UNIVERSITÀ DEGLI STUDI DI MILANO

SCUOLA DI DOTTORATO IN
Informatica

DIPARTIMENTO DI
Informatica

CORSO DI DOTTORATO
Informatica
XXVI° Ciclo

From Small-Worlds to Big Data: Temporal and Multidimensional Aspects of Human Networks

INF/01

Dottorando:
Matteo ZIGNANI

Relatore:
Dott.ssa Sabrina Tiziana GAITO
Coordinatore del Dottorato:
Prof. Ernesto DAMIANI

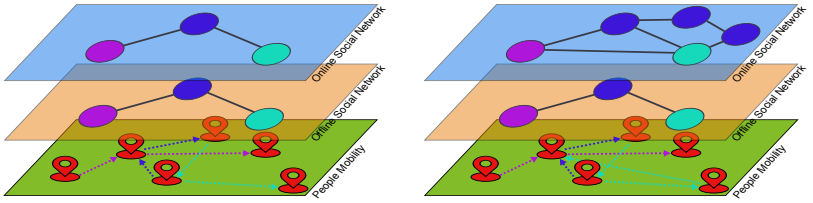
Anno Accademico 2012/2013

To my wife Alessandra

Contents

Chapter 1
Introduction

In the last decade we have witnessed a rapid growth of available data in a wide range of disciplines, from biology to computational social science. Generally, in a few years we went from small-world datasets, describing small realities, to big data. Among the different tools for the analysis of such huge amounts of data, *network science* is proving to be one of the most promising. The power of the approach lies in the efficient organization of the data, in term of memory and time efficiency of some basic operations (neighbors or proximity queries), and how they interact. Specifically, network science focuses primarily on capturing the static structure of the network formed by the constitutive objects of the system. The "linked" perspective enables the emergence of a series of general properties common to real networks in different fields, from the fat tailed degree distribution to small-world characteristics and modular structures.



(a) Scenario at time t (b) Scenario at time $t + \Delta t$

Fig. 1.1 A schematic representation of the interplay among the different dimensions a person may act on. At the first level people move, at the second people interact face-to-face and in the third they communicate through online social media. From time t to time $t + \Delta t$ the mobility patterns of the user C change maybe due to the new relationship s/he has established in the online world.

Due to the exponential diffusion of email, mobile devices and especially online social services these advances have become more apparent in the study

of the human behavior and the human networks, intended as the networks where the vertices represent individuals and the links some interactions. The results obtained by network science on this type of data are extremely interesting although the understanding of human behavior is still far from being complete and many questions remain open. Indeed, how people behave is influenced by a combination of different factors which derive from the diverse dimensions where they are embedded. For instance, the friendship acceptance in an online social network might depend on the strength of the acquaintance, if the people have met in some physical places, if they share the same interests, or if someone through other media has informed the person who might accept the request, that the requesting account is fake or malicious.

In general we can gain more knowledge about human behavior by assuming a stratified reality where individuals act as bridges cutting across the various levels and where their decisions are based on information coming from each dimension. The reference scenario of this thesis is depicted in Fig.1.1 where we consider three fundamental dimensions: human mobility, offline or face-to-face interactions and online social networks. In Fig.1.1(a) how people move and the place they visit result in real-world interactions [1] on the 'offline' plane, interactions which in turn could be reported in the online world.

Time further complicates this already complex scenario. In fact, human beings and the world they live in are dynamic systems which evolve for reasons still not understood in depth. In Fig.1.1(b) we can note that the closure of the triangle among A, B and C changes the C's mobility behaviour who visits the place shared by A and B. Meanwhile, C and B establish new connections in their online dimensions. In general we have to deal with a complex interplay among different levels which results in an interrelated set of human networks whose connections are still unexplored. Obviously, in no way do we aim to achieve with this thesis an all-encompassing human behaviour model. Nevertheless, we begin to consider the interplay of a small subset of dimensions which may also impact many aspects in the realm of computer science: e.g. mobility, real-world social interactions, online social media growth.

Peoples mobility has greatly improved of late thanks mostly to faster transport systems and better commuter facilities and strategies. Along with the widespread diffusion of GPS-equipped mobile phones, this has resulted in a large amount of mobility data. The research community has made a considerable effort to explore and grasp mobility patterns on the basis of intrinsic human behavior, how people move and interact during their daily life. While early works focused on the analysis of individual user trajectories, investigating for instance their most visited places [145], their pattern regularities [146] or on the opposite the detection of outlier locations in their movements [55], some recent studies have highlighted the strict correlation between the social space and the physical places where people move and interact. On

[1] In this thesis by the term real-world we mean the offline sphere. In this context we do not debate if one dimension is more important and informative than other dimensions, or if offline interactions are more complex or stratified [127] w.r.t. to virtual (online) ones.

the one hand individual mobility patterns shape and impact the social network, as links are often the consequence of spatial proximity in shared social locations. On the other hand nowadays people live in symbiosis with their favorite social media, seizing the opportunities provided by these services to enlarge their social circles. For example, business services such as LinkedIn create new professional opportunities which may lead to a change of jobs, while social-oriented networks, like Facebook, offer interactions that may inspire participation in potentially behavior-modifying activities and/or events. These observations make the interplay between individual or group mobility and social media a key factor in the comprehension of human behaviors.

As pointed out by the reference scenario in Fig.1.1 real-world phenomena may require rich representations where the entities under study may relate to one another for different reasons. In the case of real-world social networks a person can establish many kinds of relationships with hundreds of different people. Not all these people can be considered friends, some of them are relatives, colleagues or merely acquaintances. In this scenario we can add a further level of complexity if we also take into account interactions which occurs in online social networks. In this case the chances for enriching the social sphere increase. In fact, not only do people maintain relationships explicitly, for instance by creating a "friendship" link or by "following" a few friends, but two users may be connected implicitly thanks to a common set of of interests without sn explicit physical interaction. For example they share the same groups or the same media, or they like the same photos or status; however while in the offline sphere the sharing of a common interest imlpies a physical contact, online interaction may remain only virtual. Moreover, different ways of interacting may reflect different types of relationship or different values of the same relationship. In all these cases, considering only one kind of interaction is no longer enough;, therefore we need a more powerful representation that captures the multiple connections between any pair of nodes. Multidimensional networks [16, 100], *i.e.* an ensemble of networked layers in which the nodes may be connected through different relations, they seem to be the best candidate for modeling and analyzing the interplay among the different levels of interaction. Due to their expressiveness, they have been adopted in different problems (such as link prediction [139], community detection [18] or information diffusion and cascading [58, 161]), always outperforming the classic approaches on the aforementioned problems. Nevertheless multidimensional networks have yet to be fruitfully exploited. For instance, they could be integrated in the study of human mobility as tools for capturing the reasons behind the movements or for understanding how people interact face-to-face or via social network services.

Besides the different type of relationships, social systems are characterized also by a high dynamic which reshapes their internal structure. For instance, online social networks like Facebook or Twitter keep growing. Moreover their topologies keep changing because of the new services and features they introduce or due to how people behave when managing their own social capital.

These phenomena advocate a deep understanding of the mechanisms that determine the growth of social media and that govern how they evolve over time. Mastering these mechanisms not only allows us to generate synthetic models of social networks but also make the systems more predictable and controllable. In addition, in a big picture where elements are highly interconnected and influenced by events external to the system, temporal information is suitable for detecting and quantifying the consequences of occurrences. For instance by tracking the new links created after an event, we would be able to evaluate its impact on the social structure, its influence and the duration of its effects. Growth and evolution are not the only dynamical aspects. In fact interactions among people keep changing. For example, every day we meet or contact different people, but our interactions are not permanent and last briefly. From a modeling perspective, this means that links might pop up and then disappear. So we ask how the interaction between pairs of users varies in time. Earlier results have shown that interactions in a few social media shift away from the classic Poisson model, even though we are unable to determine if this represents a universal feature or if it is limited to some particular media.

In this thesis we explore this complex scenario, in particular identifying and facing three main open topics:

- **Human Mobility and Social Interaction: Patterns and Models**. The interplay between mobility and social networks occurs in places where people manifest a common interest. In order to study this superposition, first we need to extract these social locations from the available dataset and report them into a framework able to reproduce a structure comparable real-life ones. Specifically we need to face the following questions. How can we measure the social value of places? Do shared places give people the same chance to create and maintain a relationship? How can we embed this information into a model able to properly reproduce a structure comparable to real ones?
- **Multidimensional Networks for Offline/Online Social Interaction**: Through (possibly mobile) social service venues, people start to interact by communicating through different types of media. Indeed, every person is embedded in real and virtual social networks. Do these structures coincide or do they reproduce different ways of interacting? We must answer these questions if we want to use data from the online world to make claims about real social behavior.
- **Growth and Temporal Evolution in Online Social Networks**: Online social services keep growing thanks to new users and links; densifying and shrinking. The mechanisms and their relative evolution are still unclear. Moreover, these systems are not isolated; they co-evolve with other surrounding systems. The role of the available temporal information is essential to the understanding of how users enlarge their social circles and of the effects of external events on them.

1.1 Human Mobility and Social Interaction: Patterns and Models

Understanding individual mobility patterns is of fundamental importance. People move driven by many factors, from a particular business agenda to the need to socialize and interact with others. Consequently, mobility takes a wide lead as one of the main indicators of human behavior. Nevertheless, actual comprehension of human mobility patterns would not be possible with the continuous availability of large-scale datasets recorded by different technologies, from dollar bills registration to mobile phone calls and GPS devices all the way to location-based services. Most of the seminal works on human mobility patterns have been produced by Barabasi and his team, as they were the first to bring the big data opportunity [60, 145, 146]. By mining a large amount of mobile phone trajectories they changed the paradigm in modeling human mobility. They shifted away from classical Brownian motion, characterized by a Gaussian displacement distribution, and focused rather on the Levý process [25], which makes far distances more likely. They observed, in fact, that many mobility features are characterized by a heavy tail distribution. For instance, the distribution of displacements Δr between two consecutive user positions, along with the distribution of visiting time (the interval Δt a user spends at one location), are well approximated by a truncated power-law.

All these early studies centered on the mobility features of a single individual, meanwhile neglecting any type of interaction among people. Only recently scientists begun looking at the interplay between human mobility and social networks, questioning how human mobility patterns shape and impact our social relationships, and how our social sphere drives our movements. The reasonable assumption is that spatio-temporal proximity makes social proximity more likely [33]. Indeed, most of the social links are a consequence of visiting the same places, which often correspond to communities forming around a shared interest. These shared social foci and face-to-face interactions, represented as overlap of individuals trajectories, are expected to have a significant impact on the structure of social networks. In fact, by analyzing a large dataset from Flickr, Crandall *et al.* [38] found that even a very small number of co-occurrences can lead to greater probabilities of a social linking/interaction.

In **Chapter 2** we explore in depth the strict relation between human mobility patterns and the social structure induced by the superposition of different people's trajectories. In particular we adopt a data-driven approach at a fine-granularity scale that exploits publicly available GPS traces gathered in different urban and metropolitan areas. We analyze these trajectories having in mind the following issues:

- **Shared place extraction**. A preliminary step towards the comprehension of the interplay mobility/sociality is a *trajectory processing which*

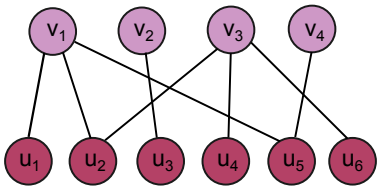


Fig. 1.2 Bipartite network users (red circles)/geo-communities (purple circles).

efficiently extracts shared social foci from raw GPS points. In fact we assume that we have no further social information (labels, semantics, etc...) about the users and the places.

- **Plausible social structure from mobility**. *Given two individuals who share multiple geographical places to what extent can we generate a plausible social structure with the real interactions?* The question implies the need for a proper modeling tool that can express the relation between users and shared foci.

In facing these issues, we have made the following contributions:

- As a first step we formalize the concept of interesting places for a single person, which we call *geo-location*, and the concept of shared social locations, which we call *geo-communities*. On the basis of these definition we develop a time efficient procedure for extracting geo-locations and geo-communities from real GPS traces.
- Through the concept of geo-community we model the human mobility adopting a bipartite graph $G = (U, V, E)$. As reported in Fig.1.2, the presence of a person $u \in U$ in geo-community $v \in V$ is indicated by the link (u, v). Thanks to this graph representation we can generate a social structure that is plausible w.r.t. the real interactions by adopting some graph projections based on classical random graph models. Finally we evaluate the obtained random projections by comparing the typical features of complex network analysis with real-world ones.
- We analyze the human mobility patterns from the geo-location/geo-community points of view, confirming the most common properties such as visiting time and displacement even at the finest scale of granularity. We also introduce some new features like the classification of the most significant places by the visiting frequency and the choice of the next destination as a function of the rank on time or distance.

Above and beyond the social implications, mobility can be exploited to spread local information through the leveraging of the face-to-face interactions. In this scenario contents can opportunistically move on the people's network for their own purposes by using human mobility as a transport mechanism. These opportunities have led to the idea of Opportunistic Network.

Opportunistic networks [130] are highly dynamic, infrastructure-less networks whose essential characteristic is a possible absence of an end-to-end communication path. These kinds of networks gain increasingly in importance as mobile carriers and devices become equipped with short-range radio capabilities. A fundamental issue in the design of these opportunistic services is the evaluation of the underlying protocols under practical and real mobility conditions. Mobility simulation seems to be the best method for evaluating situations with a variable and high number of nodes. Therefore, it becomes necessary to develop mobility models which can simulate different aspects of human mobility. Several mobility models that overcome in expressiveness and complexity the classical Random Waypoint [29] have been proposed recently. Some carefully reproduce many mobility patterns found in real traces yet focus on individual behavior and neglect all social aspects; others consider sociality as the only factor influencing human movements and are not able to reproduce the spatial features [7]. Despite recent efforts, the main research problem in mobility models is still open.Namely it concerns *the need for a simple (few parameters, easy setting) and realistic type of mobility that should be validated with traces from different scenarios and that reproduces the main spatial, temporal and social properties reported in literature [111].*

In **Chapter 3** we present a new human mobility model for urban and metropolitan areas based on the statistical analysis of the human mobility patterns addressed in Chapter 2. The model essentially relies on the graph in Fig.1.2 and on the node movement among geo-communities. We assume that the choice process of the next location is generated by a finite time homogeneous Markov chain where the states are the geo-communities linked to the moving node and where the transition probability is a function of the rank on their distances. Within a geo-community we adopt a particular variety of nonuniform random waypoint model similar to the Lévy walk. Furthermore we validate the model by comparing the generated synthetic mobility traces and the contact properties with real measurements from three different viewpoints (single node, pairwise and contact graph). Finally we show that the model reproduces the temporal, spatial and social features of human mobility.

Chapters 2 and 3 are based on the following pubblications:

- Matteo Zignani, Sabrina Gaito and Gian Paolo Rossi. 2013 "Extracting human mobility and social behavior from location-aware traces", *Wireless Communications and Mobile Computing*, 13(3):313-327
- Michela Papandrea, Matteo Zignani, Sabrina Gaito, Silvia Giordano and Gian Paolo Rossi. 2013. "How many places do you visit a day?". *In Proceedings of International Workshop on the Impact of Human Mobility in Pervasive Systems and Applications* (PerMoby '13).
- Matteo Zignani. 2011. "Human mobility model based on time-varying bipartite graph" *In Proceedings of World of Wireless, Mobile and Multimedia Networks* (WoWMoM '11).

- Matteo Zignani. 2012. "Geo-CoMM: A geo-community based mobility model". *In Proceedings of the 9th Annual Conference on Wireless On-demand Network Systems and Services* (WONS '12).

1.2 Multidimensional Networks for Offline/Online Social Interactions

While the classic network science approach calls for a single connection between the elements of the system, especially in social networks a relationship may develop due to different motivations or it may depend on the context where an interaction occurs. Nowadays, because of the widespread diffusion of online social networks, people have more possibilities to interact and increase their social capital. In fact, not only do people maintain common relationships deriving from physical proximity, through friendship links for instance, they may connect with one another in other ways (e.g. by sharing activities, participating in groups, appearing in photo albums, etc.). As a consequence of these potentialities, today's social media no longer mirror offline sociality; rather, they are drifting towards a highly connected, multidimensional social graph whose friendship connectivity is becoming flat and unstructured. This macrotrend poses some limitations in adopting online social networks as an unbiased sample of the human social behaviors in the physical space.

The above scenario sparks specific research interest on the interplay between offline and online sociality, by which we mean respectively the interaction that occurs in the physical world and through online social services. In particular, we are interested in resolving the following research issues:

- The flattening of the online social network is merely hypothetical. *We need to verify if social services are effectively shifting away from offline interactions; similarly, we must measure to what extent they coincide or overlap.*
- Expanding our neighborhood through online social networks is surely much easier, since the creation of new connections can be unilateral. This way a person may easily modify the importance and centrality of his/her role, unlike in the offline context, where maintaining many relationships is quite difficult. *Consequently, we need to quantify and measure if a person plays a central part in both networks or if the different behavior s/he adopts influences her/his role on the social media front.*
- In addition to acting and moving on different layers, people represent the link among levels. Thus *we must understand if the information about a dimension can be transferred from one level to the next and also to what extent the transferral modifies our knowledge about a single layer.* For instance, if we enter photo co-tagging info on a group layer in Facebook we can grasp whether or not the group members really meet in some place.

The challenges posed in seeking to answer these questions are of both experimental and theoretical natures. On the one hand, although along with extensive literature large datasets describing online social networks have recently been made available, the research community has very few opportunities to compare the datasets of offline encounters and online relationships re the same group of individuals. On the other hand, we need modeling tools and measure to quantify the changes that occur when moving from one level to another.

In **Chapter 4** we face both problems and offer the following contributions:

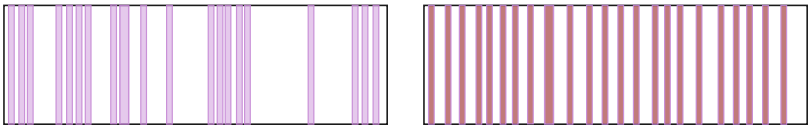
- We collect data describing the offline sociality of a set of volunteers in a time span of one month and then we integrate them with relevant data about their online sociality extracted from Facebook. On the contact set we make a covariate analysis to highlight the main dependences. In particular we focus on the influence of the location context on the meeting properties (people involved, type of the relationship).
- We describe each layer of the group's sociality through the associated complex network and then we superimpose them. First we analyze the networks separately, searching for global similarities such as small-worldness or modularity and then we make a systematic comparison of the social layers. We show that the overlapping degree is low; in fact, the Facebook and real-life contact sets vary significantly. The two networks do not coincide; moreover, we find no one-to-one correspondence between the communities extracted from them. We also face the centrality problem, discovering that node centrality is not a universal feature. In fact, node centrality is not linearly transferred across layers and consequently a person's popularity is most likely to change in different networks.
- Finally, we introduce a *unified complex network* which allows us to study the effects on the online community structure of transferring offline information in the online world.

To the best of our knowledge, this is the first research project addressing the above-mentioned issues.

Chapter 4 is based on the following publications:

- Sabrina Gaito, Elena Pagani, Gian Paolo Rossi and Matteo Zignani. 2012. "Sensing multi-dimensional human behavior in opportunistic networks". *In Proceedings of the Third ACM International Workshop on Mobile Opportunistic Networks* (MobiOpp '12).
- Sabrina Gaito, Gian Paolo Rossi, and Matteo Zignani. 2012. "Facencounter: Bridging the Gap between Offline and Online Social Networks". *In Proceedings of the 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems* (SITIS '12).
- Matteo Zignani, Sabrina Gaito and Gian Paolo Rossi. 2013. "Online and Offline Sociality: A Multidimensional Complex Network Approach". *Complex Networks and Their Applications*. Cambridge Scholars Publishing, UK (In press).

1.3 Growth and Temporal Evolution in Online Social Networks



(a) Bursty sequence (b) Poisson sequence

Fig. 1.3 1.3(a) Example of a bursty sequence of events; periods of high activity are interleaved with long pauses.1.3(b) Example of a Poisson sequence where the average rate of the events is constant.

The reference scenario in Fig.1.1 shows that online social networks are not isolated systems; rather, they co-evolve with other dimensions [168]. This phenomenon has been widely studied in literature, for instance in election results forecasts [54, 152], in news broadcasting on different media [91, 113], in the relation between trending topics and events or in information cascading [131, 119]. Nevertheless, most of these works focus on external events linked to processes which spread on a static topology. The study of the effects of external events on the network topology evolution is more challenging task due not only to the substantial lack of temporal data about online social networks but also and above all to how the temporal data have been analyzed. Usually researchers look at network growth from the system viewpoint describing how it evolves in function of the number of nodes or links. This paradigm based on the so-called logical time proves ineffective if, for example, we need to related the users' behavior w.r.t enlarging their network with other human dynamics occurring on systems close to the OSN. The physical time is more suitable for reaching our goal as it simplifies the connection between temporal trends (observed in the network) and external events. Moreover, understanding network dynamics in the context of physical time is a critical first step towards a predictive approach in the infrastructure management of online social services. In particular, dynamics involving edge creation have direct implications on strategies for resource allocation, data partitioning and replication [133, 163, 151].

In adopting the physical time approach our primary research question is:

- *Which processes characterize the evolution and growth of online social networks? How do people expand their circle of friends? Can we apply recent results re human dynamics to this kind of network?*

This last question is quite fundamental in human dynamics, as a positive answer will proves a human temporal behaviour common to many dimensions.

In fact recent studies have shown that human dynamics in many contexts are best described by periods of rapidly occurring events interleaved with long periods of inactivity, commonly denoted as *bursts* [10, 154, 76, 82] (see Fig.1.3). In **Chapter 5**we answer the above questions, introducing the following contributions:

- First we highlight the fact that the link creation process in online social networks is bursty. We do so by analyzing a large timestamped dataset describing the initial growth and evolution of a Chinese online social network.
- To understand the structure of the edge creation process we propose a new methodology able to detect bursts, their internal structure and the transitions between the different phases a node experiences. This approach is new, since known techniques for the analysis and detection of burst events focus on locating a burst when it occurs, whereas they do not consider events inside the detected temporal window.
- Finally, we apply the detection of bursty phases on the social network and find that all nodes exhibit similar patterns over time. In particular, they are characterized by an intense burst of activity shortly after joining the network. The initial burst is followed by weaker bursts over time, each consisting of an acceleration phase, followed by a longer period of slowly vanishing deceleration.

The discovery of highly bursty patterns paves the way for new generative models that not only capture graph dynamics in terms of phases of node activity but also describe such events w.r.t. physical time.

A new perspective in considering online social services is emerging, according to which they are considered as contagious networks and no longer as social ones. By the term contagious network [143] we mean networks that grow by adding new users through the exploitation of social ties external to the system per se. Temporal information on network growth, *i.e.* on the diffusion of the contagion, can be exploited both to study how the process spreads and to improve our understanding of the mechanisms that take place during the contagion. Since contagious networks are intrinsically dynamic, the timing aspects represent a natural way to infer the social ties in the external system. In fact, reasonably enough, the stronger the tie the faster the potential link will materialize. To shed some light on how networks evolve and how external events shape them, in **Chapter 6** we analyze the dynamics of the basic microscopic building blocks of online social networks: namely, *dyads* [156] and *triads* [62]. We summarize the resulting contributions as follows:

- Within a physical timing microscopic approach, we propose two new metrics: the link creation delay and the triangle closing process delay. The *link delay* is the time used by two users of a network to create a link between them. This accounts for the time a potential friendship waits to become actual. Instead the *triangle delay* accounts for the time used by three users

to all become friends. These new metrics enable us to study the dynamics of creation of dyads and triads on a sample of the Facebook graph and highlight network behaviors that would remain hidden when not-timing measures are adopted.

- Our main results on dyads are the following. Link delay is very low, especially in the early stages of a social network, accounting for the fact that if two persons wish to establish a friendship relation they do so promptly. Moreover, we show that link delay is independent of the precise date when people join the network, so highlighting the purely social feature of link delay. Finally, we find that link delay is relevant in interaction analysis, for it is a good indicator of the level of interaction between nodes.
- Regarding triads we introduce an algorithm for the extraction of the temporal triangle. This allows us to monitor and detect sudden changes in the triangle formation behavior, possibly related to events external to the network. In particular, we show that the introduction of Facebook's "People You May Know" (PYMK) functionality had a disruptive impact on the triangle creation process in the network. By analyzing the triangle delay properties we show that triad formation is very fast, accounting for the fact that if two persons have friends in common and are willing to form a relationship they will do the latter promptly. Yet, this new metric highlights slightly different types of behavior that depend on how long the network has been in operation. In fact, in the bootstrap phase of Facebook dataset the triangle formation dynamics are faster w.r.t. the Facebook consolidation phase.

Chapters 5 and 6 are based on the following publications:

- Sabrina Gaito, Matteo Zignani, Gian Paolo Rossi, Alessandra Sala, Xiaohan Zhao, Haitao Zheng and Ben Y. Zhao. 2012. "On the bursty evolution of online social networks". *In Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research* (HotSocial '12).

1.4 Conclusions

In this thesis we have studied some human networks at different dimensional scales and temporal granularities. We move from small datasets about the interaction of human-carried devices to large-scale graphs that evolve over time (e.g. Renren and Facebook), as well as from human mobility trajectories to face-to- face contacts. The main goal of our work is to gain a deeper understanding of human behavior by exploiting the close interplay among mobility, offline relationships and online interactions. One of our primary contributions is the formalization of elements linking the different levels. Re human mobility and social interaction, the definition of shared locations in terms of

denser overlap regions and their modeling by an affiliation network represent the basic building blocks for generating social structures that are plausible w.r.t. the characteristics of the graph of the contacts. Our modeling approach also results in a mobility model, which name Geo-CoMM, that reproduces the main spatial, temporal and social features observed in real contact and mobility traces. As regards the relation between offline and online sociality we attribute the problem to the multidimensional network framework. In this framework we easily assess the overlapping degree of the sociality levels from a local (neighborhood) and a modular point of view. Finally, we give a definition of some temporal local metrics which measure the time it takes for a link to materialize and the time it takes for a triangle to form. These metrics have been proven useful in predicting possible interactions via online social network and evaluating the effects of some external events on network growth.

In addition to the formalization, we adopt several tools from different research fields to obtain some results about human behavior in human networks. In the field of human mobility we exploit a time efficient clustering algorithm to find the shared locations in a group of users. This allows us to confirm, on fine spatial datasets, the properties of human mobility patterns already found by analyzing coarse-grained traces from mobile phone records. In studying offline/online sociality we heavily exploit rank correlation methods, which allows us to assess that a node changes its centrality according to the level he is currently acting on. Obviously, we adopt most of the tools made available by network science. We use community detection algorithms to study the modular properties of the contact graphs for both synthetic and real traces. By comparing the obtained results we validate the mobility model on the modularity property and on the features of the link connecting node inside and outside the extracted communities. Community detection also plays a fundamental role in the comparison of online and offline groups and in evaluating how information from the offline world impacts the way users can be divided into communities.

Finally, this thesis provides the analysis of unique datasets. The Renren dataset collects data on the initial bootstrap phase of the network and allows us to study its microscopical evolution as well as user dynamics in the creation of new links. We exploit the absolute time associate to each link to measure the dynamic aspects of basic growth mechanisms such as triadic closure. By adopting the physical time approach we are also able to connect changes in the triadic closure process to external events like the introduction of new services and features. The second dataset we gathered provides information about the offline/online interactions of a group of people enriched by contextual information. Above and beyond the offline/online relation, these data can contribute to the discovery of patterns involving the contexts and different contact features.

In addition to the previously mentioned publications, while working on my PhD I contributed to the following papers:

- Sabrina Gaito, Christian Quadri, Gian Paolo Rossi and Matteo Zignani. 2012. "THINPLE - the New Online Sociality is Built on Top of NFC-based Contacts," *In Proceedings of Wireless Days* (WD '12).
- Corrado Monti, Alessandro Rozza, Giovanni Zappella, Matteo Zignani, Adam Arvidsson and Elanor Colleoni. 2013. "Modelling Political Disaffection from Twitter Data". *In Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining* (WISDOM '13).

In the former paper we describe THINPLE, a mobile application aimed at building a trusted online social space where the growth of a persons sociality is fueled only by contacts in the physical world. Our work is based essentially on the idea expressed in Chapter 4 and on the search for a technology able to detect willing contacts represented by Near-Field Communication (NFC). In the latter one, we study the behavior of the Italian Twitter community with respect to political disaffection, *i.e.* a lack of trust in Italian politicians, a common attitude amplified by the recent Italian political situation. Moreover, we show that the bursty characteristic of this phenomenon is highly correlated to political news re various scandals and the economic crisis.

Chapter 2
Extracting human mobility and social behavior from location-aware traces

2.1 Introduction

Recent faster transportation methods have made people mobility very common for both businesses and daily life. In addition, advances in communication technologies, data analysis and smart infrastructures are enabling to streamline the transportation strategies, simplifying connections and shortening the commuting times. Together these two aspects result in a high mobility degree for many people, making places part of the people's life. Apart from the reasons governing the need of mobility, the movements from one location to another provoke contacts and opportunities to socialize. As a consequence there is a two-way relationship between how people move and how they socialize, interact and create the social structure. This way a person, and the devices s/he holds, is embedded into two layers which are strongly dependent and that co-evolve together; namely, mobility and sociality.

For these reasons in recent years we have seen a growing interest within the wireless networks research community about the concepts of location and community which are proving key points in designing new communication paradigms. In particular most of these techniques compute similarity indexes among carried devices based on the history of the visited locations or exploit the community structure to forward information, or to offload the network infrastructure. In all these applications a location is loosely meant as a point of human interest, while a community is popularly conceived as a group of people sharing common interests, needs and behavior. Having some knowledge about locations, such as what makes them unique and special and how (and what kind of) people patronize and enjoy them, is increasingly important for designing and deploying road side hot spots, city wireless networks, ad hoc and opportunistic networks. Moreover, many emerging mobile computing services are being proposed which are location-, profile- and community-aware [158], [73], [94], [52], [40].

The deployment of these networks and services would greatly benefit from, and in many cases actually requires, basing their design on reliable and quantitative knowledge and predictions of some relevant information, such as which locations are enjoyed by people in their daily life and how people aggregate within communities. This advocates the deployment of tools which combine geographical data mining techniques and social network analysis to extract some universal features of the 'socio-mobility' of the individuals. These tools should also be able to reconstruct or infer the social structure given some information about people's mobility. For example, given the history of the visited location of a population, we might recreate a plausible social network.

Unfortunately, so far neither the concept of location nor the concept of community have been univocally defined. Therefore, their specific features and, more importantly the relationship between the two concepts, are still not fully understood and exploited either. It is commonly accepted that people move during the day among locations and communities. Nevertheless, locations are just intended as points in a simulation area. Their characteristics, like size, distribution, number and population density and how people choose the next destination are still not known in full. Similarly, the concept of community has found different definitions in the literature depending on the particular context. In the sociological approach, for example, a community is defined as a group of people sharing some interests; there is the location aware definition in which a community is a group of people meeting at a given location; and the computational science approach where, loosely speaking, a community is a well connected set of people with fewer connections outgoing the group.

In this unclear picture we approach the problem from the most basic of starting points, namely by analyzing those real datasets of human mobility traces which provide detailed geographical information. To this purpose, we use two GPS datasets recently published on the data repository CRAWDAD [137] (NCSU dataset) and within the Microsoft Geolife Project [132]. On the basis of these elementary and commonly available data, in this chapter we provide a few relevant contributions. First we derive a deep understanding of the term "location" and at the same time of the notion of community strictly related to it. Second, we merge the two concepts into what we call geo-community. By proceeding from real spatial data rather than from an a-priori reasoning, we are able to quantitatively describe geo-communities and infer the probability distributions of most of the features of human mobility. These behaviors are relevant to build realistic mobility and aggregation models on which to design new communication paradigms as shown in Chapter 3. Finally, not to lose social implications, we present and validate a graph-theoretic method to derive people sociality from geo-communities.

The chapter is structured along the different steps of the method we propose to process the datasets. First of all, we describe how the points (locations) of interest, what we call *geo-locations*, can be extracted from GPS

traces. Geo-locations are places that attract and interest a person for a few reasons. Then we describe how we discriminate GPS points visited while moving, *i.e.* during the person's movement phase, from the points where the person has spent a relevant amount of time. In order to identify the real locations of a larger group of persons, which name *geo-community*, we propose a clustering method to aggregate individual geo-locations. To clarify let consider the following situation. Let's take a train station, for example. Many people go and spend at least a few minutes there, so making it an individual geo-location for each of them. But to extract the location of the train station we must properly overlap all of them in a suitable way. The idea of geo-community we propose is strictly connected to these locations. People mobility is motivated by the need to go somewhere, more than by some social needs. Social relationships are created by patronizing the same place: people who hang out at the same location have the potential to establish a social relationship or not, depending on whether or not they meet. But for the aims and purposes of deploying network connectivity and services, they belong to the same geo-community. So, a geo-community is both a point of attraction for people, including train stations, workplaces, pubs, etc., and the set of people who go there, maybe at different times. The formal definition of geo-community and the methodology to extract them from GPS traces are described in Section 2.3.

At this point, by statistically analyzing the available GPS traces, we study the aggregated properties of the main quantities involved in human movements. We report results on mobility features considering both the detail level of the geo-locations and the geo-community granularity. The NCSU dataset has been analyzed from the geo-community viewpoint as it records the mobility of a population in a small area. On the contrary, the GeoLife dataset is more suitable for a geo-location perspective because it spans a long period but it gathers data about the mobility in an area too large to extract places shared by many people. In both approaches our main results concern the pause times in a place, the distance covered inside and among points of interests and the choice of the next destination. The latter point concerns how the distance influences the choice of the next geo-community during the daily movements. Furthermore, in the geo-location setting we introduce a classification of the points visited by people which allows us to define two general mobility profiles based on the number of visited locations and the pause time. The details about the analysis are presented in Section 2.5.

