Università degli Studi di Pisa

Dipartimento di Informatica
Dottorato di Ricerca in Informatica

Ph.D. Thesis Proposal

Evolution in social networks

Giulio Rossetti

Supervisor	Supervisor
Dino Pedreschi	Fosca Giannotti

December 4, 2012

Contents

1 Introduction **5**

2 State of the art **7**

2.1 Network Science 7

2.1.1 Network Models 7

2.2 Multidimensionality 11

2.2.1 Multidimensional Measures 12

2.2.2 Internetworking Scenario 13

2.2.3 Hub Analysis 14

2.2.4 Trust Networks 14

2.3 Time and evolution 15

2.3.1 Link Prediction 16

2.3.2 Information Propagation 17

3 Open problems **19**

3.1 Structural Issues 19

3.1.1 Multidimensional Node Ranking 19

3.1.2 Multidimensional Link Strength 20

3.2 Topological Issues 21

3.2.1 Community Kernel 22

3.2.2 Multidimensional Community Discovery 22

3.2.3 Community Proximity 23

3.3 Evolutive Issues 23

3.3.1 Community Evolution 24

3.3.2 Link Prediction and Network Archeology 24

3.3.3 Trust and Privacy 25

4 Proposal **27**

Bibliography **31**

Chapter 1

Introduction

In a society that relies heavily on technological systems every human activity can be seen as a source for huge amount of data ready to be collected and analyzed. Social networking sites, telephone providers as well as a wide range of other private, and public, companies collect every day massive quantities of data and use them to profile users habits. Companies that provide online services (i.e Facebook with its 955 million members, Twitter, Foursquare, Google and so on) study the preferences of their users, in order to shape their offerings leveraging on the results of such analysis, each competitor trying to capitalize the desires of its loyal customers. This knowledge, as well as the data from which it is extracted, represents today the most valuable asset that any organization needs to collect, understand and be able to exploit.

In recent years a new term has begun to appear frequently in scientific literature and has been brought to the attention even of non-experts by the media: *Big Data.* The scientific excitement expressed by this, overused, expression led to several questions: when the data should be considered “big”? Is the term “big” related only to the size of the data or it expresses also the complexity of the processes necessary for their analysis and understanding? Is “big” an intrinsic or an extrinsic property of data?

Certainly, the size of a dataset plays a primary role on its categorization as Big Data but the chance to augmenting its semantics trough the adoption of orthogonal sources, now easily reachable on the web, represent the key to understand the real weight of the scientific revolution that we are experimenting. Nowadays, social networks can be integrated by cellphone calls, GSM and email logs data as well as by descriptive data: the dynamics of users interaction, in this complex scenario, allow researchers to detail, in a fine-grained way, patterns and behaviors otherwise unidentifiable.

Those huge datasets are often unstructured and not easily understandable, for these reasons scientists feel the need to find a simple way to model them. Network theory is one of the tools that, in the past decade, was used to this purpose.

Participating in an online social network, making phone calls, online shopping as well as sending emails or co-authoring a paper are common activities that establish a connection between different entities: exploiting such connections in order to impose a structure to otherwise unstructured data make possible the investigation and understanding of complex dynamics expressed by those simple actions. Networks are the natural way to express a wide category of data but, in this context, they can’t capture all the semantics expressed by this evolving multi-domain: a way to overcome this issue is to extend the measurements done so far proposing approaches that, moving from the classical network science, take care to include and make coherent all the novel informations provided as well as their correlations. Multidimensional networks that gather together multiple dimensions of interaction between people, taking care of the spatio-temporal information related to them, are actually one of the most promising models proposed by the scientific communities so far.

Starting from the current research on complex networks, in this thesis we propose to investigate spatio-temporal models in a two fold manner: firstly, we want to improve the understanding of network evolutions, investigating how new edges arise and how communities take place and evolve. Secondly, we aim to use such results as baseline to study multidimensional contexts, and develop novel measures and algorithms capable to better describe the dynamic nature of the real world complex networks.

The rest of the thesis proposal is organized as follows. Chapter 2 presents an overview of the techniques and models in the fields of complex networks and evolutive analysis. Chapter 3 introduces the current open problems in literature. Finally, in Chapter 4 we describe the work that we would like to realize in the depicted contexts.

Chapter 2

State of the art

In this section are introduced some of the results obtained by the scientific communities that relates to the topics of the proposed work.
In order to provide an organized general view of the ideas established so far, the main results introduced are grouped by their research category.

2.1 Network Science

The science of networks is a discipline that examines the interactions, and interconnections, among diverse entities: this mathematical approach, in its generality, makes possible to analyze with the same instruments networks of very different nature (i.e. social, technological, biological, physical...).
The mathematical structure used to model networks is the graph. A graph, often specified as $G = (V, E)$, is defined as a set of nodes (also called vertices) V that are connected by links (also called edges), belonging to the set E. Given this simple definition the applicability to a wide set of real world scenarios appear bright.
Graph Theory first appear in 18th century when the mathematician Euler solved the famous Seven Bridges of Königsberg problem. In his 1736's paper [25] Euler shows how relationships between simple graph measures, such as node degree and node cardinality, could be exploited in order to tackle everyday problems. Since then a huge amount of works have used the graph embedding, where possible, in order to impose structure to real life tasks and exploit the knowledge given by the topology observed. Example of well-known problems that are formulated as graph tasks range from Operational Research (i.e. shortest path, minimum spanning tree, maximum flow...) to the complexity theory (i.e. Traveling Salesman Problem, Graph Isomorphism...). The field of graph theory continued to develop up to the end of the 20th century thanks to interdisciplinary interest. Those increasing attentions have led to the need of analyze and categorize network models; real networks often express very distinctive topologies, been able to understand them and to propose valid patterns of network growth is one of the central problems discussed so far.

2.1.1 Network Models

Here are reported four important models in order to show how peculiar characteristics can appear analyzing different kinds of networks: Random Graph, Small World, Scale Free and Forest Fire model. Each model provide, as we will be able to notice, a generative formulation that makes possible to build synthetic datasets that respect characteristics observed in real world networks. None of the proposed model is able to capture the exact evolutive pattern of any real networks but they can be seen as proxies that can be used in order to classify different typology of networks.

Random Graphs

One of the most famous model produced in the 20th century is the one known as *Random Graph model*.
Introduced, independently, by Solomonoff and Rapoport [80] and by Erdös and Rényi [24] the random graph theory is one of the breakthrough that had led to an initial analysis of topologies and characteristics expressed by complex networks. A random graph, $G_{n,m}$, is defined as a set of n labeled nodes connected by m edges chosen randomly, with probability p, from the $\frac{|V|(|V|-1)}{2}$ possible edges. The given definition make possible the existence of several different graphs for the same chosen value of n and m: all those graphs belong to a probability space in which every realization is equiprobable.
Since its first formulation major outcomes for this model were showed by studies of several mathematicians: the tendency of random graphs to have small diameter and a degree distribution that follow a Poisson distribution; the similarity among the diameter value and the average path length; the presence of giant component if the mean degree $\langle k \rangle = pn < 1$ and, after a phase transition for $\langle k \rangle > 1$, the presence of isolated trees.
This model, born in 1959, was one of the first attempt to describe, through the graph theory, social and communication networks. The assumption that real networks have to shown random topology was subverted by numerous measurements done at the end of the last century on real networks: the observation of nonrandom topology became a trigger for the study of alternative models.

