choices the potential audience of a user's comments is larger than the set of user's followers. By commenting someone else's update in Friendfeed (as well as in Facebook) a user moves herself into a semi-unknown place populated by semi-unknown users composed by all her friends and her friends' friends. Within this perspective Friendfeed can be considered both a microblogging service that allows you to share short thoughts and information with a network of friends and, at the same time, a purely conversational space where you can chat or discuss in a semi-protected environment mainly composed of your friends and their contacts.

In addition to a descriptive analysis of communicative practices taking place on Friendfeed we moved a first step toward a series of comparative analysis of SNS use in different cultural contexts. The analysis of the used language and the cluster analysis of the Italian posts gave us the opportunity to suggest some preliminary considerations. The Italian network of Friendfeed users seems to address his communication mainly to a local/national audience. The high level of language fidelity that has been observed suggests this. The average Italian Friendfeed user writes mainly in Italian and this could suggest the nationality of his *perceived* audience. The network here seems to be very closed on a Geo-cultural basis and connections with different international networks seem to be very rare.

The cluster analysis (Figure 3 and Table 3) shows the existence of two groups of users within the Italian Friendfeed network. A small highly dedicated users group coexists with a larger but less active group. This suggests the high level of flexibility that Friendfeed allows. A double use of the medium is confirmed and the difference seems to be mainly based on the amount of communication that users produce. **Weak users** (group 1 in Table 3) are characterized by a couple of entries every day with few comments. This pattern of use seems to be the most classical use of microblogging saw as a way to update your online status and to share short thoughts with your friends. The **heavy users** group shows a completely different scenario. The level of average daily entries rises up to more than ten entries per day with a large level of comments done and received and even a greater level of likes done or received. These data suggest what we could define as a continuous stream of sharing that resembles the idea of life-stream. This comes together with a high level of conversational use of the service itself suggested by the high level of comments and likes and by the reciprocity of these. For this group of users Friendfeed seems to be a perfect platform which is able to host fruitful conversations starting from the sharing of a large quantity of information. Obviously, as we are going to discuss further in the conclusions,

a qualitative analysis of the entries produces by the two groups is required in order to better understand the differences pointed out by the cluster analysis.

5 Conclusions and research perspectives

This paper presented the first results of the #SIGSNA research project. Aim of the #SIGSNA project is to develop a comprehensive analysis of social interactions that take place in the Friendfeed SNS. In this paper we showed a first descriptive analysis of the whole network that pointed out the existence of a large variety of uses inside the SNS. In addition to that, we used a language identification system to make comparative analysis of SNS uses in different cultural contexts. As a preliminary investigation we have also performed a cluster analysis of the Italian Friendfeed network, with which we have identified the existence of two different clusters of users: weak users and highly dedicated users.

The #SIGSNA project is characterized by the large database of entries that has been collected during the sampling period. Due to the large dimension of the database and to the high quality of the collected data the presented results have to be considered just as a first bite of the whole research that is still in progress. At the same time, our analysis is highlighting some computational limitations of traditional social data analysis tools, which cannot deal with the large amount of information produced by SNSs — in particular, traditional and text clustering algorithms implemented into widely used statistical tools could not be applied to the whole network, which presents hundreds of thousand users, millions of arcs, and millions of text posts. These limitations will drive the development of scalable techniques for the analysis of large and complex networks, which are necessary to deal with the size of current real social datasets.

References

1. Cosenza, V.: Osservatorio facebook http://www.vincos.it/osservatorio-facebook, retrieved on August 31, 2009.
2. danah boyd, Ellison, N.: Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication **13**(1) (2007)
3. Siiback, A.: Online peer culture and interpretative reproduction on children's social networking profiles, the good the bad the challenging. In: COST conference proceedings. (2009)
4. Hardey, M.: ICT and generations constantly connected social lives, the good the bad the challenging. In: COST conference proceedings. (2009)
5. danah boyd: Taken Out of Context: American Teen Sociality in Networked Publics. PhD thesis, University of California-Berkeley, School of Information (2008)
6. magnani, m., Montesi, D.: A unified approach to structured and XML data modeling and manipulation. Data & Knowledge Engineering **59**(1) (2006)
7. Celli, F.: Slide: Simple language identifier in perl. Technical report, Language Interaction and Computation Lab, University of Trento (2009) http://clic.cimec.unitn.it/fabio/.
8. Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)