So far, our method has been driven by the spatial issues of human mobility. We proceeded backward, from locations to communities. Sure enough, we are aware of the relevance of social relationships in deploying most of the emerging mobile computing services. To this purpose, we propose a methodology to capture human social aspects from spatial data instead of imposing them as a priori knowledge or using the contact graphs. Social relationships naturally emerge by describing people and locations through an undirected bipartite graph also called *affiliation network*. By analyzing its projection on the indi-

viduals set, for each person we can deduce the existence and the strength of his social interactions. This kind of representation enables us to make explicit the social relationships which develop when people spend time in the same location. We use the bipartite network to infer such connections, creating a one-mode projection from the two-mode bipartite form. This methodology and the results obtained comparing different projections and real traces are introduced in Section 2.6.

2.2 Related Work

In recent years many authors have analyzed human mobility patterns from different perspectives. Most of the examined features regard the spatial and the temporal dimensions of the movement. More recently the properties of the contacts induced by mobility have been taken into account.

To gain a better understanding of the dynamics involved in mobility, many experiments, based on different detecting technologies and performed in various locations, have been conducted. Most of them have been made available in the public repository CRAWDAD [137]. Among these datasets we focus on GPS-based traces as they allow us to precisely determine the geographical positions of the contacts. As for the latter aspect, other technologies suffer from some limitations; for instance GSM cell-tower data are constrained by the service area of the cell resulting in a coarse granularity of the records, whereas wireless networks cover smaller regions but their positions and the localization of the attached devices is quite difficult. Recently positional data from location-based social network LBSN (Foursquare, Gowalla,Brightlife) have been collected and analyzed although the reliability and the universality of the obtained results may heavily depend on the propensity of the users to share their positions [166].

Significant location extraction. Part of our work, which involves data mining of geographical data, has been devoted to detect the significant locations of a user from GPS-based traces. Many authors have suggested different extraction methods [8, 64, 80, 169, 170, 171].

Ashbrook *et al.* [8] have proposed a two-step method to infer the significant locations. In the first step, the loss of the GPS signal is used as an indicator of interesting locations because it likely corresponds to buildings or indoor points. In the second step these points are clustered into locations using a variant of the k-means algorithm. In the clustering procedure, round clusters with a given radius are initially placed at k chosen points, and iteratively they move to a denser area, until no further increases in the point density is observed. Depending on the length of the radius the significant locations may shift from cities to single buildings. Since the loss of GPS signal serves as the main clue to identify significant locations, main buildings are found. However other types of interesting locations where the signal keeps to be collected, such

as outdoor places, may be lost. Furthermore rather than detect locations with an arbitrary shape, they retrieve only circular locations. On the contrary, we apply a clustering method able to find arbitrary shape clusters. Moreover, they neglected any social implications of the significant locations and did not mind the analysis of the movement patterns.

Hariharan and Toyama [64] proposed an approach that uses time information to distinguish significant places. From the raw traces they identify contiguous sequence of GPS points within a distance d and for a period t adopting a variation of an agglomerative clustering algorithm. They called these areas 'stays'. However, their algorithm is computationally expensive because the identification of a stay requires the distance between all pairs of coordinates within a specified time window to be computed after every new location measurement. In our approach we choose a more computational efficient algorithm neglecting the temporal information as traces have been recorded with a fixed sample rate.

Kang *et al.* [80] proposed a method, suitable at resource-limited mobile device, that computes incrementally significant locations . Their time-based approach clusters the stream of incoming location coordinates along the time axis and drops those clusters where few time is spent. In particular the algorithm compares each new GPS point with the previous coordinates in the current cluster; if the stream of coordinates is far away from the current cluster then it detects a new location. Authors validate their algorithm with localization data inferred from RF(radio frequency)-emissions of known base stations. As the main goal of the method is the portability on mobile device authors did not investigate the trajectories of multiple users and their social features.

Zhou et *al.* [171] proposed a density- and join-based clustering algorithm called DJ-Cluster to infer significant locations. The dense points are those with at least certain number of other points lying within a distance of their neighborhood. Relaxing the DBSCAN conditions on reachability, the clusters are formed from a set of dense points, which are density-joinable, *i.e.* the neighborhood of the dense points share a common point. A further preprocessing procedure, which removes GPS points corresponding to limited movements, is introduced to improve the performance of the algorithm. The experimental results indicate great improvements in terms of both recall and precision over those obtained from the k-means algorithm.

A similar approach has been adopted by Zheng *et al.* [169], [170]. They applied a density based clustering algorithm (OPTICS [6]) to extract significant locations in order to infer transportation modes and to predict users' preferred locations. With respect to these works, we propose a more general definition of stay-location that enables us to consider temporal reappearances at the same place.

While the previous works concentrated on detecting the significant locations of a single individual, we mainly focus on the detection of the social places combining the density-based clustering approach of the latest works

with an analysis of the visiting time. In particular while the previous authors concentrate on the individual behaviors we extend their idea taking into account the superposition of the traces of many people.

Statistical analysis of mobility. Spatial mobility patterns have been analyzed in different disciplines from physics to pervasive computing. Works from the physicists' community focus on concepts from statistical mechanic and thermodynamic. Their main goal is to identify what kind of diffusion process is able to better reproduce the human mobility. For these reasons they analyze the displacement and the length of the movements, searching for evidences of sub- or super-diffusive processes. On the contrary works from computer science focus more on human mobility properties, that can be exploited in the deployment of different services from opportunistic networks to link prediction in LBSN.

In their seminal work Brockmann *et al.* [25] investigated human traveling statistics by analyzing the circulation of banknotes in the United States. Based on a huge dataset of over a million individual displacements, they found that the distribution of the traveling distances decays as a power law, indicating that trajectories of bank notes are similar to Lévy flights. Secondly, they showed that the probability of staying in a confined region (pause time distribution) is characterized by a long tail leading to a sub-diffusive process.

Gonzalez *et al.* [60] also focused on distances covered by people. In particular they analyzed mobile phone users for a six-month period in a large area. They found that the distribution of the distance between two consecutive calls is well approximated by a truncated power-law. Moreover each individual tends to return to a few frequented locations with high probability.

Rhee *et al.* [138] were the first at dealing with the statistical properties of human mobility using GPS traces. By analyzing the NCSU dataset they reported that bursty hot spot sizes play an important role in causing the heavy-tail distribution of distances in human walk. They show that visit points are clustered and that pause time distribution in hot spots follows a truncated Pareto. With respect to these works we use a smaller spatial scale in extracting hot spots and then analyze distances after dividing them in intra-distance and inter-distance (distance inside and among interesting locations). Lee *et al.* [89] do not consider these distinctions as they do not correlate spatial and social features. Moreover, unlike them, we find slightly different distance distributions.

A recent study posed some doubts about the power law distribution of the distance as a universal feature of human mobility. In fact Noulas *et al.* [118] focused on human mobility patterns in a large number of cities. Mobility data have been retrieved from mobile location-based social services. They first observed that mobility, when measured as a function of distance, does not exhibit universal patterns instead they obtained more general results for all cities considering another variable. Precisely they discovered that the probability of transiting from one location to another is inversely proportional

to a power of their rank, *i.e.* the number of intervening opportunities between them.

Other works investigate characteristics other than distance. For instance Song *et al.* [146] studied the predictability of human trajectories derived from the estimated entropy of the mobile phone data. The predictability is centered around 93% over a large population, independently of the size of the area covered by individuals' mobility or other demographic factors. Likely, the high predictability is obtained based on low resolution positioning data since the average size of a 'location' is roughly 3 km^2. For higher resolution positioning data such as the GeoLife dataset, Lin and Hsu [96] showed that an high predictability is still present at fine spatial/temporal resolutions. However they observed an invariance between the predictability and spatial resolution, *i.e.* we can not obtain a high prediction accuracy and spatial precision simultaneously.

Kim *et al.* [84] used AP log data to extract information about users' movements and pause times but they did not care about location distances in computing users' transition probabilities. They found that pause time and speed distributions follow a log-normal distribution and that the directions of movement follow the direction of popular roads and walkways on the campus showing a symmetry across 180 degrees.

Most of the datasets and of the studies on mobility refer to direct contacts between users induced by their movements. They cover several environments and span different periods. For example, data recording contact in conferences [72], [31], on campus [84], [106] or between people affiliated with a department [90], [44] are available. The features of these traces, such as inter-contact and contact duration distributions, number of neighbors in different periods, periodic reappearances in the same places and cyclic contacts, have been explored in several studies. Most of them confirm some widespread features as the power law with cut-off distribution of the inter-contact time or a near-perfect regularity in the visit of one or more locations. Some of these characteristics will be deepened in Chapter 3 for the validation of the mobility model.

On the one hand our results confirm, at a higher temporal and spatial granularity, some of the statistical patterns the mobility captured by mobile phone call records. On the other hand we find that the distribution of the displacement between the visited points of interest is better approximated by a lognormal distribution. Far from entering in the Pareto/lognormal debate, we only underline that different results may be due to the different scales in the analysis. In fact power-law has emerged in coarse-grained datasets, while lognormal in our fine-grained datasets. Furthermore we analyze the mechanism of the choice of the next destination considering the space as the main factor that impact the decision in urban areas.

Graph-based mobility analysis. Complex network analysis [116] has recently been proposed in the analysis of mobility and human contacts in shared places. In this framework contacts between users are modeled by a

undirected graph, called *contact graph*, with an edge representing past encounters. From the contact graph we can infer many properties concerning the type and the strength of contacts and of social relationships. This brings us to assign to a node a relevance value by measuring some centrality metrics such as degree centrality, betweenness centrality and eigenvector centrality [116]. Many features of the social structure, such as the human tendency to group in communities, have repercussions on the contact graph. As a matter of fact in recent years many community detection algorithms - based on different ideas of community - have been developed [47] and have been applied on social networks to extract real communities, even overlapped [122], and to analyze the social interactions that occur in them.

Complex network and graph theories have not just been applied on the analysis of the social structure induced by the people's mobility but also on its modelling. In fact individuals' mobility is described as a heterogeneous graph, where a set of nodes represents the places visited by people and another disjoint set contains the individuals themselves. A place is connected with a person if s/he visits it. Links could be enriched with other features, such as the duration of transitions and the frequencies. For example the graph representation has been adopted by Zheng *et al.* [170]. They have combined a tree-based hierarchical graph (TBHG) for modelling individual's location histories and a user-location graph to rank the location. The hierarchy in the TBHG is obtained from the single person's locations in a agglomerative manner. For instance smaller clusters are formed on the lower level of the tree, and larger clusters grouping locations of the layer below are on the higher level of the tree. The ranking of the location is computed by running an HITS-like inference model on the user-location graph, where the location set is given by a level of the TBHG of the users' population.

While the previous work adopted a graph representation of the users' mobility it focused on inferring location properties neglecting any social implications. A work more closely related to ours is Onnela *et al.* [120]. In the paper authors constructed a social network of individuals from mobile phone data. They divided the individuals into groups according to a community detection algorithm and compared the found groups with the geographical position of their members. In particular they studied the relationship between topological positions and geographic positions of their members, finding a lack of correlation between the two variables.

With respect to the last works, we adopt an opposite point of view on the problem. In fact we begin from the shared geographical areas and then we generate a plausible social network which is comparable with the real one. In reaching these results a graph-based approach whose aim is to express the relation users/share-places and not the association between a single user and the hierarchy of his visited places as done in [170].

2.3 Dataset

In this section we present the two datasets adopted in our analysis: NCSU and Microsoft GeoLife. Both collections are publicly available and represent the major GPS resources for the understanding of human mobility at high resolution spatial and temporal scales. After describing the datasets, we introduce some pre-processing heuristics which simplify the geo-location and the geo-community extraction. Moreover the aforementioned scale allows us to relate the spatial aspects of mobility given by the geo-locations with the socials aspects given by the possible contacts. Contacts that mainly occur inside the shared social foci given by the geo-communities.

2.3.1 NCSU Dataset

NCSU dataset is made up by 5 groups of traces, each gathered in a different environment. Data cover two campus (NCSU and KAIST), Midtown Manhattan at New York, Orlando Disney World and an American state fair. Traces have been collected by GPS device with a 3 m accuracy. As reported by the authors [138], devices recorded GPS points every 10 s but the released trajectories have a 30 s granularity resulting from a pre-processing procedure to reduce the GPS signal noise. For privacy concerns trace format is not compliant to any standard but each trajectory point is defined by the triple (t, x, y), where t is the time elapsed from the beginning of the recording, while x and y are the offset on the x- and y-axis from a reference point. Both the absolute time of the recording beginning and the GPS coordinates are unknown, so a direct mapping between the detected interesting locations and the real places is impossible. Despite these limitations meaningful statistical results are still possible to be obtained.

As concerns the cardinality of the dataset, NCSU campus traces collect data about 20 randomly selected Computer Science students. KAIST trajectories involve 34 students living at the university dormitory. New York data have been gathered by a group of 12 volunteers living in the Manhattan borough. For the state fair dataset authors choose 18 experimenters who collected data for at most 3 hours, while at Disney World we have 19 traces corresponding to visitors of the entertainment complex who spent there a couple of days.

As regards the temporal coverage in Table 2.1 we report the average duration μ_d, the standard deviation σ_d, the maximum and the minimum duration of the collected trajectories. Data from campuses, New York and Disney World span the daily activities, while state fair traces are limited to the period during which experimenters are inside the park.

Site	users	traces	μ_d	σ_d	$\max_d$	$\min_d$
KAIST	34	91	12	5	23	4
NCSU	20	35	10	5	22	2
New York	12	39	8	6	23	1
Disney World	19	41	9	3	14	2
State fair	18	19	3	1	3	1

Table 2.1 General properties of the groups of traces in the NCSU dataset. μ_d, σ_d, $\max_d$ and $\min_d$ respectively indicate average, standard deviation, maximum and minimum of the duration (in hours) of the traces contained in the datasets.

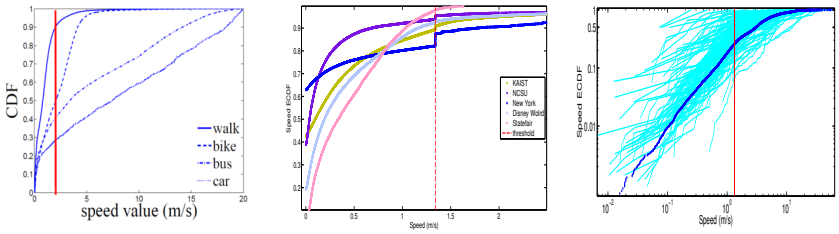
2.3.2 Microsoft GeoLife Dataset

In the analysis of the mobility from the single user viewpoint (geo-location), we also use a very large GPS dataset recording the movement of 178 people in a period of over 4 years (from April 2007 to October 2011). It was collected during the GeoLife Project and released by Microsoft Research Asia [170]. People participating to the experiment are students, government staff and employees from Microsoft and several other companies equipped with GPS loggers or GPS-phones. Overall the dataset provides 17,621 trajectories with a total distance of 1,251,654 kilometers and a total duration of 48,203 hours. With respect to other datasets collected in a limited area or in a particular context, Microsoft GeoLife dataset offers a high heterogeneity. As a matter of fact, it contains a broad range of users' outdoor movements, including both everyday routines imposed by working and free time activities.

Besides having been conducted over a long period of time and involving a high number of users, this dataset is interesting also for its temporal and spatial fine granularity. Specifically, 91% of the GPS trajectory are recorded in a dense representation, *i.e.* every 1∼5 seconds or every 5∼10 meters per point. This allows us to precisely capture significant locations associated to the different activities a user undertakes.

If on the one hand the dataset is very rich, on the other side it exhibits a high level of fragmentation, especially regarding features as the effective duration of the trajectories, the data collection period and the number of trajectories per user. Indicatively, more than half of the trajectories span less than one hour, while about 60% of users collected data for less than a month. Furthermore the dataset covers a large portion of the Earth from Europe to USA to Asia. Nevertheless it is not a problem since most of the data are located in Eastern Asia, in an area corresponding to the region around Beijing. We limit our analysis to GPS data collected in this area, as our main goal is to characterize the most visited places in an urban area.

2.3.3 Dataset pre-processing



(a) (b) (c)

Fig. 2.1 2.1(a) Cumulative speed distribution of different transportation modes. The red bar indicate the value of the walking speed. The figure is from [97]. 2.1(b). Cumulative distribution function of the speed for the groups of traces in the NCSU dataset. The red dotted line indicates the threshold we use, which corresponds to a step in the distribution shape. 2.1(c). Cumulative distribution function of the speed in GeoLife dataset; the red line represent the aggregated sample, while cyan lines correspond to the distribution of each user.

Although GeoLife and NCSU represent the most reliable dataset publicly available, they were not collected to extract interesting places and thus we need to pre-process trajectories in order to find the most meaningful ones to our goal. The need of a pre-processing phase is dictated by the dataset bias which favors movement, while we are interested in people still in their important locations. In particular we aim at densifying trajectory points corresponding to the pause phase by a filling heuristic, while removing points belonging to users' movements. Some of them have been applied only to the GeoLife dataset as in NCSU points have been continuously recorded and traces do not present gaps.

Indoor filling. Mobility data collected by GPS devices present gaps because GPS signals are often disrupted inside buildings or due to the "canyon" effect in narrow roads. This represents a big problem, especially if we are interested in detecting the most visited places of a user. In fact, in many cases, buildings or other indoor locations represent most of the significant locations visited by a person during the day. To overcome the problem given by missing records [96], so to avoid an underestimation of the number of places, we apply the following simple rule. When the ending and beginning GPS points of a gap are within a distance of 35 m and the gap duration is greater than 5 min, the user is judged as residing at the same location during that time. This rule also supplies for the situation where the individual enters into a building, or where the individual turns off the GPS device in an indoor place. Practically, we add as many GPS points equal to the entry point as the duration in seconds of the gap. After the trajectory reconstruction phase, we noticed a big

increment of points, anyway limited by the threshold imposed on the gap duration. Note that the 'indoor filling' heuristic has been applied only to the GeoLife dataset.

Movement phase reduction. We can separate human mobility into two phases; a *static phase* where a person spends some period of time in a location, and a *movement phase* where he moves towards a place of interest. Since people manifest their interest by staying in a particular location, we focus on the static phase. Therefore, we apply a filter aimed at leaving out data which describe the movements among the locations a user visits, thus reducing the number of points to analyze. This way we consider the periods in which a user stays still in a place, assuming that users manifest their interests by spending an amount of time at these places. In order to extract the pause periods and their related GPS points from the whole individual trace, we apply a simple heuristic to the users' speed. If two points p_i and p_{i+1}, with timestamp indicated by $t(p_.)$, do not satisfy

$$\frac{\|p_{i+1} - p_i\|}{t(p_{i+1}) - t(p_i)} \leq \Delta \tag{2.1}$$

then we delete p_{i+1} from the original trace, since it belongs to the movement phase. Analyzing walking mobility data, we set the threshold to the very low value of $\Delta = 1.3m/s$, according to the fact that we observe that human walking speed is about $4-5km/h$ $(1.1-1.4m/s)$. It seems a reasonable value as generally, in a location, people do not reach the maximum speed. This way, we capture points where a person is still or is moving very slowly inside a small area. The good choice of threshold is also confirmed by the cumulative speed distribution function shown in Fig.2.1. Fig.2.1(a), taken from Lin *et al.* [97], reports the speed distribution of different transportation modes. Choosing a value lower than the threshold marked in the figure, we can remove most of the movements not attributable to walk. Also the step we observe in the speed distribution in Fig.2.1(b) suggests that Δ is a critical point where people likely change the transport. The result of the speed filtering process is a sequence of points that forms the trajectory $S = ((p_1, t(p_1)), ..., (p_n, t(p_n)))$, where $t(p_i) \in \mathbb{N}$ is a timestamp and $p_i \in \mathbb{R}^2$, to which we apply the geolocation extraction methodology proposed in Section 2.4. From now on with the term trajectories we intend the GPS points belonging to the static phase.

Users' selection. In GeoLife the point reduction has also effects on the number of users and the number of days, per user, from which we can extract the places of interest. The reduction is mainly due to the fact that GeoLife dataset has been built for the transportation prediction task, and, as a consequence, it favors movements. To overcome the GeoLife bias, we classify the users considering two properties: the period (in hours) a single day trace spans and the number of days the single user traces cover. In particular, for each user, we only consider the daily traces that record more than h hours. On these tracks we count the number of users that have more than

	10d	**15d**	**20d**	**25d**	**30d**
1h	49	39	30	26	24
1.5h	40	34	25	25	22
2h	36	26	25	24	17
2.5h	33	25	22	17	16
3h	28	24	**21**	16	15
3.5h	27	22	19	15	14
4h	25	22	17	13	10

Table 2.2 In the table we report the number of users with more than d days data spanning more than h hours per day. In bold we highlight the number of users for d and h values that represent a good trade-off between the generalization and the statistical significance.

d days of data. In particular, for all the users of the dataset, we filter out all the days of sampling (data collected within the 24 hours, going form 00:00 AM until 11:59 PM) which have $h \leq 3$ hour of sampling. All the remaining days are considered *relevant days*. After this first processing, we filter out all the user which collected less than 20 *relevant days* of data ($d = 20$): the resulting number of users is 21, over the total number of 178 users. We apply these values for the users filter parameters, in order to optimize the trade-off between the importance of having a large number of users, to be able to generalize our analysis; and the need to deal with sampled data which do not correspond to the sole movement phase. For example, only by increasing of one hour the threshold h we obtained a number of users that is not enough to our goal (10). Although the dataset used is namely a collection of trajectories capturing the movement phase, hence only a reduced subset of collected data fulfills our requirements, we are able to obtain initial but meaningful results. Besides, note that the resulting dataset almost completely spans the original GeoLife period.

Summarizing, after the pre-processing, we are able to analyze two spatial and temporal datasets with the highest resolution. This enables us to infer with a great precision the social contacts and map them on the physical space. The adoption of these datasets highly reduces the bias and the drawback of mobile call dataset where the result precision is greatly determined by the network infrastructure (coverage area of the antennas) and social interactions can be only hypothesized. On the other side the small sample makes the results lose their universality w.r.t mobile phone data, calling for larger GPS trace collections.

2.4 Geo-location and geo-community

In this section we formally define the two main concepts which our modeling tool rests upon, geo-location and geo-community, and we propose a methodology to extract them from real GPS traces. The idea of geo-location is close related to the mobility behavior of a single user while the concept of geo-community mimics those places shared by many people where individuals interact. In this respect we have to distinguish between individual and aggregated traces. By *individual trace* we mean a GPS-point sequence associated to a single node; on the other hand, by *aggregated trace* we mean the whole set of individual traces.

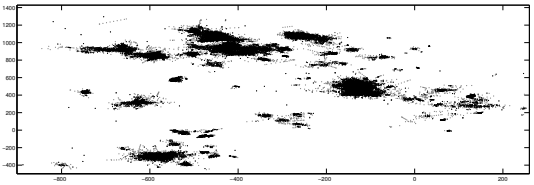


Fig. 2.2 GPS points from an area of the KAIST campus. Points are not uniformly spread but densify around some regions.

By graphically inspecting most of the trajectories disregarding the temporal dimension, we observe that GPS points do not spread uniformly, but tend to gather in a few limited areas (see Fig.2.2). This characteristic agrees with the intuition that, during the day, people tend to gather in places that attract them. These places could be offices, restaurants, bars and coffee shops, or in general any location they spend time in for a while.

Consequently our first goal is the formalization of what we mean by places that capture the person's attention and interest. In particular first we neglect the temporal dimension exploiting the quite constant sampling rate characterizing our datasets and then we reintroduce it to extract important locations. Finally we aggregate the interesting places of the population.

The movements among places where people stop are the basic assumption of our approach. So basically we start our formalization by the static phase, *i.e.* the trajectory S resulting from the pre-processing in Section 2.3.

Definition 2.1. Let S be a trajectory and $L = \{L_1, \ldots, L_k\}$ a partition of $\{p_1, \ldots, p_n\}$ s. t. for each $Li \in L$, L_i is maximal w.r.t. the property that for each $p_u, p_v \in L_i$ exists a sequence $(p_u = p_w, ..., p_{w+j} = p_v)$ of points in L_i, s.t. $\|p_{w+k} - p_{w+k+1}\| \leq \delta, k = 0, ..., j-1$ for a fixed δ. A **stay-location** is an element of L.

Informally, a stay-location is an area where the density of the points belonging to a trajectory is higher than in other areas. So stay-location quantitatively captures individual's tendency to remain in some places, no matter

the time he spends. Moreover it is easy to see that stay-location definition is closely connected to the idea of single-linkage cluster where two points are linked if their distance is less then the threshold δ.

Stay-locations are merely denser area but we need a way to account for the interest expressed by a person. We can exploit the temporal information to infer the interest, in particular we can assume that a person's interest towards a place is proportional to the time s/he spends there. This hypothesis leads to the definition of *geo-location*.

Definition 2.2. Let S be a trajectory and $L_i \in L$ a stay-location. L_i is a **geo-location** if in S there exists a subsequence $((p_i, t_i), \ldots, (p_{i+k}, t_{i+k}))$ such that $p_{i+j} \in L_i$ for $j = 0, \ldots, k$ and $t_{i+k} - t_i \geq \phi$.

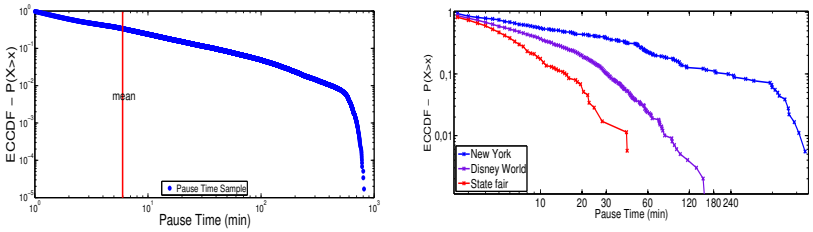
Loosely speaking, a geo-location is a stay-location where an individual spends at least a fixed amount of time (pause time) during his movements. We must underline that we do not consider the sum of the pause times in a stay-location; rather, we consider the single values. For example, if a person spends three minutes in three different occasions in a stay-location, then we do not consider it a geo-location if the threshold ϕ is greater than $3min$.

Up to now, we focus on the places visited by a single user, nevertheless our main goal is the study of the interplay between mobility and sociality. This way we need to define what we intend by social places, *i.e.* locations where people can interact and express their social relationships.

Definition 2.3. Let $GL = \bigcup GL_i$, where GL_i is the set of the geo-locations of the individual trace i. Let OGL be a subset of GL closed under transitivity of the overlapping relation O, s.t. $GL_i, GL_j \in OGL, GL_i O GL_j$ iif $convexhull(GL_i) \cap convexhull(GL_j) \neq \emptyset$. We define a **geo-community** as $convexhull(OGL)$.

The social aspects of geo-community are captured by the overlap of geo-locations which are not disjoint. In particular, by means of geo-community, we want to capture those areas shared by a few people albeit not necessarily at the same time. This is the reason why we overlap geo-locations extracted from different individual traces. For example, if we have two circular geo-locations, extracted from two different traces, we join them even if they overlap only partially. Moreover the geo-communities encompass those regions in which people manifest a particular interest as the overlapping concerns the geo-locations. This way areas like the stops near a semaphore or a bus stop are not considered although they are common to many people.

Geo-location extraction from individual traces. In our analysis we examine both individual traces and aggregated traces. Individual traces are used to infer geo-location, as well as the criteria people use in deciding upon their next geo-location; instead aggregated traces and the aggregation of geo-locations are used to extract geo-communities and study the interplay about social interactions and mobility.

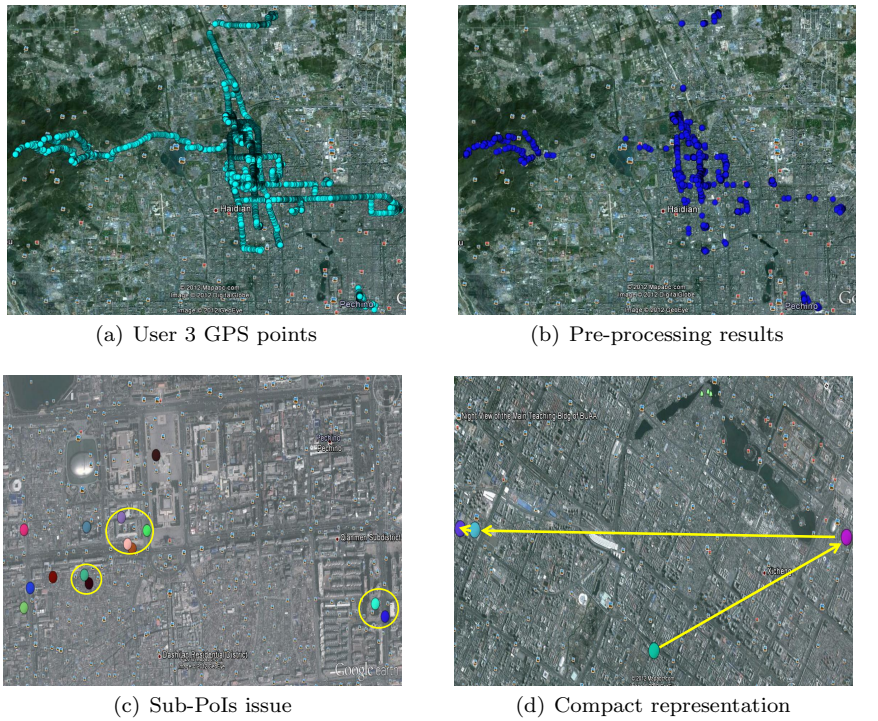


(a) GeoLife dataset (b) NCSU dataset

Fig. 2.3 ECCDF of the aggregated pause times in the stay-locations.

Here we provide a methodology that extract geo-locations and geo-communities from a pre-processed dataset as in Section 2.3. As already noted, the definition of stay-location is close to the concept of cluster. Therefore, to find stay-locations we apply a density-based clustering [87] algorithm DBSCAN [45]. In particular, according to the definition, from the trajectory S we consider only the point p_i, discarding the associated timestamp t_i. We choose DBSCAN because it does not need in input the number of clusters to be found and allows us to detect arbitrary shaped clusters. As DBSCAN parameters we use $\sigma = 25$ mt and $\epsilon = 2$ neighbours (σ represents the maximum distance such that two points are considered neighbors, while ϵ is the minimum number of neighbors that a node must have to be considered in a cluster). This way we adopt a stricter definition of cluster than in [138], where authors set $\sigma = 100$ mt as the limit for the connection between nodes and adopted a less robust to noise clustering method. Due to the fine granularity we use a fine grain because we can better analyze movements through small locations such as restaurants, offices, coffee shops and bars. Moreover the possibility of extracting places at a high detail allows us to better correlate close social interactions and locations.

Stay-locations represent the first step in extracting geo-locations. As a matter of fact there are many stay-locations where an individual stays for a short amount of time. These stay-locations are transit-locations and represent small pauses in the movement towards destination, so we need to filter them out. The choice of a proper threshold represent an issue since the distribution of the pause time changes among the datasets, as we can observe in Fig.2.3. This observation obliges us to make a sharp decision about ϕ. Indeed in the following analysis of the datasets we set the threshold $\phi = 5$ min, which roughly corresponds to the median of the pause distribution in stay-locations in most of the datasets. Once fixed the pause time threshold, we remove noisy points not belonging to geo-locations. The thresholding on the stay-location pause time have two important effects: *i)* we drastically reduce the number of stay-locations, as shown in Table 2.3, and *ii)* we can infer which are the main destinations, the geo-locations, during daily movements. The stay-location



(a) User 3 GPS points

(b) Pre-processing results

(c) Sub-PoIs issue

(d) Compact representation

Fig. 2.4 Geo-location extraction methodology applied to the user 3's trajectories. In 2.4(a) we plot all the recorded points (raw data). In 2.4(b) we show the points resulting from the application of the pre-processing phase. In 2.4(c) the sub-locations that have to be grouped in the real geo-location (yellow circle). In 2.4(d) a compact representation of user 3's mobility during a single day.

reduction strongly manifests in the GeoLife dataset, where on average 90% of stay-location disappears, while in the other dataset stay-locations are halved.

The thresholding results in the meaningful geo-locations, however, especially in the GeoLife dataset, we observe situations, such that presented in Figure 2.4(c) where we have many sub-locations of the same general geo-location. This phenomenon might be due to the indoor filling heuristic and the small ϵ threshold. To overcome this empass we run a second passage of DBSCAN with a larger ϵ on the centroids of the sub-locations detecting the real points of interest.

Aside from finding geo-locations, the above methodology has the capability to express human mobility as a compact trace that summarizes the transitions between important locations and the users' pause time in them. In a more formal way, we can extend $\mathcal{TAS}$ notation introduced in [57]. A $\mathcal{TAS}$ can be defined as a couple $T = (S, A)$ where S is a sequence $\langle s_1, \ldots, s_n \rangle$ with temporal annotation $A = \langle a_2, \ldots, a_n \rangle$. In our case A represents the temporal

transition or the distance covered between two consecutive elements in S, *i.e.* the sequence of visited geo-locations. In order to introduce the pause time in $\mathcal{TAS}$, for each geo-location $i \in S$, we add the entry time t_i^{IN} and the exit time t_i^{OUT}. Accordingly, we obtain

$$S = \left\{(s_1, t_1^{IN}, t_1^{OUT}), \ldots, (s_n, t_n^{IN}, t_n^{OUT})\right\} \tag{2.2}$$

where $\forall_i t_i^{OUT} > t_i^{IN} > t_{i-1}^{OUT}$ for $i = 2, ..., n$. For example the following extended $\mathcal{TAS}$ $(\langle(1, 34, 67), (2, 78, 84)\rangle, \langle 9\rangle)$ represents a person who remains in the geo-location 1 from time 34 to 67, he moves to the geo-location 2, where he stops for 6 time slots, taking 9 time slots.