Small World

In a paper of 1967, the sociologist Stanley Milgram propose an experiment [62], that became very popular thanks to a play [36] by John Guare, known nowadays by the name of “Six degrees of separation”. In his work Milgram decided to verify the if small-world experience (an unknown person we meet knows a person we know) is related to some real phenomenon or is only a simple anecdotes. For his experiment the sociologist asked Midwestern volunteers to send packages to a stranger in Boston knowing only his name and profession: packages could not have been sent directly to the recipient, but should have been delivered only by exploiting the personal contacts. The results reported shows that, on average, the number of intermediaries needed to complete the chain was 5.5[1].
This findings led, in 1998, Watts and Strogatz [84] to describe a peculiar kind of networks as *small-world networks*, in analogy with the small-world phenomenon. In a paper where they analyzed the neural network of the worm *C. Elegans*, the collaboration graphs of film actors and the western U.S. power grid, they found that these networks shows an average small path length and a high clustering coefficient. The latter observation were in contrast with the characteristic low clustering coefficient shown by random graphs and reminds to

[1] As matter of facts this value was lower if the contacts were chosen among people that belongs to the same profession of the final recipient, a little bit higer if chosen following only geographical proximity.

the more rigid structure typical of regular lattices.
From this observation Watts and Strogatz decided to propose a model (they called it Small-World model) that interpolate between the regular structure of lattices to the random ones. The methodology used to perform such interpolation was to, starting from a circular lattices in which each node is connected to k neighbors, rewire with probability p an edge at a time avoiding duplicate edges and self-loops. In a variant of this model [69], that does not contemplate rewiring, few random edges were added to random nodes of the lattices in order to lower the average path length (recreating in this way the small-world phenomenon).

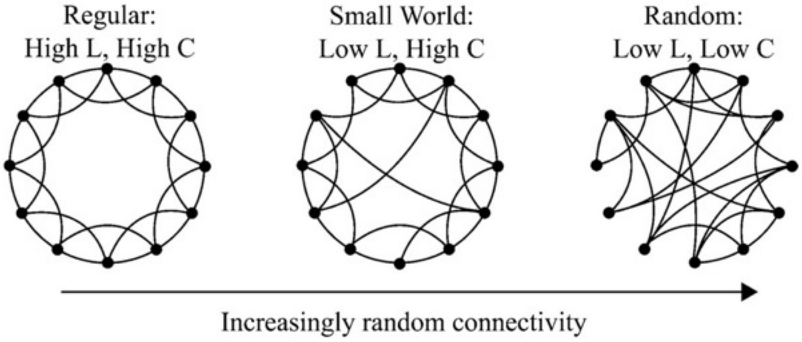

Figure 2.1: Network topology from regular lattice to random network at different p rates

Varying p we can observe variation on the network structure: for $p = 0$ we have a regular lattice with high clustering coefficient, for $p = 1$ we have random networks that exhibit small-world properties and have low clustering coefficient. In the interesting interval $0 < p < 1$ fall all the networks that even having high clustering coefficient shown a small geodesic distance among the nodes. Watt and Strogatz analyzed, given the value of p, the clustering coefficient $C(p)$ and the characteristic path length $L(p)$: varying the value of p in the latter interval they discover that there are an high number of cases for which $L(p)$ is almost small as L_{random} and at the same time $C(p) >> C_{random}$. This results is given by the introduction of random shortcuts that connect nodes otherwise distant in the original lattice (as shown in 2.1).
The highly clustered nature of networks was guessed in social context by Granovetter [36], who proposed an excellent sociological interpretation: the network behind our society consists of small, fully connected circles of friends connected by *strong ties*. The rewired edges introduced by Watts and Strogatz represent the *weak ties* that connect the members of these cliques to their acquaintance, who have strong ties to their own friends.
The small-world effect and the clustered nature of real networks was discovered in a wide range of natural and artificial structures like the Internet [86, 74] , the WWW [5, 2], in email contacts networks [64], cellular networks [41], scientific collaboration networks [10, 67], in the neural network [84, 39], in the Messenger contacts network [53]. These findings suggest that results of the sociologist Milgram and Granovetter are pervasive in networks, in nature and technology.

Scale Free

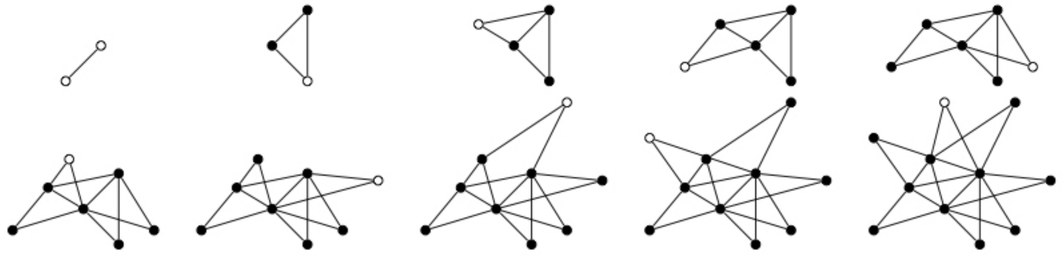

Figure 2.2: Network growth as proposed by Barabási-Albert model. Starting from a dyad (top-left) at each itaration a new node, hilighted in white, and two new edges were added following preferential attachment. Older nodes becames quikly hubs as time goes by.

An aspects not covered by the previous models regards the existence of *hubs* among the nodes: we define an hub as a node that is highly connected than the others. Hubs exist in most of the network today object of analysis: a well-connected user in Facebook, a celebrity in Twitter, a major airport (as the international ones in Rome, New York or Frankfurt) a well-known scientist in a collaboration network are all examples of such particular kind of nodes.
In order to study this property we use the distribution $P(k)$, the fraction of the nodes with degree k. The simple Erdös-Rényi model shows a Poisson distribution for $P(k)$, but in many real networks $P(k)$ appear ti be highly skewed and to decay much more slowly than a Poisson, mostly following a power law $P(k) \sim k^{\gamma}$. These networks are called *scale-free*, because they have not a scale, that is a characteristic node.
Although this distribution is not universal in networks [6], it is very common and it is observed in several artificial and natural networks, such as the WWW [5], Internet backbone [26], metabolic reaction networks [41]. For co-authorship networks of scientists [67, 11] the degree distribution is fit better by a power law with an exponential cutoff, for the power grid of western United States it is an exponential distribution, for a social network of Mormons in Utah, $P(k)$ is a gaussian [6].
In 1999 Barabási and Albert [9] showed that this heavy-tailed degree distribution is the result of networks continuos expansion by the addiction of new vertices. The two scientists speculate that new nodes preferentially attach to ones that are already well connected in such a way that "richer nodes became richer". On top of this two assumption, the growth over the time of networks and the preferential attachment, they proposed the *Scale Free model* (also known as Barabási-Albert model). Starting at a time t_0 with a low number of nodes at every time step we add a new node with m edges that link to m different vertices already present in the network: the probability of choosing a node in order to establish a link is proportional to its degree. After t time steps the model converge to a random network with $t + m_0$ nodes and mt edges (as shown in 2.2). This approach led to a scale-invariant state with the probability that a node has k edges follow an heavy tail distribution (called Power Law) that shows exponent $\gamma = 2.9 \pm 0.1$.
Given the high impact caused by this model, more sophisticated variants were proposed during the last decade, models that include the effects of adding a rewiring links, allowing nodes to age so that they can no longer accept new links, or varying the form of preferential

attachment.

Forest Fire

Once discovered the Scale Free and Small World properties, several models were proposed in order to describe, and build, networks that show both those desirable characteristics. One of them, probably the most famous, is the *Forest Fire model.* Proposed by Leskovetc [55], this growth model try to introduce community structures, that appear in almost all the social-like real networks, into the generative process. In particular, the authors would like to capture the shrinking diameters that have been observed in complex networks, the fact that real networks tend to have heavy-tailed distributions for in- and out-degrees and a Densification power laws which estimate that networks become denser over time. The latter observation is strictly related to the community structure, previously taken into consideration by other models (i.e. copying model [46, 48]), expressed by a wide typology of networks.
In order to perform this task, Leskovetc define a basic version of the model in the following way: nodes arrive one at a time and form out-links to some subset of the earlier nodes; to form out-links, a new node v attaches to a node w in the existing graph, and then begins burning links outward from w, linking with a certain probability to any new node it discovers. This process could be intuitively seen as the strategy by which an author of a paper identifies references to include in the bibliography. He or she finds a first paper to cite, chooses a subset of the references in that paper (modeled here as random), and continues recursively with the papers discovered in this way. Depending on the bibliographic aids being used in this process, it may also be possible to chase back-links to papers that cite the paper under consideration. Similar scenarios can be considered for social networks: a new computer science graduate student arrives at a university, meets some older students, who introduce him/her to their friends, and the introductions may continue recursively.
Adding edges in this way, each node connecting only with entities that are closer to its center of gravity, the community structure arise very quickly.

