Goals of the research:
In recent years we witnessed the growing awareness of the fact that human communications and social interactions are built on a stratified structure. Today, a variety of techno-communication channels - including online social networks, mobile phone calls, short messages (SMS) and e-mails - provide an intricate bundle of interactions that is overlaid on real life relationships enabled by individuals' spatial proximity. Multidimensional or multiplex networks adequately describe and formalize this complex scenario. In this emerging field, the specific topics of my research are the following:

- **Multidimensional networks from mobile phone data**: Mobile phone data (CDRs) gathered by the billing system of mobile phone operators represent one of the most interesting subject of computational social science. In fact they capture the interactions expressed by voice calls, text messages (SMS) and the user's position when these events occur. Many studies on calls and mobility of mobile phone users have validated at a large-scale previous results relying on a small numbers of volunteers. Nonetheless, most of these studies limit their analysis to a single dimension, while an all-around vision is still missing. The understanding of the interplay among the different dimensions characterizing mobile phone datasets, i.e. calls, SMSs and mobility, is the primary goal on this topic, and the analysis of how people maintain their relationships through diverse communication media represents the first step. To this aim we focus on the following research questions, whose answers require a multidimensional analysis of the mobile phone networks:
  a) Do people indistinctly use voice calls and SMSs to establish and keep strong and weak ties, or do they differentiate the communication media according to the type of the relationship?
  b) Do groups who exclusively communicate by one communication medium exist inside users' ego-networks?
  c) What are the relations among the communities detected on the different mobile phone layers? Do they coincide or are they totally different?

The above goals may be further extended whether we include into our analysis the spatio-temporal dimension given by the spatial proximity. Specifically, in mobile phone data, we denote the spatial proximity as the co-presence under the same cellular tower and we measure it by the co-location rate. By comparing the co-location network with the communication networks, we ask whether only 'virtual' communities, which interact by calls or SMSs only, exist, or whether groups with a high co-location rate correspond to communities into the communication networks. The last matter is strictly correlated to the interplay between co-presence and the interactions expressed by the communication media. Can we use the information given by call and SMS to improve the prediction about the next cell visited by a user? What medium is more predictive of a potential co-presence?
• **Multidimensional networks from online social networks**: We adopt the multidimensional network approach to analyze how much online social networks that share a set of individuals are similar/dissimilar. In fact, nowadays people have at their disposal a wide selection of online social services, each having its own peculiarity. This way the adoption of multiple social services by the same person is becoming increasingly spread and, in particular, people exploit and combine their favorite social media according to their needs (Twitter for news and marketing, Flickr o Instagram for photo-sharing, LinkedIn for business or YouTube for video-sharing). The opportunity to merge the information generated on different social media allows us to get a better understanding of the online behavior of the users registered to multiple social networks.

In this research field a data collection containing both the identities of the same person into different social media and her/is publishing activities does not still exist. Consequently, our first goal is to collect such a dataset. The strategies to reach the above goal are:

  o **Data gathering from social media aggregators**: Social media aggregators allow users to combine their own identities into a single profile by gathering their online activities from different social platforms. Technically, the aggregation is enabled by APIs provided by social networks, so that people may have control on the data the aggregation platform can access. Data gathered by crawling these platforms represent a feasible solution for the data collection issue.

  o **Profile matching**: The sampling of a high number of public profiles from a well-known online social network and the search of their presence in other social media represents the second strategy. To reduce the number of profiles resulting from the search we may adopt profile matching algorithms based only on the public information returned by the API of the social media. We focus on profile matching algorithms based on binary classifiers since they are able to scale on big datasets. However, their performances have been scarcely evaluated or heavily depend on the chosen training set. For this reason our goal is to collect a benchmark dataset containing the profiles on different social network associated to an individual. We plan to exploit this dataset to improve the algorithm performance in terms of reduction of false positive profiles. The resulting matching algorithm will be applied to the evaluation of the overlapping of the neighborhoods that belong to the same user in different social media.