Summarizing, in Fig.2.4, we apply the overall methodology applied to a single user of the GeoLife dataset. From the raw GPS data in Fig.2.4(a), which cover most of the Beijing area, we remove the movements as explained in Section 2.3 obtaining the blue trajectory S in Fig.2.4(b). On the blue points we run DBSCAN and then we apply a thresholding on the pause time in stay-locations. The obtained geo-locations are further clustered by a second run of DBSCAN with a larger σ as shown in Fig.2.4(c). Finally we can compress the movement among the main interesting destinations providing the extended $\mathcal{TAS}$ reported in Fig.2.4(d), where we omit the entry and the exit times.

Adopting this compact representation in the following section we can analyze some properties of human mobility and of the interesting locations human beings visit during their daily movements.

Site	μ_{stay}	σ_{stay}	med_{stay}	μ_{geo}	σ_{geo}	med_{geo}
KAIST	16	6	15	7	2	6
NCSU	9	4	8	4	3	4
New York	13	10	10	4	3	4
Disney World	32	12	30	16	6	16
State fair	18	4	18	8	2	9
GeoLife	3194	2531	2457	221	185	172

Table 2.3 Statistics on number of stay-locations (first three columns) and geo-locations (last three columns). μ, σ and *med* denote the average, the standard deviation and the median of the visited stay/geo-locations per user, respectively. In the GeoLife row, values are not averaged on the number of days as the duration of the traces is too heterogeneous.

Geo-community extraction from aggregated traces. After the geo-location extraction on individual traces, we associate to each user a set of important locations, representing the main destinations which get the user's interest during the daily mobility. Therefore aggregating these information is useful at finding how many people share the same geo-location.

According to the definition of geo-community we need to overlap the geo-locations of different users. The superimposition of the geo-locations has been obtained by applying a further DBSCAN run on the points p_i belonging to

geo-locations and not to the whole dataset points. Limiting the application of DBSCAN to a small subset of the original raw GPS implies an high reduction of the computational complexity. Since an aggregated trace contains many border points between clusters, especially in the NCSU dataset, we apply DBSCAN using a different set of parameters ($\sigma = 10$, $\epsilon = 4$). Through this setting we are able to better deal with border and outlier points. In particular the availability of a larger quantity of points inherent to the regions of interest allows us to adopt a strict proximity distance and to extract denser overlapped areas.

By means of the clustering process we can infer geo-communities and compute the most shared locations. By computing the last quantity on our dataset we note that many geo-communities contain few elements. This is due to the small number of users in the dataset. We delete these communities from our geo-community analysis, adopting a threshold value of 5 different elements per community. The resulting geo-communities are reported in Table 2.4 where we summarize for each group in the NCSU dataset the number of geo-communities visited by at least five persons. Note that we do not report the results about GeoLife geo-communities because none of the shared locations has been visited by at least five people. This way, from now on, the mobility features in GeoLife will be computed in the geo-location setting, while in the NCSU dataset we describe movements adopting the geo-community viewpoint.

In general the overall methodology allow us to infer geo-communities that, from a geographical point of view, represent the regions of interest of a population while, from a social prospective, identify groups of people who raise a community since they have a place in common. As a geo-communities represent a framework to study contact behaviors derived from people's mobility as we will see in Section 2.6.

	KAIST	NCSU	New York	Disney	State fair
# geo-com	11	4	3	38	12
mean dim	30000	5850	4300	5723	3150
mean dens	0.0016	0.0017	0.0016	0.0022	0.0026

Table 2.4 Results of geo-communities analysis. The first row contains the number of geo-communities. Mean dimension (m^2) and mean density (number of users/m^2) are reported in the second and third rows.

2.5 Spatial and temporal dynamics

Various measurement studies of human mobility traces have discovered significant statistical patterns. In particular the most meaningful results concern

the distance covered between consecutive visited places and the time spent in them. The investigation of these quantities have been made possible thanks to a series of studies on the trajectories of several mobile phone users collected by mobile carriers from different countries. A first research result concerns the distribution of displacements Δr between user's positions at consecutive calls; a quantity that has been considered as a good approximation of the distance between consecutive points of interest. Gonzalez *et al.* showed that Δr is well approximated by a truncated power-law characterized by a probability density

$$f(\Delta r) = \frac{\beta - 1}{\Delta r_0^{1-\beta} / \Delta r_{max}^{\beta-1}} \left(\frac{1}{\Delta r} \right)^{\beta} \tag{2.3}$$

with $\beta = 1.75$, $\Delta r_0 = 1.5$ km and the cut-off value Δr_{max} varying in different experiments. This equation form, found in other mobile phone datasets with slightly different parameters, has suggested that human motion follows a truncated Lévy-flight. However, some restrictions on the process randomness have been observed in [96, 145]. In particular they noted a 'go-back-home' effect, *i.e.* most people usually travel in close vicinity to their preferred location (home), while a few frequently move towards far destination. In a social perspective these results could indicate where a person would socialize, but they disregards the temporal information about how many time he might stand at a place. Song *et al.* [145] have took into account this variable and measured the distribution of the visiting time or *pause time* (the interval a user spends at one location). The resulting distribution follows a truncated power-law with an exponent $\beta = 0.8$ and $\Delta t_{max} = 17$ hours, a value reasonably connected with the daily activity period of human beings.

As already noted the previous analysis and results have been obtained from mobile phone call dataset, however this type of collection impose some limitations and approximations given by the detection technology (antenna coverage area). A typical example is the Δr_0 value which indicates that the power-law distribution characterizes movements 1.5 km longer. What occurs within a 1.5 km radius is unknown because the data can not provide this information.

In this section we try to fill this gap and we investigate human mobility patterns at the high spatio-temporal resolution given by the GPS trajectories. In particular, exploiting the geo-location/community extraction methodology provided in the previous section, we can refine the statistical analysis about the displacement between consecutive points of interest (no more between regions covered by antennas) and about the pause time. Furthermore we provide a classification of the interesting places by the users' frequency visit. Such classification allows us to characterize the user mobility and to divide users in routine and globe-trotters. Finally we investigate the influence of the distance in the choice of the next destination finding a common probability distribution based on the rank induced by the distances from the possible next destinations.

From a methodology point of view, we apply the same inference procedure for each variable we investigate. In particular we consider as possible models nine distributions of the exponential family [78]: exponential, Weibull, normal, generalized Pareto, Pareto, tapered Pareto [79], upper-truncated Pareto [74], lognormal and Pareto-lognormal [136]. To find the best parameters setting for each distribution we use Maximum Likelihood Estimation (MLE), adopting a closed form when available or a numerical maximization otherwise. Moreover we adopt two measures to evaluate the goodness-of-fit and to choose the right model: the log-likelihood and the Akaike's Information Criterion (AIC)[3, 28]. In particular the AIC compares models from the perspective of information entropy, as measured by Kullback-Leibler divergence. The AIC for a given model is computed as:

$$-2\log(L(\mathbf{s})) + 2k \tag{2.4}$$

where $L(\mathbf{s})$ is the likelihood of the sample given the parameters returned by MLE and k is the complexity (number of parameters) of the model. AIC values for different models are used to identify the best model, indeed smaller values of the criterion are better.

2.5.1 Movement displacement distribution

The approach adopted in the extraction of the geo-locations/communities enables us to focus on two types of movements: inside a geo-community and among geo-communities. This way we can model the human mobility by two independent processes. First a person chooses his destination and once he reaches it, he moves inside it. Note that from now on we equivalently use the term jump length, flight length, displacement and distance.

Jump length inside geo-communities. The first characteristic we examine is the distribution of distances inside geo-communities. This way we statistically characterize human micro-mobility, *i.e.* how people move inside places. As a common definition of distance covered inside a place does not exist, we measure the displacement between two consecutive pauses in the trajectory. Two points $(p_1, t(p_1))$ and $(p_2, t(p_2))$ are considered in a pause period if $||p_2 - p_1||/(t(p_2) - t(p_1))$ is less than a threshold close to 0. We apply this definition and the following analysis only to the NCSU dataset, as the application of the indoor filling heuristic on the GeoLife heavily bias the flight length. Pause definition simplifies the computation of the displacement since it reduces to calculate the distance between two consecutive points, leaving out the pauses, and filtering the admissible distance. In fact for each site, we perform the MLE over the x-axis range between 10 meters and the maximum value of the sample. We observe that the Pareto distribution is the best model in almost all the cases, as we see in Table 2.5.

	KAIST	NCSU	New York	Disney	State fair
Pareto	**5690**	**1213**	402.8	3075	**308.3**
tapPareto	5692	1215	402.6	3075	310.3
UTP	5692	1216	**397.7**	**3070**	310.9

Table 2.5 AIC values for the flight length distribution computed on the different trace groups in the NCSU dataset. We report the values related to the best three distribution fitting.

Moreover if we examine the parameters of the tapered Pareto distribution we note that η [79] related to the exponential cut-off is close to 0. So simplifying $\eta = 0$, we obtain a pure Pareto distribution. In general we can conjecture that movements inside geo-communities are distributed according to a power-law. From the modeling point of view these observations imply that inside geo-communities we can apply a Lévy process to generate plausible movements.

The obtained results are novel in the literature on human mobility pattern analysis, as ,thanks to the high resolution of GPS trajectories, now we are able to extract and deal with the micro-mobility.

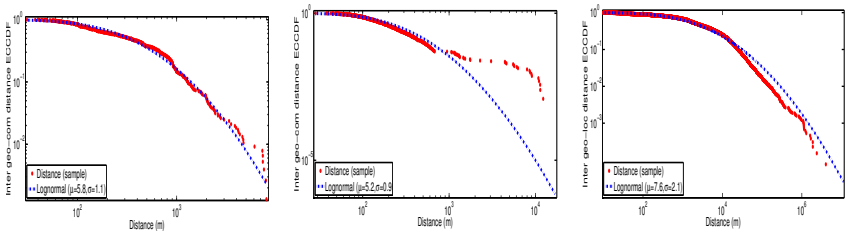
Jump length among geo-communities. Here we analyze the second component that characterizes the human movements according to our approach, namely movements between different geo-communities. This quantity represents the most comparable measure to the displacement as computed by Barabasi and his research group. This way we are able to verify if human movements captured by mobile phone calls are statistically similar to those one detected by the GPS technology. As sample distance between points of interest we take the euclidean distance between geo-community centroids for single trace. The analysis outcomes are presented in Table 2.6. In this case we do not find a common distribution over all sites. Nevertheless, we must underline that the NCSU campus, New York and state fair datasets have few elements (≤ 170), leading to unsupported estimates. In fact for these sites we found that distances are distributed as a tapered Pareto or an exponential. When the sample is sufficiently numerous, we find that the displacement among consecutive geo-communities is distributed according to a lognormal, as shown in Fig.2.5. In the figure we can note that the parameters in the GeoLife set differ from those characterizing KAIST and Disney World. In fact the $\sigma = 2.2$ indicates that in GeoLife people are more likely to move towards farther locations w.r.t. individuals in the other datasets.

Lognormal distribution of inter-distance is important because it implies a more gradual decay slope, with respect to the sharp exponential cut-off presented in [138] or in [60]. We might say that people generally prefer short paths between geo-communities, but take long jumps more frequently than observed up until now. Moreover these results, combined with the other ones in literature, confirm that human movements are far from being the result of a Brownian process. Nevertheless we show that the way we infer points of

interest and the trace granularities given by the detection technology impact on the statical properties of the human mobility. This advocates the collection of new trajectories to further confirm or reject the results in literature.

	KAIST	NCSU	New York	Disney	State fair	GeoLife
Exponential	11295	2263	2713	11344	**1980**	275832
Lognormal	**11191**	2251	2627	**10672**	1984	**256103**
Tapered Pareto	11231	**2230**	**2565**	11011	1982	269278
UP.Trunc.Pareto	11410	2243	2568	11020	2010	265837

Table 2.6 AIC values for inter-distance distribution.



(a) KAIST (b) Disney World (c) GeoLife

Fig. 2.5 Inter-distance fitting distribution. We only report KAIST Disney World and GeoLife results since they are the largest datasets.

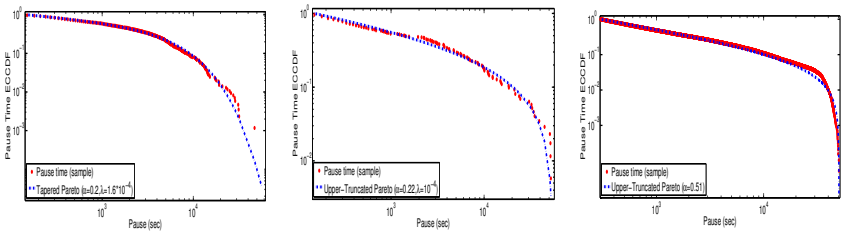
2.5.2 Pause-time distribution

Pause time distribution plays an important role in making the mobility more or less diffusive, in fact the more likely are long pause times, the less will be the possibilities of moving. Pause time also influences the social aspects. Indeed the more a person waits in a place, the more likely he will meet other people, but conversely it will have less opportunities of meeting new people which attend other places. As far as pause time distribution is concerned, for any site, we easily retrieve pause time values from the $\mathcal{TAS}$s associated to each user. More precisely given a $\mathcal{TAST} = (S, A)$ for each geo-community $i \in S$ in the sequence we compute $t_i^{OUT} - t_i^{IN}$ putting it into the sample set. From our datasets we find that the pause times is fat tailed, confirming the results in [60, 25]. Fig.2.6 shows the CCDF of some pause-time distributions extracted from our traces. In nearly every case the best model is given by the

tapered Pareto, as shown in Table 2.7. The only exceptions are New York and NCSU.

	KAIST	NCSU	New York	Disney	State fair	GeoLife
Lognormal	15396	3252	3398	1332	2651	344847
Pareto	15592	3224	3311	1308	2568	334846
Tapered Pareto	**15219**	3196	3313	**1295**	**2556**	334678
UP.Trunc.Pareto	15377	**3173**	**3285**	1301	2558	**332908**

Table 2.7 AIC values for pause time distribution.



(a) KAIST (b) NCSU (c) GeoLife

Fig. 2.6 Pause time distribution in KAIST, NCSU campus and GeoLife. In each figure, over the sample, we report the best model and the parameters estimated by MLE.

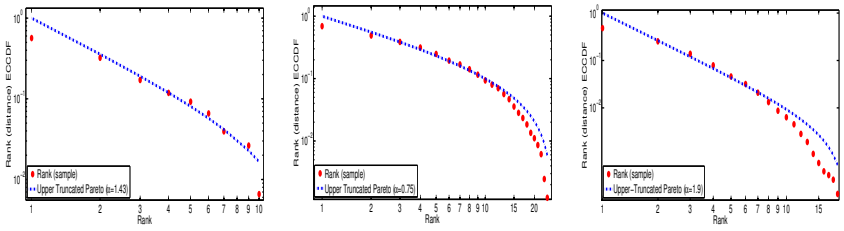
2.5.3 Choosing the next geo-community

An important question about human mobility concerns the criteria people use to choose the next geo-community. This decision can be influenced by many factors, since every individual has different cost functions and factors. Among the various criteria, we choose distance between geo-communities as the main factor influencing how people decide. To study the influence of distance in the choice of the next geo-community destination, for each geo-community, we order the remaining ones depending on the distance between the respective centroids. Then, for each visited geo-community in a user's trajectory we register the position, in the ordered sequence, of the next geo-community. At the end of the process, we obtain a sequence of integer representing the rank of the geo-community at the moment of the choice. Looking at the best AIC values (Table 2.8) for each dataset, we can note that the rank distribution follows an upper-truncated Pareto as shown in Fig.2.7. This distribution implies that people frequently move towards geo-communities nearby, yet sometimes

decide to move toward distant ones. Moreover the fact that we find a common distribution in different datasets suggest that in urban area the distance between the points of interest play a key role in the choice of the destination.

	KAIST	NCSU	New York	Disney	State fair	GeoLife
Pareto	2533	341	365	3333	500	22201
Tapered Pareto	2459	343	368	3267	492	22203
UP.Trunc.Pareto	**2433**	**340**	**364**	**3242**	**481**	**22131**

Table 2.8 AIC values for the distribution on the rank of the next geo-community induce by the distance. Bold values represent the best model.



(a) New York (b) Disney World (c) GeoLife

Fig. 2.7 Distribution on the choice of the next geo-community

2.5.4 Location classification by visit frequency

The wide period spanned by the GeoLife dataset allows us to compute how frequently geo-locations are visited by the different users. The evaluation of the frequency is strictly related to the importance of a geo-location since we suppose that a user, who often reappear in a place, like it. According to this hypothesis, we can classify the geo-locations visited by a user, based on their relevance, in order to infer the most important places he visits and characterize his mobility behavior. The *relevance* [124] or *visit frequency* of a certain geo-location L_i has been calculated on the mobility history of each user, and it is defined as:

$$relevance(L_i) = \frac{d_{\text{visit}}(L_i)}{d_{\text{total}}} \tag{2.5}$$

where $d_{\text{visit}}(L_i)$ is the number of days a geo-location L_i has been visited (one or more times per day) by the user and d_{total} is the total number of days, collected by the user. The relevance of a certain geo-location is, according to the formula, the percentage of days the user visits this location, over the total number of days of sampling.

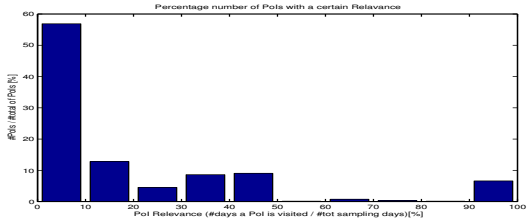
According to frequency visit values, we show that the geo-locations associated to each user can be grouped in 3 classes:

- **Mostly Visited geo-Location (MVL)**: locations most frequently visited by the user. We can easily infer their semantic meaning, and associate them to home, work place, gym.
- **Occasionally Visited geo-Location (OVL)**: locations of interest for the user, but visited just occasionally.
- **Exceptionally Visited geo-Location (EVL)**: PoIs unlikely visited more than very few times.

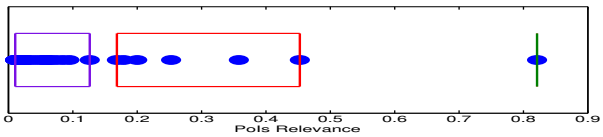
The evaluation of the distribution of the aggregated relevance allows a straightforward identification of these three classes. In Fig.2.8(a) we show a cumulative characterization of the geo-locations identified for all the users in the dataset. In the figure we report the average number of geo-locations belonging to the corresponding relevance interval. From the figure, it is easily visible that, on average, 57% of the geo-locations visited by a user can be bring into the EVL class: this means that more than half of the geo-locations seen by each user will be hardly visited for multiple times. Conversely 6.7% of geo-locations assume a MVL behavior. This fact gives an idea of the limited number of locations which are visited by each user almost daily.

Finding class of relevance. The identification of the upper and lower bounds for each of the three classes is strictly related to every single user; in fact it depends on the user's mobility style. As a consequence, class bounds cannot be fixed *a priori* but claim at an automatically detection algorithm able to adapt to the single user mobility pattern. In particular, we adopt an unsupervised approach which groups the geo-locations of a single user according to the values of geo-location relevance and maximize their separability. The clustering algorithm we choose is the k-means with $k = 3$ which corresponds to the number of geo-location classes. To avoid the problem related to the initial choice of the centroids, we run 10 replicas of k-means with different initial seeds and choose the partition that minimizes the within-cluster sums of point-to-centroid distances, thus maximizing the separability. In Figure 2.8(b), we show the result of the k-means clustering on a sampled user. The EVL class (purple box) covers the range from 0.01 to 0.12, the OVL (red box) spans the range from 0.16 to 0.46 and the MVL class (green line) contains only one geo-location with relevance 0.82.

For each filtered user, we apply the k-mean algorithm to classify the related PoIs in three main classes of relevance (2.5.4) and over these classes we study three main features: *(i)* the number of PoIs which reside within each class of



(a) PoIs Relevance



(b) Three classes of relevance in a sampled user

Fig. 2.8 Aggregated and single user classes of PoIs based on the relevance. In the x-axis we identify 10 bins of relevance spanning form 0% to 100%, where each of them has a width of 10%.

relevance, *(ii)* the percentage of the daily time spent in each class and *(iii)* the average time of the visits to the PoIs of the classes.

In Fig.2.9 we represent the number of geo-locations associated to each class of relevance, per user. In the upper plot we can notice the large difference in the number of EVLs, with respect to the geo-location belonging to the other two classes of relevance. This is an evidence of the fact that the user always visits new locations, but only few of them are visited regularly. In the lower plot, we zoom on the OVL and MVL classes. Here the number of OVLs and MVLs is limited and their average values are 4.19 and 1.76, respectively. As expected, each user has a very small number of preferred locations (MVL) which are visited daily (e.g., home, work place), and a higher but still limited number of location of interest (OVL) which are visited with a lower frequency but regularly (e.g., gym, favorite restaurant, parent's house).

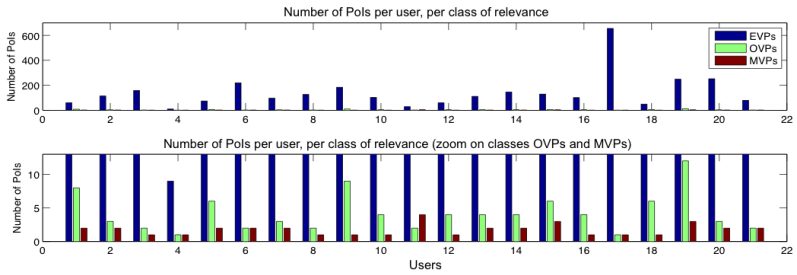


Fig. 2.9 Number of PoIs per class of relevance

As concerns the relation between the average visit time and the classes of relevance, we notice that for all users, the average visiting time to EVLs is very limited and on average lower than one hour. On the contrary the average visiting time for OVLs and MVLs is quite heterogeneous and depends on the user's mobility style. Indeed some users tend to spend a long time in their MVLs, other users instead, use to have very long visits to the OVLs.

Another interesting aspect emerging from the geo-location classification is the percentage of the visit time allocate for each class. In particular we observe two opposite behaviors: the 'stay-at-home' and the 'globetrotter'. The typical 'stay-at-home' user spends most of his time in the MVLs, and less than 9% in the EVLs. As opposite, a 'globetrotter': the percentage of time allocates to the MVLs is below 10% while the remaining time is spent in the OVL and EVLs , even if the average time spent in each EVL is still significantly smaller than the average time spent in each MVL.

2.6 Social aspects of geo-communities

Up to now we focused on the statical features of the human mobility expressed by places which capture the interest of a group of individuals. Nevertheless we still have to analyze the network which describes the social interactions occurring inside the geo-communities. In this section we concentrate our attention on the social aspects of our modeling approach. First, we validate the basic assumption that people express their sociality inside geo-communities, *i.e.* we justify that the overlap of the geo-locations of different users is able to include the points where social interactions occur. Second we formalize the association between users and geo-communities through a graph-theoretic approach. In particular, we exploit the well-known idea of affiliation network or bipartite graph to generate a social graph plausible with the one induced by the real interactions among the users. More precisely we extend the concept of one-mode graph projection [116] combining it with some random graph model. Finally, we validate our approach showing a simple method to tune the projection parameters to generate social interactions similar to the real ones. In general we argue that the geo-community bipartite network approach could represent a fruitful modeling tool to study the interplay between mobility and sociality, as we will also see in Chapter 2.

2.6.1 Contact positions

By analysing spatial traces we can obtain all the information about meetings typical of direct contact traces such as inter-contact time, contact duration and contact frequency. In addition, from GPS trajectories we can extract

another important piece of information: the position of the contacts. Localization allows us to quantify how many interactions occur inside and outside geo-communities. This way we can validate the concept of geo-community intended as the shared social foci where the interactions given by the spatial proximity manifest.

In order to localize the contacts, we define two users to be in contact if their contact areas overlap. With the term *contact area* we mean a circle centered in the position of the node and characterized by a radius r that we set to 5 m. Furthermore the contact position is defined as the pair of GPS coordinates of the users in contact. By taking advantage of the assignment of the GPS points to the geo-communities given by the clustering algorithm, we can decide whether each of the two points belongs to a geo-community. This way we can correlate contacts and geo-communities by checking if at least one of the points is inside a geo-community.

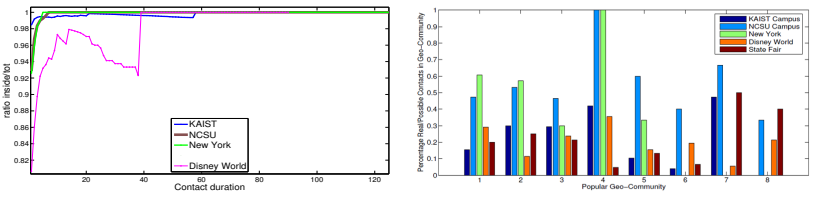
A further advantage of this approach can be achieved by studying the relation between contact durations and the percentage of contacts inside geo-communities. This way we also consider the strength of the social interactions, not only their positions. To study the relation between geo-communities and strong interactions, we compute the ratio between the number of contacts inside geo-communities and the total number of contacts with a given duration, for each allowable duration value. Results are shown in Fig.2.10(a) and can be summarized as follows:

- **KAIST**: $98,4\%$ of all contacts occur in geo-communities as 99% of the contacts lasting more than 1 min.
- **NCSU**: $92,8\%$ of all contacts take place in communities as well 99% of those that last more than 2 min.
- **New York**: this dataset presents results very similar to the previous one.
- **Disney World** and **State Fair**: due to the different nature of these places (fun fair), respectively $80,6\%$ and $65,2\%$ of the entire contacts occur inside geo-communities, however 90% of the contacts lasting respectively 2 and 8 min, happens inside them.

Generally, for each dataset, we can always find a contact duration value such that all contacts take place inside geo-communities. The obtained results validate the concept of geo-community as a shared social place where the interactions given by the spatial proximity occur.

2.6.2 Geo-communities bipartite graph and its projections

From the perspective of modeling GPS mobility data by creating a smart complex network, the concepts of geo-community and of individuals (belonging to them) can be represented by a large bipartite graph. This kind of net-



(a) (b)

Fig. 2.10 2.10(a): Ratio between the number of contacts inside geo-communities and the total number of contacts with a given duration as a function of the contact duration. 2.10(b): Rate of actual contacts on the number of possible contacts computed on the eight most visited geo-communities in each dataset.

work appears in many settings such as social [156] and co-citation networks [116] or market basket analysis [134]. Some examples, in which the relations between individuals and communities could be explored, are given by the conference-author network which contains information about the number of papers published by each author in each conference. On this kind of networks, for example, we can infer communities or calculate proximity measures between nodes. Therefore, the bipartite network approach is very meaningful in graph mining and, in our case, it is further enriched by an additional information: the geographical position. These emergent properties might be treated in a formal manner by describing the entire process in terms of an undirected bipartite graph. An undirected bipartite graph $G = (U, V, E)$ is a graph whose vertices can be divided into two disjoint sets, U e V, such that every edge connects a vertex in U to one in V. In our case, we define U as the set of the geo-communities and V as the set of nodes. Consequently, E can be defined as

$$E = \{(v, u) | u \in U, v \in V \wedge v \text{ visits } u\}$$

A bipartite graph can be also represented by an incidence matrix $\mathbf{B}$. If n is the number of users and g is the number of geo-communities, then $\mathbf{B}$ is a $g \times n$ matrix such that $B_{ij} = I_j(i)$, where $I_j(i) = 1$ if the user i belongs to the geo-community j and 0 otherwise.

Bipartite graph projections. Bipartite graph representation enables us to infer all the possible contacts resulting from the sharing of common communities between nodes. In order to reach this goal, we can perform a one-mode projection from the two-mode bipartite form. The simplest projection on the nodes can be obtained by constructing an undirected graph $G_{proj} = (V, E')$ where

$$E' = \{(v_1, v_2) | v_1, v_2 \in V, \exists u \in U, (v_1, u), (v_2, u) \in E\}.$$

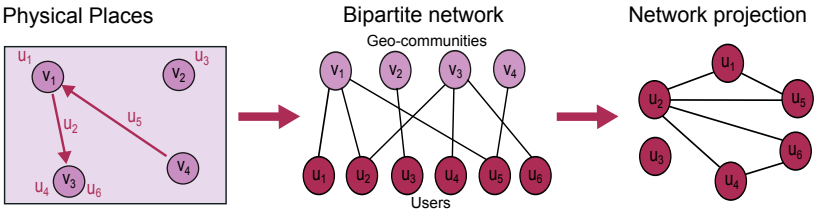


Fig. 2.11 Unweighted projection on the node set of a bipartite graph. In the bipartite graph, constructed from the traces on the physical space, purple circles represent geo-communities and the red circles are the users. An undirected link in the bipartite represents the membership relation, while in the projection graphs a link represents a social relationship.

When we form such a projection, each geo-community in the bipartite graph results in a k-clique of vertices in the projection, where k is the number of users belonging to a geo-community. Thus the overall projection is the union of g cliques (Fig.2.11). It is easy to see that, in general, this construction discards a lot of information present in the structure of the contacts. For example, two users could belong to the same geo-community but visit it on different occasions, so a link between them cannot exist. However, already using this type of projection, we can find an upper bound on the number of contacts between nodes that can be taken within the geo-community, as shown in Fig.2.10(b). Such upper bound identifies those potential contacts that may occur only because of sharing a geo-community. In an optimistic way it represents a possible prediction on both contacts and social relationships among people.

Many direct contact traces have been studied in literature, by applying different strategies in the construction of the contact graph [73], [69], [141]. To the best of our knowledge, the properties of the contact graph induced by the overlap of GPS traces have never been investigated.

As our main purpose is to analyze the structure of the contacts detected only within geo-communities, we discard all coordinates of the contacts that happen outside geo-communities. To maintain uniformity, in our analysis we apply the same definitions and the same contact radius presented in Section 2.6.1. We associate to each contact its duration and for each pair of nodes (i, j) we compute the aggregated duration d_{ij}(sum of the contact durations). After that, we define the contact graph through the matrix $\mathbf{C}$ where

$$C_{ij} = \begin{cases} 1 & \text{if } d_{ij} \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

We set the threshold $\theta = 2$ min in order to delete rare and brief contacts that do not add information to the relationship between nodes.

From the graph $\mathbf{C}$ we extract the vertex degree distribution, the mean betweenness centrality, the diameter, the mean shortest path and the mean

clustering coefficient. The values of these indexes are summarized in Table2.9. We can consider KAIST, Disney World and State Fair to be a small-world as their mean clustering coefficient is higher than those computed on Erdös-Rényi graph [126] whose mean shortest path is similar to that of **C**. As regards NCSU and New York we find more connected components; therefore there exist groups of nodes that do not interact.

Site	$\bar{d}$	$\bar{b}$	diam	msp	$\bar{c}$
KAIST	34	59	4	1.64	0.6
Ncsu	4.6	47	5	2.4	0.22
New York	2.4	4.2	4	1.48	0.27
Disney World	7	87	6	3.2	0.62
State Fair	6.2	16	4	1.82	0.52

Table 2.9 Mean degree ($\bar{d}$), mean betweenness ($\bar{b}$), diameter (diam), mean shortest path (msp) and mean clustering coefficient ($\bar{c}$) computed on real contact graphs.

By comparing the contact graph **C** to the bipartite graph, we can evaluate the existing relation between real and potential contacts. In particular, for each geo-community we compute the ratio p_u between the number of edges between vertices belonging to it and the maximum number of edges given by the unweighted projection of the same geo-community. By means of this index, we introduce a new type of projection. Let us assume, in the bipartite graph, to assign to each vertex $u \in U$, representing a geo-community, a value equal to the previous ratio p_u. Fixed an $\alpha \geq 1$, each geo-community is projected onto a subgraph built according to the Erdös-Rényi model of parameters n, equal to the number of vertices belonging to the geo-community, and $p = p_u^{\alpha}$. As for the other projections, we join subgraphs obtaining a *proportional projection*.

To enrich the projection, we develop a new way of projecting based on a preferential mechanism. This approach is justified not only by many results in literature [73, 68] showing the presence of a so called "hub node", but also by the number of neighbours distribution inside geo-communities that we observe in real traces. Our way of projecting builds the subgraphs of their geo-communities by using two kinds of construction. If k, the number of users belonging to a geo-community, is less than a fixed threshold η, then we apply an unweighted projection creating a k-clique. Otherwise we build a Barabasi-Albert graph [11] on the vertices inside the geo-community. After subgraph constructions we join them in an unique graph obtaining a *preferential projection*. The use of a threshold can be explained by observing that in real traces, when geo-community cardinality is small, everybody meets.