2.2 Multidimensionality

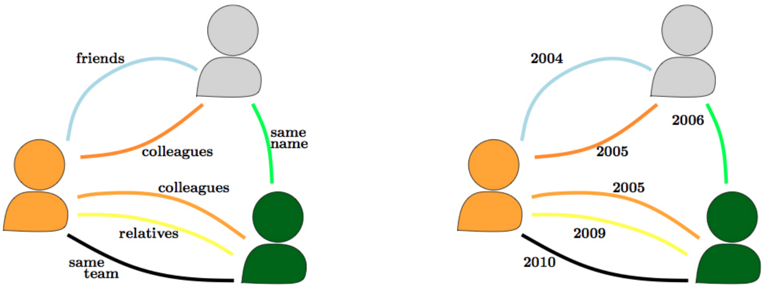

Figure 2.3: Examples of multidimensional networks.

In recent years, complex networks have been receiving increasing attention by the scientific community, also due to the availability of massive network data from diverse domains, and the outbreak of novel analytical paradigms, which pose relations and links among entities, or people, at the center of investigation. Inspired by real-world scenarios such as social networks, technology networks, the Web, biological networks, and so on, in the last years, wide, multidisciplinary, and extensive research has been devoted to the extraction of non trivial knowledge from such networks. Most of the networks studied so far are monodimensional: there can be only one link between two nodes. In this context, network analytics has focused to the characterization and measurement of local and global properties of such graphs, such as diameter, degree distribution, centrality, connectivity - up to more sophisticated discoveries based on graph mining, aimed at finding frequent subgraph patterns.

However, in the real world, networks are often multidimesional, i.e there might be multiple connections between any pair of nodes (see Figure 2.3). Therefore, multidimensional analysis is needed to distinguish among different kinds of interactions, or equivalently to look at interactions from different perspectives.

Dimensions in network data can be either explicit or implicit. In the first case the dimensions directly reflect the various interactions in reality; in the second case, the dimensions are defined by the analyst to reflect different interesting qualities of the interactions that can be inferred from the available data. Those complex systems are referred also as multislice [63], networks with explicit dimensions are named multiplex, and, often, temporal information is used to derive dimensions for the network where other semantics are not available.

Examples of networks with explicit dimensions are social networks where interactions represent information diffusion: email exchange, instant messaging services and so on. An example of network with implicit dimensions is an on-line social network with several features: in Facebook, while the social dimension is explicit, two users may be connected implicitly by their like on posts of shared friends, by their favorited pages or groups they belong to. Moreover, different dimensions may reflect different types of relationship, or different values of the same relationship.

Given the novelty of the analysis performed on this complex scenario here we report some introductory works, mostly done by the Knowledge Discovery and Data Mining Laboratory, that can be considered actually the state of the art of this field.

2.2.1 Multidimensional Measures

Most real life networks are intrinsically multidimensional, and some of their properties may be lost if the different dimensions are not taken into account. In other cases, it is natural to derive multiple dimensions connecting a set of nodes from the available data to the end of analyzing some phenomena.

In order to study this complex scenario a framework that extends the classical graph theory is needed. Reasoning on multidimensional networks seems clear that the usual graph model is not enough to represent all the available information. Luckily, graph theory provides us a more flexible model, called *multigraph* that suits our needs: a multidimensional network could be seen as a labeled multigraph, $G = (V, E, L)$, defined by a set of nodes V, a set of edges E, a set of labels L that represent dimensions of a network. Each edge $e \in E$ is now defined by the triple (u, v, d) where $u, v \in V$ are nodes and $d \in L$ is the dimension to which the edge belong. At the same time the same edge could occur in multiple dimensions.

In their work "Foundations of Multidimensional Network Analysis" [13] Berlingerio et al. proposed and evaluate, on real datasets, a wide spectrum of measurements that are able to capture the interplay among dimensions and to overcome some limits that made the classical monodimensional measures unsuitable in this complex scenario.

Moving from the simple observation that given a node u of a multigraph its degree does not correspond to the cardinality of its neighbors' set, $|\Gamma(u)|$, is clear how even the basilar concepts of neighborhood has to be revised in a multidimensional settings. Several variants of Neighborhood functions were proposed[2] and a wide set of connectivity measures, with different grade of exclusivity, both at local and global level were introduced.

Experiments shown that the provided measures are able to capture interesting informations making of this framework one of the first for the multidimensional networks analysis.

2.2.2 Internetworking Scenario

The rapid development of the number and the size of Online Social Networks (OSNs) makes the analysis of Social Internetworking Scenarios (SISs) extremely challenging. In a SIS, a user can join multiple OSNs and two users can interact with each other even though they joined different OSNs and did not know each other. While OSNs have been extensively studied in the last years, the most peculiar aspects of Social Internetworking Scenarios have not been yet investigated, especially from the Social Network Analysis perspective.

Many researchers started to collect large amounts of data from OSNs and to apply techniques of classical Social Network Analysis on them. The results they obtained are numerous and extremely interesting: they are based on the intuition that a strong correspondence between the user behavior in an OSN and the structural properties of the corresponding graph is likely to exist.

An important aspect to take into account is that nowadays users tend to spread their activities among more OSNs and, often, to show a different behavior in different OSNs [20]. As a consequence, different Social Networks could be seen as interconnected thus resulting in a global graph whose structural features are very different from those of each single Social Network seen as a graph. This complex topology represents the Social Internetworking Scenario, where a user can join multiple OSNs and two users can interact with each other even though they joined different OSNs and did not know each other. Only few commercial attempts to implement Social Internetworking Systems have been proposed so far (i.e. Google Open Social and Friendfeed).

Despite the great attention given by scientists towards Social Networks, Social Internetworking Scenarios have been little investigated in the scientific literature also due to their young age. Some papers focus on cross-folksonomies [40, 82], i.e. they analyze the tagging behavior of users in multiple folksonomies and try to relate information about these behaviors. Other ones, such as [1, 81], assume users can join heterogenous social systems (e.g., a folksonomy and a blogging platform in [81], or Social Networks like Facebook, and social media, like Flickr, in [1]). Their goal is to aggregate user information in such a way as to build a global user profile.

[2] Aside from the complete list of neighbors of a node we could be interested in all the neighbors reachable through edges belonging to a set of dimensions ($Neighbors_{Set}(u, D)$) as well as the ones reachable *exclusively* by edges belonging to that set ($Neighbors_{Xor}(u, D)$).

Reasoning about this multilevel structure is easy to apply the theory formulated to analyze multidimensional networks (where the dimensions are the different OSNs analyzed).

2.2.3 Hub Analysis

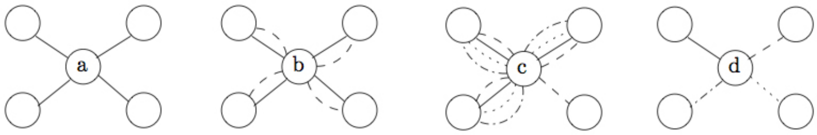

Figure 2.4: Examples of possible different configuration for a multidimensional hub with four neighbors.

One topic of research in the field of complex networks that has received considerable attention from the scientific community, since the introduction of the Barabási-Albert scale-free model [9], relies on finding and analyzing hubs, i.e., nodes with a large number of neighbors. Hubs plays an important role for the network connectivity: given the pawer law that regulate the degree distribution of a wide range of social networks, the presence of few power nodes with high degree assure the resilience at random failure of vertexes or edges.
On the other hand, those nodes, that are easily identifiable, represent objects that are valuable for the analyzed structures: targeted attacks to those entities may cause the network to fall apart.
Most of the networks studied so far are monodimensional: in this setting the concept of hub has been widely studied, and is at the basis of many important applications, ranging from analysis of the structure of the Internet to web searches, from peer-to-peer network analysis to social net- works, from Viral Marketing to analysis of the Blogosphere, from outbreaks of epidemics to metabolic network analysis [4, 9, 30, 41, 44, 52, 60, 78]. Moving from those works, recently, the scientific community have begun to take into consideration, as a natural evolution, the multidimensional settings.
In [14], Berlingerio et al. study multidimensional hubs verifying their meaningfulness and the role that they play for the connectivity of a multidimensional networks. Since, as seen in the previous section, the degree of a node do not correspond anymore to the cardinality of its neighbors' set a wide range of different configuration now fall in the classical definition of hubs (as shown in Figure 2.4).