In general the opportunities offered by a dataset that includes the online identities associated to the same individual are multiple. In this work we focus on the comparison among the different ranking induced by several network centralities in order to check whether the user’s importance maintains across the social networks. Furthermore, we exploit the contents published on different media to explore the similarity of the identities associated to a user.

• **Link strength prediction in online social networks**: Multiplexity does not only involve the profiles in online social networks but also the relationships they capture. In particular the awareness of the fact that the presence of a link provides only a coarse indication of the nature of the relationship is growing. For this reason the interactions occurring on the links are becoming an important research topic since they allow to better define the nature and the role of the links. In this research field our main goal is the development of a predictor of the number of interactions after a fixed period that will be applied at the creation time of the link. Due to the temporal constraint, the predictor must only rely on topological and temporal features not related to the interactions on the link; unlike other proposed methods which exploit past interactions on the link. We plan to apply the predictor on two problems: influence maximization and data replication for the storage dynamic networks. As for the influence maximization problem, the prediction of the link strength is fundamental since the chosen nodes mainly depend on the link weights. As concerns the second issue, the predictor should drive the replication of the vertices in case of a predicted strong tie. In fact vertices belonging to a strong tie will be replicated into the right shard or they will be placed in the same shard. We evaluate the
performances in terms of number of replicas and number of communications towards nodes in other shards, assuming a query sequence typical of a real-time news feed application.

Results of the research:

Multidimensional networks from mobile phone data: We analyzed the communication and the co-location networks of a million subscribers for a two-month period in the metropolitan area of Milan through a collaboration with the mobile operator H3G. We show that the two single layers describing on-phone interactions, SMSs and calls, are macroscopically similar as far as regards the connected components, but they are microscopically different. In fact, the two single networks partially overlap, since many users use a communication medium only (call or SMS). User ego-networks perceptibly enlarge in the multiplex network, confirming that both communication media are needed to get a complete vision of the users behavior captured by a mobile phone dataset. Second, we introduce the notion of multidimensional link reciprocity into the set of metrics for multiplex networks. We show that interactions by mobile phone are much more reciprocal, and thus social, than what could be speculated if only calls were considered. Nevertheless, reciprocity is much lower than observed in online social networks and in the Web.

As third contribution we add a third layer given by the spatial proximity obtaining a denser graph than on-phone communication graphs and we find that people communicating by phone are more likely to be in spatial proximity w.r.t. individuals who do not interact through any mobile medium. In particular, interactions by SMSs are more predictive of spatial proximity than calls. However a correlation analysis between the degree of co-location (proximity) and the strength of the communications between co-located pairs let emerge a novel result: in a metropolitan area like Milan, people who are strongly spatially close do not need to frequently communicate, as observed in other studies. At most the frequency of communications increases when people not frequently share some locations. Fourth, we study the correlation between the different centralities of the mobile phone users in each layer. The results confirm that the network, which merges on-the-phone communications and spatial proximity, is made of loosely coupled layers in terms of degree and strength.

Finally, we investigate the impact of multiplexity also at mesoscopic scale by performing a community detection analysis on each layer and on the multiplex network. It comes out that communities at different layers do not match. Among them, the SMS communities are more representative of groups of people sharing the same interest than the call ones, being call-based communities weaker. The previous results have been obtained by introducing three indexes (cardinality, coverage precision and coverage ratio), whose combination evaluates how a community is approximated by a set of other communities.

Multidimensional networks from online social networks: We collected two datasets to reach the aforementioned goals. The first dataset contains about 19,000 profiles and their activities on several social media. We developed a crawler to retrieve data from the social media aggregator Alternion, since it allows users to decide which information about their social sites – including profile information and contents - can be made public. Unlike other datasets in literature, this collection is the largest in terms of number of profiles and the availability of the posting activities. Moreover time-series of the activities of the users across today social media can be extracted. The second dataset collects Facebook and Google+ profiles of about 8,000 people. We gathered this collection through a crawling of one million public profiles on Google+. When links to other profiles in different social media were available, we stored the matched profile and their information retrieved by APIs.