We can observe that both preferential and proportional projections are parametric models; the first depends on the threshold η and the number of edges m, while the second relies on α. Therefore, we can exploit this characteristic with the aim of finding parameter values such that their projection

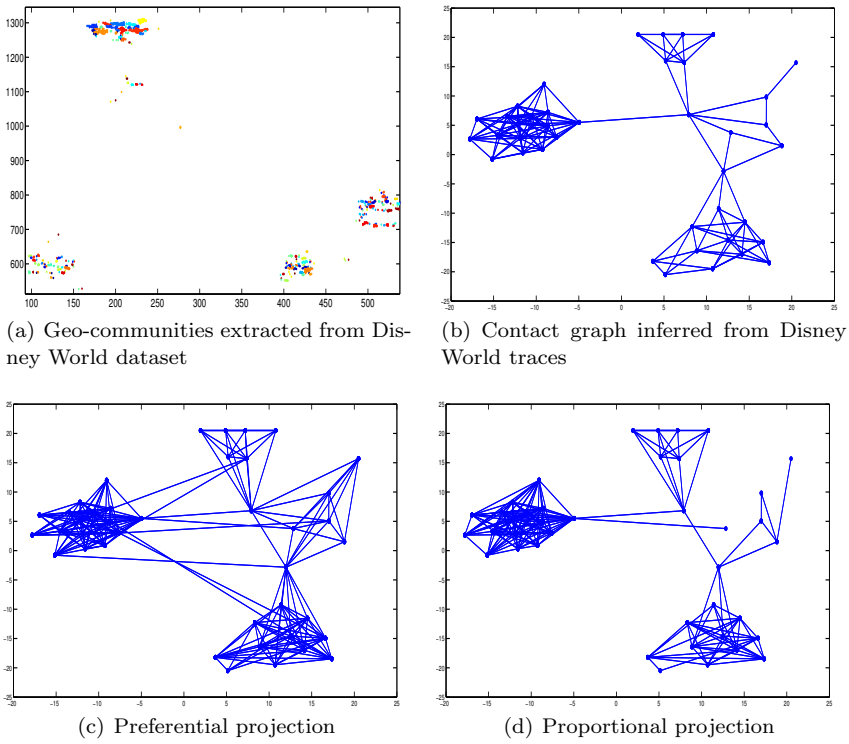
reproduces at best contact graph properties. In particular we need to find the parameter values which maximize the similarity between the synthetic contact graph and the real one. Computing the similarity between two graphs G_1 and G_2 is a well known task and various approaches has been proposed, from the isomorphism problem and its generalizations to the search for the maximum common graph. Among the different methods we apply the algorithm proposed by Blondel *et al.* [20]. In this method an element in the graph G_1 and an element in G_2 are similar if their respective neighborhoods similar. So iteratively the method constructs a similarity measure between any two nodes in any two graphs, expressed by a similarity matrix $S(n \times m)$ where $S(i, j)$ denotes how i and j are similar. Once computed the matrix S, we can assess a measure of similarity between the two graphs by computing a maximum weight matching on S. To resolve the matching problem, we apply the Hungarian algorithm, while the overall similarity is the sum of matched similarity values. As concerns preferential projection, we perform a grid search, *i.e.* a bidimensional numerical search over the parameters space, to identify the pair of values whose projection better approximate the real one. Specifically, for each pair of parameters, we generate 100 samples of the projection and we take the average graph similarity. The best parameter choice corresponds to the pair which maximizes the average similarity. We apply the same method to proportional projection considering a mono-dimensional parameter space.

By analyzing the resulting best projections on different datasets, we can say that the preferential projection quite closely reproduces many characteristics of KAIST contact graph ($\eta = 2, m = 27$), whereas the proportional projection better applies on situations where there are several connected components or the graph structure is highly modular, *i.e.* there exist non-overlapped communities with few elements. For example, as we can see in Fig.2.12, proportional projection captures many features of Disney World dataset by reducing the number of paths connecting different communities.

To summarize, the bipartite graph and its projections well fit with our approach where nodes move in shared areas (geo-communities) and the sharing of these areas by different nodes generates contacts and social interactions. By applying the best projection on the bipartite graph, we can also infer and reproduce contact graph properties being able to generate synthetic contact traces.

2.7 Conclusion and future work

In this chapter we explore in depth the close relation between human mobility and the social structure induced by the superposition of different people's trajectories by adopting a data-driven approach on publicly available GPS traces collected in different urban and metropolitan areas. We have defined the notion of geo-location and geo-community which are operational in de-



(a) Geo-communities extracted from Disney World dataset

(b) Contact graph inferred from Disney World traces

(c) Preferential projection

(d) Proportional projection

Fig. 2.12 In this figure we synthesize our method. From the extracted geo-communities 2.12(a) and the bipartite graph form, we build two different projections 2.12(c), 2.12(d). We compare them to the contact graph 2.12(b) inducted by overlapping GPS traces. In this dataset, we can see how geo-communities position reflects on the high modular structure of the related contact graph.

scribing in a unique framework both spatial and social aspects of human behavior. On the basis of these definitions we develop a time efficient procedure for their extraction from GPS traces. By statistically analyzing the available GPS traces, we show how the main quantities involved in human movements are distributed. Our results on mobility features considering both geo-locations and the geo-communities viewpoint confirm most of the observations reported in literature even at the finest granularity. In addition to the classic properties we investigate how distance influences the choice of the next destination and we introduce a classification of the geo-locations which results in a categorization of the users in two mobility profile. Finally through the concept of geo-community we model the human mobility adopting a bipartite graph Thanks to this graph representation we can generate a social structure that is plausible w.r.t. the real interactions, by adopting some graph projections based on classical random graph models. The obtained random

projections have been evaluated by comparing the typical features of complex network analysis with real-world ones. The modeling approach have the merit for reporting the mobility in a graph-theoretic framework making the study of the interplay mobility/sociality more affordable and intuitive.

Chapter 3
Geo-CoMM: A geo-community based mobility model

3.1 Introduction

In recent years we have observed an exponential diffusion of mobile devices equipped with short-range wireless network interfaces (Near-Field Communication NFC , Bluetooth, Wi-Fi). These devices can be exploited to create an opportunistic connectivity for complementing both the wired and cellular, licensed network infrastructures. This scenario gives a strong background for the development of the so-called Opportunistic Networks (ONs)[37]. In ONs users are the carriers and messages are forwarded and eventually delivered to the destination upon the unstable links generated by the encounters with other users. Such a contact-based infrastructure and the performance of the algorithms running on it are highly influenced by human mobility. Therefore the deployment of novel mobile services requires understanding the behaviors in people's movements to exploit the patterns in human connectivity.

The fundamental issues in the design of new opportunistic services are two: *i)* the extraction and the analysis of human mobility pattern and *ii)* the evaluation of the underlying protocols under practical and real mobility conditions. In Chapter 2 we have faced the first one, here we concentrate on the realization of synthetic mobility traces. Two different evaluation approaches can be adopted to the purpose: the first is based on traces obtained from real environment, the second is based on synthetic traces generated from a mobility model. An ever-growing amount of datasets, based on different detecting technologies and performed in various locations, is currently available. However, these experiments involve a small number of nodes, are environment specific and cover a limited area (at most campus or urban area). Moreover, because of their lack of flexibility, the validation and the evaluation of the protocols are not general, but specific to the mobility sample. As a consequence, mobility simulation seems to be the best method to evaluate situations with a variable high number of nodes. Thereby, it becomes necessary to develop mobility models which can simulate different aspects of human mobility. As

we will see in Section 3.2 several mobility models have been recently proposed that overcome in expressiveness and complexity the classical Random Waypoint [29]. Some of them carefully reproduce many mobility patterns found in real traces but they focus on individual behaviors neglecting any social aspects; others consider sociality as the only factor influencing human movements and are not able to reproduce the spatial features [7]. Despite the recent efforts, the main research problem in mobility models remains still open, *i.e.* the need of simple (few parameters and easy setting) and realistic mobility model that should be validated with traces from different scenarios and able to reproduce the main spatial, temporal and social properties reported in literature [111].

The reproducibility is still a great issue in the mobility model literature, as most of the models have been validated only on the aspects they want to reproduce neglecting the other important features. In particular, in literature some key features have been indicated and categorized into spatial, temporal and social ones. Spatial properties indicate features which are extracted from the trajectories of mobile carriers. The main ones include the distribution of the length of human jumps and the radius of gyration. In particular the jump length has a huge impact on how messages are spread across the network since nodes that move over longer distances become bridges between far destinations increasing the coverage area. Among the temporal features, pause time (or waiting time) and return time are the most important. The pause time, *i.e.* the period that a user is in a specific place, influences the probability of making new contacts and their duration. Instead the return time, *i.e.* the tendency of humans to regularly return to the locations they visited before, makes the human mobility highly predictable [146]. As regards the so-called social properties, pairwise inter-contact time, *i.e.* the time intervals between consecutive contacts of a pair of nodes) and pairwise contact duration are very important. Indeed, the first impacts the distance between two consecutive forwarding opportunities and consequently the delivery delay, while the second restricts the quantity of the data that can be transmitted during each encounter. Recently, mainly because of the combination of routing protocol and community detection techniques, the structural properties of the contact network have become important. So the classical 'pairwise' paradigm has to be extended by including concepts from the social network analysis.

In this chapter we propose three main contributions in the field of human mobility model and in its feature validation. First we provide a new mobility model, we call Geo-CoMM, which lies on and exploits the human mobility patterns we analysed in Chapter 2 and the idea of geo-community. Essentially the model follows the location-based paradigm, where users are assigned to a subset of places. Here the assignment is done according to the affiliation network introduced in Section 2.6. In the model we consider two levels of mobility: among the geo-communities and inside a geo-community. In driving the movement among the geo-communities we adopt a distance based metric for the choice of the next location to overcome the unrealistic

star model, in which a node returns home every time it wants to change location, and to reproduce the jump length distribution. Inside a location, we introduce a truncated Lévy Walk in order to respect the statistics observed on real traces. Furthermore we consider some variety in human behaviors; in particular, we introduce the "jumper nodes" category, that is people moving quickly and almost randomly, which act as bridges between communities. The introduction of jumper nodes is justified by the observation [60] that in mobility traces a movement heterogeneity coexists with individual Lévy trajectories and by analysis of mobility patterns among the categories of visited location provides in Chapter 2.

Second, we provide a general framework that makes explicit the social structure behind the preferred-location based mobility models. The framework represents an extension of the user/geo-community bipartite graph where the users' mobility impacts the activation of the connection user/geo-community. Reporting Geo-CoMM in this framework we can highlight its social characteristics and in general reply to the criticism about the incapacity of the preferred-location model to reproduces the social features. In fact the social structure naturally emerges from the model without needing in input any social graph or social overlay.

Third, unlike most of the human mobility model reviewed in Section 3.2, in Section 3.5 we validate Geo-CoMM on spatial, temporal and pairwise connectivity features showing that it reproduces the main statistical properties observed in real traces. In light of the previous contribution about the social aspects of the model, we also focus on the properties of the social graph induced by the contacts. In particular we provide the characteristics of contact graph of some traces and we show that the model can properly reproduce them.

Summarizing, the obtained results show that Geo-CoMM can be adopted as a good generator of synthetic traces reproducing time and spatial probability distributions and social features of the human beings.

3.2 Related Work

Mobility models. Several mobility models have been recently proposed that overcome in expressiveness and complexity the classical Random Waypoint [29]. We can roughly divide these models into two categories: location driven and social driven models. In a location driven approach, the main goal is to reproduce individual movement behavior as observed in real mobility traces. One of the first proposal in this class of models is the ORBIT model [56]. ORBIT is a framework for the modeling a preferred-location models where nodes selects a spots and moves between them based on a customizable behavior. Realistic distribution of re-appearance frequencies is then achieved indirectly by defining probabilities of transitions between places, while other

properties to be reproduced need an estimate of the parameters on real traces. For example, this category contains SLAW [89], a model based on Lévy walk and on the minimization of the covered distance, that captures the main statistical distributions on relevant quantities as flight length, pause time, inter-contact time and speed observed in GPS-based traces. Another example is TVCM [71] that focuses on reproducing temporal (periodical reappearance) and spatial (location preference) regularities, two properties extracted from Wi-fi LAN-based traces. In order to capture preferential locations, it forces a node to visit its preferred location more often than other locations. In addition, TVCM defines time periods during which a node moves towards its preferred location with higher probability. This way, it reproduces periodical reappearance at the same location. Most of these models carefully reproduce many mobility patterns found in real traces; nevertheless they mainly focus on individual behaviors. In these models it results difficult and fictitious correlating the social features to the spatio-temporal variables. So much so that social ties (social graph, social overlay) are to be introduced to reproduce the features of real contact graphs. SWIM [13] is another approach based on location preference. The model assigns to each agent a so called home, which is a randomly and uniformly chosen point on the plane. The agent then selects a destination for next moves depending on the weight of each site, which grows with the popularity of the place and decreases with the distance from the home (in this way the model captures power-law distribution of the jumps). The popularity of a location, however, does not depend on the personal but on the overall preferences, and it is calculated as the number of other people encountered last time the agent visited the place.

By contrast, social driven models assume that social relationships heavily influence the mobility of people. In general their goal is to simulate mobility in order to reproduce the social interactions described by a social network. One of the first social driven model is the community based mobility model (CMM) [112]. In CMM nodes belong to a primary community and by a rewiring process they may have links to external communities. The mobility of nodes is induced by probabilities defined by community attraction. In order to overcome a gregarious behaviour, that is, all users tend to follow the first node that leaves the community where it is located, Boldrini *et al.* have proposed HCMM [22] which extends CMM, by introducing a home-cell in which nodes tend to spend most of their time. In SIMPS [24] nodes are continuously in motion and can assume two states. In the socialize state, nodes move in the direction of social relationships, while in the isolation state they try to escape from neighbours which they do not have social relations with. Switching process is controlled by a feedback loop so that each node needs for sociability to remain constant. In Heterogeneous Human Walk (HHW) [164] human mobility is based on heterogeneous centrality and overlapping community structure typical of social networks. HHW constructs a k-clique structure of overlapping communities built on common statistical properties, extracted by several real social networks. The idea on which these models

are based is that people move in order to respect a social relationship but, recently, Leskovec *et al.* [33] have shown that short-ranged travel is periodic both spatially and temporally and not effected by the social network structure.

Connectivity properties. Connectivity properties have been extensively studied in the context of mobile ad-hoc networks research. As messages are forwarded from node to node when they meet, the time between two consecutive contacts impacts overall delay, while the duration of the contact bounds the size of the data that can be exchanged. Moreover a contact between two devices implies that the corresponding users are close to each other, promoting mobile devices as a proxy of human movements. Thus, by extension, we take the inter-contact time and the contact time as measures of how frequent and how long two users spend time together. Specifically, we define the contact time between two mobile devices as the time interval during which two devices are in a radio range of one another, while the inter-contact time is the length of the time interval between two consecutive encounters. Chaintreau *et al.* [31] showed that the distribution of inter-contact time has a power-law nature over a wide range of values from few minutes to half a day. Later, Karagiannis *et al.* [81] extended this result suggesting that the power-law decay should be complemented with an exponential cut-off. Although less attention was dedicated to the duration of contact times, Hui *et al.* [72] showed that this quantity is distributed according to a power-law with exponential cutoff.

A recent work that is orthogonal to the above classification is the work by Hossmann *et al.* [69]. They have found that, regardless of the modeling approach to human mobility, the contact graph (*i.e.*, the graph whose vertices are the nodes of the network and whose edge weights are given by a combination of contact frequency and aggregate contact duration) generated by most synthetic models differs from that obtained from mobility traces.

This chapter presents a new mobility model, named Geo-CoMM, that attempts to overcome these limitations by considering social aggregation as spatio-temporal element. This is achieved by exploiting the concept of geo-community, which has been treated in Chapter 2, as a place visited by a subset of users who potentially raise up in a community.

3.3 The Geo-CoMM Model

Using the results in Chapter 2 on the analysis of GPS-based traces, we propose the notion of geo-community as the starting point of our modeling approach. We consider nodes moving among geo-communities where they spend some amount of time. In our model nodes are driven by several quantities, whose probability distributions were analyzed in [172], including the factors that influence the choice of the next destination. In particular, we only focus

on distance and do not consider any social factor, a feature not yet statistically studied to the best of our knowledge.

In our analysis it emerges that geo-communities have different shapes. This is due to the fact that they approximately reproduce the shape of buildings or other shared geographical areas. In our model we simplify these characteristics by considering geo-communities of circular shape. Thus, a geo-community is described by a radius r and by a center point c. In order to maintain a link with the geo-community density analysis in Chapter 2, we reasonably set r=*70* m. This way we consider geo-communities of the same order of magnitude of the average dimension computed on real traces.

In order to reproduce different scenarios, we consider a squared simulation area of 100 km^2, thus allowing the comparison of simulation results with GPS traces. Over this area we spread geo-communities. We extract their centers from a fixed spatial probability distribution, whose aim is to model different urban conditions, ranging from uniform to normal buildings distribution. Coherently with the definition of geo-community and given a fixed radius r, we also impose a non-overlapping constrain of the circle.

In order to capture and reproduce workday mobility patterns, we focus on a fixed temporal range. In particular, nodes are considered active in the time period $[8:00+u_{start}, 20:00+u_{end}]$, where u_{start} and u_{end} are uniformly distributed on $[-60, 60]$ sec.

So far, we have considered only those characteristics which regard the geographical environment and the temporal constrains. A preliminary step is the definition of the initial statement, that is the assignment to each node v of a list $l(v)$ of geo-communities among which it is allowed to move. More specifically, in the assignment process of each node, we generate a random permutation of the geo-communities and we take the first $k+m$ elements, where m follows a normal distribution, $m \approx N(0, \sigma)$. This allows to obtain two parameters controlling the node to geo-community association: k e σ. During this process we do not assume any preferential geo-community, as other models do [22],[71]; we only consider the first geo-community in the permutation as the starting one. Inside this geo-community we chose a random uniform point as initial point of the movement.

For the sake of better understanding mobility and modularizing our model, we split human walks in two phases: movement inside geo-communities and movement among them.

Inside a geo-community. Inside a geo-community nodes move following a Lévy Walk. According to this model, the distances covered by a single node are distributed according to a power-law. This simple model reflects the requirement of capturing some properties about distances that came out from our analysis in [172]. In particular, we observed that transition lengths follow a Pareto distribution. In order to restrict the movement within a limited area (geo-community) we adopt the following procedure to generate waypoints. At each step, we generate two random variables: an angle $\theta \approx U(0, 2\pi)$ and a distance $r \approx Pareto(\alpha, x_{min})$. These two variables define a movement vec-

tor **v** that we add to the node position. We discard the destination point if the node leaves of the circle. We repeat the procedure until the new point is inside the geo-community area. As we consider a limited area which forces the algorithm to discard distances, the distance distribution of our model properly follows a truncated Pareto distribution. In fact we can observe a truncation effect in the distance CCDF.

Among geo-communities. In order to model node movements among different geo-communities and to introduce a mechanism to choose the next destination, we attached to each geo-community a list of the remaining communities sorted by their geographical distances. After list creation, each node has all the informations to move in the environment. For example, let us suppose that a node just entered a geo-community i. It has to spend an amount of time in this community and then it has to decide which is its next destination, using as metric the geographical distance.

Formally, the movement among geo-communities is described by two probability distributions. The first one, denoted by $P_{pause}(t)$, describes the pause time and it follows a tapered Pareto distribution, whose CDF is given by

$$F_{tap}(t) = 1 - \frac{t_{min}}{t}^{\beta} exp\left(\frac{t_{min} - t}{\theta}\right) \tag{3.1}$$

where $t_{min} \leq t < +\infty$. This choice is based on the fact that a person usually spends a lot of time in a given set of locations, and sometimes makes little breaks in other locations (breaks during the movement, coffee-break in the nearest bar and so on). This attitude is also supported by results in [89, 172, 60].

In the definition of the movement among geo-communities and in the choice of the next goal, a fundamental role is played by the distribution $P_i^{move}(j)$, which denotes the probability that a node in the geo-community i has to move towards the geo-community j, which belongs to the ordered list associated to community i. We set $P_i^{move}(j)$ according to a power-law with exponential cut-off distribution; in particular we choose a truncated zeta distribution. The aim is to capture the tendency of a person to move towards closer destinations and sometimes to jump to far locations. This characteristic has been analyzed by several works on different traces [60, 138].

The choosing process for a given node can be modeled by a finite time-homogeneous Markov chain, summarized in Fig.3.1, where the states are the geo-communities that can be visited by the node and $Pr(X_{n+1} = j|X_n = i) = P_i^{move}(j)$. Each element on the diagonal of the associated transition matrix M is equal to 0 (no self-loop) and M is primitive.

Unlike other models [22], [71] where the movement is based on a Markov chain, we do not follow a "star model". This is motivated by the need of avoiding a node to return to a fixed location before moving to another.

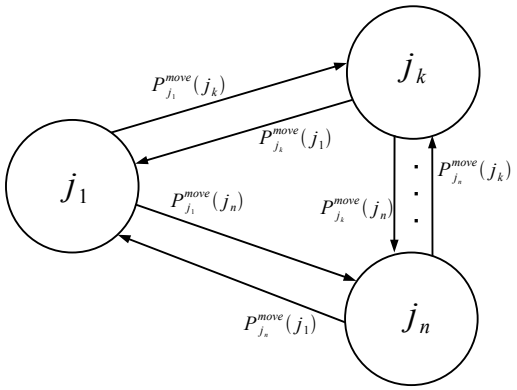


Fig. 3.1 Example of the Markov chain transition matrix driving the choice of the next geo-community.

3.3.1 Node classes

The human attitude to visit closer locations well suits to different situations, especially when considering working day activities. Nevertheless, in real-life, we can observe that a few people, for instance for professional reasons, do not follow the previous choice (as also evidenced in Chapter 2). In particular, they happen to "'jump"' among different locations because of scheduling constrains that do not involve location distances. This kind of people represents, in a contact prospective, random neighbors and their presence could be exploited to speed up information spreading between different communities.

With the aim of introducing in our model the previous profiles, we consider three types of nodes:

- **standard node**: these nodes mimic the standard situation described by pause times following a power-law with exponential cut-off distribution and preference to closer geo-communities.
- **quick standard node**: these nodes are very similar to standard, but they tend to stay in geo-communities for shorter time.
- **jumper node**: nodes who have not a preferential mechanism but they choose uniformly their next destination. They remain in a geo-community for a short time.

In Section 3.5 we evaluate how the presence of jumpers and quick standard nodes influences contact behavior and statistics.

3.4 Social aspects of the model

In this section we introduce a general framework that could incorporate many mobility models presented in Section 3.2 and report the different approaches into a unique formalization which extend the idea of affiliation network used in Chapter 2 by adding the temporal dimension. Here first we recap the basic assumptions and concepts on the basis of the social and the preferred-location modeling approach, then we provide the definition of time-varying bipartite graph. Finally we bring the modeling methods into the framework and what validation problems arise.

Social-based mobility models rely on the assumption that the social graph and its community structure highly impacts on human mobility, therefore people move among the graph communities, which should correspond to meeting places. Usually in real social network such groups are overlapped, because a user may belong to different social circles. In these model the social dimension is given explicitly as input so that output movements heavily depend on the social graph and on its properties, *i.e.* hierarchical and modular structure, scale-free degree distribution, nested modules. The approach is depicted in Fig.3.2(a) where from the social network we apply a community detection algorithm [1] to extract cohesive groups.

On the contrary preferred-location models do not make any assumption on the social structure of the users, but they hypothesize that the social relationships are determined by the sharing of some meeting places. So people do not move because they want to socialize but they are social because they move. In these models the input is made up by the assignment user/location and by a decision rule in the choice of the next destination. A generalization of a preferred-location model is reported in Fig.3.2b) where we indicate the places in the physical space assigned to two different users and the choice mechanism ϕ.

As we can note both the approaches reckon on an association mechanism. In the social-based model the assignment results from the community detection algorithm which divides the users among the social groups, while in the preferred-location case the association is explicitly asserted as input. Anyhow we can exploit the affiliation network $G = (U, V, E)$ (see Chapter 2) where $u_i \in U$ are the users and $v_j \in V$ represent the social or physical places. This way the affiliation network unifies the representations adopted in both methods. In order to add the temporal dimension to the framework and consequently introduce the dynamical aspects, we extend the affiliation network adding a time function which activates/deactivate the user/place assignment.

Definition 3.1. Given an undirected bipartite graph $G = (U, V, E)$ where U is the set of the user and V is the set of social/physical places, we define the **time-varying bipartite graph** $G_t = (U, V, E, \Psi)$ where the function

[1] In the light of the results about real social networks [162], the best choices are community detection algorithms for weakly [122] or strongly [160] overlapped communities

$\Psi : E \times \mathbb{R} \rightarrow \{0, 1\}$ is the link activation function. $\Psi((u, v), t)$ indicates if a user u is visiting the location v at time t.

The values of the link activation function, which takes as argument the links incident to (u), determine the sequence of the places visited by the user u. Consequently the shape of the function Ψ is fundamental in the determination of the movement. Obviously different models correspond to diverse activation functions. For instance, in Geo-CoMM, it will be based on the pause-time distribution and on the Markov chain.

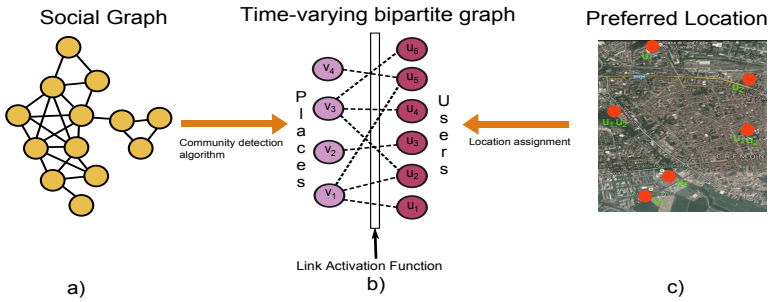


Fig. 3.2 a) Social-based models receive as input the social network of the users and then they apply a community detection algorithm which associates each node to a subset of social groups. c) Preferred-location models position the locations in the spatial dimension and the assign each node to a subset of physical places. b) The assignments in the social and location approaches we can be generalized to a time-varying bipartite graph where the movements are driven by the link activation function.

In general, as highlighted in Fig.3.2c), the time-varying bipartite graph represents the point of convergence of the social and preferred-location modeling approaches. Starting from the common formalism we can proceed in two ways depending on the spatial information we have about the set V.

If localizations of the places are unknown, as in the social-based approach, we need to map the social places in physical location complying with the constrains imposed by the distances and the movement times. This represents an issue in the approach since the mapping into physical location influences the link activation function. Therefore a possible solution is to directly incorporate the mapping into the definition of the function as occurs in HCMM or SWIM.

On the other hand, in the preferred-location case, we are able to make explicit the social relationships resulting from the sharing of common locations. In fact given a place $v \in V$ and time slice Δt, by the activation function we can retrieve the users u, s.t. $\Psi((u, v), \Delta t) = 1$ [2], and form a clique with them. Repeating the procedure for each location we obtained the social interactions among the users during the period Δt as shown in Fig.3.3. Applying the same

[2] Abusing of the notation, $\Psi((u, v), \Delta t) = 1$ denotes that the link is active for a time Δt

method for each non-overlapping time window, we are able to infer the evolution of the social interaction caused by the movement of the users among the shared locations. As depicted in Fig.3.3, we can reconstruct the overall social graph by eliminating the temporal dimension and flattening the graph, *i.e.* we compute the union the graphs corresponding to each time windows. In the described method the link activation function and the location assignment have a key role. The computation of the cliques heavily depends on the Ψ function, as the co-presence of the nodes assigned to a specific location relies on the simultaneous activation of the association user/place. The assignment is equally important as we need to guarantee that location are shared by at least two users to have some possibilities to interact.

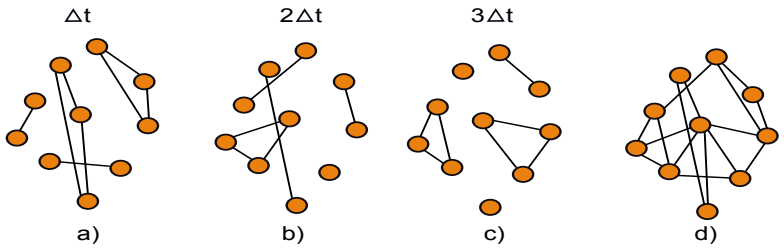


Fig. 3.3 a-c) Graphs of the relationships induced by sharing the same location at different non-overlapping time windows. The union of the graphs in a), b) and c).

Summarizing we have reported the main approaches adopted in human mobility modeling into a general graph-theoretic framework based on the concept of time-varying bipartite graph. In this framework we are able to explicit the social structure induced by the mobility of nodes among preferred location, as in Geo-CoMM. This way we can reply to the criticism about the inability of preferred-location models to capture the social characteristics of the human mobility.

3.5 Evaluation

In this section we consider most of the features analyzed in literature. In particular we analyze synthetic mobility traces and the properties of their contacts from three different viewpoints, comparing them to real measurements. The first one is the one-node perspective and regards mobility features of a single node as flight length, spatial regularities and pause time distribution. The second viewpoint involves pairwise properties such as inter-contact time, contact duration and contact frequency, while in the third perspective we take into account the structural properties of contacts. Following the ap-

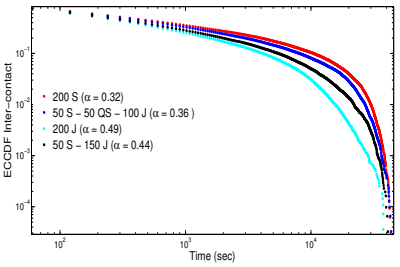
proach described in [66], we construct a weighted contact graph W, where each node is a vertex and a link weight w_{ij} represents the degree of the relationship between nodes i and j. In order to compare the contact graph of our model with those inferred from some real traces, we apply the same assignment methodology proposed in [66]. Hence we consider two contact features: contact frequency and aggregate contact duration and we combine them to assign weights to links.

Our interests include both global topological properties, such as degree distribution, clustering coefficient or diameter, and hierarchical and local properties, such as the presence of communities and their properties. In particular, we need to define the degree and weight distributions inside a community and between different communities. In fact, such distributions play a fundamental role in the capacity of spread information inside and across communities. For example, if several heavy links between two communities exist, we could balance the traffic load exploiting all of them. To test these properties we have developed a custom simulator in MATLAB. We compare traces generated by the simulator to some direct-contact traces and GPS traces available in CRAWDAD repository. In particular, we present results for Infocom05 [72] and NCSU [138] datasets.

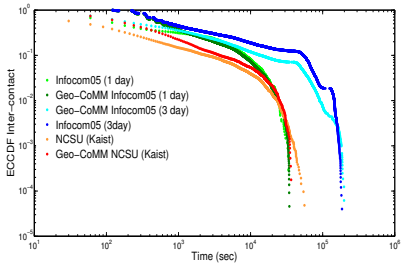
In the analysis of different human mobility traces we can identify some standard metrics and properties that a realistic mobility model should capture:

- *inter-contact time*, *i.e.* the interval between two successive contacts of a pair of nodes. It seems an established result that the aggregated CDF is distributed as a power-law up to specific time (half a day), thereafter followed by an exponential decay [81].
- *contact duration*. As shown in [138, 72, 49] also contact duration follows a power-law distribution with exponential cut-off.
- *number of neighbors*. Given a contact radius r, we say that nodes are in contact if their contact circles overlap. In analyzing the numbers of neighbors we can choose to consider or not a temporal dimension. In the first case, we evaluate the neighbor fluctuation along the day, while in the second case we aggregate the number of contacts. Some results about the number of neighbors have been presented in [72, 44].
- *flight length*. By flight length we mean the distance covered by a node between two successive pauses. As regards this metric, [60] and [138] showed that it follows a power-law distribution with exponential cut-off. In [172], flight length has been analyzed both inside a geo-community and among geo-communities, finding similar results.
- *spatial regularity*. Gonzalez *et al.* [60] showed that each individual has a significant probability to return to a few highly frequented locations. In particular the probability $P(R)$ of finding a node at a location of rank R, where R represented the R-most visited location, is approximated by R^{-1}.

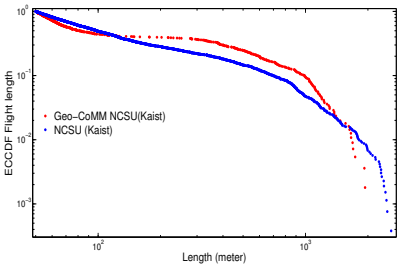
3.5.1 Simulation results on one-node and pairwise properties



(a) Inter-contact distributions for different types of node (S=standard, QS=quick standard, J=jumper).



(b) Comparison between Infocom05(3 days), Infocom05(1 day) and NSCU inter-contact and synthetic inter-contacts generated by Geo-CoMM.



(c) NCSU e Geo-CoMM flight length distribution.

Fig. 3.4 Simulation results on the temporal and spatial properties.

In the following, we investigate some aspects of Geo-CoMM in comparison to real traces, hence the simulation parameters were chosen in order to fit such real settings. As regards NCSU traces, it results quite easy to infer parameters because of the previous statistical analysis based on geo-community [172]. Hence, if not stated otherwise, we consider a squared area of 5000×5000 m^2 and 13 geo-communities of 70 m of radius, uniformly distributed in the area. The assignment parameters are $k = 6$ e $\sigma = 0.7$. In the definition of the Pareto Waypoint we set $\alpha = 3$ and $x_{min} = 10$, while pause time and the next geo-community distributions are set to $\alpha_{pause} = 0.2$, $x_{min}^{pause} = 2$h and $\alpha_{next} = 0.97$. Furthermore, as in NCSU traces, we consider 90 nodes.

1) Inter-contact: We compare Geo-CoMM inter-contact distribution to the one computed on Infocom05 trace. The number of nodes and the days of simulation are the same. We cut the simulation area (200×200 m^2), considering that the experiment was hosted in a hotel, and we set 10 geo-communities.

We also set the radius of geo-community, considering conference room dimensions, to 20 m. In order to mimic people seat in conference room, we limit the Pareto Waypoint mobility considering $x_{min} = 0.5$ and $\alpha = 0.01$. The remaining parameters have not been changed. The results of this experiment are shown in Fig.3.4(b). In this figure we also consider Infocom inter-contact distribution in a single day and traces taken from NCSU dataset. In all of the three traces, Geo-CoMM proves to properly approximate real inter-contact distributions.

In Fig.3.4(a) we show how the presence of jumpers influences the inter-contact distribution. We can observe that jumper nodes do not modify the shape of the distribution (a power-law with exponential cut-off), but they change the shape parameter. In particular, the increasing of the number of jumper nodes has the effect of raising the shape parameter. This implies that the presence of jumpers node reduces the inter-contact time between nodes.

2) Contact duration: In our evaluation we consider contact duration distribution and we observe that it follows a power-law distribution as in real traces [72],[138]. In particular, we check if jumpers also modify this distribution. The presence of jumper nodes do not affect the contact duration distribution. More precisely, if we increase the number of jumpers, the shape parameter slightly decreases.

3) Spatial regularity: Gonzalez *et al.* [60] and Hsu *et al.* [71] have shown that individuals have a preference for a small number of locations, while visiting all other locations only few times. In order to evaluate this property in our model, we consider a single node and a running period of 25 days. As described by the model, every day a node starts from the same location, so we do not consider it. Geo-CoMM exhibits spatial regularity properties, in particular a node prefers some locations. We can explain this characteristic by considering that the decision process tends to favor closer destinations.