2.2.4 Trust Networks

An interesting, and simplified, subset of multidimensional networks are the signed ones.
Social interaction on the Web, as well as in the real life, involves both positive and negative relationships[3]. While the interplay of positive and negative relations is clearly important in many social network settings, the vast majority of online social network research has considered only positive relationships [68].

[3] In some OSNs people can establish connections to indicate friendship, support, or approval but they also can express disapproval of others, or disagreement or distrust of the others' opinions

A number of papers have begun to investigate negative as well as positive relationships in online contexts: example spread from the analysis of Wikipedia users' vote for or against the nomination of new administrator [18], users on Epinions expression of trust or distrust of each others [37, 61]; and participants on Slashdot that declare others to be either "friends" or "foes" [17, 49, 50]. Even hyperlinks on the Web can be used to indicate agreement or disagreement with the target of the link, though the lack of explicit labeling in this case makes it more difficult to reliably determine this sentiment [73].

In [54] Leskovec propose a way to predict sign of unlabeled edges of such kind of networks. The idea proposed by the authors is to apply the *Social Balance theory* based on the common principles that "the friend of my friend is my friend", "the enemy of my friend is my enemy", "the friend of my enemy is my enemy" and (perhaps less convincingly) "the enemy of my enemy is my friend". Concretely, this means that if w forms a triad with the edge (u, v), then structural balance theory posits that (u, v) should have the sign that causes the triangle on u, v, w to have an odd number of positive signs, regardless of edge direction - just as each of the principles above has an odd number of occurrences of the word "friend".

Given the unnecessary reciprocity of the relationship between each pair of users belonging to a social network - if a trusts b, the inverse relationship is not necessarily true - all the problems studied so far can be expressed and generalized to more complex scenarios (introducing multiple levels of trust and analyzing how relationships change over time).

A recent work [8] propose the analysis of more complex patterns not covered by the social balance theory. In their work the authors extract frequent pattern from a signed network and use them in order to predict the sign of an edge, on the target network, whose surroundings are matched.

2.3 Time and evolution

One of the main unresolved problems that arise during the data mining process is treating data that contains temporal information. In such cases, a complete understanding of the entire phenomenon requires that the data should be viewed as a sequence of events. Temporal sequences appear in a vast range of domains, from social science, to medicine and finance, and the ability to model and extract information from them is crucial to made inference on the behaviors of the systems that are object of the investigation. The ultimate goal of temporal data mining is to discover hidden relations between sequences and subsequences of events. Discovery relations between sequences of events involves mainly three steps: the representation and modeling of the data sequence in a suitable form (i.e. timeseries attached to the nodes or edges of a graph that represent the interactions among nodes or the variation in time of the degree of a specific node); the definition of measures between sequences; and the application of models and representations to the actual mining problems (i.e. predicting new links or understand how relationships change over time).

One of the models often adopted in order to manipulate evolving data is provided by the timeseries: this simple structure were used for several different kind of studies ranging from the discovery of association rule [23, 38, 71], classification processes [42], unsupervised clustering approaches [43], to prediction tasks [27, 32, 57, 85].

The topics proposed so far cover a subset of static analysis that could be performed on complex networks. The world we live in is constantly evolving: as time goes by relationships change, links are substituted by new ones, new collaboration arise and old ones fall apart. Freezing the network structure in time is certainly very useful in order to observe, study and understand its peculiar traits but is not enough if we are interested in the dynamics that regulates its life, growth end evolution. Observing the mere static structure of an online social network, for instance, is not completely adequate if we want to reason on diffusion processes that could take place: often in such kind of structure most of the links, even if present, have to be considered "non active" because no social interaction occur between their endpoints.

In this section we introduce two problems studied so far by the scientific community that relates to dynamic networks: *Link Prediction* and *Information Propagation.*

2.3.1 Link Prediction

Reasoning on networks evolution one of the first interesting question to answer is: is there any rule that regulate the rising of new edges? or similarly, there exists couples of nodes that are most likely to establish a connection than the others?

As we have seen before 2.1, a wide set of models were studied with the aim to understand and reproduce real networks traits: all those models are generative, and reproduce the processes of networks growth over time. Here we are interested in a slightly different problem: we know the nodes of our network (we assume that no other nodes could be added in successive time steps) and want to study the probability that two of them became neighbors. Suggest new friendships on a social network, co-autorships on a professional network or interesting products in an online-market are certainly facilities that online services need to offer to their users.

A formal definition describe the classic Link Prediction problem as the problem of identify, given a snapshot of a network G at a time t_0, the top-k edges that are most likely to appear among its set of nodes, at a time t_1. The prediction is restricted to those nodes that are not connected by edges during the first observation.

Predicting a new link correctly is certainly an hard task to accomplish[4]: for this reason several approaches were proposed on order to study this evolutive aspect of complex networks, trying both supervised and unsupervised methodologies. Unsupervised approaches are based only on local (neighborhood-based), or global (path-based), topological aspects relative to the couples of nodes for which its needed a prediction: among those approaches [75] presented a solution based on the preferential attachment principle, while [3] and [66] introduced models based on the quantitative characteristics of common neighbors. A survey on unsupervised approaches is proposed by Kleinberg [56] that empirically compare many different models.

Those methodologies, given their simple nature, shown performances that led to, at most, 10% of correct predictions: given the complexity of the problem, this value that could seem very low is actually a very good result.

To improve this result, supervised approaches that exploit not only the topology of the network but even semantic informations attached to the nodes (and edges) were proposed. Two supervised approaches are the ones in [16, 15], where the first one allows also for the prediction of new nodes. In order to extend the semantic information provided by

[4] Most real networks, for instance the social ones, are quite sparse and the set of all possible edges that can be potentially predicted is $\frac{|V|(|V|-1)}{2}$ (if we considerer the case of an undirected graph).

the network in [58] were presented a link prediction framework that uses multiple data sources, while [65] proposed an analysis through the use of some graph proximity measure and weight of the existing links.

2.3.2 Information Propagation

Even when we consider networks structure as static objects, not allowing creation and removal of edges and nodes, there are interesting problems that are strictly related with temporal analysis. One of the most famous is the *Information Propagation Problem.* Users in a social network communicate (i.e. in Facebook they can publish on their wall, express their approval trough likes and tags on photos and posts of their friends), exchange informations (i.e. in a professional network news on positions openings) search and promote topics and trends (i.e. hashtags in Twitter). Some of these actions performed by users on a social network could cause a cascade phenomena: the aim of works on information propagation is to understand how this phenomena are related to social influence, how information spread recursively from users to their friends and how this process could define tribe or communities.
Two main elements that regulate those kind of processes have been studied:

- *Time:* how rapidly actions became viral? Is there a temporal threshold (or number of hops) for which the diffusion process ends? What is the distribution of temporal intervals among "hops" expected for an information to stop its spread?

- *Causes:* every action lead necessary to a cascading effect or some kind of topological features are needed?

Some works that tries to answer the first set of questions are [12, 31] where application of graph mining and annotated sequences are used in order to discover frequent patterns (both in graphs and in sequences that represents flows of communication) and then to obtain a rich description of the information propagation process.
The topological feature that makes some kind of nodes "Power nodes" capable of a more pervasive spreading of information are analyzed in [33]. Another work [59] aims to represent the information spread as an heat diffusion process.
Analyzing diffusion processes on a trust network is possible to model not only the spread of positive opinions about a specific person (trust), but also the negative ones. In order to register the information spread and/or to identify leaders from a set of timestamped actions, seems that the temporal analysis represent a mandatory step. However, in literature the impression is that our knowledge about this phenomenon is still far from complete.

Chapter 3

Open problems

In the previous chapter we have highlighted several studies that covers topics in complex network science, multidimensionality and temporal analysis. In this chapter we propose a short, and obviously non exhaustive, resume of open problems that we consider interesting for those fields.