The analysis of the Alternion dataset allows us to quantify how common the usage of multiple social media is. We observe that on average a user is simultaneously registered on 5 social media sites, while 73% of active users publish on at least three social networks. We also verify if statistically significant correlations between the centralities of the users in different sites exist; in fact the presence of these correlations measures whether or not popular users in one media maintain their popularity across media. By a rank correlation analysis we found that a strong positive correlation does not exist, i.e. a single user might be an hub on one system and loose part of its hubbiness on the other. We also apply the above methodology to
investigate how often users publish contents in different social media. Specifically we aim at verifying whether users who post a lot and often on a social platform, are equally active in other platforms. Unlike the above discussion on popularity, there is a more evident positive correlation between the posting activities across social sites. Finally, the Alternion dataset provides information about the usernames an individual adopts, indeed we evaluate how users are coherent in the choice of the username by computing several string similarity indexes. We found that people are not likely to change their username, even if this behavior depends on the pair of social media we consider. For instance the change of the username between Pinterest and Google+ is more evident w.r.t. the average behavior.

Through the second dataset we investigated the profile-matching problem by exploiting public information returned by API only. As first result, we implemented most of the matching solutions reported in literature in order to point their limits out and to compare our solution. In fact the previous solutions return a high number of false positive profiles, especially when they compare profiles with a high degree of homophily. To reduce the number of false positive we acted on the construction of the training dataset, since we adopt a machine learning approach to solve the problem. The methods in literature randomly build negative instances (‘no match’ class), while we introduce negative instances with a high degree of homophily and similarity. The classifier trained on the new training set got 0.92 in F-score using a random test set and 0.79 with a high homophily test set. Since the classifier reduces the number of false positive, we have applied it to evaluate the overlapping of the neighborhoods of the same person in Facebook and Google+. Specifically we used the algorithm to match the neighbors. We found that the average overlapping degree is quite low (Jaccard coefficient equal to 0.2), highlighting that many users exploit different social media to establish relationships with different individuals.

**Link strength prediction in online social networks:** As regards the last goal, up to now we developed a binary classifier that detects interacting links (> 4 interactions per year) at the creation of the link. We trained the classifier on a public Facebook dataset that includes the temporal annotated relationships and the interactions among the users. We exploited topological (common neighbors, degree, clustering coefficient) and temporal (link delay, triangles with low delay) features and we obtained 0.71 as F-score on the Facebook dataset. At present, we are evaluating the performances of a data replication mechanism based on the classifier, meanwhile we implemented the most recent algorithms that solve the influence maximization task to evaluate the effects of the prediction on the choice of the seed set.

**Other Activities:**
- Teaching:
  - 48 hours “Reti di Calcolatori e laboratorio” (corso di Laurea Triennale in Comunicazione Digitale)
- Co-Advisor:
  - Master thesis: ‘Metodi e strumenti per il rilevamento ed analisi di community in grandi dataset telefonici’.
  - Master thesis: ‘Profile matching in online social networks’.
  - Master thesis: ‘Metriche temporali per la predizione delle interazioni fra utenti su reti sociali’.

**First year Publications:**
• Zignani, M., Dimauro, M., Gaito, S. and Rossi G.P. People across social media sites: usage and behaviors. *International Conference on Computational Social Science (IC3S*) , 8-11 Jun 2015.

La presente relazione, non contiene dati sensibili e dati giudiziari di cui all’art. 4, comma 1, lettere d) ed e) del D.Lgs. 30.6.2003 n. 196. Si autorizza la pubblicazione della relazione annuale sul sito web del Dipartimento.

Firmato(in Stampatello)   NOME :............................................. COGNOME :...........................
Il Responsabile Scientifico

(Firma)

L' Assegnista di Ricerca

Matteo Zippiani

(Firma)