4) Flight length: In Fig.3.4(c) we show a comparison between real trace flight length taken from NCSU dataset and distance resulting from simulation of Geo-CoMM using a standard setting. We can see that the synthetic trace well approximates the real one. Geo-CoMM seems also good in capturing the flight length distribution among geo-communities extracted from GPS traces.

3.5.2 Contact graph of the comparison real traces

So far we have considered the pairwise properties of the contacts, but recently some studies have attempted to go beyond these statistics. In particular, they have also considered the structural properties of the network of contacts. To study these patterns different compact and manageable representations have been presented. A common process is to aggregate the whole sequence of contacts in a weighted graph. In our work we follow the approach of Hossmann *et al.* [66]. In their method, the authors consider a set S of pairs (f_{ij}, t_{ij}) where

f_{ij} and t_{ij} respectively represent the frequency and the aggregated duration of the contacts between i and j. The weight assigned to the edge (i, j) is the projection on the principal component of S. Applying this projection, we obtain a single value combining the frequency and the aggregate duration and we can manage the structure of the contacts with a weighted simple graph.

Previously we showed that the synthetic traces produced by our model well reproduced the pairwise statistics. Now we want to verify that these traces have a contact structure similar to the real one. To achieve this goal we analyze the contact graphs of the datasets Infocom05, Infocom06 and NCSU. Regarding to the last dataset, we consider the traces collected in the KAIST campus because they contain the largest number of nodes. Once built the contact graph, we examine whether it possesses a small-world structure. First, we calculate the density $D = |E|/|V||V-1|$ obtaining the following values:

$$D_{I05} = 0.46, \;\; D_{I06} = 0.48, \;\; D_{Ka} = 0.21$$

Proceeding in the analysis of the properties characterizing small-world structure, we consider the average shortest path length and the average clustering coefficient. By applying the same threshold mechanism on the contact graph presented in [66], we calculate the previous metrics on the resulting binary graph. Such threshold allows us to control the graph density of the binary graphs that we investigate. By using the density values suggested in [66], we obtained disconnected graphs. This fact suggests that a backbone, connecting all the nodes, does not exist. In light of this observation we find the density values corresponding to the emergence of a giant component. Results are summarized in Table3.1.

	0.01	**0.02**	**0.03**	**0.04**	**0.15**	**0.20**	**0.30**	**0.38**
Infocom05	0.23	0.32	0.36	0.40	1.78	1.64	1.40	1.22
Infocom06	0.07	0.10	0.27	0.31	-	-	-	1.21
KAIST	0.21	0.35	0.38	0.44	1.72	1.57	-	-

Table 3.1 Average clustering coefficient and average shortest path length considering different densities. - (missing value) indicates that the related binary graph is disconnected or the density exceeds the maximum value.

Regarding the average shortest path length, we can observe that, when the giant component appears, we have very low values suggesting a compact graph. Concerning the average clustering coefficient we can say that the scenarios taken into account are less clustered than other datasets analyzed in literature, however it increases enlarging the graph density. This phenomenon suggests that less frequent or short duration contacts raise the transitivity of the relationships. After analyzing local clustering, we consider the global modularity of the contact graphs. In particular we extract community applying the Louvain community detection algorithm which attempt to maximize the modularity function Q. The number of communities and the related Q

values are reported in Table3.2. As regards modularity we can see that in the conference datasets (Infocom05 and Infocom06) it takes low values, especially in Infocom06 the Q value is close to a random community assignment. We can observe a difference between campus and conference scenarios also analyzing the community sizes. In particular, in the conference datasets we observe that most of communities are small, while in the Kaist dataset half of the communities contain more than 10 people.

Trace	# Comm	Modularity Q
Infocom05	13	0.18
Infocom06	16	0.09
KAIST	7	0.40

Table 3.2 Number of communities and modularity Q

The subdivision of users in communities allows us to analyze the distribution of the weights inside and between communities. Regarding the first variable, for each dataset, we calculate the median of the intra-community and the global weights. As expected, the median of the intra-community weights is greater than the global one. We also plot the CCDF of the global and intra-community weights in Fig.3.5. We observe that the distributions differ between the datasets. The straight line in log-linear scale implies a distribution close to exponential of the KAIST weights. For Infocom06 the plot suggests a truncated Pareto or log-normal shape, while in Infocom05 we can conjecture a truncated Pareto distribution for the global weights and an exponential distribution for the intra-community weights. In order to evaluate the connectivity of a node globally and inside the community which it belongs to, we calculate the *normalized global degree* defined as

$$\frac{1}{|V|}\sum_{j\in N} w_{ij} \tag{3.2}$$

and the normalized community degree

$$d_{c_i} = \frac{1}{|c_i|}\sum_{j\in c_i} w_{ij}. \tag{3.3}$$

By calculating the median of these two variables, we see that it is much higher for community degrees. In Fig.3.6 we report the CCDF of the normalized degrees. We can visually suggest that distributions could follow an exponential law.

So far we have considered some of the features of the internal structure of the communities, however we can analyze how weights are distributed between communities. To achieve this goal, we investigate the inter-community weights w_{ij} such that $c_i \neq c_j$ and $i > j$ as we deal with an unweighted graph.

If we plot the CCDF of the inter-community weights (see Fig.3.7) we can observe two types of distribution. In the Infocom05 dataset, the distribution seems to follow a trend with two regimes, while in the remaining ones the distributions are close to an exponential.

3.5.3 Contact graph of the model

By applying the standard setting on 100 nodes, we analyze some properties of the contact graph induced by synthetic traces. In particular, we focus on some global properties regarding node degrees and link weights. We also analyze some metrics concerning the structural properties of the contact graph. The approach allows to study node communities and the degree and weight distributions inside and between them. This last property can influence the sharing of information across communities. In the following we present some results:

1) Small-world properties: We calculate the contact graph density extracted from synthetic traces and we observe that density $D_{Geo} = 0.15$ is similar to the real one ($D_{Ka} = 0.21$). Regarding to the average shortest path length we see that the giant component emerges considering a density equal to 0.05. We have similar average path lengths values (1.68) to the density value corresponding to the appearance of the giant component in the real graph. As concerns average clustering coefficient we obtain the following values $cc_{0.01} = 0.15$, $cc_{0.02} = 0.27$, $cc_{0.03} = 0.34$, $cc_{0.04} = 0.37$. These values are very similar to the real ones.

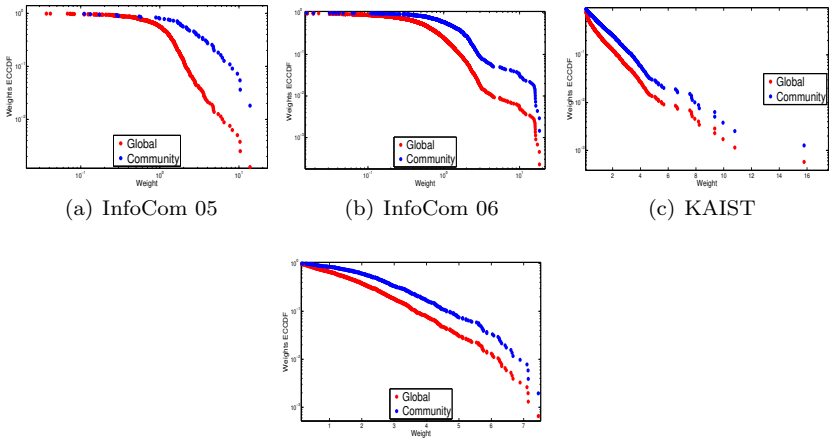
2) Communities and modularity: We obtain a modularity $Q = 0.32$ close to the real one (0.4) and most of the communities containing more than 10 people as in the real scenario.

Global and intra-community weights: We plot these metrics in Fig.3.5(d). The straight line in log-linear scale implies a distribution close to exponential. Previous we have made the same observation for the KAIST dataset. So, synthetic traces can reproduce this property.

4) Normalized global and community degrees: If we consider Fig.3.6(d) we can see that traces produced by Geo-CoMM follow the same trend of the real ones. In the evaluation of node degrees, we run Geo-CoMM considering only 100 jumpers node and we observe that the node degree distribution follows a log-normal distribution and the node degree values are smaller than in all standard setting.

5) Inter-community weights: The plot of the CCDF of the inter-community weights in Fig.3.7(d) visually suggests that this quantity closely follows the same trend of the real traces. Consequently, even regarding inter-community weights, Geo-CoMM reproduces the properties typical of real traces. Concerning this property, we compare all standard setting with a 70 standard-70

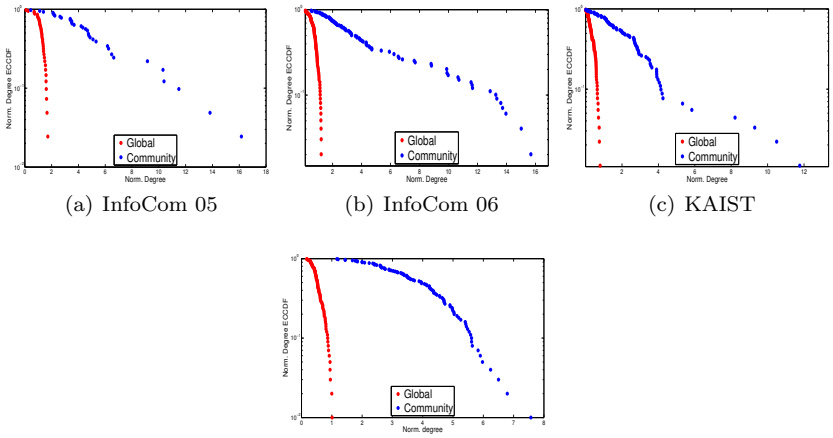
jumper setting and we find that the introduction of jumper nodes tends to uniform inter community weights and to reduce a little bit weight values.



(a) InfoCom 05 (b) InfoCom 06 (c) KAIST

(d) Geo-CoMM

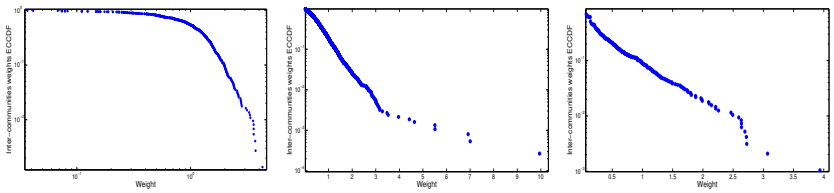
Fig. 3.5 Weight ECCDFs, global and intra-community weights. 3.5(a) and 3.5(b) are in log-log scale, 3.5(c) and 3.5(d) in log-linear scale.



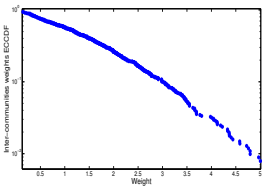
(a) InfoCom 05 (b) InfoCom 06 (c) KAIST

(d) Geo-CoMM

Fig. 3.6 ECCDFs for normalized global and community degree in log-linear scale.



(a) InfoCom 05 (b) InfoCom 06 (c) KAIST



(d) Geo-CoMM

Fig. 3.7 ECCDFs for inter-communities weights. 3.7(a) is in log-log scale, 3.7(b), 3.7(c), and 3.7(d) in log-linear scale.

3.6 Conclusion and future work

In this chapter we provide a new mobility model, which names Geo-CoMM, which lies on and exploits the idea of geo-community. Essentially the model follows the location-based paradigm, where users are assigned to a subset of places. The users assignment essentially relies on an affiliation network on top of which we run a finite time homogeneous Markov chain where the states are the geo-communities linked to the moving node and where the transition probability is a function of the rank on their distances. The Markov chain drives the movement among geo-communities, while within a geo-community we adopt a particular variety of non-uniform random model similar to a Lévy walk. Furthermore, we provide a general framework where the social structure behind the preferred-location based mobility models emerges. The framework provides a temporal extension of the user/geo-community affiliation network, in which the users' mobility impacts the activation of the connection user/geo-community. Reporting Geo-CoMM in this framework we can highlight its social characteristics, in fact the social structure emerges from the model without imposing any social graph or social overlay. Finally we validate Geo-CoMM on spatial, temporal and pairwise connectivity features showing that it reproduces the main statistical properties observed in real traces. We also focus on the social graph induced by the contacts. In particular we provide the characteristics of contact graph of some traces and we show that the model can properly reproduce them.

As future works, we plan to introduce a synchronization mechanism in the time-varying bipartite graph, as up to now the choice made by node are

independent from one another. The synchronization might be obtained by maximize an utility function which includes social and spatial variables. In particular,to search for an optimal activation function we plan to act on the transition matrix of the Markov chain, changing the choice probabilities. Furthermore as concerns the validation aspects, we need more traces to further validate the model.

Chapter 4
Multidimensional Complex Network for Online and Offline Sociality

4.1 Introduction

If online social networks (OSNs) were to mirror the offline sociality of individuals, they would be able to reflect offline relationships and unveil the social behaviours that impact on online sociality. Unfortunately, there is a growing belief that today's social networks are quickly shifting away from their original goal because of marketing, fake accounts and profiles not related to human beings; and, by contrast, sparking fears about the fact that they are drifting towards a highly connected, unstructured and flat social graph. The in-depth understanding of relationships between online and offline sociality, beyond being a key issue of Human Sciences, would produce the practical effect of promoting OSNs to the status of best platform for the effective delivery of mobile computing services (e.g. recommendation systems, advertising, content dissemination, crowd sourcing, social discovery, etc.). In fact, the online deployment of these services would benefit offline social knowledge, for instance, improving trustworthiness of a service, tailoring it according to the target's interests, leveraging context information or predicting the impact on OSN structure of an upcoming event. On the opposite end, it would definitely help in deploying improved mobile services which leverage their online life features.

The above arguments concern a specific research interest whose main goal is to understand the relationship between the two faces of sociality. This is of interest at all investigation scales. At microscopic level, for example, the user can judge his/her role in real society and in the online world, by understanding how his/her centrality measures change from the first to the second one, thus allowing him/her to improve his/her popularity in the networks. At mesoscopic level, it is interesting to deepen how groups and communities change in the two worlds. Besides, at each level, the two networks can be merged to get a global overview of user's sociality which allows us to exploit

the knowledge of the social features in one of the two worlds to enhance the comprehension of the other one.

As we will see in Section 4.2, the challenges in answering these questions are both experimental and theoretical. On the one hand, although large datasets describing online social networks have recently been made available together with an extensive literature, datasets concerning offline encounters are few. Moreover they combine both explicit and extemporaneous contacts because of the incapability of the short range radio technology currently adopted to discriminate the willing contacts. As a consequence, the research community has very few opportunities to compare the datasets of offline encounters and online relationships of the same group of individuals. On the other hand, while the modeling of a single layer of sociality has been successfully faced by means of complex network theory, the merging of interrelated complex networks still presents theoretical aspects to be investigated.

This chapter offers a first complete contribution to face these challenges. To address the data lack problem, in Section 4.3 we describe data from an experiment we ran, that enables us to answer the above challenges by exploring the intimate relation between online and offline sociality of a group of students. Data describing the offline sociality of a set of 35 volunteers were purposely collected in a time span of one month and then integrated and compared with relevant data about their online sociality extracted from Facebook. The dataset has been further enriched by recording contextual and temporal information of the encounters, in order to complete the online friendships with more than the pure real friendship information, thus allowing us to deepen the impact of real life contexts on people's online relationships. The influence of contextual information, besides acting on the online level, has a deep impact on the offline dimension itself. For instance it has been shown that contact features may change from a context to another. So before delving into the relation between offline/online sociality, in Section 4.4 we highlight the role of contexts and their effects on the classic encounter properties, *i.e.* duration, frequency and the type of relationship of the involved people.

Re the theoretical aspects, in Section 4.5 we model the interleaving among the information about offline and online sociality through a multidimensional complex network, *i.e.* an undirected multigraph, where each dimension accounts for one kind of sociality. From the overall model we extract the separated networks referring to the different social worlds: the contact and the Facebook layers. The dimension separation enables us to discover interesting features about the single layers and the relation between people's online and offline sociality. In Section 4.6 we show that the overlapping degree is low; in fact, the sets of Facebook and real contacts are quite different. Secondly, by comparing the ranking induced by several centrality metrics, in Section 4.7 shows that node centrality is not a universal feature. In fact, node centrality is not linearly transferred across layers and, as a consequence, the people's popularity is most likely to change among the different networks. Finally,

by exploiting two ways of projecting the network dimensions into a single graph (see Section 4.5) we achieve very interesting results on the relationship between online and offline sociality and the related role of contexts in the communication shortest path and in the community structure. More precisely in Section 4.8, besides providing a comparison between the Facebook and the contact communities, we show how the introduction in the Facebook network of external information about the real-world interactions among the active users can reshape the modular structure.

In general the resulting work presented in this chapter is one of the first effort in understanding the fundamental relation between the different aspects of the real-world or offline sociality and the online social activities. The novelty resides not only on the tool adopted, *i.e.* multidimensional networks, but also on the rich released dataset. Moreover, the obtained results could have practical implications especially on those methods and approaches which rely on online information to deliver service on the offline world.

4.2 Related Work

Online and offline dataset. While there is a very extensive literature on online social networks, research on offline sociality and how it relates to online friendships is still in its infancy.

As already seen in Chapter 3, the first works that deal with automatic detection of the expression of offline sociality, *i.e.* meetings or contacts, are referable to the DTN and opportunistic network literature. Many experiments collecting contacts and mobility traces have been developed in this research field. Contact detection methodologies and the obtained connectivity traces can be categorized according to the technology adopted in recognizing the encounters and to the granularity of the sample rate (seconds, minutes, hours or days) and the coverage area.

The most coarse-grained method to obtain contact or proximity traces is to exploit wireless systems (GSM or WLAN) and monitor the connection of the devices being traced. If two devices are connected to the same cell or AP, then users are assumed to be in contact. The proximity accuracy is limited by *i*) the wide area covered by the system antenna; *ii*) the density of the access points and *iii*) the correlation between device signal strength and its distance to the access point, which does not always hold. Collected data employing this methodology and their analysis can be find in Henderson *et al.* [65], McNett and Voelker [106], Hsu and Helmy [70], although none of them take into account the online connectivity of the traced users.

A second approach, that aims at overcoming WLAN drawbacks, is the use of the Bluetooth and other short-range ad-hoc wireless protocols that sniff for other mobile devices around them. This way fine-grained offline sociality traces can be acquired. All other devices in range are seen as a contact. In

fact, while Bluetooth resolves the coverage area issue, it introduces noise in data as it also records extemporaneous, opportunistic and unwilling contacts. Moreover device inquiry and service discovery drop down the connection success rate; for instance Pietilainen *et al.* [129] report to identify only 50% of the neighbouring devices. Bluetooth contact detection was adopted in Hui *et al.* [72], Chaintreau *et al.* [31], Natarajan *et al.* [114], Mtibaa *et al.* [109] and Su *et al.* [148]. Some improvements in the discovery service have been reached by employing other short-radio technologies or ad-hoc protocol. For example, in the SocioPattern project, Panisson *et al.* [123] developed a badge equipped with a RFID chip to detect face-to-face contacts. Such approach reduces the number of unwilling and opportunistic contacts as the human body acts as a natural shield against radio waves. On the other hand Gaito *et al.* [50] developed a wireless device able to detect and record very short contact events with a temporal granularity comparable with GPS traces [173]. Although short-range radio technologies offer a better reliability in capturing offline social interactions, none of the previous experiments merges online and offline interactions. A special mention should be made to the Reality Mining [43] and the Social Evolution [99] projects, that enrich proximity data with rich personal behaviors such as phone calls, text messages, location and personal survey about work, free-time and social relationship. Despite the rich multidimensional datasets, also in these project the online sociality and interactions have been disregarded. In general the analysis of such traces has shown that there is some correlation between mobility and social connections. However these studies fail to reveal which nodes would actually experience an encounter during which they could communicate.

Some experiments have attempted to collect data on offline and online social relationships. Their main goal is to exploit these data for purposes of designing opportunistic routing algorithms that take into account online sociality. The first one, described in [19] gathers contacts detected by an ad-hoc wireless device, called I-Mote, and the Facebook graph restricted to the participants. A similar approach was performed in [129], where they adopt Bluetooth as contact detection technology and create a new online social network lying on a opportunistic middleware known as MobyClique. Both experiments suffer from limitations due to the detecting technology: they detect proximity and not an encounter between willing parties. Furthermore they employ ad-hoc or new technologies not so widespread in the population.

A small step forward came from the experiments in [67] and in [150]. In [67] the authors developed a Facebook application where a small group of experimenters reported their daily face-to-face meetings with a subset of their Facebook friends. This way, however, only relationships among Facebook friends can be analyzed, so leaving out all friends in real life who are not Facebook contacts. Barrat *et al.* [12] [150] developed the Live Social Semantics platform that collects and integrates data about people from their online social networks and tagging activities , their publications and co-authorship

networks and their offline face-to-face contacts with other attendees collected by means of RFID sensors.

Last experiments suffer from an intrinsic limitation given by the sampling mechanism they adopt. As users need a specific application or sensor to detect online or offline contacts, they only collect data among the experiment, a typical drawback given by the so called "snowball sampling" in the sociology literature. That results in difficulties when we compare the online/offline layers, as the sampling mechanism on a layer influences how the other layer is sampled [67]. The approach adopted in our work overcome these limitations because it independently samples the two dimension. In fact with respect to [19, 129] we only detect willing contacts without using particular ad-hoc devices [12] Besides providing a reliable sample of a group, the collected data are rich and provide many additional information for the comprehension of the online/offline relation and the role of contexts on the way people interact [46].

Multidimensional networks. In the last decade a great effort in modeling many problems and natural systems by complex network theory has been made. Most of these works assumes only a single layer of connectivity among the constitutive elements of the system. Nevertheless the real world is more complex and elements interact on different layers, *i.e* there might be multiple connections between any pair of nodes. In the OSNs world that means a user could be registered to many social networks and interacts with her/is friends' circle exploiting different channels. For instance s/he can share her/his photos on Instagram, organize an event in Facebook and maintain the relationships with workmates on Linkedin. In this scenario a multidimensional analysis is needed to distinguish among different kinds of interactions, or equivalently to look at interactions from different perspectives.

The study of the superposition of networks originates from social sciences although a complete framework for multidimensional network analysis is still missing. As a matter of fact there is not a unique word for identifying this kind of networks. For instance terms as multiplex network, multi-layered network, multidimensional network, interconnected network are considered almost equivalent [1]. Basically in the multidimensional network literature two definitions have been proposed. Berlingerio *et al.* [17, 16] proposed a definition of a model for multidimensional networks, with a repertoire of measures able to characterize the local relationships among different dimensions. Magnani and Rossi [100] introduced a different model, called *ML-model*, to represent an interconnected network of network layers, where users belong to and interact at the same time. In addition they also extended classical graph measures to deal with multidimensionality. They applied the model to a real dataset extracted from micro-blogging sites (Twitter and Friendster). Based on the ML-model, the same authors propose a formation model which introduces cross-effect among the layers in the link creation process [102].

[1] For now on we make ours this equivalence.

A great effort has been directed to the introduction of new measures and metrics that encompass the different dimensions. In [63], Hao *et al.* introduced a measure of influence of a single layer on the others in studying the interaction between multiplex community networks. Brodka *et al.* [26] focused on the neighborhood properties introducing cross-layer clustering coefficient, cross-layer degree centrality and different kind of degree centralities, each of one extending the monodimensional counterpart. The same authors [27] investigated the shortest path properties in multidimensional network developing two diverse algorithms for the search of shortest paths in multilayered social networks. About the graph-distance, Magnani and Rossi [103] defined a new concept of distance that takes into account the different dimensions where a link between a pair of nodes exists, without converting them into a homogeneous unit. Based on Pareto efficiency they define the Pareto distance, of which geodesic distance is a particular case. Though the Pareto distance is a set of paths, it makes possible the extension of the betweenness centrality in multidimensional network.

Many of the classical applications and problems have been extended to the multidimensional case. For example, in [110] Mucha *et al.* developed a community detection method on multiplex and multi-scale networks. Szell *et al.* [149] studied correlations and overlaps among different kinds of networks by analyzing the social networks of a massive multiplayer online game. Rossetti *et al.* [139] deal with link prediction in multidimensional networks extending the classical Common Neighbors and Adamic-Adar predictors and deriving new temporal and multidimensional ones. Gomez *et al.* [58] study the diffusion processes that take place on multidimensional networks. By the introduction of a supra-Laplacian matrix and its spectrum, they show that the emergent physical behavior of the diffusion process when considering layered networks is far from trivial ranging from sub-diffusive to super-diffusive behaviors. Finally, Magnani *et al.* [101] investigated a real-world multidimensional network that combines different kinds of online and offline relationships. They also present an analysis tool that discovers the existence of hidden patterns correlating different representation layers.

Together with Magnani *et al.* [101], this work represents the first effort in understanding the relation between real-world and online relationships by means of a multidimensional approach. The formalization we adopt represents an hybrid between Berlingerio's and Magnani's definitions as we generally define our scenario as an undirected multigraph from which we extract the layers we analyze. Despite the previous works, we propose a method to transfer the information of a layer into another one and measure the effects of this transferring. Although we do not have other comparison results, we observe phenomena on the centrality measures similar to those observed by Magnani and Rossi [100] on micro-blogging services. Besides centralities we also focus on the community structure of both layers, showing that not al-

ways communities in online social network are represented offline and vice versa.

4.3 Online and offline dataset

The main issue in comparing the offline and online activities of people is the availability of data combining these two aspects. In fact nowadays two distinct classes of dataset exist: one describing a usually huge set of online activities, such as Facebook, Twitter, LiveJournal, Slashdot, Instagram, the other detecting co-location or contacts between devices that not really describe the actual social relationship. The real problem is connecting these two worlds offering a complete vision of the whole social sphere of the people. To overcome these limitations we develop a client-server architecture that records and detects the online and offline social dimensions of the experimenters. The approach we adopted overcomes intrinsic limits of the methodology that only captures the encounters among Facebook friends. In fact, we also record the encounters between strangers and between familiar strangers, *i.e.* people whom we do not interact with but who we frequently see. This way we are able to collect a rich dataset containing the Facebook and the real-life relationships of a group of people. From now on, we consider contacts, encounters and meetings as synonymous with one another, each an indication of some social relationship.

4.3.1 Client-server application

The data acquisition about the encounters, the relevant contextual information and the Facebook friendship graph was performed by means of a simple Client-Server application whose architecture is described in Fig. 4.1. The design and the development of the required components (client and student-server) have been assigned to a class of undergraduate students of the Computer Science course at the University of Milano. During the experiment, each involved student used the desktop Client to record and manage his/her daily encounters in the personal storage (student-server) together with the personal Facebook friendlist, retrieved by a dedicated Facebook application accessing Facebook Graph API. Data formats are checked during the insertion operation. At the end of the project, all personal records in each student-server were automatically collected in the main Server where they were merged in order to build the social graph of the experimenters.

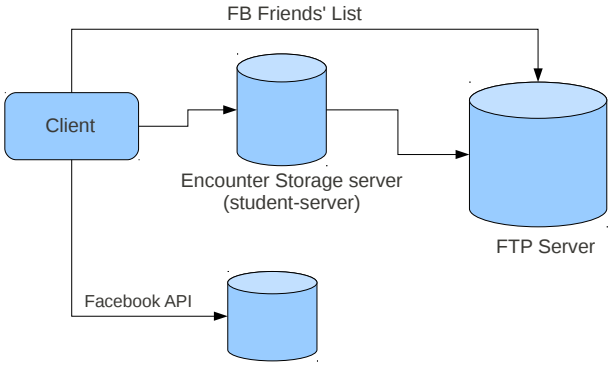


Fig. 4.1 Application architecture: A user records his contacts on the student-server by a simple ad-hoc protocol. At the end of the experiment a Facebook application retrieves the user's friendlist and all the information are sent to the FTP server.

4.3.2 Dataset description

Each encounter record reported by a student provides the following information:

- *Name and Surname*. The identity of the person met. The surname is optional because it might not be known.
- *Facebook name*. The value of the field *name* associated to the object *User* in the Facebook API. This information is optional.
- *Date*. In MM-DD-YY format. Letting the users record temporal information allows us to achieve a variable reporting interval and thereby avoid the problem of daily and persistent reporting [67]. For instance, if a user forgets an encounter, s/he can record it the day after.
- *Duration*. This field represents an estimate in minutes of the encounter duration. If a user meets a person many times, s/he has to report all separate estimates. This way a temporal granularity greater than a global daily aggregation can be achieved. Because of the recording specification, we can assign to each pair of users at a generic day t two quantities; i) d_t, the sum of the duration of the contacts and ii) f_t, the number of encounters.
- *Type of relationship*. Users select the type of the relationship by choosing from a predefined taxonomy including: friend, acquaintance, stranger (*i.e.*, never met before), relative and other. To make data comparable with other works [44, 2], categories recall the taxonomy proposed by Nicolai *et al.* [117]. During the kick-off meeting, we asked to repeat this information

for each contact to capture, in combination with the feature 'Location context', the encounter semantic in a specific context.

- *Strength of the relationship.* The strength of a relationship has been classified by choosing among the scores: *high*, *medium* and *low*. For example, if a user classifies a contact as Type 'friend' and Level 'high', it means that s/he met one of his/her best friends.
- *Location of the meeting.* This variable can assume the values: *home, work, university, sport, free time* and *other*. The values are tuned on the target (university students) and should cover the main locations of a meeting. With the term 'free time', we refer to a large set of locations including pub, restaurant, voluntary associations, etc...

To form the experiment team, we gathered more than 70 students from different courses and different years. After the project presentation 35 out of 73 students volunteered. They were required to develop their own client [2] according to specifications, as well as participate in the experiment. In this type of experiment, initial motivation is essential for obtaining rich and consistent datasets. During the experiment lifetime, the 35 students met 1115 other peo-

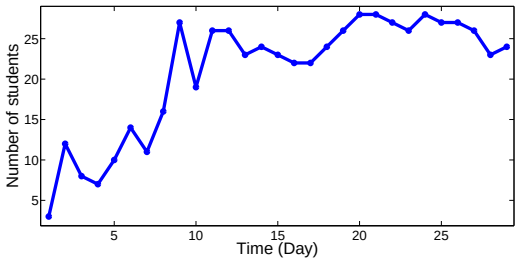


Fig. 4.2 The number of distinct participants, for each day, who reported at least a meeting.

ple, while the corresponding total amount of Facebook friends is 10291. Even from these few data we can observe that a large majority of the Facebook friends never met during the experiment. The experiment lasted four weeks, from December 13, 2011 to January 10, 2012. At the end of the first week almost 25 students completed the development of their application so their reporting phase initiated before Christmas. A few students, with the highest level of motivation, regularly maintained a paper record of encounters during the application development and reported it on their servers as soon as the client was ready. In Fig.4.2, we report the number of students that recorded their daily contacts. We can observe that a stable condition is reached after 9-10 days only and that some 25 out of 35 participants constantly reported their data during the remaining 20 days. That accounts for a student motivation that remains quite constant during the experiment lifetime without

[2] The development of the client was part of a exam project.

downfall in the production of contact events. At the end of the experiment, we recorded 3713 encounters: 257 at university, 1907 at home, 133 at work, 1231 in free time and 185 in sport activities.

To provide a more detailed daily view, in Fig.4.3 we show the daily number of encounters per different contexts. As expected, during the holidays home and free time meetings prevail with the extreme of traditional Christmas dinner (family context) and New Year's Eve (free time context). By contrast, the accounts of the remaining contexts show an opposite trend. These differences are very interesting because they allow to study how sociality and encounters change with changing external conditions.

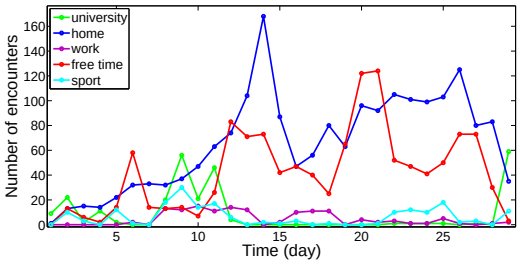


Fig. 4.3 Context trends: number of encounters within a given location context, for each day.

4.3.3 Technical Issues and Limitations

A few technical issues about managing and cleaning up the experiment dataset deserve examination in greater detail.

The need to compare offline and online social networks advocates a policy to map the set of encounters of each person onto her/his Facebook ID. The Facebook policy can help us by stating: *"We require everyone to provide their real names, so you always know who you're connecting with."* [3],although some users, even in our dataset, ignore this advice. As a consequence, we mainly exploited the Facebook Graph API to get the user Facebook ID. This kind of request is based on public information (such as, the ID and the full name) and does not require any user's authorization.

Nonetheless, we had to deal with many different conditions as to the available data. Of course, when the encounter record contains the Facebook name, the mapping is simply obtained by querying the Facebook Graph. When the fields "'name"' and "'surname"' are used, the query might return namesakes. In this case we operate as follows: if one of the friend lists of the people met is

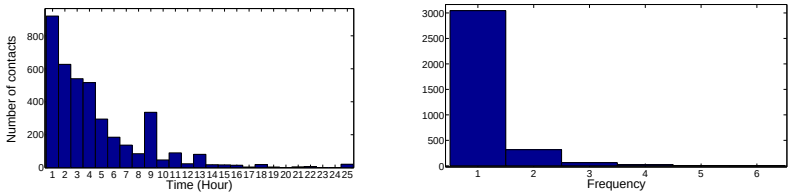
[3] Facebookpage:https://www.facebook.com/help/?faq=112146705538576

public, we search the encounter name person and extract her/his ID; if both lists are private, we try to find the most likely profile leveraging the public information. When only the person's name is available, we do not perform any mapping (5% of the nodes).

Errors might arise because students happen to be unable to pay attention to details about daily encounters. To enable some statistical adaptation, we estimate the magnitude of these errors by evaluating the one-sidedness of the recorded offline friendship, *i.e.* when all the records of a relationship are registered by only one person. We calculate that bilateral relationships happen on 90% of links, accounting for the reliability of the experimenters.

Finally it is interesting to observe that the approach we adopt overcomes intrinsic limits of the methodology that only captures the encounters among a subset of Facebook friends [67]. In fact, we also record the encounters between strangers and between familiar strangers, who usually are not in the Facebook friends' set.