In order to make the exposition structured, three main fields were identified and detailed with examples of open issues: *Structural*, *Topological* and *Evolutive analysis.* For each field we will provide a short description, along with some considerations about what is already done and what is currently under development.

The selected issues are to be intended as an introduction to the main themes of the thesis proposal that will be discussed in Chapter 4.

3.1 Structural Issues

The first set of open problems relate to the analysis of the structures that compose complex networks.

As said before (in 2.2) networks are often multidimensional: such extended information about the interactions among nodes could reveal interesting patterns that are precluded to the analysis of monodimensional networks. In order to show how this enrichment of the graph model could be exploited we propose here two problems: *Multidimensional Node Ranking* and *Multidimensional Link Strength.*

3.1.1 Multidimensional Node Ranking

Nowadays people are digitally involved throughout their life and there is not a clear separation between personal and professional network. If we consider a professional network, real as well as online, each member can be defined by its skills and role. Obviously different connections could occur on such networks (think about a co-authorship network in which two authors are related if they have published a paper together in a specific venue): how can we rank nodes in a multidimensional multi-skill context?

In order to identify and understand the position of a person in the networks, based on its skills or knowledge, we need to find a way to perform some ranking. The richness of the data and the hidden knowledge demand for a multidimensional and multi-skill approach

to the node ranking problem: the definition of a ranking algorithm on networks, able to capture the role of different kinds of relations and the importance of different skill sets.
Multiple kinds of relations in a network are usually modeled with multidimensional networks. Different dimensions intuitively lead to different skill exchange among the nodes. To the best of our knowledge, no network-based ranking algorithm is able to handle multidimensional networks today. Further, the most famous ranking algorithms [72, 79, 51] usually provide a monolithic ranking, without differentiating among different possible characteristics that can have different importance in different nodes. HITS [45] is one of the few making this distinction, but it is limited to only two different types of importance measures, not an arbitrary number of different attributes, making it impossible to use HITS for real world skill ranking scenarios.
Given the diverse nature of all the possible relationship that occur in a network different skills could be reachable in many different ways. For instance we could need to identify someone that have skills in biology but none of our contacts posses those competence: a classic approach would rank equally all our neighbors while a multidimensional and multi-skill ones could try to exploit the path that, passing trough our acquaintances, led to someone that have such kind of knowledge because of his work or studies. Being able to identify the typology of query we need to answer (i.e. we need someone who is competent in biology, rather than in fencing or cooking) and the dimension most suitable to reach such information could led to a ranking algorithm that can diversify the results showing not only the optimal answer (if present) but even rank all the contacts that could be used as bridge to reach such information.

3.1.2 Multidimensional Link Strength

In the last few years, the advent of OSNs has completely redefined the way we conceive our social relationships, creating the sensation of having broken the constraints of time and geography that limited peoples social world. In these virtual environments establishing new friendships is immediate and effortless, so it is reasonable to think that the number of our social bonds could approach to infinite, removing the social boundaries of our modern, technological era. However, what social networks have allowed us to do is to build massive networks of weak ties: acquaintances and non intimate ties we use all the time to reach out persons, business requests, speaking engagements, or ideas and advice. Despite such enormous quantity of acquaintances, recent works have revealed two major aspects of both online and real social networks:

- people still have the same circle of intimacy as ever,
- the formation of friendships is strongly influenced by the geographic distance, breaking down the illusion of living in a “global village”.

People tend to interact intensely with a small subset of individuals, carrying out a social grooming in order to maintain and nurture strong, intense ties. Strong ties are the people we really trust, people whose social circles tightly overlap with our own and, often, they are also the people most like us. Although such trusted friendships are not so important in the spreading of information [70], new ideas [19], or in finding a job [34], they can affect emotional and economic support [77, 29] and often join together to lead organizations through times of crisis [47]. Unfortunately, the majority of social media do not incorporate tie strength in the creation and management of relationships, and treat all users the same: friend or stranger, with little or nothing in between. A first attempt to take into consideration the social role of a friendship was done by Facebook and Google+ by the introduction of the "circles". Users can use circles as a way to organize their contacts in a sort of address book, creating different groups for relatives, work colleagues, close friends and so on. However, such conceptual organization of contacts does not provide quantitive information about the real strength of the ties, but only the nature of relationships between users and the context in which they take place. For example, the presence of a user in the circle of work colleagues does not necessarily imply the existence of a strong tie, and does not provide any explicit quantification of the importance of the relationship.

How to define a tie strength measure that is capable to discriminate between intimate ties and mere online social contacts?

Actually, it does not exist a formal, unique and shared definition of tie strength, and literature has often provided very personal interpretations of Granovetters intuition: "he strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding) and the reciprocal service which characterize the tie" [35]. The most frequently used measurements of tie strength in social networks are based on the number of conversations between users [7], or, in the mobile phone context, on the duration of calls [70]. However these common approaches suffer two major shortcomings. Firstly, the number and intensity of conversations strongly depends from user to user, making it difficult to understand which of these conversations are dedicated to intimate relationships. Secondly, they do not take into account that strong ties must be powered by a form of social grooming, that is mainly based on geographical proximity and face-to-face contacts.

Giving those assumption it is interesting to analyze (maybe using in a SIS[1] dataset) if bridges that emerge through different OSNs could be considered "strong ties". Do the connection over different online services implies some sort of strength relationships? Is there some kind of backbone that connects users among different OSNs? Could we use this information to build a better proxy that resemble the real social network?

The strength of a tie is related not only to the number of interaction among the users that are connected. There are more complex and meaningful informations that define a strong link among two people: the geographical distance, the similarity of their mobility habits, the different typology of services by which they share informations and, also, the trust or distrust they express by their actions.

Trying to depict a coherent definition of ties strength that ensemble all this aspects could led to a better understanding of the social context expressed by complex real networks.

3.2 Topological Issues

One of the basic topological bricks of social networks are the communities. The concept of "community" in a (web, social, or information) network intuitively depict a set of individuals that are very similar, or close, to each other, more than to anybody else outside the community [21]. This has often been translated in network terms into finding sets of nodes densely connected to each other and sparsely connected with the rest of the

[1] *Social Internetworking Scenario*, as defined before in 2.2.2

network.
The presence of big OSNs datasets have made possible, in the last decade, an analysis of such structure. Here we introduce issues that are not yet treated extensively in literature and that, in our opinion, could be valid ideas for a deeper study: *Community Kernel*, *Multidimensional Community Discovery* and *Community Proximity.*

3.2.1 Community Kernel

Observing the social structure proposed by OSNs we can easily notice that the neighborhood of a single node is often composed by a huge number of entities. Not all this acquaintances have to be considered as "real" ones: in online services, where as notice before, the possibilities to meet new people comes for free, most of the links does not represent relations existing in real life. OSNs are only proxies biased by the absence of all those constraint (cultural, geographical and so on) of which we all have experience. All the problem introduced in the previous chapter deal with this observation simply considering those online networks as object "by them self" completely foreign w.r.t. the reality that they are modeling (i.e. an online social network is seen as not necessary related to the real social network). Perhaps this assumption is too binding: those two structure could shown, even if dissimilar at a first glance, some common traits that need to be investigated.
Several studies of social anthropology show that the social connections of a person are organized according to circles of increasing size and decreasing average tightness. Friendships in inner circles are stronger, and shows a higher level of trust. The size of these circles increase approximately according to a factor 3 up to the Dunbar number (equals to 150): a value that we can see as a sort of cognitive capacity limit for humans. If we base our analysis on the Dunbar's theory of intimacy, the first question we have to ask ourself analyzing OSNs' ego-networks[2] is: is there a way to extrapolate the real neighborhood of a node from those proxies?
Our interest is to find a methodology aimed to clean network's datasets from those links that are not relevant (i.e. unused or even fake friendships), a way to discriminate between real and strictly-online connections We have seen how the identification of strong ties could be an interesting problem: now we want to move forward our investigation asking ourself if strong ties discovered on an OSN represent "real" friendships or if there are other ways to improve the representation provided by this proxy removing "extra" links.
One possible strategy is to define something like a *Kernel Community*: a subset of nodes that belong to a community, extracted by an OSN, that is most likely to represent an analogue structure in the real social network. One of the major peculiarity of such communities is that they have to be "stable" (i.e. the flow of interactions that occurs among the nodes that are part of them are resilient at the network's evolution). Finding *Kernel Communities* of an OSN is a way to refine the data for subsequent analysis and could allow us to transfer our results in a smoother way from synthetic networks to real ones.