4.4 Encounter general features



(a) Contact duration

(b) Contact frequency (number of contacts per-day)

Fig. 4.4 Temporal properties of the encounters.

In this section we analyze the encounter dataset, correlating the different features involved. We begin to analyze separately the recorded quantities presented in Section 4.3.2, then we make a correlation analysis between the most interesting variables.

Contact duration and frequency. In Fig.4.4, we analyze the two classical properties in DTN literature: frequency and duration of the meetings. In Fig. 4.4(a) we compute the number of contacts whose duration belongs to a given time interval; in particular we set one hour bins. We can see that the most likely duration is within a hour and that the 80% of the contacts last at most 4 hours. In general we can observe that the contact durations in this experiment is generally longer than in other data collection campaigns

in literature. This result is almost surely biased by the data collection procedure, as people are more unlikely to remember short contacts during the day. Fig.4.4(b) reports the histogram of f_t computed for each day t and for each user pair involved in at least one meeting. We can note that most of the meetings happen only once a day.

We can deepen the relation between frequency and duration by analyzing their correlation. In Fig.4.5(a) using an heat map, we report the connection between the frequency f_t and the aggregated duration d_t, computed by aggregating for each day t and for each users' pair involved in at least a contact. For instance, if we consider the most likely frequency (one contact per day) we observe that the values assumed in the interval $1-4$ hours are quite similar, suggesting this type of contacts are homogeneous, not characterized by a particular duration.

Type and level of relationships. As concerns the type of contacts, most of them involve friends (36%) and relatives (42%) while acquaintances (10%), strangers (2%), workmates (4%) and others (4%) cover the remaining part. The results confirm the expectation, in fact during holidays people tend to socialize with people different from colleagues. In particular within the Christmas holidays a person usually (at least in Italy) spends more time with her/his family and his/her friends. Similar expected results concern the strength of relationship. In this case we find that 66% of the meetings occurs between people with a high level of relationship, while the remaining 34% uniformly divides between medium and low connections. Also for these quantities the sampling methodology based on hand reports could bias the results toward the detection of strong relationships as it results easier to record and remember meetings between people that we think closer to us. From our point of view, this represents a strength as we are more interested in the set of closer friends of each person and fits with our goal of quantitatively checking if an online social network as Facebook reflects our offline relationships.

Duration in location. In Fig.4.5(b) we show how long people meet in a given location. In general we observe that the location context highly influences the duration of the contacts happening inside it. In particular it is evident that in the workplace relationships last longer than those in other locations. This result diverges from other contact detection experiments [50], where the duration has been reported to be smaller. With respect to other environments, we observe that the duration distribution is quite similar at university and in free time activities where the 70% of the meetings last more than 1 hour. As for sports, the results meet our expectations because this type of activity usually lasts at least an hour. Counter-intuitively most of the encounters in the family environment last less than an hour.

Frequency in location. As concerns frequency, the results, shown in Fig.4.5(c) are less meaningful as most meetings happen only once a day. In general, also due to this properties, the results satisfy what we expect, *i.e.* at work and in sport the frequency 1 is predominant. In fact it is difficult to do a sport activity with the same people more than once a day while at

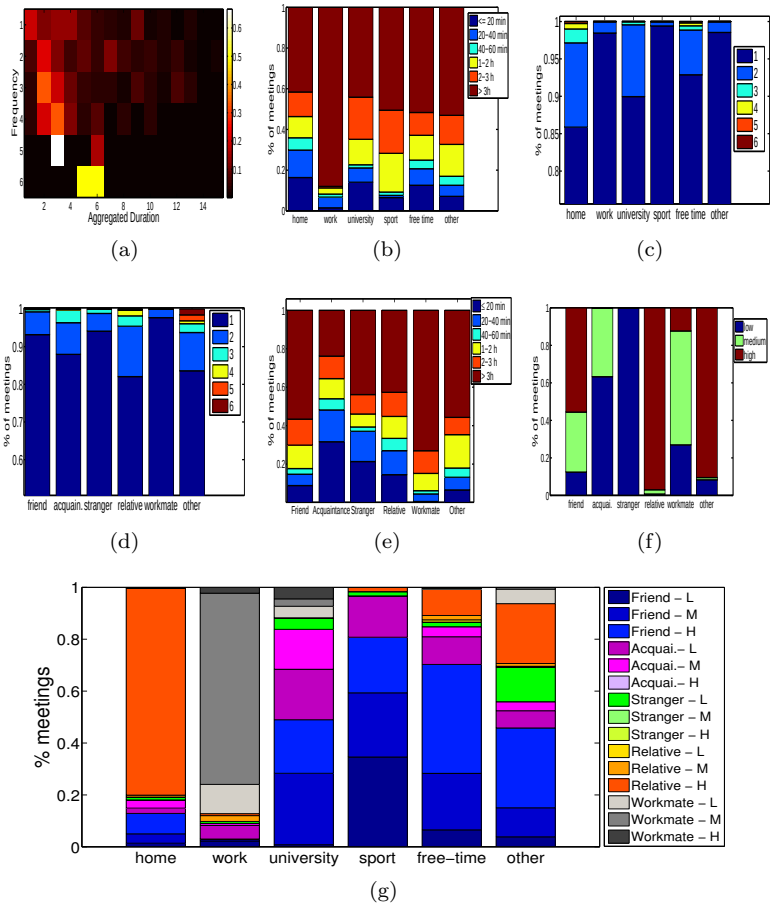


Fig. 4.5 Correlation analysis on the encounter features.

work the frequency is correlated to the long contact duration. One could hypothesizes that experimenters consider as valid contact the co-presence or the co-location of the colleagues. From Fig.4.5(c) we can infer where the most frequent encounters happens. Most frequent contacts are likely to happen at home and in free time locations. This fact has an easily explanation if we consider that during holidays the alternation among these kind of context location is more frequent. For example: lunch out with friends, dinner at home with family, night/evening at a disco or pub. Such kind of agenda could also explain the duration distribution shown in Fig.4.5(b); in fact one could reasonably assume that a meeting happens only at lunch or dinner time.

Type and strenght of the relationships in location context. Locations and the associated context play a fundamental role in determining

what kind of relationship exists between the people involved in the meeting. In Fig.4.5(g) we plot for each location context the percentage of the relationships divided respectively in the three levels of strength. This way we summarize not only the types of relation but also their strength. Graphically we indicate the relationships with the same color tone and the strength in a category with different intensities. For example the 'friendship' relationship is blue and 'low' friends are darker that 'strong' friends. This allows us to discover some interesting patterns. As regards the "home" context, of course, the most frequent encounters with the strongest strength happen between members of the family, although it is not so rare to meet friends, especially with strong ties. Analogous results have been obtained for the "workplace" context. Most of the encounters happen between work mates, whose majority has been classified with a medium strength. The highest percentage of encounters between friends occurs in sport contexts. Inside the friendship relationship, the strength is equally divided among the categories. These observations have practical consequences, in particular on social-based forwarding protocols or information diffusion algorithms. In fact we can exploit this context to capture a large piece of strong ties. A quite similar distribution of the relationships and strengths characterizes the free-time context where the percentage of acquaintances raises and we could meet relatives (brothers or sisters, perhaps). Surprisingly it is less likely to meet strong friends in free-time locations than in sport. The most interesting results concern the university environment, where we can observe the higher percentage of acquaintance and different levels of friendship from medium friends to medium acquaintances. In fact in the university setting different factors come into play, ranging from friendship to building "strategic" acquaintances, so a wide range of relationships is plausible. In general we find that the type of the relationship among the met people strongly depends on the context of the location.

In Fig.4.5(g) strengths and relationships are spread throughout the different contexts; to obtain a more general picture of their correlation in Fig.4.5(f) we aggregate over the location contexts. In the "friend" category over half are contacts between strong friends, although 15% occur between weak friends. As to "acquaintance" we observe an opposite result where obviously it proves difficult to find a strong relation of acquaintance, more likely recorded as weak friends for the students. The other relations have been already commented as the strong correlation between home/relative and work/colleague.

Duration and type of relationship. In Fig.4.5(e) we show how duration and type of relationship relate to each other, in particular for each type of relationship we report the distribution of the encounter durations. Some results are quite obvious. For example most of the meetings among friends last more than an hour, or among colleagues the duration of contacts highly correlates with the contact duration at work. The same considerations could be extended to the home location and the relative relationship. More interesting results concern the "acquaintance" and the "stranger" relationships. In the firs case, more than 50% of the contacts between acquaintances last

less than an hour, implying that usually we do not spend much time in maintaining acquaintance relationship. The "stranger" case is quite surprising as one expects that a person does not stay with a stranger for long, instead duration distribution is similar to the "relative" relationship.

Frequency and type of the relationship. Finally in Fig.4.5(d) we show the relation between frequency and relationship categories. The results are very similar to those observed in Fig.4.5(c). The most evident difference involves the "other" category that collects the most frequent meetings.

4.5 Network definition

In this section we provide some definitions to formally describe the offline and online environment surrounding a group of users. We start from a general definition which encompass all the dimensions and then we propose a formalism to extract the network relative to a specified dimension. We specialize the definitions for the Facebook and the contact cases. Finally we provide a projection of the multidimensional network into each single layer in order to mimic the transferring of information among the levels.

Although the number of dimensions considered in our dataset is limited, here we introduce the definition of edge-labeled multigraph which can cover many multidimensional situations. For the sake of clarity, we only consider undirected network without any label on vertices.

Definition 4.1. A **edge-labeled undirected multigraph** is a tuple $\mathcal{G} = (V, E, D, l)$ where V is the set of vertices, $E \subseteq V \times V \times D$ with D the set of dimensions or layers and $l : E \to S$ is a general mapping which assigns a element $s \in S$ to an edge (u, v, d); where S is a generic set.

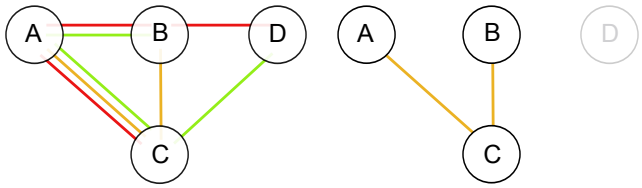
Given a edge-labeled undirected multigraph, we may need to extract only a particular dimension or to consider separately each dimension. For instance we would compare the properties of different dimensions or evaluate the importance of a vertex in a specific dimension. This way we provide the definition of d-network layer or d-network dimension.

Definition 4.2. Given a edge-labelled undirected multigraph $\mathcal{G} = (V, E, D, l)$ and $d \in D$ we define the **d-network layer** G_d as the graph $G_d = (V_d, E_d)$ where $E_d = \{(u, v) \in V \times V | (u, v, d) \in E\}$ and $V_d = \{u, v \in V | (u, v) \in E_d\}$

The previous definitions represent a tentative of unification of the approach presented in Berlingerio *et al.* [16] and Magnani and Rossi *et al.* [100]. From the first we take the definition of multigraph and we add the function l which assigns values to the dimensions while from the second one we resume the concept of network layers but we change the perspective. In [100] they apply a composition of the layers introducing a matrix to map the nodes among

the levels, what they called *pillars*, while here we apply a top-down approach which does not need any mapping matrix on the vertex sets.

An example of the modelled situation can be seen in Fig.4.6. In Fig.4.6(a) we report an undirected multigraph where the labels on edge have been omitted. In the network we have 3 users linked on three dimensions represented by the different line styles. In Fig.4.6(b) we extract the network layer corresponding to the orange dimension. As we can note the vertex D disappears since it is linked to the other nodes on the remaining layers.



(a) Undirected multigraph (b) Orange network layer

Fig. 4.6 4.6(a) The undirected multigraph with three dimensions (red,orange and green) and four nodes. 4.6(b) The orange network layer, the node D in gray does not belong to the network.

The multigraph and the network layers definitions can model situations with many dimensions, although for our purpose we need to consider few layers: the online dimension represented by the Facebook connectivity and the contact layer. So we can simplify the multigraph model by imposing $D = \{f, c\}$ where f and c respectively indicate the Facebook and the contact dimensions. Furthermore we define the function l only for the dimension c, such that it assigns to an edge (u, v, c) the number of contacts between the nodes u and v. Finally, in order to distinguish the experimenters from the other users we introduce the set V_s which represents the students involved in the experiment.

Now, starting from the multigraph of the experiment, we can define the different objects we analyse and compare.

- We define the f-network layer $G_f = (V_f, E_f)$ as the **Facebook graph**, where E_f represents the link set retrieved from the student friend lists, *i.e.* $(u, v) \in V_s \times V_f$ belongs to E_f if and only if u and v are Facebook friends .
- We define the **contact graph** $G_c = (V_c, E_c)$, where E_c represents the link set retrieved from the contact record of the students. Specifically, $(u, v) \in V_s \times V_c$ belongs to E_c if and only if u and v experience at least one encounter during the experiment.
- We extend the simple contact graph G_c to the **weighted contact graph** $W_c = (V_c, E_c, w_c)$ by adding a weight function $w_c : E_c \to \mathbb{R}$. In the

rest of the chapter we consider the function $w_c((u,v))$ that assigns to each edge (u,v) the number of contacts between u and v, although we could substitute the number of contacts with their aggregated or average durations. W_c represent an extension of the c-network layer, where we maintain the information of the mapping l.

- Let finally be $W_{fc} = (V_c \cup V_f, E_{fc} = (E_c \cup E_f), \phi, w_{fc})$ the **merged graph**. The link labelling function $\phi : E_{fc} \to \{0,1,2\}$ is defined as:

$$\phi((u,v)) = \begin{cases} 0 & (u,v) \in E_f - E_c \\ 1 & (u,v) \in E_c - E_f \\ 2 & (u,v) \in E_f \cap E_c \end{cases} \tag{4.1}$$

that is, ϕ indicates if two nodes have a relationship only on Facebook, only in real life, or both. The link labelling allows to efficiently calculate the overlapping of the network layers and of the node neighbourhood. While w_{fc} is defined as

$$w_{fc}((u,v)) = \begin{cases} 1 & \phi((u,v)) = 0 \\ w_c((u,v)) & otherwise \end{cases} \tag{4.2}$$

For sake of simplicity we define the merge graph in terms of the network layers, although the definition can be rephrased starting from the multigraph, too.

The w_{fc} definition depends on the dataset we analyze, in particular, as we do not have information on the Facebook link weights, we assign the contact weights only when possible. Some issues have been faced in other works on community detection in multidimensional network [110, 14] or link prediction in multigraph [165], where authors reduce multilayered network into a monodimensional one. Preferring contacts has the side effect of making less flat the Facebook including correlated external information.

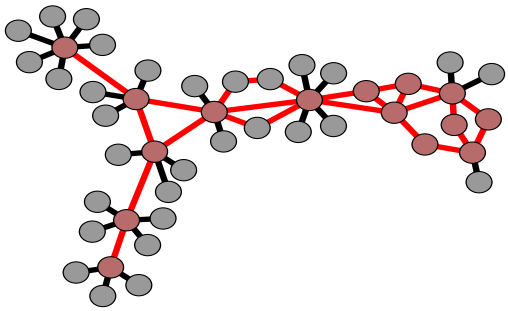


Fig. 4.7 Example of an inner graph. The red subgraph represents the graph of the nodes with degree greater than 1. The "'leaves"' of the graph are gray nodes.

Besides, we introduce the notion of *inner graph* $I_{V_1}(G)$ indicating a the 2-core graph of G. The 2-core graph is the subgraph induced by the set of nodes $V_1 \subseteq V$ with degree greater than 1, deleting peripheral nodes, *i.e.* the leaves of the graph. The inner graph will be adopted in the following section when we need to *i)* limit the snowball sampling effect, *i.e.* star structures around the nodes and *ii)* to reduce the size of the graph if necessary. In Fig.4.7 we show both these effects as red nodes represent a small subset of the overall nodes and we remove peripheral vertices.

4.6 Network description and overlapping

In this section we address the following question which advocates a comparison between the network layer G_f and W_c.

Question 1: How do offline and online social networks relate to one another and to what extent do they overlap?

As a first step towards the answer, first we compare the basic properties of the encounter and then online social networks and the we quantify to what extent students' neighborhood on Facebook graph G_f and on the contact graph W_c overlaps (see Fig.4.8(a) and Fig.4.8(c)). The graph G_f, shown in Fig.4.8(a), consists of 10,326 nodes and 10,864 edges, while the weighted contact graph W_c is made up of 1,150 nodes and 1,201 edges. The graph is shown in Fig.4.8(c), where the thickness of an edge is proportional to its weight. The high number of nodes in both networks derives from the multiple star structures associated to each node. They originate from the sampling mechanism of the experiment, which adopts a snowball sampling of the classroom network. The stars are composed by nodes in the ego-network of the experimenters who are not known by te other participants. As for some metrics, only the network of student nodes and their overlaps composed by external people known by at least two experimenters are interesting, we have introduce the inner graph $I_{V_1}(G)$. We visualize the inner graphs of G_f and W_c in Fig.4.8(b) and Fig.4.8(d). All students are present in the inner graphs since they all have a degree greater than one. Obviously the number of nodes is considerably lower, 446 and 65 respectively, while the number of link 1,153 and 116. It is interesting to see in Fig.4.8(b) and Fig.4.8(d) the number of persons who share a relationship with more than one participant in the experiment. While on Facebook they are 411, in real life they are only 30 of these people.

We first measure, for each node u, the quantity

$$\Gamma_2(u) = \{v \in V_c \cup V_f | (u,v) \in E_{fc}, \phi((u,v)) = 2\} \tag{4.3}$$

which enumerates how many Facebook friends a person u has met during the course of the experiment. In particular we analyze the distribution of

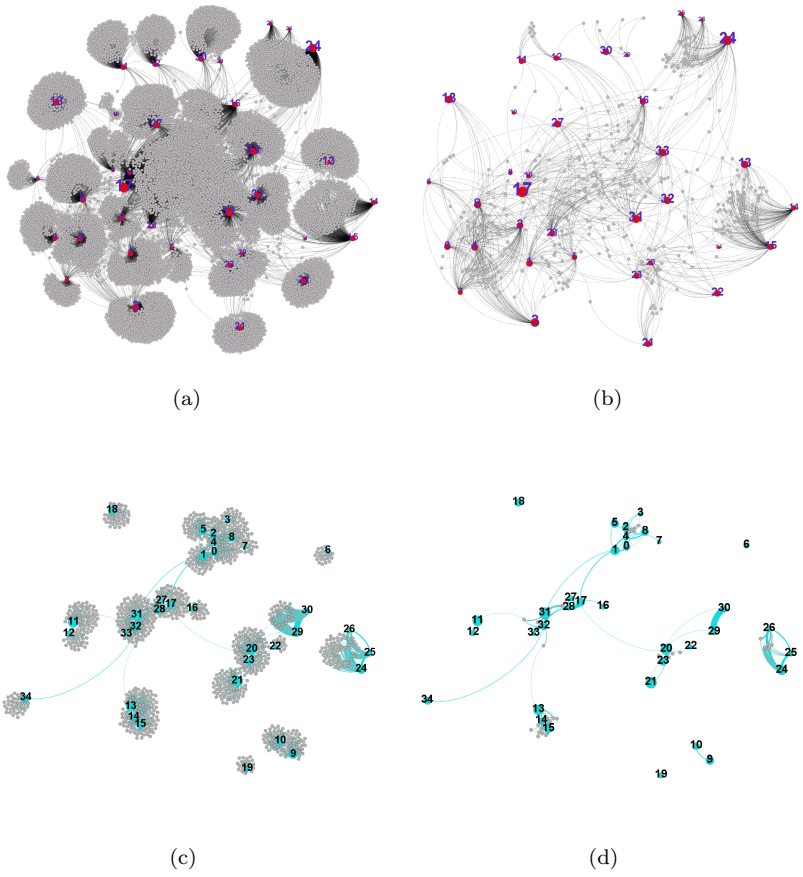


Fig. 4.8 4.8(a) Facebook graph G_f: red nodes represent the experimenters. 4.8(b) inner graph $I_{V_1}(G_f)$; red nodes represent the experimenters. 4.8(c) weight contact graph W_c: link size is proportional to its weight, cyan nodes represent the experimenters. 4.8(d) inner graph $I_{V_1}(W_c)$, cyan nodes represent the experimenters.

$\Gamma_2(u)/|\Gamma_f(u)|$, where $\Gamma_f(u) = \{v \in V_f | (u, v) \in E_f\}$. Results indicate that on average only 4% of the Facebook friends were met and, apart from few nodes, percentages oscillate between 0% and 10%. So people actually met represent a small subset of the Facebook friends.

So far we first considered the neighborhood on Facebook and we find those really met. Now we take into account the opposite direction. We examine the people involved in the encounters, looking at factors such as *i)* how many have no Facebook account, *ii)* how many are on Facebook but are not friends with one another and *iii)* how many are on the social network. The first quantity is defined as

$$\Gamma_{\overline{f}}(u) = \{v|(u,v) \in E_{fc}, v \in (V_c - V_f)\} \tag{4.4}$$

given a vertex u. We find that the average number of people met who were not on Facebook is 18. In particular we observe that for a third of the students 50% of the meetings involve people not on Facebook. As concerns point *ii*, we compute $|\{(u,v)|u,v \in V_f, \phi((u,v)) = 1\}|$ discovering that 75 people met with someone having a Facebook account but who is not as yet one of their Facebook friends. For point *iii*, for each u, we measure $\Gamma_2(u)/\Gamma_c(u)$, finding that on average 45% of contacts involved Facebook friends. In general the Facebook connectivity is far from capturing all the relationships a person is involved in. Thus G_f and G_c are complementary in describing a person's sociality. Moreover, results obtained for *i* and *ii* have also practical implications, for instance in opportunistic routing algorithms which leverage the knowledge of OSN connectivity of the users carrying devices.

An important measure used in many friend recommendation algorithms is the number of common neighbors (overlap), or its variants and extensions, between two nodes [1]. Usually common neighbors-like measures represent a similarity measure for nodes. In fact the higher the overlap the more the nodes are likely to share the same interests and the same features. Results follow the trend of the neighborhood properties, in particular in the Facebook graph we find 411 common neighbors, but only 15% of them (54) were met during the experiment. This observation makes us wonder about the universality of the common neighbors as a general similarity measure and about its effectiveness when it is employed in real life recommendation systems [98].

4.7 Structural analysis and network layers correlation

In this section we deal with the topological properties of the graphs under investigation, in particular first we analyze the classical macroscopical properties, as connected components, diameter and density and then we explore the small-world characteristics and introduce the concept of contextual path. Contextual path is an extension of the classical shortest path distance enriched by the context information on edges. Its analysis shows that contexts play a role in the optimization of the information flow.

In the second part of the section we face the correlation of the two layers, in particular our research question is

Question 2: Is a person's popularity uniform, *i.e.* more or less the same in all social dimensions, and how do different centrality metrics account for people's popularity in this multidimensional network?

In answering we analyze different centrality measures from degree to betweenness and we find that node centrality is not a universal feature. In fact, node centrality is not linearly transferred across layers and, as a consequence,

the people's popularity is most likely to change among different network layers.

4.7.1 Connected Components

By Fig. 4.8(a) we explore the structural properties of the Facebook graph G_f of the classroom. The whole graph is connected, *i.e.* each node pair is connected through a path. Despite it seems a quite trivial result, also in the light of other large scale studies of Facebook [153], the experimental environment and the sampling mechanism could result different effects. For instance, as the class is composed by students from different years, links among these diverse subgroups could be unlikely. Moreover the network has a very low density equal to 0.012, so a giant component would be unlikely. As we indicate in the following section, this property derives from the presence of few nodes that act as a bridge between different groups in the class. Analogously, in Fig. 4.8(c) we explore the structural properties of the contact graph G_c of the classroom. We can promptly note that G_c is not connected; there are, in fact, 6 components. This produces a less connected scenario in comparison to the Facebook one; so in the Facebook layer, links not related to the physical meetings make the network more connected. The giant component of G_c is composed of 914 nodes and characterized by a low density (0.014). We highlight that the remaining components contain at most eight students, forming groups marginal to the class.

4.7.2 Small world properties

The study of several real networks has pointed out the existence of bridges linking different far regions [116]. The small-world property is also associated with an high density of the neighborhood of the vertices, *i.e.* the clustering coefficient C_u of the node u. To see if our network presents a small-world phenomenon, we analyze the average clustering coefficient C and the average path length L. L is the number of hops in the shortest path averaged over the pairs of nodes, while C is the average of C_u. The clustering coefficient C_u is defined as the fraction of edges that exist between all the possible links connecting the neighbors of u. According to the seminal work of Watts and Strogatz [157] we are observing a small-world graph if L is similar to L_{rand} (characteristic shortest path of a random graph with n nodes and average degree k equal to the real one) and $C >> C_{rand}$. We perform the following computation only on the inner graphs $I_{V_1}(G_f)$ and $I_{V_1}(W_c)$, because the star structures in the corresponding graphs would artificially decrease the average clustering coefficient.

Facebook. Comparing the above quantities, we can see that our network is a small-world one as $L = 3.55$ and $L_{rand} \approx \frac{\ln(n)}{\ln(k)} = 3.65$ while $C = 0.73$, which is much greater than $C_{rand} \approx \frac{k}{n} = 0.012$. In general the Facebook network contains nodes that are high clustered and a few shortcuts that reduce the distance between the nodes. In Fig. 4.11(b) the role of shortcut is played by the central area around the node 17, which links the different high clustered groups. In fact, as many paths pass through the center of the graph, the most likely distance is 3.

Encounter. As for small-world properties, the induced subgraph is quite cryptic to classify. In fact the average clustering coefficient $C = 0.764$ is greater than the expected one in a random version, *i.e.* $C_{rand} = 0.053$, while the average path length $L = 4.03$ is greater than $L_{rand} = 3.33$ and the diameter is equal to 7. So, as shown in Fig. 4.12(b), the structure presents highly clustered regions (explaining the high C) connected through few links (explaining the path features). In particular, we can observe a sort of backbone, comprised naturally by the nodes with the highest betweenness centrality.

4.7.3 Contextual path

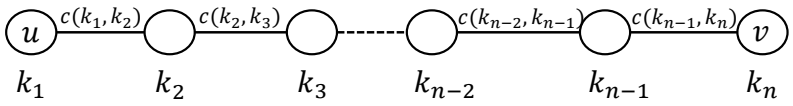


Fig. 4.9 Contextual path

The availability of information on the edges of the contact graph allows us to introduce a different way to evaluate the importance of the minimum shortest paths between two nodes. In particular we can exploit the information given by the context of the place where the meeting occurs in evaluating the influence of a path in the diffusion of messages. The idea is based on the hypothesis that a message can more easily propagate if the contexts of the links it passes through are the same or similar. For example, let us consider two nodes, A and B, who meet at the university and exchange a message about an exam. It is more likely that this message propagates through links incident with A and B which have the same context, *i.e.* the university. In general we can say that the more context switches there are in a path, the harder it is for a message to be propagated on it.

In order to quantify this property let us consider the general path (u, v) of length $n-1$ in Fig. 4.9 where each edge (k_i, k_{i+1}) has been labeled with context value $c(k_i, k_{i+1})$. Taking two successive edges (k_{i-1}, k_i) and (k_i, ki_{i+1}), we can say that the probability that a message coming from $i-1$ will be pass to $i+1$ is higher if $c(k_{i-1}, k_i) = c(k_i, k_{i+1})$, and much lower if the contexts of the two edges are different. Let α and β be the above probabilities, we define the contextual weight CW of the path $p = (u = k_1, \ldots, k_n = v)$ as

$$CW(p) = \prod_{t=2}^{n-1} (\alpha - \beta)\mathbb{I}\{c(k_{i-1}, k_i) = c(k_i, k_{i+1})\} + \beta \tag{4.5}$$

where $\mathbb{I}\{A\}$ is the indicator function of the event A. Thus the contextual weight CW of a path is α^{n-1} if the contexts do not change along the path and goes down to β^n as contexts change along the path.

We study the above property on the contact graph to verify how the positioning of the contexts in the network structure influences the contextual weight of the shortest paths. In particular we compare the contextual weight of the contact network with a null model where contexts are randomly assigned to each edge. The resulting distribution in the null model is a binomial distribution on the number of context switches. As the network diameter is 7, the parameters are $m = n - 2$ and $p = 1/6$, which corresponds to the probability that $c(k_{i-1}, k_i) = c(k_i, k_{i+1})$, as we are in k_i. We compute the different distributions measured on all the minimum shortest paths between each pair of nodes. In Fig. 4.10 we show the resulting distributions, in particular we group contextual weights based on the path length. As we can see, with the increase of path length the differences between the null model and the real values go up. This implies that, in a group, contexts are not randomly put on the shortest paths, *i.e.* the location context of the meetings favor in some way the propagation of the information, assuming that it diffuses following shortest paths.

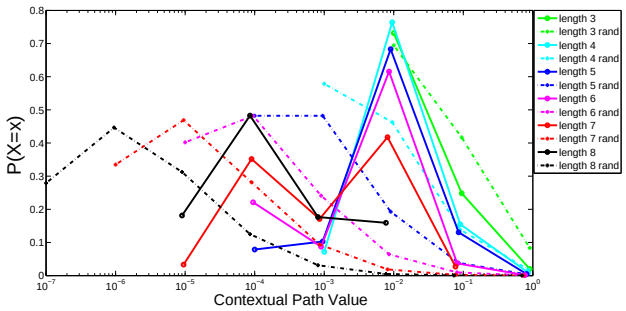


Fig. 4.10 Solid lines correspond to the measured contextual weights, while dot lines represent the distribution of the randomized ones. We report the distributions of the paths with length greater than 2.

In the following we proceed with the analysis of the structural properties of G_f and G_c but we also focus on the correlation analysis between the degree centrality, the betweenness centrality and the eigenvector centrality of the vertices in the sociality layers. This way we can answer to question 2.

4.7.4 Degree Centrality

The simplest centrality measure in a graph is the degree. We take into account two kinds of degree depending on the network we analyze. The first type, called *total degree*, is the usual definition applied to graph $G_{(.)}$, while the second, denoted as *inner degree* is computed on the corresponding $I_{v_1}(G_{(.)})$. We compute the above quantity both on G_f and on the unweighted contact graph G_c and on its weighted extension W_c. Obviously, in W_c we apply the usual definition of strength. The last metric allows us to measure the popularity of a person not only by the number of friends s/he has but also on the basis of how often s/he meets them.

Facebook. Observing the Fig. 4.11(a), relevant to the Facebook graph,

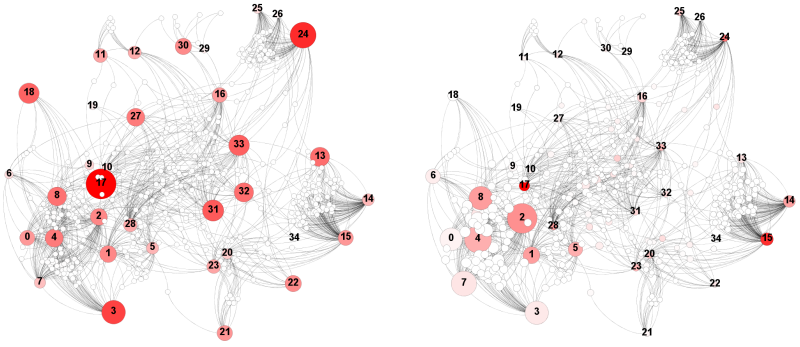


Fig. 4.11 4.11(a) Facebook graph: size and color (from white to red) of the nodes are respectively proportional to the total and the inner degrees. 4.11(b) Eigenvector and betweenness centralities: size and color nodes are proportional to their eigenvector centrality and betweenness centrality, respectively.

we obtain different behaviors involving the same nodes. In the figure the size and the color of the nodes are respectively proportional to the inner and total degree. For example, node 18 has the highest total degree (787) while its inner (44) is lower w.r.t. the other nodes. To quantify the agreement between Facebook importance and classroom importance we perform a rank correlation analysis [83]. Rank correlation analysis allows us to test if the ranking

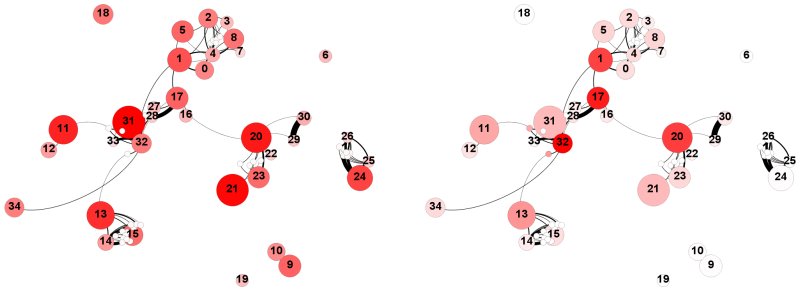


Fig. 4.12 4.12(a) Unweighted contact graph: size and color (from white to red) of the nodes are respectively proportional to the total and the inner degrees. 4.12(b) Eigenvector and betweenness centralities: size and color of the nodes are proportional to their eigenvector centrality and betweenness centrality, respectively.

induced is similar in the different degrees or not, also assessing how well the relation between two variables can be expressed by a monotonic function. This kind of correlation analysis is more general than the usual linear correlation and is able to capture more complex functions. As a rank correlation method, we compute the Spearman's rank correlation coefficient ρ on the ranking induced by total and inner degrees on V_s respectively. The more $|\rho|$ becomes closer to 1, the more the relation between the two quantities can be expressed by a monotonic function. Comparing G_f and $I_{V_1}(G_f)$ we obtain $\rho = 0.4$, which indicates that the two degree measures induce different rankings. So, some nodes, relevant for example in G_f, lose their importance in $I_{v_1}(G_f)$. An explanation of these changes is rooted in the numbers of common neighbors in the inner graph $I_{v_1}(G_f)$. In fact, nodes with a high total degree and a small inner degree have few common neighbors and share few connections with other nodes in the subgraph. Generally, the above results suggest that Facebook popularity is not uniform among groups a person belongs to and so people with a high overall importance may not be popular in a specific community (the class). We find that students have an average degree equal to 312, higher than the 190 reported by Facebook[4] and a 0.8-quantile equal to 447. As for the inner graph $I_{v_1}(G_f)$, the average student degree is 35, while the 0.8-quantile corresponds to 52. These measures suggest that all students are socially active online.

Encounter We analyze the degree distribution of G_c and W_c to highlight the number of people met and the number of contacts per person. In particular we focus only on the degree of the students, since for the other nodes we

[4] Facebook page:https://www.facebook.com/help/?faq=112146705538576

have incomplete information. The degree results are presented in Fig. 4.12(a). In the figure the node color is proportional to its degree computed on W_c, while the size is proportional to the one computed on G_c. On average the number of people met by each participant is 37 and the average number of encounters is 125. As suggested by the figure and by the Spearman coefficient $\rho = 0.6$, a clear relation between the degree and its weighted version does not exist; actually, there are many nodes having a high degree yet a color that indicates a medium-low weighted one. This explains why maintaining many close friendships proves difficult.

Correlation We also compare the different degrees of the student nodes in Facebook and in the unweighted contact networks. By analyzing the Spearman coefficient matrix, we find quite heterogeneous results. For example, the Facebook total degree quite positively correlates with the inner degree in the contact graph, while, at the same time, it has no correlation with the total degree in the contact weighted graph. Generally, we have shown that the centrality measure given by the degree does not maintain the rank, so that popularity in Facebook does not correspond to the same popularity in the encounter networks.

4.7.5 Eigenvector Centrality

An extension of the degree centrality is the eigenvector centrality [23]. The eigenvector centrality relates the node importance to the importance of his neighbors. Instead of awarding vertices just with the cardinality of their neighborhood, this quantity gives each node a score proportional to the sum of the scores of its neighbors. In particular it may be large either because a vertex has many neighbors or because it has important ones. We calculate the eigenvector centrality defined by

$$x' = \frac{1}{\lambda_1} \sum_j a_{ij} x_j \tag{4.6}$$

where A is the adjacency matrix of the graph and λ_1 is the largest eigenvalue of A. In Fig. 4.11(b) and Fig. 4.12(b) we report the eigenvector centrality computed respectively on G_f and G_c, the size of the vertices is proportional to its centrality, allowing an intuitive comparison.

Facebook In Fig. 4.11(b) we can see how eigenvector centrality acts on the bottom left of the graph; in fact, node 2 gains its centrality from its numerous neighbors and conversely spreads its value among them. In this respect, comparing Fig. 4.11(a) and Fig. 4.11(b), we can see that the degree centrality is different from the eigenvector centrality: specifically node 17 has a high degree, yet is connected to nodes with low importance.

Encounter As to this measure, we calculate the eigenvector centrality

considering $I_{v_1}(G_c)$ and $I_{v_1}(W_c)$ on each component of the relative graph. In particular, in the weighted case we apply the general centrality proposed in [128] which still corresponds to the leading eigenvector of the adjacency matrix, with matrix elements being equal to the edge weights. The meaning of this measure is quite similar to the one in a citation network. In fact, if we use the frequency encounters as link weights, eigenvector centrality would then give people high rank in either of two cases: when they are met by many others and if they meet frequently with a few others. Weights play a fundamental role in comparing the ranking induced by the two measures; in fact, analyzing only the two most numerous components, we find opposite results. In one case we observe a strong monotone increasing relation between the weighted and the unweighted centrality ($\rho = 0.8$), while in the other we observe a substantial lack of correlation between the variables ($\rho = -0.3$). These results depend on the distribution of the weights: in one case the highest weights are among central nodes, in the other case the opposite is true.

Correlation If we consider both Facebook and encounter networks, we find results which show a substantial lack of correlation among the eigenvector centralities of the student nodes computed on the different graphs. In fact for each pair of centralities involving the Facebook and the contact graphs, we obtain correlation values near to zero. Also in this case, these findings claim the observation that eigenvector centrality is not linearly transferred across layers.

4.7.6 Betweenness Centrality

A different concept of centrality is betweenness centrality. It captures the extent to which a node is on paths between other nodes. We may formally define the betweenness of the node i as

$$b_i = \sum_{s,t \in V} n_{st}^i / g_{st} \tag{4.7}$$

where n_{st}^i is the number of shortest paths from s to t passing through i and g_{st} is the total number of shortest paths from s to t. The betweenness measures the amount of information passing through each vertex, if it follows the shortest path. Therefore, nodes with high betweenness may have a high influence due to a sort of control over the information passing among nodes.

Facebook Betweenness values are depicted in Fig. 4.11(b) where the colour of the vertices is proportional to their betweenness. As expected, the betweenness values are different from the other centralities. In particular, node 17 gains the maximum betweenness in that it acts as a bridge among the different areas of the graph.

Encounter In Fig. 4.12(b) we report the values of the betweenness centrality computed on the simple and weighted inner graphs. Here the weights are equal to the inverse of the value returned by w_c up to a scaling factor which affects paths passing through strong links. Comparing the relative values, we can observe how the introduction of the weights changes some node centralities. In particular, weights enhance the probability that information passes through some paths. For example, if we consider an unweighted graph and two minimum paths between two nodes, the probability that a message follows one of them is an even split. In the weighted case, the path might be unique concentrating all the probability on it. We can observe this phenomenon on node 14, where in the unweighted case the betweenness is distributed between 15 and 14, while strength forces paths to pass through 14. As reasonably expected, weight introduction changes not only the betweenness values but also the student node ranking. In fact, Spearman coefficients measured on each pair of betweenness show uncorrelation between the different centralities. Therefore, even on the social dimension (offline sociality) a node can assume different relevance depending on the features of the network we are considering.

Correlation In comparing the two social layers by the Spearman coefficient matrix, we find results in accordance with those presented in the above paragraphs. In fact betweenness centrality values on the student set are almost uncorrelated. This further corroborates the fact that betweenness centrality does not transfer monotonically too.

In summary we quantify if nodes maintain their centralities across the two layers finding that the importance of vertices may change and depends on the level under investigation. These observations suggest that layers are loosely correlated and could be intended complementary in describing people's sociality. Furthermore these results impact on the modeling of the phenomenon, as it advocates multilayered network models that have to take into account the weak correlation.

4.7.7 Merging the complex networks

In this last section we analyze the merge graph W_{fc} shown in Fig. 4.13. Our main goal is to blend the two social layers in a unique network and check if student nodes, in this merged scenario, maintain their centralities or the merging modifies the ranking among nodes. In the following analysis we compare total and inner degrees, eigenvector centrality and betweenness centrality on the graphs G_f, G_c, W_c, W_{fc}, G_{fc} (unweighted version of the merge graph) and their induced subgraphs. In general G_f influences the most of the measured centralities on the merge graph because of its denser and

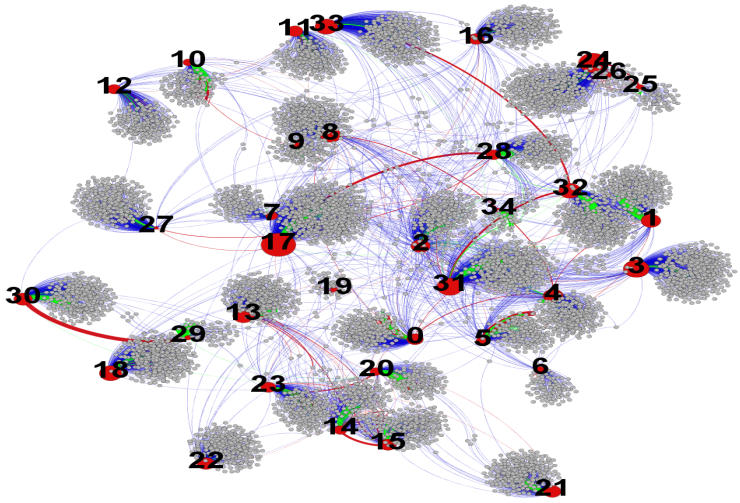


Fig. 4.13 Merge graph W_{fc}: link color indicates the value of the link labeling function ϕ: 1 (blue), 2 (green) or 3 (red). The link size is proportional to its weight assigned by the function w_{fc}.

more compact structure. That happens because the weight function affects links corresponding to encounters.

As the correlation values of the total degrees and the inner degrees are similar when compared with the other variables, in the following we only report the results of the total degrees. First we observe a strong correlation between the total degree on G_{fc} and G_f as the number of Facebook friends is much higher w.r.t. the encounter one. Otherwise the total degree on W_{fc} correlates with the total degree on W_c and G_c as the bias introduced by w_{fc}.

The eigenvector centrality measured on W_{fc} and G_{fc} has a particular meaning as it mixes the contribution of the degree and the connectivity of the two social layers. Furthermore in the weighted case it depends on the attitude a person has to connect with other important nodes through the strong links given by the contacts. By analyzing results we obtain that the eigenvector centrality on $I_{V_1}(W_{fc})$ positively correlates with that on $I_{V_1}(G_f)$. An interesting result concerns the centrality on W_c, in fact it correlates with G_f and unexpectedly negatively correlates with W_c. This shows that the Facebook connectivity differently allocates the centrality portion given by the strong links. We observe the same effect in the betweenness centrality where in G_{fc} and W_{fc}, it correlates with the betweenness measured on G_f.

As the strong influence of the Facebook graph we investigate how weights act on the different centralities of student nodes only on the Facebook relationships. We compare F_c and F_w, *i.e.* the induced subgraph of W_{fc} containing links such that their ϕ function equals 0 or 2. By comparing the Spearman

coefficient of each centrality, we find a strong correlation between the total strength and the total degree. This fact can be explained by the low degree of overlapping between contacts and Facebook friends. We came up with an opposite result as to eigenvector centrality. For this measure we observe a low negative correlation ($\rho = -0.33$). So the weight insertion drastically changes the importance a person acquires in the network. Last we consider the betweenness centrality on G_f and W_f. In this case we obtain a value ρ equal to 0.6. This finding implies that the ranking does not drastically change, although, also in this case, some nodes consistently increase or decrease in the ranking.

In general we observe that the merging of the two social networks induces a different ranking on the set of the student nodes w.r.t. the ranking in each layer. Furthermore we observe that the Facebook structure and the weights inferred from contacts play a fundamental role in making the centralities always different.

4.8 Community

Almost all networks show the tendency to group in clusters that reveal structure or social circle (e.g. friends, colleagues, family). In particular, we find communities intended as groups of nodes such that there are many links within them and few between. As the richness of our dataset, the community structure represents one of the most interesting feature, as we can compare the results obtained by community detection to a ground truth. In this way we can test *a)* if the communities we found actually correspond to real-life ones, *b)* which communities are present both in the online/offline networks and *c)* we can assign a possible context to the online communities.

Community detection is one the most studied task in complex network literature. In the last years many algorithms and quality functions have been proposed to identify communities. Among the plethora of community detection algorithms [47], we choose the Louvain algorithm [21] whose aim is to maximize the modularity [35], defined as

$$Q = \frac{1}{2m} \sum_{i,j} \left(a_{i,j} - \frac{k_i k_j}{2m} \mathbb{I}\{c_i = c_j\} \right) \tag{4.8}$$

where k_i is the i's degree and c_i is the community the node i belongs to. Despite the limitations of modularity [48], we use Louvain because, as shown in Fortunato [47], it has good performance and the size of the network mitigates the drawbacks in finding small communities, especially in the G_c case.

Facebook. The results obtained are shown in Fig. 4.14 where each community is represented by a different color. Despite the fact that our

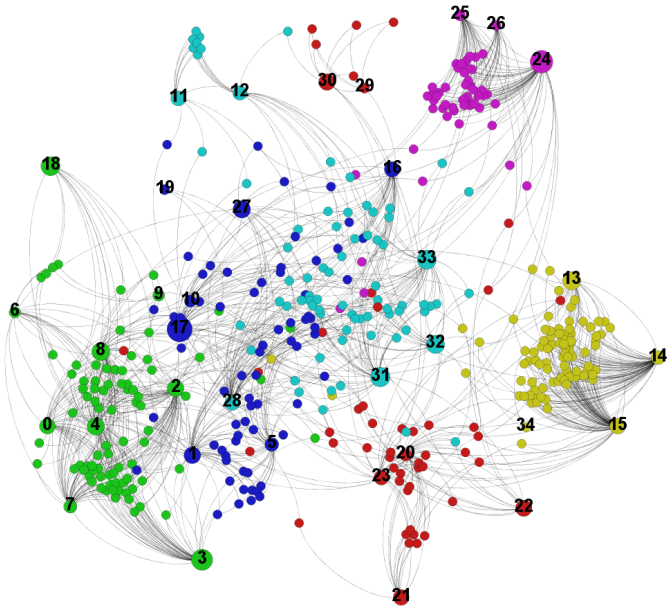


Fig. 4.14 Facebook graph: communities detected by the Louvain algorithm

sample consists of only one single classroom, the network is highly modular ($Q = 0.67$) and composed by 6 groups of different sizes. This feature may be rooted in the age heterogeneity of the class or in the missed links between nodes that are not students, not detected because of the sampling mechanism.

Encounter and context. We run the Louvain community detection algorithm for weighted graphs, which aims to maximize the modularity as defined in Equation 4.8 where a_{ij} is substituted by the weight w_{ij}. Actually the computed modularity is equal to the high value 0.778. As we can see in Fig. 4.15 the algorithm finds 5 communities in the giant connected component.

Exploiting the context information, we enrich the found communities in the contact graph; in particular we verify if a given community corresponds to a certain context or if it embraces different locations. Furthermore, we can test which context acts as a bridge among the communities. As concerns the last aspect, we observe that the 'university' context is the glue of the different communities, as all the bridge edges belong to this context. This fact implies that not all contexts are suitable to spread information among the contact communities. We analyze, for each group, the intra-edge context distribution, *i.e.* the distribution of the context of the links connecting people that belong

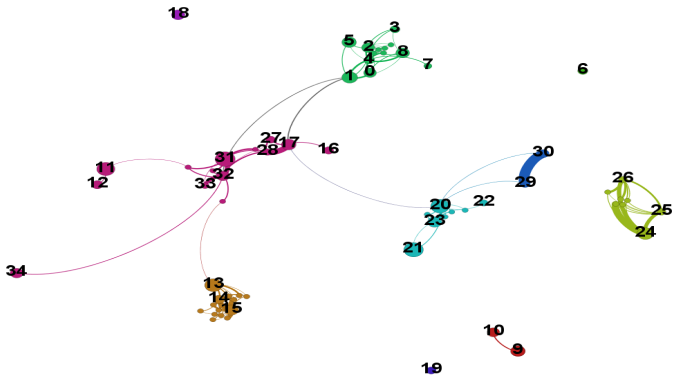


Fig. 4.15 Weighted contact graph: communities detected by the Louvain algorithm

to the same community. We find that, with the exception of a community, each community is characterized by a well-defined context. The contexts and the high modularity of the network could be the cause of the results about the contextual weight distribution, *i.e.* the greater contextual weights of the paths with respect to a null model. In fact, as contexts are quite similar inside communities, the increase of the cost is given by the possible context switch when a path passes through different communities.

Merging. After obtaining the communities in each level, we overlap them to find if some communities in Facebook (Fn) correspond to some real communities or if an online community contains more offline communities (Cn). In order to quantify the level of inclusion and overlapping of the communities we compute the cardinality of the intersection of the sets involved. In Table 4.1 we report for each pair of communities (Fn, Cn) the number of nodes in their intersection. Inspecting the columns, we see that some offline communities are split between different Facebook communities as in C1 and C7 cases. Other ones (C8 and C10) are totally contained in online communities, while the Facebook communities F4, F1 and F3 contain more offline communities. In general we find that not all the detected online communities correspond to offline ones, thus suggesting that also at a community level the two graphs are not completely overlapped. In particular the link densification in the online level compacts and joins groups that are separated in the offline level.

The encounter information can be exploited, not only to compare nodes but also to reciprocally enrich the relative topology. In particular, in the following we show how to assign weights in Facebook based on the type of

Table 4.1 For each pair of communities extracted from the Facebook graph ($F1, \ldots, F6$) and the weighted contact graph ($C1, \ldots, C10$) respectively, we report the cardinality of their intersection.

	$C1$	$C2$	$C3$	$C4$	$C5$	$C6$	$C7$	$C8$	$C9$	$C10$
$F1$	6	0	1	1	0	0	0	0	1	0
$F2$	0	0	0	0	0	0	2	13	0	0
$F3$	2	1	0	0	0	0	3	0	1	0
$F4$	0	0	0	0	2	8	0	0	0	0
$F5$	0	0	0	0	0	0	0	0	0	8
$F6$	0	0	0	0	0	0	6	0	0	0

relationship and the frequency of the contacts, how this assignment impacts on the network modularity and on the community structure. The type and the degree of the recorded relationships can be exploited to generate weights between two Facebook friends who meet in the real life ($\phi = 2$). In fact we can rank the meeting type according to an arbitrary order based on relationship importance. For example, we consider friends and relatives almost on the same level (strong ties), while acquaintances are less important than friends. Inside each category we use the degree property as our ordering factor. Besides relationship ranking, we also consider the number of meetings. So if ranks are equal we prefer pairs with numerous contacts. Namely, given a ranking function $r : relationship \times degree \to \mathbb{N}$ to an edge (u, v) belonging to the encounter graph, we extend the weight $w_{fc}((u, v))$ in the merged graph as:

$$w_{fc}((u, v)) = r(rel(u, v), degree(u, v))w_c(u, v) \tag{4.9}$$

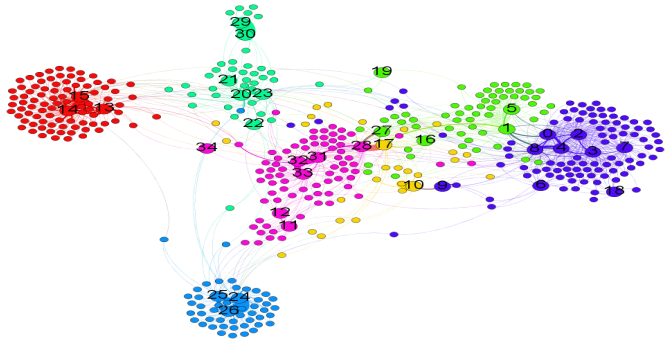
where rel, $degree$ and w_c respectively return the relationship, the degree and the number of contacts of the pair (u, v). Therefore link weight depends on the ranking function and on which relationship we consider important for purposes of our analysis. In the following, we determine friendship and relative relations to be important.

The extension of the weight function, which includes both relationship and tie, has an impact on the community structure of the network. Thus we show in Fig. 4.16(a) and Fig. 4.16(b) the communities found by the Louvian algorithm in the weighted graph (Fig. 4.16(a)) according to the extended weighting function and in its unweighted version (Fig. 4.16(b)). As we can note, the number of communities are different, since in the weighted merge graph we detect 8 clusters while in the unweighted merge one we see 7 groups (the algorithm has been executed many times and with different resolution parameter obtaining the same general results). In particular, except for the quite stable red, cyan and blue communities, we observe many modifications in the node communities shifting from Fig. 4.16(a) to Fig. 4.16(b). The bright green community breaks in two parts because of the strong tie between the nodes 29 and 30 in the brown community. The same phenomenon happens in the bottom part of the magenta where the nodes 12 and 11 separate into a diverse community. A significant change involves the yellow community,

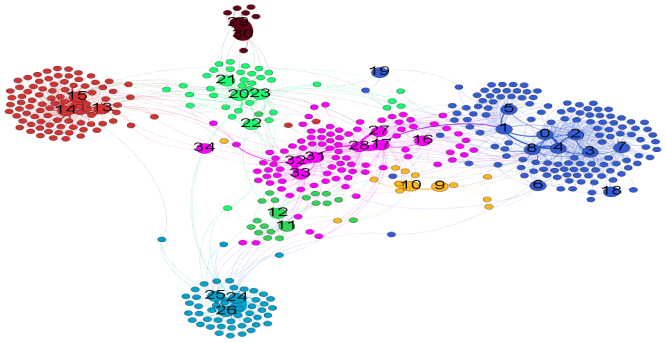
which consolidates around the strong tie between nodes 9 and 10 and loses the nodes on its frontier. The most significant modification regards the disappearance of the green community in the Fig. 4.16(a) as its members split up between blue and magenta communities. Such changes also impact the modularity values, since the weights increase the Q value (from 0.65 to 0.78). This results in a more modular graph. In fact weights may reduce the influence of weak ties in the modularity computation. In general, we observe that the reinforcement of friendship and relative links changes the community structure and highlights the two backbone communities of the classroom (blue and magenta) containing most of the nodes.

4.9 Conclusion

This chapter represents a first effort to provide a complete overview of the close connection between online and offline sociality. The completeness of the dataset, enriched with contextual information, and the proposed multidimensional complex network allowed us to deeply understand how the characteristics of users in the distinct networks impact each other. Our work shows how offline friends are completely different from Facebook ones, so confirming and worsening the general intuition that online social networks have shifted away from their original goal to mirror the offline sociality of individuals. In this general framework, it becomes apparent that social features such as user popularity or community structure do not transfer along social dimensions, as confirmed by our correlation analysis of the network layers and by the comparison among the communities. Finally, contextual information has been revealed to be a key factor in understanding people's offline and online sociality. In real-life we highlight how the location contexts impact the properties of the face-to-face contacts and the structure of the communication pathways by introducing the idea of contextual paths. On the one hand the location context influences the duration and the type of relationship between the persons who come into contact. On the other hand the shortest pathways in the contact layer tend to reduce changes of context. Finally, in the online social sphere, we show that introducing external information the users' role and their groups may change.



(a)



(b)

Fig. 4.16 4.16(a) Communities detected on the inner merged graph considering its unweighted version. 4.16(b) Communities detected on the inner merged graph with weights assigned according to Equation 4.9.

Chapter 5
On the Bursty Evolution of Online Social Networks

5.1 Introduction

The rapid growth of online social networks (OSNs) has created numerous technical challenges for the providers that supply the hardware and software infrastructure behind these web services. As one example, the creation of social links between users dramatically change demands on social network infrastructures in terms of access, storage and computation. Depending on the specific configuration of backend servers, for example, changes in the social graph can affect how data is partitioned across clusters, or how much replication is necessary to sustain low query response times.

However, very little is known about how social network dynamics correspond to actual clock time. The large majority of prior work on OSN analysis has focused on analyzing, mining, and modeling static topologies or static snapshots of dynamic processes. Only recently have researchers begun to study dynamic processes in social networks, most often by analyzing how classical graph metrics such as degree, connected components, and shortest paths change over time. This has led to models of underlying processes such as densification and shrinking diameters [93]. These models describe how graphs change and how edges are created with respect to a logical clock, *i.e.* a homogeneous sequence of events.

But how do these events match up to events in real time? Can we better understand how edge creation events relate to each other, and can the occurrence of such events be predicted with respect to a physical clock? This work is an initial effort to answer some of these questions, but analyzing one specific temporal property of *burstiness* in edge creation. Our work is motivated in part by models of human dynamics adopted in a wide range of disciplines, from economics to communications. Recent studies [154, 85] have shown that human dynamics are best described by periods of rapidly occurring events interleaved with long periods of inactivity. Thus we ask the question: *Is link creation in online social networks a bursty process?*

In this chapter, we provide an initial answer to this question, by analyzing an anonymized temporal trace of edge creation events over a period of a year in a large Chinese online social network. This online social network has more than 200 million users, and our analysis of its dynamics shows that edge creation is a highly inhomogeneous and bursty process. We then ask two follow-up questions: a) *Given a high level bursty structure, does an inner substructure exist, and how can it be characterized*; and b) *How can we detect both the whole burst and its internal phases?*

Understanding the internal structure of edge creation bursts can shed light on the underlying user process, *e.g.* is the user gradually enlarging her circle of friends or has she discovered a new cluster of her offline friends. Known techniques for the analysis and the detection of burst events (gamma-ray, text mining, stock market) focus on locating a burst when it occurs, but they do not consider events inside the detected temporal window. Thus we propose a new methodology able to detect bursts, their internal structure and the transitions between the different phases a node experiences. We perform a second order analysis on the link creation process by computing, for each node, the acceleration of the degree time function to characterize the burst structure.

Finally, we apply our acceleration metric and the detection of bursty phases on our dynamic graph. We find that all nodes exhibit similar patterns over time, characterized by an intense burst of activity following their joining the network. The initial burst is followed by weaker bursts over time, each composed of an acceleration phase, followed by a longer period of slowly vanishing deceleration.

The discovery of highly bursty patterns paves the way for new generative models that not only capture graph dynamics in terms of phases of node activity, but also describes such events with respect to physical time. In addition, burst analysis can reveal further insights into the formation and liveness of individual users, communities, and provide a basic and useful metric to characterize and compare different dynamic behaviors.

5.2 Related Work

Time evolving OSN Snapshots. While static features of OSNs are well studied, works on dynamics of online social networks are still ongoing. Among all Leskovec *et al.* in [93] detected two important properties on dynamic OSN data: graph densification, *i.e.* the average degree increases, and shrinking diameter. Several different social graphs has been studied in order to capture the growth of components and communities. Palla *et al.* [121] investigated the time dependence of overlapping communities and Berlingherio *et al.* [15] detected clusters of temporal snapshots of a network, interpreted as eras of evolution. Authors in [104, 168, 4] studied the dynamics of discon-

nected components. Finally Backstrom *et al.* [9] investigated the structural features which influence people in joining communities and their growth process. Alternatively, the per node dynamics was studied in [92] where the authors captured the evolution of key network parameters, and evaluate the extent to which the edge destination selection process subscribes to preferential attachment. As concerns acceleration, in [42, 107] the authors considered an overall network size growth as a global property and they modelled this global acceleration for the purpose of predicting the next network stage.

Interdisciplinary Study of Human Dynamics. In [10], Barabasi observed that the timing of human activity is inhomogeneous and bursty, disputing the previous hypothesis that human activities are randomly distributed in time. The inhomogeneity idea was extended in [154] and validated on few networks such as a Hungarian news portal, e-mails, library activities in a University and a trade transactions. Similarly, McGlohon *et al.* [105] analyzes the activity burstiness of blogs using entropy plots, and shows non-uniformity and self-similarity of the number of posts time sequences. Furthermore, Jo *et al.* [77] observes that temporal patterns are inhomogeneous or bursty even in mobile phone calls.

More recently, several studies examined user interaction in online social networks, both visible in-network interactions [159, 34], as well as invisible activity such as profilebrowsing [75, 142, 115]. However, these studies did not consider user interactions in the context of absolute time.

Finally, [85] developed a burst detection algorithm and observed that the appearance of a topic in a stream of documents, such ase-mails or research papers, has a bursty behavior. To the best of our knowledge, burstiness as a metric has never been investigated in the context of online social network dynamics.

5.3 Timestamped OSN Dataset

A major obstacle to studying OSN dynamics is the difficulty of obtaining detailed data traces. Our study uses an anonymized dataset that contains the timestamped creation of all users and edges in a large social network in China. This network has functionality very similar to Facebook, and has more than 200 million users.

Our anonymized trace was shared by the social network provider. All information on users and edges have been anonymized to protect user privacy. The trace is a complete dynamic graph that describes the evolution of the social network graph as a sequence of timestamped edge creation events. The first edge created in the network dates back to November 2005. We track the complete evolution over a period of one year from November 2005 to December 2006, including the creation of the first 60000 nodes and 8 million edges created.

5.4 Bursty nature of link creation

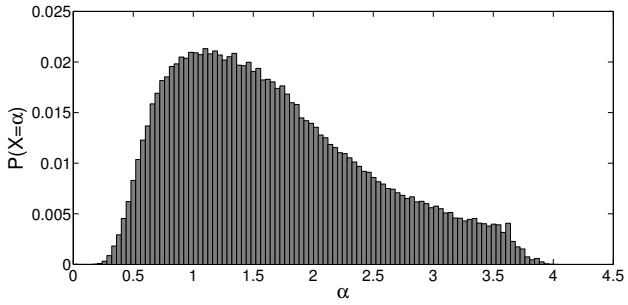
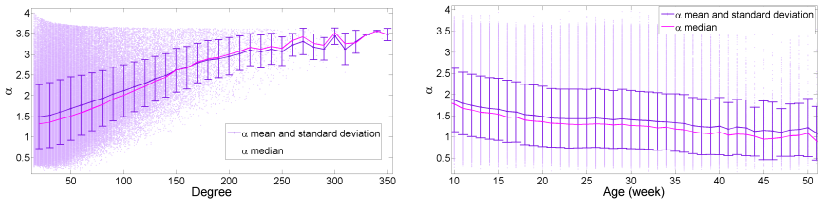


Fig. 5.1 Distribution of the scale parameter α, which characterizes the inter-event time distribution between consecutive link creations for a single individual. α values have been grouped in bins of length 0.05. Values past the peak around 1 decrease much more slowly with respect to the left side.



(a) (b)

Fig. 5.2 a) Mean, median, and standard deviation (error bar) of α as function of the final degree. Mean and median increase with the final node degree. To compute these values, we group node degrees in bins of 10, and consider their relative α values. b) Mean, median and standard deviation (error bar) of α as function of node age measured in weeks. Mean and median decrease very slowly with node age.

Bursty behavior has been observed in various contexts such as WWW traffic patterns [39], emails exchanges, and in general human behavior [154]. But it has not been studied in the context of online social networks. In this section, we study the link creation process as the growth of the neighborhood of each single node, and show that the linking activity of online social networks users is characterized by temporal bursty patterns.

To prove the burstiness of link creation, we consider for each user the event time series where an event represents the creation of an edge incident to the considered node. On each time series, we apply the technique proposed in Vasquez *et. al.* [154] and extended by McGlohon et. al. [105], both

based on the inter-event time distribution between consecutive events for a single individual. If the edge creation process is a Poisson-like process, *i.e.* homogeneous, then the inter-event time distribution should be an exponential distribution. On the other hand, a bursty arrival process is characterized by a power-law distribution where many short time intervals, each corresponding to intensive activities forming a burst, are separated by relatively fewer but longer periods of low or zero activity.

Results. In order to distinguish if the process is homogeneous or bursty, we fit the inter-event time data per node in our dataset using MLE (Maximum Likelihood Estimator), and select the model with the minimum AIC (Akaike Information Criterion). As a representative of the power law distribution family, we choose the Pareto with exponential cutoff $P(t) = t^{-\alpha}exp(-t/\lambda)$, and use the exponential distribution $P(t) = \mu exp(-\mu)$ to describe the inter-event time Poisson process. Finally, to avoid the impact of outliers, we remove from consideration users who have too few events, *i.e.* nodes with final degree less than 15 (median degree).

Our results show that the Pareto distribution with exponential cutoff exhibits the minimum AIC, meaning that almost all users in our dataset manifest a bursty behavior in link creation. In addition, the Kolmogorov-Smirnov (K-S test) validates the selected hypothesis for almost all users (86% of the population). These measurements offer direct evidence that at the level of a single individual, there is a heavy-tailed activity pattern, also found in other datasets [50].

Having shown that individuals add links in a temporal bursty manner, we analyze the similarity of the bursty process across users, by computing the distribution of the scale parameter α determined separately for each user. As shown in Figure 5.1, α values are scattered around a peak at 1, with an heavy tail in the right side. This partially corroborates the results found in [154], which showed a single group of users with very similar behavior described by the Gaussian distribution of α centered at 1. However, the heavy tail suggests that users in online social networks cannot be easily grouped in a single category, but have quite different behaviors in adding links.

To understand the reasons behind the observed heavy tail, we take into account two factors: the degree and the age of a node, *i.e.* how long the node has been in the network (in weeks). In Figure 5.2, we show the relationship between the scaling parameter α and the two variables, i.e. degree and age of a node. Figure 5.2(a) shows that the mean α value increases with degree. This fact suggests that nodes with higher degree contribute more to the right tail. This means that, although all the nodes manifest the same bursty behavior, nodes with higher degree have more closely spaced bursts. On the contrary, Figure 5.2(b) shows that age does not influence the right tail, since the mean value in each age slot is close to the mean of the α distribution in Figure 5.1, and remains quite constant for different age values. The small decrease is due to the fact that older nodes have a greater chance to undergo long periods of inactivity.

In summary, we showed in this section that users follow a bursty process in creating links, where bursts occur more frequently in nodes with high final degree. In the next section, we will describe our new proposal for using acceleration and deceleration as metrics to analyze graph dynamics.

5.5 Degree Acceleration

Bursty phenomena have been studied in different areas of human activities, such as clicks or queries in search engines [125]. However, these previous investigations focused on bursts resulting from aggregate actions, such as group of users that manifest a common interest at a certain time. These burst detection algorithms are not suitable to investigate per node time series based on their link creation events, or substructures inside bursts.

In this section, we propose a new methodology that identifies different phases that make up the bursty nature of the link creation process, and detects when bursts occur. We also identify the role played by each phase during the bursty process. From a high level, we observe that the alternation of activity/inactivity phases determines the burstiness of the event trace. In addition, bursts of activity have a typical internal structure, composed by a rapidly increasing slope and a gradually decreasing phase possibly interleaved by a plateau. An example is shown in Figure 5.3.

Degree Acceleration. Inspired by studies in physics and neuroscience on highly dynamic systems [32], we investigate the phases in bursty processes and detect bursts by measuring significant increments and decrements of new links formed per node. A burst begins when link formation activity rapidly increases, and ends following a decreasing phase. By leveraging the concept of acceleration, it is possible to easily identify and quantify significant changes in link creation activity. Let $d_i(t)$ be the degree of node i at time t, *i.e.* the total number of links incident to node i at time t, and let Δt be the time granularity that interleaves each $d_i(t)$ measures. We can then compute degree acceleration as:

$$a_i^d(t) = \frac{d_i(t) - 2d_i(t - \Delta t) + d_i(t - 2\Delta t)}{(\Delta t)^2} \tag{5.1}$$

By computing degree acceleration, we can observe the initial start of bursts ($a_i^d >> 0$) and a burst's decaying phase ($a_i^d << 0$). An example is shown in Figure 5.3. Note that acceleration captures two types of steady state conditions: a period of consistently high activity representing the plateau inside an activity burst (after an initial acceleration phase), and a steady state of low activity outside of activity bursts.