3.2.2 Multidimensional Community Discovery

The problem of finding communities in complex networks is very popular among network scientists, as witnessed by an impressive number of valid works in this field. A huge survey

[2] An *ego-network* is a network centered upon a single user, the *ego*, that includes all its neighbors, *alter*, and the links among them

by Fortunato [28] explores all the most popular techniques to find communities in complex networks.
Analyzing the structure of multidimensional networks we have already noticed that even basilar definitions (i.e. the strict correspondence between neighbor number and degree) fall apart. As was observed for the multidimensional hub analysis in [14] slightly different kind of structure could express different meaning within the network: even for multidimensional communities this assumption holds. The first questions that we need to ask are: what is a multidimensional community? which exactly is its social interpretation? In every day life we are used to interact with other people in different contexts: we could share work informations, play sports together, live in the same building. Sometimes those contexts are mixed. As an example consider an undergrad student that has for roommate a friend who's play fencing in its own club and attend to some common classes. In this situation the two friends share, with high probabilities, acquaintances in all the three common groups to whom they belong. Ignoring the different dimensions of the network, a community discovery algorithm is likely to discover a huge monolithic community that incorporate the three different groups destroying, in such way, part of the informations useful to discover its real semantic.
If taken into consideration, those informations could allow the discovery process to identify correctly the three overlapping communities (each one of them clearly defined by its own meaning). Moving forward this approach could led to the discovery of hierarchical communities: unveiling all the sub-communities that repeat themselves over different network's dimensions is the first step in order to define a metric that estimate the strength of those structures (similarly to what was done for social ties). Obviously this analysis must be integrated with the *Kernel Communities* in order allow a better interpretation of the community discovery's results.

3.2.3 Community Proximity

Often the datasets available (i.e. GSM and GPS ones) does not explicitly provide a social layer that interconnects the agent observed. Studying mobility data we are able to discover sensitive location (workplaces, houses, schools...) for the set of users analyzed. Given those information and imposing time constraints we are able to build a social co-occurrency graph over the users in order to simulate, and trying to infer, social relationships.
An issue that arise following this approach regards the validation of the social network obtained: is it a good proxy for the real, and unknown, social network? Can we define some measure to assess the similarity of users visiting habits?
Answering positively those questions led to novel approaches to solve mobility problems. Often, in order to optimize the resources available (car, gasoline, time), car-sharing scenario where proposed. The data used as input for car-sharing simulators relies on mobility patterns: if we will be able to provide as well the, induced, social graph the methodologies nowadays adopted could be improved and a more realistic analysis could be made.

3.3 Evolutive Issues

In 2.3 we have introduced the field of "Temporal Analysis" and seen how its instruments began to be applied to understand the evolutive behavior of complex networks. Here we want to propose some problems, or variants of well-known ones, that in our opinion are open for a deeper study. All the definitions given so far in this chapter (i.e. ties strength, skill-based ranking, kernel community) could, and need to, been extended in order to take into account the evolutive nature of networks. Even all the problems introduced in the first chapter can take an advantage by a reformulation that exploit the results of temporal analysis.

3.3.1 Community Evolution

As time goes by people tends to modify their social relationships: travels, job change, rising of new interests are only few examples of the causes that led to a perturbation (and in some case even to the ending) of the interactions and connections. Obviously stronger ties, the few lifelong connections each of us have, are more capable to withstand and adjust themselves in order to overcome those changes. Weaker ties often fall apart.

This dynamic evolution is mostly evident on the social tissue represented by communities. We have described a *Kernel Community* as a structure that, somehow, define the most stable part of a community: this definition, that take care of the time as well as the strength, could be seen as a network *invariant*. Once such stable structures were identified, and described in their semantics, further questions arise: can we identify a person that is part of a community and is likely to abandon it? what kind of feature is needed to possess to join a specific community? can we infer the destiny of a community? will a community hold, even increase is numbers, or is doomed to disappear?

Obviously the data produced by OSNs are not expected to describe the evolution of a real world community. Online the timing for the born and ending of a relationships are quicker than the ones we are used to in our daily experience: nonetheless we can speculate that the mechanisms that regulate evolutions are, to some extent, alike.

In an online social website the result provided by such analysis can be used in order to discover emerging trends (i.e. the growth over time of the Coke's addict community) and plan marketing strategies targeted on those users that are more likely to be interested in joining a particular group or, on the other hand, aimed to retain hesitant users.

This kind of analysis could be extended to more complex communities (i.e. multidimensional ones) and could be a valid support to a wide range of tasks not only related to the analysis of social networks.

3.3.2 Link Prediction and Network Archeology

In 2.3.1 we have introduced the classic formulation of the Link Prediction problem.

This problem tries to capture the future evolution of a network analyzing its past: the vast majority of the works on such topic, however, considers the past of a network static. During the process of predicting new connections is easy to produce an huge number of False Positive (i.e. edge predicted to appear in the future, maybe even with high probabilities, that do not): analyzing how links age is a way to reduce this issue and, as a consequence, to improve the predictive performances.

Working with social networks (and other kind of complex networks) the idea of model their edges as static entities appears an oversimplification: a friendships established long time ago that were not renewed by successive interactions has, intuitively, lesser value when used as input for a prediction algorithm than ones that are more recent or "active". For this reason Link Prediction approaches have to be revised and extended taking care, at least, of this two key points:

- networks have to be represented by a graph (or multigraph, in the case of multidimensional analysis) whose edges are annotated with temporal informations (i.e. timeseries of weighted interactions, strength or other measures);

- all the possible edges of a network must be considered as candidate for prediction (not only the ones that are not present in the training set) because multiple interactions could occur over time between the same couple of nodes.

A preliminary analysis of link prediction in a multidimensional and temporal scenario were studied in [76]. In this paper were analyzed some basic unsupervised approaches, extensions of well-known monodimensional measures (i.e. Adamic-Adar, Common Neighbors). Moving from those approaches that exhibit poor performances (even if exploiting temporal information led to significant improvements) is very interesting analyze if is possible to achieve more precise results in a supervised fashion. Recent works [83] exploit mobility information within classical network measure in order to enhance the predictive power of link prediction approaches. This methodology, based on semantic enrichment of the underlying data structure, is exactly the way we need to follow when we try to overcome such kind of complex social network problems. The history of interactions among users is the key to understand the network's dynamics and try to predict its future and the role that the users will play as time goes by.
Even if the time flows only in one direction our analysis could investigate not only the possible future but even the past of networks. Imagine that we have access only to partial data of a network (i.e. a terrorist network) and that we want to discover missing links, edges that exists but are not provided (or that we are not aware of). Network archeology can be seen as an application of link prediction strategies with the aim of uncover hidden or missing links. Analyzing timings, patterns of communication and informations flows for a set of user could be possible to uncover relationships.

3.3.3 Trust and Privacy

OSNs often propose to their users the possibilities to define privacy policy in order limit the access to their data. On Facebook, for instance, is possible to specify the visibility of each kind of contents that a user decide to share (wall posts, likes, photos, tags) to certain group of friends. This settings often relate to the trust that elapses among each couple of (not necessarily) connected users. However, different kinds of contents needs different levels of privacies exposing less or more sensitive informations: in a dynamic scenario, where interactions took place and friendships evolve, the relationship between trust and privacy necessarily change over time.
Considering a multidimensional evolutive context (in which every dimension represent a peculiar typology of interaction) an interesting problem is to understand how trust propagate over the network and how this impact the privacy choices made by the users. Being able to identify the relation that occur among these two entities, how they feedback each other, is the first step to build a model that tries to suggest, or adjust, privacy rules based on the real interaction expressed by the single users.

Chapter 4

Proposal

As stated in the opening chapter, online generated data are the most valuable assets that can be used to understand and make inference upon interests and habits of users. All those informations need to be structured in order to let data scientists analyze them: for this reason a model, and a theory, that balance the understandability and completeness is needed. Graph theory, within temporal and other semantic extensions, allows to depict a clear picture of the relations existing among the studied entities and to unveil informations otherwise unknown.

My proposal is to investigate how well-known approaches to network problems can benefit from the introduction of temporal informations and the adoption of the more complete model offered by multigraphs in order to improve their performances; at the same time I am interested to investigate to what extent this representation could lead to a better understanding of the surrounding reality.

Analyzing static data often lead to incomplete, even wrong, interpretation of the observed phenomena. Imagine to be a tennis player in the process to hit the ball that have just crossed the net: if able to observe only a single snapshot of the ball motion you are not capable to define where, and when, it is likely to fall and you can't decide how to move yourself in order to hit it. Looking at the whole evolving picture (or at a reasonable subset of frames) your task became easier and, as consequence, the chance of hitting the ball increases. The same observation holds when looking to network structures as well (for instance in social graphs). The more we are able to enrich the model semantically (time, mobility, connection type, users preferences. . .) the clearer is the picture that we obtain. Observing, for instance, the process that describes the birth and growth of communities could be useful to understand the temporal patterns that express the life of a social group and identify what are the ties that are likely to fall apart or led to the rising of new realities. Here the aim is to improve existing studies on complex networks and develop more complete and novel ones able to capture those informations. This revised typology of complex network analysis could be seen as a novel *Social Data Science* that tries not only to observe frequent regular patterns on the model, but also to propose interpretation guided by the extended semantics of the founded results. Spatio-temporal analysis and data enrichment are the keys to overcome the gap that exist between the real and the online world: in a society that relies on online services more and more, this new science could offer a privileged point of view on the evolution of its habits.

During my thesis I would like to study how structural informations (i.e. multidimensional tie strength and node ranking) and topological ones (i.e. communities and their semantic extensions) could be used in order to perform an evolutive analysis of social networks. Obviously, how to compute those informations must be considered as an issues to be explored in order to build a solid theoretical basis for the study of evolution in social networks.
During this first year I have worked on a subset of the open problems previously introduced: in particular the following papers were produced[1]

- Structural Analysis:
  - Giulio Rossetti, Luca Pappalardo, Dino Pedreschi *"How well do we know each other?: Detecting tie strength in multidimensional social networks"*, ASONAM CSNA 2012 IEEE, Instambul
- Topological Analysis:
  - Michele Coscia, Giulio Rossetti, Fosca Giannotti, Dino Pedreschi *"DEMON: a Local-First Discovery Method for Overlapping Communities"*, KDD 2012 ACM SIGKDD , Beijing
- Evolutive analysis:
  - Giulio Rossetti, Michele Berlingerio, Fosca Giannotti, *"Scalable Link Prediction on Multidimensional Networks"*, ICDM DaMNET 2011 IEEE, Vancouver
  - Giulio Rossetti, Michele Berlingerio and Fosca Giannotti, 2011. *"Link Prediction su Reti Multidimensionali"*, SEBD 2011, Maratea

Moving from this works I plan to carry on the proposed investigations as follows:

- Structural Analysis:
  Consolidate the metrics builded for the analysis of tie strength in multidimensional scenarios and extend them in order to capture the strength of more complex structure (ego-networks, communities). Study how multidimensionality affect the node ranking problem.
- Topological Analysis:
  Analyze the roles of nodes within a community in order to identify key entities (community kernel). Evaluate the impact of multidimensionality on the community discovery process moving from the community definition given in DEMON[22].
- Evolutive analysis:
  In social networks structural and topological features changes over time: classic problems, as the link prediction one, can be improved if we understand how, when and

[1] Another research paper to which I have contributed during this first year, external to the topics introduced in this thesis proposal, was:

- Mirco Nanni, Roberto Trasarti, Giulio Rossetti, Dino Pedreschi *"Efficient distributed computation of human mobility aggregates through User Mobility Profiles"*, UrbComp KDD 2012 ACM SIGKDD, Beijing why, these changes take place. Observing how network's structure, at the right level (perhaps the meso-level of communities), change over time I plan to build mechanistic models that could help us understand how different events affect a social network over time (i.e. community life-cycle, rising of new edges, diffusion of informations, trust-privacy issues. . .)

Due to its nature, the proposed work is affected by the availability of datasets that express the adequate level of information needed for the analysis. In order to overcome this issue I plan to make use of some interesting datasets that are available to my research group, the KDD Lab at ISTI-CNR of Pisa[2]. Some examples of network datasets available (or collectable) are the Facebook, Flickr, Last.fm, Foursquare and Twitter datasets containing social informations, habits (i.e. weekly information regarding listened music in Last.fm), and interactions among users of the services. Other datasets of interest for the proposed analysis are call-graphs of Telcos (at the moment not available locally) as well as gsm and gps datasets. All those datasets, especially the ones gathered from OSNs, can be mixed together in order to build a more complex and exhaustive view of the (online) social life of a specified set of users.

[2] Knowledge Discovery and Data mining Laboratory, http://www-kdd.isti.cnr.it/

Bibliography

[1] F. Abel, N. Henze, E. Herder, and D. Krause. Interweaving public user profiles on the web. In *UMAP*, pages 16–27, 2010.

[2] L. A. Adamic. The small world web conference. In Berlin: Springer Verlag, editor, *Proceedings of the European Conference on Digital Libraries*, number 443, 1999.

[3] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[4] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman. Search in power-law networks. *CoRR cs.NI/0103016*, 2001.

[5] R. Albert, H. Jeong, and A. L. Barabasi. Diameter of the world wide web. *Nature*, 401:130–131, 1999.

[6] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of behavior of small world networks. In *Proc. Natl. Acad. Sci. USA 97*, pages 11149–11152, 2000.

[7] A. Vespignani B. Goncalves, N. Perra. Modeling users' activity on twitter networks: Validation of dunbar's number. *PLoS One*, 2011.

[8] G. Bachi, M. Coscia, A. Monreale, and F. Giannotti. Classifying trust/distrust relationships in online social networks. In *SocialCom*, 2012.

[9] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. In *Science*, volume 286, page 509, 1999.

[10] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica*, (331):590–614, 2002.

[11] A. L. Barabasi, H. Jeong, R. Ravasz, Z. Neda, T. Vicsek, and A. Schubert. On the topology of the scientific collaboration networks. *Physica*, pages 590–614, 2002.

[12] M. Berlingerio, M. Coscia, and F. Giannotti. Mining the information propagation in a network. In *SEBD*, 2009.

[13] M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale, and D. Pedreschi. Foundations of multidimensional network analysis. In *ASONAM*, pages 485–489, 2011.

[14] M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale, and D. Pedreschi. The pursuit of hubbiness: Analysis of hubs in large multidimensional networks. *J. Comput. Science*, 2(3):223–237, 2011.

[15] M. Bilgic, G. M. Namata, and L. Getoor. Combining collective classification and link prediction. In IEEE Computer Society, editor, *ICDMW*, pages 381–386, 2007.

[16] B. Bringmann, M. Berlingerio, F. Bonchi, and A. Gionis. Learning and predicting the evolution of social networks. *IEEE Intelligent Systems*, 25(4):26–35, 2010.

[17] M. J. Brzozowski, T. Hogg, and G. Szabó. Friends and foes: ideological social networking. In *In Proc. 26th CHI*, 2008.

[18] M. Burke and R. Kraut. Mopping up: Modeling wikipedia promotion decisions. In *In Proc. CSCW*, 2008.

[19] R. S. Burt. Structural holes and good ideas. *American Journal of Sociology*, 2004.

[20] S. A. Catanese, P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Crawling facebook for social network analysis purposes. In *In Proc. of WIMS*, pages 52–59, 2011.

[21] M. Coscia, F. Giannotti, and D. Pedreschi. A classification for community discovery methods in complex networks. *CoRR*, abs/1206.3552, 2012.

[22] Michele Coscia, Giulio Rossetti, Fosca Giannotti, and Dino Pedreschi. Demon: a local-first discovery method for overlapping communities. In *KDD*, pages 615–623, 2012.

[23] G. Das, H. Mannila, and P. Smyth. Rule discovery from time series. In *KDD*, 1998.

[24] P. Erdos and A. Renyi. On random graphs. *Publicationes Mathematicae*, 6:290– 297, 1959.

[25] L. Euler. Ad geometriam situs pertinentis. 1736.

[26] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *ACM SIGCOMM Computer Communication Review*, 29(4), 1999.

[27] E. Fama. Efficient capital markets: a review of theory and empirical work. *Journal of Finance*, 1970.

[28] S. Fortunato. Community detection in graphs. *Physics Reports*, (486):75–174, 2010.

[29] J. H. Fowler and N. A. Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over, 20 years in the framingham heart study. *BMJ*, 2008.

[30] D. W. Franks, J. Noble, P. Kaufmann, and S. Stagl. Extremism propagation in social networks with hubs, adaptive behavior. *Animals, Animats, Software Agents, Robots, Adaptive Systems*, 16:264–274, 2008.

[31] F. Giannotti, M. Nanni, and D. Pedreschi. Efficient mining of temporally annotated sequences. In *SDM SIAM*, 2006.

[32] C. Giles, S. Lawrence, and A. C. Tsoi. Noisy time series prediction using recurrent neural networks and grammatical inference. *Machine Learning*, 44, 2001.

[33] A. Goyal, B. W. On, F. Bonchi, and L. V. S. Lakshmanan. Gurumine: A pattern mining system for discovering leaders and tribes. In *ICDE*, pages 1471–1474, 2009.

[34] M. Granovetter. Getting a job: A study of contacts and careers. *University Of Chicago Press*, 1974.

[35] M. S. Granovetter. The strength of weak ties. *America Journal of Sociology*, Volume 78(6):1360–1380, May 1973.

[36] J. Guare. *Six Degrees of Separation: A Play.* Vintage Books, New York, 1990.

[37] R. V. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *In Proc. 13th WWW*, 2004.

[38] G. Guimarães. The induction of temporal grammatical rules from multivariate time series. In *ICGI*, 2000.

[39] S. Horita, K. Oshio, Y. Osama, Y. Funabashi, K. Oka, and K. Kawamara. Geometrical structure of the neuronal network of caenorhabditis elegans. *Physica*, (298):553–561, 2001.

[40] J. Iturrioz, O. Diaz, and C. Arellano. Towards federated web2.0 sites: the tagmas approach. In *In Proc. of the International Workshop on Tagging and Metadata for Social Information Organization*, 2007.

[41] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, (407):651–655, 2000.

[42] E. Keogh and M. Pazzani. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In *Knowledge Discovery in Databases and Data Mining*, 1998.

[43] A. Ketterlin. Clustering sequences of complex objects. In *KDD*, 1997.

[44] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, (46):604–632, 1999.

[45] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 1999.

[46] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: Measurements, models, and methods. In *International Conference on Combinatorics and Computing*, pages 1–17, 1999.

[47] D. Krackhardt and R. N. Stern. Informal networks and organizational crises: An experimental simulation. *Social Psychology Quarterly*, 1988.

[48] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, , and E. Upfal. Stochastic models for the web graph. In *In Proc. 41st IEEE Symp. on Foundations of Computer Science*, 2000.

[49] J. Kunegis, A. Lommatzsch, and C. Bauckhage. The slashdot zoo: Mining a social network with negative edges. In *In Proc. 18th WWW*, 2009.

[50] C. Lampe, E. Johnston, and P. Resnick. Follow the reader: filtering comments on slashdot. In *In Proc. 25th CHI*, 2007.

[51] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (salsa) and the tkc effect. *Computer Networks*, 2000.

[52] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *ACM Conference on Electronic Commerce, ACM*, pages 228–237, 2006.

[53] J. Leskovec and E. Horvitz. Worldwide buzz: Planetary-scale views on an instant-messaging network. Technical report, Microsoft Research Technical Report MSR-TR-2006-186, 2007.

[54] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *WWW*, pages 641–650, 2010.

[55] J. Leskovec, J. M. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD*, pages 177–187, 2005.

[56] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM*, pages 556–559, 2003.

[57] H. Lu, J. Han, and L. Feng. Stock price movement prediction and n-dimensional inter-transaction association rules. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 1998.

[58] Z. Lu, B. Savas, W. Tang, and I. S. Dhillon. Supervised link prediction using multiple sources. In IEEE Computer Society, editor, *ICDM*, pages 923–928, 2010.

[59] H. Ma, H. Yang, M. R. Lyu, and I. King. Mining social networks using heat diffusion processes for marketing candidates selection. In *CIKM*, pages 233–242, 2008.

[60] A. S. Maiya and T. Y. Berger-Wolf. Online sampling of high centrality individuals in social networks. In Springer, editor, *PAKDD*, volume 6118 of Lecture Notes in Computer Science, pages 91–98, 2010.

[61] P. Massa and P. Avesani. Controversial users demand local trust metrics: an experimental study on epinions.com community. In *In AAAI*, 2005.

[62] S. Milgram. The small world problem. *Psychol*, pages 60–67, 1967.

[63] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J. P. Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, (328), 2010.

[64] R. Muhamad, P. S. Dodds, and D. Watts. An experimental study of search in global search networks. *Science*, pages 827–629, 2003.

[65] T. Murata and S. Moriyasu. Link prediction of social networks based on weighted proximity measures. In IEEE Computer Society, editor, *Web Intelligence*, pages 85–88, 2007.

[66] M. E. J. Newman. Clustering and preferential attachment in growing networks. *PHYS.REV.E*, 64, 2001.

[67] M. E. J. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA 98*, pages 404–409, 2001.

[68] M. E. J. Newman. The structure and function of complex networks. In *SIAM Review*, volume 45, pages 167–256, 2003.

[69] M. E. J. Newman and D. J. Watts. Renormalization group analysis of the small-world network model. *Phys. Letters*, pages 341–346, 1999.

[70] J. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi. Structure and tie strengths in mobile communication networks. In *Proc Natl Acad Sci*, pages 7332–7336, 2007.

[71] B. Özden, S. Ramaswamy, and A. Silberschatz. Cyclic association rules. In *ICDE*, 1998.

[72] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Technical Report 1999-66, Stanford InfoLab, 1999.

[73] B. Pang and L. Lee. Opinion mining and sentinment analysis. In *in Foundations and Trends in Information Retrieval*, 2008.

[74] R. Pastor-Satorras, A. Vazquez, and A. Vespignani. Dynamical and correlation properties of the internet. *Physical Review Letters*, 2001.

[75] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles. Winners don't take all: Characterizing the competition for links on the web. *PNAS*, 99(8):5207–5211, 2002.

[76] G. Rossetti, M. Berlingerio, and F. Giannotti. Scalable link prediction on multidimensional networks. In *ICDM Workshops*, pages 979–986, 2011.

[77] C. Schaefer and J. C. Coyne. The health-related functions of social support. *Journal of Behavioral Medicine*, 1990.

[78] X. Shi, B. Tseng, and L. Adamic. Looking at the blogospheretopology through different lenses. In *ICWSM*, volume 1001, 2007.

[79] A. Sidiropoulos and Y. Manolopoulos. A new perspective to automatically rank scientific conferences using digital libraries. *In Information Processing and Management*, 2005.

[80] R. Solomonoff and A. Rapoport. Connectivity of random nets. In *Bulletin of Mathematical Biology*, pages 107–117, 1951.

[81] A. Stewart, E. Diaz-Aviles, W. Nejdl, L. Balby Marinho, A. Nanopoulos, and L. Schmidt-Thieme. Cross-tagging for personalized open social networking. In *HT*, pages 271–278, 2009.

[82] M. Szomszor, I. Cantador, and H. Alani. Correlating user profiles from multiple folksonomies. In *HT*, pages 33–42, 2008.

[83] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A. L. Barabási. Human mobility, social ties, and link prediction. In *KDD*, pages 1100–1108, 2011.

[84] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, pages 393–440, 1998.

[85] A. Weigend and N. Gershenfeld. *Time Series Prediction: Forecasting the Future and Understanding the Past.* Addison-Wesley, 1994.

[86] S. H. Yook, H. Jeong, and A. L. Barabasi. Modeling the internet's large-scale topology. In *Proceedings of the National Academy of Sciences*, number 99, pages 13382–13386, 2002.