Defining Phases. While exploring the burstiness of the link creation process, we found that the growth of each node is characterized by transition phases in which users significantly change their link formation behavior. This

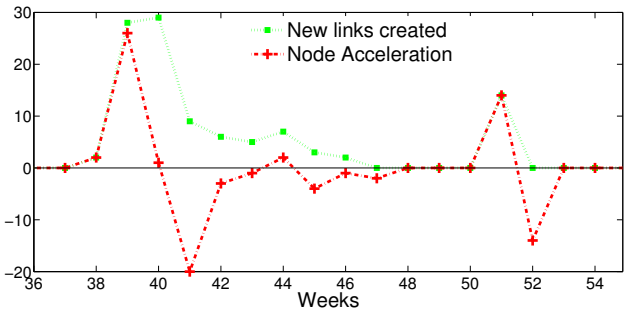


Fig. 5.3 An example of degree acceleration, computed on a single node from our dataset. The green dotted line represents the number of links created by the node each week. The red dotted line represents the acceleration computed according to Eq. 5.1. In week 39, the node shows a large acceleration, follows by a plateau. The node decelerates into week 42, when it enters a cruising phase (link creation is stable) for 4 weeks.

led us to identify four different phases that describe the patterns involved in the nodes' growth processes. The different phases can be described by defining a time-dependent state variable for each node in the system. More specifically, the acceleration phase is characterized by a large increment in creating new links, i.e. $a_i^d >> 0$, and the deceleration phase is described by a strong decay measured by $a_i^d << 0$. Then we define two intermediate phases: *cruising* and *inactivity*. The first corresponds to a steady state of a node, where the number of links created per week is almost constant. This phase can correspond both to high activity or to small oscillations around inactivity, and is characterized by at least one new edge (captured by the variable $c_i(t) = 1$) and small a_i^d values. These small a_i^d values are centered around the value $a_i^d = 0$, and are bounded with two thresholds θ_1 and θ_2. The second phase, *i.e.* inactivity, occurs when a node does not create any links for an entire time window. We formalize these four phases by introducing the function $s_i(t) : \mathbb{R} \to \{acc, dec, cruise, inact\}$ defined as follows:

$$s_i(t) = \begin{cases} acc & a_i^d(t) \in (\theta_1, +\infty) \\ dec & a_i^d(t) \in (-\infty, \theta_2) \\ cruise & a_i^d(t) \in [\theta_2, \theta_1] \wedge c_i(t) = 1 \\ inact & c_i(t) = 0 \end{cases} \tag{5.2}$$

where $c_i(t) = 1$ if and only if node i creates at least one edge at time t, otherwise $c_i(t) = 0$. Degree acceleration $a_i^d(t)$ and the related $s_i(t)$ function represent a general tool to investigate the burstiness structure, and to highlight the detailed properties of each phase.

5.6 Experimental Analysis

In this section, we characterize the link creation process by analyzing our traces using our acceleration methodology. The experimental analysis has been performed with the following settings: $\Delta t = 1$ week to avoid cyclic fluctuations in acceleration due to increase in user activities over each weekend, and cruising phase thresholds are $\theta_1 = 2$ and $\theta_2 = -2$. On each user event time series we apply the degree acceleration methodology, thus identifying bursts and their internal phases.

We first calculate the duration of each burst and the inter-burst times, *i.e.* the time between consecutive bursts. By leveraging the s_i function, we can convert a node time series in a string where a burst corresponds to the string pattern $acc^{+}cruise^{*}dec^{*}$. In this way we can extract substring and detect when a burst happens, how long it lasts and how long after the next burst appears. Results on these analyses are presented in Figure 5.4. As concerns the burst duration we can see that one half of the bursts lasts a week and most of them hold over at most 4 weeks. Therefore, inter burst times exhibit the same behavior of the inter-event time distribution; in fact bursts could also appear with a certain probability after different weeks, as the red bar plot suggests.

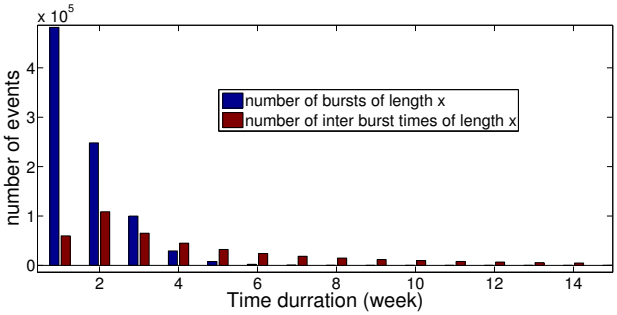


Fig. 5.4 Histogram of the duration (blue) and of the inter-burst times (red)

5.6.1 The Role of Phases

The role played by each phase along the node lifetime is a key element to understand the network dynamics, and is also crucial when designing generative models based on per-node temporal behavior. To this purpose, we consider two main aspects: (i) the time a node spends in each phase and (ii) the per-node amount of links created in the different phases.

We perform this analysis from two perspectives, by considering the aggregate behavior of all nodes, and on per-node behavior. In order to understand the role of different phases during a node's lifetime, we define ϕ^l and $\psi^l(i)$ to compute the percentage of time spent in each phase by all nodes (Equation 5.3) and by each node (Equation 5.4).

$$\phi^l_{phase} = \frac{\sum_{i \in N} \sum_{t=1}^{T} \mathcal{I}_{phase}(s_i(t))}{\sum_{i \in N} life(i)} \tag{5.3}$$

$$\psi^l_{phase}(i) = \frac{\sum_{t=1}^{T} \mathcal{I}_{phase}(s_i(t))}{life(i)} \tag{5.4}$$

where $life(i)$ represents the lifetime in weeks of a node, $\mathcal{I}$ is the indicator function and $phase = \{acc, dec, cruise, inact\}$. N indicates the number of nodes at time T, which represents the last week considered in the dataset.

The relationship between link creation and phase is quantified by ϕ^e, which corresponds to the percentage of the overall edges created within each phase, and $\psi^e(i)$, which is the link generation rate for node i in each phase:

$$\phi^e_{phase} = \frac{\sum_{i \in N} \sum_{t=1}^{T} \mathcal{I}_{phase}(s_i(t)) n_i(t)}{2m} \tag{5.5}$$

$$\psi^e_{phase}(i) = \frac{\sum_{t=1}^{T} \mathcal{I}_{phase}(s_i(t)) n_i(t)}{d_i(T)} \tag{5.6}$$

where m is the number of link at time T and $n_i(t)$ is the number of links acquired by the node i in the time slot t. The results are reported in Table 5.1, where $\psi^l_{0.8}$ and $\psi^e_{0.8}$ are the 0.8-quantile of the distributions of ψ^l and ψ^e, and are discussed below.

	acc	dec	cruise	inact
ϕ^l	0.11	0.14	0.28	0.47
ϕ^e	0.52	0.17	0.31	0
$\psi^l_{0.8}$	0.25	0.27	0.44	0.66
$\psi^e_{0.8}$	0.74	0.70	0.16	0

Table 5.1 In the first two rows we report ϕ^e (definition 5.5) and ϕ^l (definition 5.3) values for each phase. In the last two rows the 0.8-quantiles of ψ^l and ψ^e distributions.

Inactivity phase. During the inactivity phase, by definition, we do not observe growth since no links are created. However, inactivity acquires importance in the temporal dimension, because it deeply affects the burstiness. The high values of ϕ^l and $\psi^l_{0.8}$ highlight that node activities are concentrated in few and small periods; thus, for most part of their life, nodes do not influence the network dynamic evolution.

Acceleration and Deceleration Phases. Nodes spend only a small amount of their life in these phases, in particular after acceleration events, longer periods of weaker activity follow. However, the amount of links generated in these phases determines the structure of our social network. In fact, a link has very high probability, 69%, to be generated in one of these two phases, in particular 52% in acceleration and 17% in deceleration.

Cruising Phase. Cruising periods cover an important portion of nodes' lifetime. Furthermore, $\phi^e = 0.31$ and $\psi^e = 0.16$ would suggest that this phase has a role also in link creation. However, only few cruising periods have relevance in the edge growth. Indeed, it depends on whether the cruising phase is inside a burst or it corresponds to small oscillations around inactivity. A node in a burst cruising phase is creating many links, while in the other case the number of links created is irrelevant. Finally, the cruising phase has a pronounced impact only for nodes with low degree, as shown in Figure 5.5.

We have shown that acceleration (*acc*) and deceleration (*dec*) phases are those responsible of the growth and dynamics of the network, despite the fact that they represent a very small part of a node life.

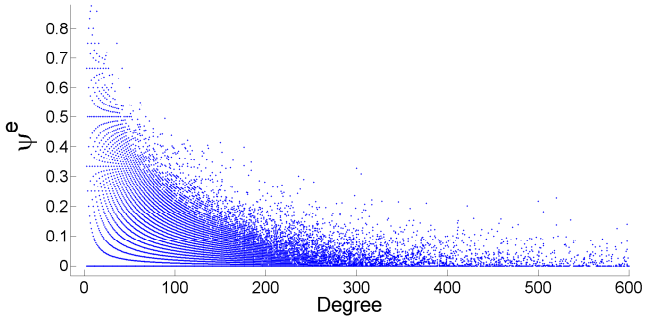


Fig. 5.5 Relationship between ψ^e_{cruise} and the degree. ψ^e_{cruise} decreases as the degree raises, so the cruising phase has a pronounced impact only for nodes with low degree.

5.6.2 Acceleration and Deceleration Features

In depth understanding of acceleration/deceleration phases reveals how users operate in the network after they join. This knowledge could be very useful to ensure efficient management of the OSN's resources. This section focuses on acceleration and deceleration by means of illustrating their importance from a network perspective; showing that they follow a power law distribution and finally investigating the impact of node aging on link creation process.

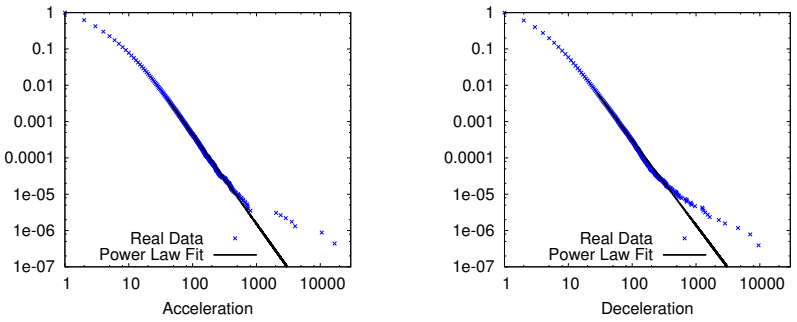


Fig. 5.6 a) Acceleration CCDF (complementary cumulative distribution function) and the resulting fitted distribution ($\alpha = 3.46$). b) Deceleration CCDF and the resulting fitted distribution ($\alpha = 3.34$).

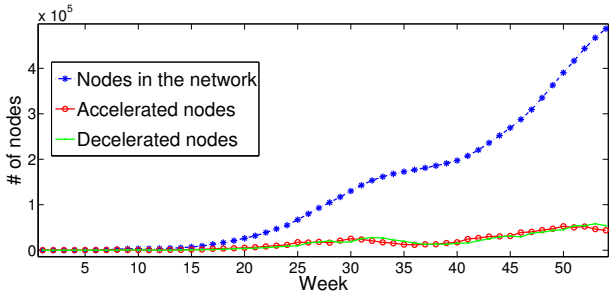


Fig. 5.7 For each week, the number of nodes in the network (network size) and the number of nodes in the acceleration/deceleration phases. In each week only a very small number of nodes are in the acc/dec phases though the network size rapidly grows.

Network perspective on acceleration/deceleration. From the network perspective, an estimate of how many and which nodes are changing the graph structure would greatly help in managing the system resources. Figure 5.7 shows that in each week only a very small number of nodes are in acc/dec phases. Specifically, at the end of the year, they roughly represent the 20% of the network nodes. These nodes can be easily identified as soon as they experience a phase transition from the inactive/cruising to the accelerated phase since their values of acceleration abruptly increase.

Acceleration/deceleration probability distributions. By applying the statistical framework proposed by Clauset *et al.* [36], we find that acceleration and deceleration distributions are power law, (Figure 5.6(a) and 5.6(b)). By considering the overall network, this result implies that half of the acceleration and deceleration events have a small size, but they are very likely to show rapid increase and decrease respectively. The upper tail of the ac-

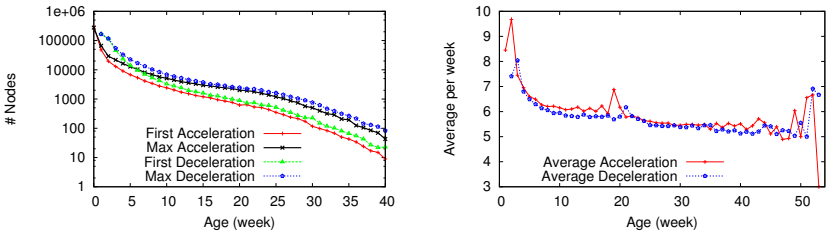


Fig. 5.8 a) shows the times when nodes experience their first acceleration, maximum acceleration, first deceleration, and maximum deceleration (y-axis on logscale). b) shows the average acceleration/deceleration with respect to node age.

celeration distribution exhibits such high values of acceleration that can't correspond to normal user. Those events are most likely associated to people with a large amount of followers or accounts for advertisement.

The impact of aging. The general behavior of a node is a sequence of acceleration/deceleration phases of constant magnitude, after an initial burst. In general, nodes wait at most for one month before initiating their activity.

We start by defining the $age(t)$ of a node u as the time elapsed between the appearance of u in the network (timestamp of the first edge incident to u), and time t. The observables whose dependence on age, need to be studied are: $n_{firstAcc/Dec}(t)$, the number of nodes showing their first acceleration / deceleration at time t and $n_{maxAcc/Dec}(t)$, the number of nodes manifesting their maximum acceleration / deceleration at time t. Finally, we calculate the average acceleration/ deceleration $avg_{Acc/Dec}(age)$.

Analyzing and comparing $n_{maxAcc}(t)$ and $avg_{Acc}(age)$ in Figures 5.8(a) and 5.8(b), we observe that most nodes enter the phase of maximum acceleration in the first week. In addition, Figure 5.8(a) shows that the activity after the first peak does not decrease as fast as its respective acceleration.

Figure 5.8(b) highlights another interesting behavior of the acc/dec phases. The average acceleration remains constant when age increases. This is consistent with what we found in Figure 5.3, *i.e.* nodes experience a big burst of acceleration in the first week after joining the network, and subsequent bursts never match the first in intensity.

5.7 Conclusion

In this chapter we investigated the bursty nature of the link creation process in OSN. We prove not only that it is a highly inhomogeneous process, but also identify patterns of burstiness common to all nodes. In terms of edge

creation, users are inactive for most of their lifetimes, and concentrate their link activity in a number of short regular time periods. To characterize node activity, we developed a new methodology based on the acceleration of degree growth, which allows us to highlight the internal structure of link creation bursts.

We believe using acceleration as a general metric to characterize network dynamics prompts future work in studying link generation mechanisms. In particular, defining different phases of edge creation hints at the possibility of characterizing users into distinctive activity levels that correlate with their likelihood of adding social links. Some preliminary results confirm this intuition: when nodes (users) first join the network, they create links based on the preferential attachment mechanism; while in later bursts, nodes seem to explore (acceleration phase) and densify (deceleration) in far regions of the graph that could correlate to spatial areas. These results open the door for new generative models [140] that consider different phases of node activity.

Chapter 6
Link and Triadic Closure Delay: Temporal Metrics for Social Network Dynamics

6.1 Introduction

Online social networks have been extensively studied in the last decade, starting from their static properties computed on a single snapshot of the network. A huge amount of works highlight interesting properties about the fat tailed degree distribution, the homophily and a positive degree correlation, the high average clustering coefficient and the small average distance length between pairs of users. In general we have witnessed the analysis of larger and larger networks which has culminated in the study of the whole Facebook graph [153]. Subsequently, the research focus shifted toward the temporal properties of OSNs with the aim of understanding dynamic patterns and mechanisms. In particular research on online social network has dealt with two main themes: the dynamic of the networks, *i.e.* how the graph evolves and the dynamics of the process running on the network, *i.e.* how information is spread and propagates. Up until recent years these two aspects were studied separately, although it is becoming clear that information spread through the network influences the users in expanding their neighborhood. Similarly, it leads to a reshaping of the entire network. For instance a tweet via Twitter might spark the interest of other users, who, intrigued, might begin following the user who wrote the original tweet. Furthermore the diffusion of information is not the only factor determining how the network structure evolves. As discussed in Chapter 1, online social networks are not isolated systems; rather, they are part of a complex ecosystem where users maintain different kinds of relationships and are linked at diverse levels, in various dimensions. Moreover in the external ecosystem the same loop between diffusion and growth may reoccur.

The above considerations give rise to a new perspective in viewing online social services: namely, the notion of contagious network is taking the place of the concept of social network. The term contagious network [143] indicates a network that grows by adding new users or links through the exploitation

of social ties external to the system itself or by including information about the presence of a user retrieved through a search or some suggestion/sharing mechanisms. Temporal information on the network growth, i.e. on the diffusion of the contagion, can be exploited not only to study how the process spreads but also to improve our understanding of the mechanisms acting during the contagion and to grasp whether or not the process has been driven also by external factors. As far as the last problem is concerned, physical time, as introduced in Chapter 5, makes a good fit with the study of the influence of external events because it represents a bi-directional mapping between events occurring on the contagious network and events external to it. Moreover, in contrast to logical time, physical time enables us to indicate when a growth mechanism acts and how long its effects last.

For the above reasons, the belief that the understanding of network dynamics must necessarily take physical time into consideration is gaining momentum. Only recently have a few works dealt with the problem of describing the evolution of online social networks, embedding the physical time of the real world where users act. They mainly adopt a macroscopic vision of network evolution by measuring the trend of some measures in each temporal snapshot [167, 159]. In this chapter we go a step further in understanding the dynamics by adopting a microscopic approach enriched by the physical time variable. In particular we analyze the true dynamics of the two fundamental network building components: dyads and triads. Within the sphere of this physical timing microscopic approach, here we are proposing two new metrics: the link creation delay and the triangle closing process delay. The *link delay* is the time used by two users of a network to create a link between each other. This accounts for the time it takes for a potential friendship to become real. From a contagious network perspective, link delay might be adopted as an efficacy measure of the diffusion of the information about a user's presence or it could indicate that a link is already present in the underlying latent network and actualizes in the online social network. More specifically, if two users are four or five hops away and the connecting link is established as promptly as possible, we may reasonably assume that the two users already know each other outside the network context. These considerations are not possible if we use logical time, since the previous phenomenon (not-local links) is not related to the actual number of nodes or links in the system. Regarding triads, we focus on the triangle closure dynamics. Triadic closure, which describes the transitivity of friendship, has been proven to be effective in modeling network evolution and in predicting future link formation. The triangle formation process can be reformulated well as a friendship information diffusion; in fact in the triangle A-B-C, the presence of A propagates to C through the links AB and BC. This propagation is not instantaneous; by contrast, it takes a little while. The *triangle delay* represents the time it takes for the three users to all become friends. In general these new metrics enable us to study the dynamics of creation of dyads and triads and to

highlight network behavior that would remain hidden if not-timing measures were adopted.

In this chapter our main contribution concerns the temporal features of links and the triangle formation measured on the two temporal annotated datasets presented in Section 6.3 (Facebook). As for link delay, in Section 6.4 we find that this quantity is very low, accounting for the fact that if two persons wish to form a friendship they do so promptly. Specifically, when a social network is still in the earliest stage of development link establishment takes place quite quickly. Moreover, link delay proves to be independent of the actual date when people join the network. This highlights purely social aspect of link delay. So we introduce and define the concept of link peerness, which denotes how the linked nodes are coetaneous. Furthermore, by comparing two kinds of locality of the edge topological and temporal we examine how a friend's presence in the network propagates. Results show that links that span farther nodes generally establish faster. Finally we show that link delay is relevant in interaction analysis, as it represent a good indicator of the level of interaction between nodes. With respect to triads, our contribution in Section 6.4 is twofold; we define and study the concepts of both temporal triangle and triangle delay. First we introduce an algorithm for the extraction of temporal triangles. This enables us to monitor the triangle formation process and to detect sudden changes in the triangle formation behavior, possibly related to events external to the network. In particular, we show that the introduction of "People You May Know" (PYMK) Facebook functionality had a disruptive impact on the triangle creation process in the network. From a microscopic perspective, we shed some light on the physical time of the triadic closure process by introducing a formal definition of the delay of the triangle formation. The triangle delay represents a normalization which accounts for the node and link arrival processes. By analyzing the above quantity we find that triad formation is very fast. This derives from the fact that if two persons have friends in common and are willing to form a relationship they will do the latter promptly. Yet, this new metric shows slightly different in the bootstrap phase of Facebook dataset which is also faster in the triangle formation dynamics w.r.t. the Facebook growth. In addition, the triangle delay allows us to identify which latent triangles have been stimulated by external events or services. Therefore through triangle delay we are able to evaluate the effects of these types of events.

6.2 Related Work

With the availability of datasets on online social networks which evolve in time, researchers start thinking about the mechanisms by which nodes arrive and links form or disappear. Among proposed mechanisms of how a link is created, triadic closure is the most basic and powerful principle to model the

evolution of social network. Specifically this principle states that if individuals with a common friend have a higher chance to become friends themselves at some point in the future [135]. This basic role of triadic closure in the evolution of social network has motivated the introduction of the clustering coefficient [157], defined as the probability that two randomly select neighbour of a node are friends with each other. Given the correlation between triadic closure and clustering coefficient a few works have faced the dynamic of the triadic closure process by analysing the temporal trend of the average clustering coefficient. By exploring the social network dynamics of a portion of the arXiv repository, Amblard *et al.* [5] shows that the average clustering coefficient is quite constant (≈ 0.5) since 1992, nevertheless they did not investigate how triangles form. A more detailed analysis of the average clustering coefficient in a Chinese social network has been presented by Zhao *et al.* [167]. Authors observed that in the early the average clustering coefficient is high, but once the network expands, it move to a smooth decreasing curve. Gonzalez *et al.* [61] have presented a detailed characterization of Google+ based on large scale measurements. By comparing the clustering coefficient distributions taken in different periods, they showed that G+ structure has become less clustered as new users joined the largest connected component over the one year period. Although the distribution comparison says nothing about how the triadic closure has acted during Google+ evolution. More insightful results about the clustering coefficient in Google+ has been proposed by Gong *et al.* [59]. In the paper the average clustering coefficient has reported to follow a three-phase evolution pattern where first it decreases then increases slowly and finally decreases again. In general the previous approaches suffer from the same averaging effect, indeed the wide fluctuations affecting the clustering coefficient of the single nodes are lost in the computation of the average. This way they obtain a measure of the whole network which poorly describes how the triadic closure process acts.

While the aforementioned studies focus on the temporal trend of the clustering coefficient to quantify the magnitude of the closure process, other works combine the snapshot paradigm with the likelihood of a link given the number of common friends. One of the seminal work has been carried on by Kossinets and Watts [86]. They analysed the fraction of links formed during two consecutive snapshots of a e-mail network as a function of the number of common friends measured on the first snapshot. They have found that two users are more likely to close a triangle if they share many friends; moreover this dependence is almost linear in the number of common friends. Even though the method does not explicitly focus on the process timing, it heavily depends on the choice of the temporal gap between two snapshots. A more detail and wider study of the triadic closure process has been conducted by Leskovec *et al.* [92]. Starting from the empirical observation that a high fraction of edges close triangles, they model how a source node decides to add an edge to some other node two hops away. By using Maximum Likelihood Estimation, they have found the most likely choice mechanism is given by a

combination of the number of common friends between the extremes and the recency of activity by the possible destination. Due to the adopted microscopical approach, this work is the most related to ours, nevertheless it focuses on the choice mechanism neglecting any temporal trend of the triangles or how much time they take to establish.

Also Mislove *et al.* [108], Viswanath *et al.* [155], Garg *et al.* [53] have treated the role of triadic closure in the evolution of online social network. In particular they analysed the proximity bias, *i.e.* the tendency of nodes to link with those nearby in the network graph. The results obtained on Flickr and FriendFeed have shown that proximity bias influences the formation of new links, making two nodes which are two hops distant, more likely to form a link. These studies only confirm the predominant role of the triadic closure but they do not deep the closure process.

With respect to the above literature in our work we investigate the temporal dynamic properties of the triadic closure process and the link formation. As for triadic closure, to the best of our knowledge this is the first study which explores how long the triangle formation lasts and the role of external events on the closure process. Moreover we also provide a link property, the link delay, which accounts for the time a link has to wait before forming and may be adopted as measure of the link strength.

6.3 Measurement Methodology

The main goal of our work is to introduce some measures that can quantify the microscopic dynamics of the growth process of online social networks adopting physical time as reference system. A main advantage that the physical time approach offers is the possibility of relating the global and local changes in the growth process to events external to the network system.

Besides the possibility of a mapping between structural events inside and outside the network, the physical time allows a proper understanding of the growth processes - for example, by highlighting coupling or synchronization effects among the nodes as in the brain network, or addressing the choice of the proper growth is acting. Moreover, we can apply physical time not only to the link creation process but also to the interactions that arise on them. This way we can show the relation between the temporal features of the links and the relative interaction properties. In other words, temporal information plays a role in defining the type of interaction.

Introducing the new measures we start by presenting the theoretical framework that we have adopted to describe the growth processes in physical time. Then we introduce the datasets on which we apply the dynamics measure. In particular we needed online social networks with fine-grained temporal information, so as to stress the microscopical dynamics of their evolution. In

our case (Facebook), we can explicitly highlight the structure dynamics and the interaction dynamics occurring on it.

6.3.1 Notation

In the following we introduce the notation we adopt to describe the growth of the whole network and its constitutive elements, as well as the user interaction properties. Formally, we represent the social network as an evolving undirected graph. Usually the network growth is represented by a sequence $< G_1, \ldots, G_T >$ where each $G_t = (V_t, E_t)$ is an undirected graph denoting the state of the network at time t having $|V_t|$ nodes and $|E_t|$ links. As we have no information about node and edge removals, the number of nodes and edges always increases in time up to the end of the measurement process indicated by T. At last, graph $G_T = (V_T, E_T)$ will contain the whole set of nodes and edges appearing during the growth.

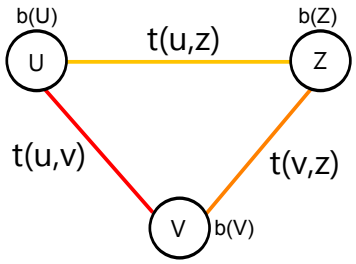


Fig. 6.1 Triangle $u\hat{v}z$ with temporal information about the edge creation τ and the birth date b of its nodes. Link colors (from red to yellow) are proportional to time appearance.

In order to analyze the microscopic structural properties and make definition more affordable, we project the sequence $< G_1, \ldots, G_T >$ into an undirected graph $G = (V, E)$ where links are time-stamped by the *time function* $\tau : E \to \mathbb{R}$ that assigns to each edge $e = (u, v)$ its creation time. Given the increasing monotonicity of the sequence, $V = V_T$ and $E = E_T$. In accordance with this framework, we use $K_t(u) = \{v | \tau((u, v)) \leq t)\}$ to denote the set of neighbors of u at time t and consequently the degree of node u at time t as $k_t(u) = |K_t(u)|$, while its final degree at the end is $k_T(u)$. As we have temporal information only about edges, we define the time of the first appearance of node u into the network $b(u)$. We call this the *birth date* assuming that $b(u) = \min(\tau((u, v)) | v \in K_T(u))$, i.e. the time of the first link incident to u.

We extend the evolving graph notation to include the interactions occurring between a pair of linked users. Interactions between pairs of nodes are

$b(u)$	time of the first appearance of the node u
$\tau((u,v))$	creation time of the link u,v
$K_t(u)$	u's neighbours at time t
$k_t(u)$	degree of u at time t
$k_T(\mathrm{u})$	final degree of u
$I((u,v))$	interaction times between u and v
$n_\iota((u,v))$	number of interactions between u and v
$d((u,v))$	delay of the link (u,v)
$d(u\hat{v}z)$	delay of the triangle $u\hat{v}z$

Table 6.1 Notation summary

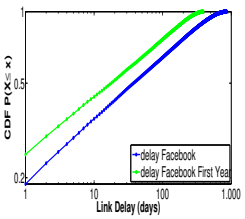
described by the function $I : E \to \mathbb{R}^*$ which returns the set $I((u,v)) = \{t_1, \ldots, t_k\}$ of times the interactions between u and v, while $n_\iota : E \to \mathbb{N}$ represents the number of interactions during the period under analysis, i.e $n_\iota((u,v)) = |I((u,v))|$.

Finally, as the main subject of Section 6.5 is the dynamic of transitivity, we denote a triangle composed by the vertices u,v and z as $u\hat{v}z$ as shown in Figure 6.1. Exploiting the time function we can assume that for each triangle $\tau((u,v)) < \tau((v,z)) < \tau(z,u))$ holds. Consequently, we have an ordering of the edges in a triangle and we lose the triangle isomorphism typical of static undirected graphs, i.e. $u\hat{v}z$ is not equivalent to $u\hat{z}v$. This property allows us to assign each triangle to a single node, which we call the *initiator* of the triadic closure.

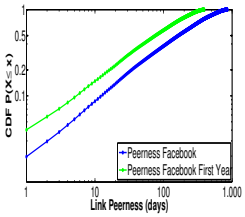
6.3.2 Dataset

The main obstacle to studying dynamics is the difficulty of obtaining detailed data describing OSN dynamics. In this study we employ an online social gathered from Facebook [155]. The dataset results from a crawling of the New Orleans Facebook network that collects the network growth of about 60.000 nodes and 800.000 links from September 2006 to January 2009. It contains the timestamped creation of all users and edges, but 4.2% of vertices and 6.0% of links, which were not considered in our analysis. As this phenomenon is very limited, the results that we obtained are still very general. This dataset has two interesting feature: *i*) it describes a growth period at an early stage of Facebook's existence; *ii*) it contains user Wall interactions between 176054 pairs of users, corresponding to 21.54% of the links in the network. This latter feature allows us to study the possible relations between the dynamic topological characteristics of the social network and the temporal properties of relative user interactions.

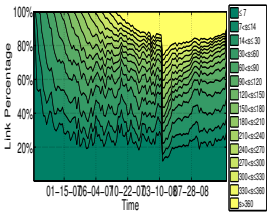
6.4 Link delay



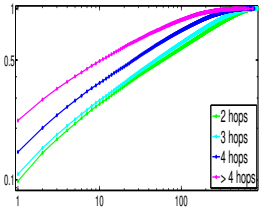
(a) Link Delay



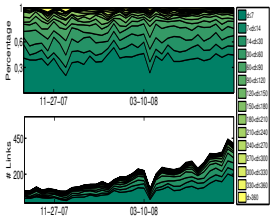
(b) Link peerness



(c) Link Delay Classes in Time (Facebook)



(d) Link Delay and distance (Facebook)



(e) Link Delay and distance > 4 hops (Facebook)

Fig. 6.2 6.2(a) and 6.2(b): CCDF of the link delay and link peerness measured on Facebook First Year and Facebook. Figures have different y-axis scale, 6.2(a) starts from 0.2 while 6.2(b) starts from 0.1 . 6.2(c): The trends of the link delay classes during the growth of Facebook. Class green intensity is inversely proportional to link delay (upper and lower bound of the classes are in days).

In this section we introduce *link delay*, a novel indirect measure of the eagerness of a tie obtained by quantifying the elapsed time between the potential establishment of a link and its real activation. A link is potentially established when all the enabling conditions are set but the link has not yet been created. When applied to social networks, the link delay is an index assessing the strength of a social relationship between individuals.

Below, first we define the policy that we have adopted to measure the time spent to establish a link between two nodes, i.e. the delay of that link. Then we evaluate the link delay properties on Facebook finding that link delay is very low, accounting for the fact that if two persons wish to get a friendship relation, they actualize their willing very fast. Exploiting the Facebook dataset peculiarity on wall posts we compare the link delay with the dynamics of node interactions. This enables us to highlight the existing relation between the link delay and the level of interaction between adjacent nodes. Although the observed behavior does not allow us to assert that the number of interactions is strictly dependent on link delay, it clearly suggests that

link delay really matters in predicting the number of interactions between adjacent nodes.

6.4.1 Definitions

We assume that nodes are free to enter the network anytime during the network lifetime. To properly mirror this assumption in the link delay definition we apply a simple normalization on the values returned from the time function to include the birth date b of a node. This leads us to define the *link delay* $d(u, v)$ as follows:

Definition 6.1. Given $G = (V, E)$ and its time function τ, the delay $d : E \to \mathbb{R}$ of the link (u, v) is defined as

$$d(u, v) = \tau(u, v) - \max(b(u), b(v)) \tag{6.1}$$

where the *max* function on the birth date implies that both nodes need to be created in the graph. $d(u, v)$ measures the elapsed time between the potential link creation time (when all conditions hold) and the real link creation time.

6.4.2 Link delay analysis

Overall statistics. Here we compute and analyze the distribution functions of link delay. The distribution of the delay of link creation can shed light on a few properties of OSN's friendship: i) how much time people take to become friends once they both join the OSN, i.e. we wonder if two users form a friendship quickly or not; ii) how link delay changes in view of the different stages captured by different snapshots. Indeed we analyze the first year of Facebook dataset where Facebook becomes a service open to everyone from a university network and its consolidation period (overall Facebook dataset). Figure 6.2(a) shows the CCDF of the link delay extracted from the first year of Facebook dataset (green line) and from the entire Facebook dataset (blue). The link delay distribution always shows a very quick shift from potential to actual link state. In fact, for all the distributions links were created within a day in $20 - 27\%$ of the cases and within a week in $32 - 50\%$ of the instances. These results highlight that as soon as a pair of nodes have the chance to get connected they do so promptly.

Beyond this common strong property, some differences can be noted among the various stages. In fact the comparison between Facebook and its first year strengthens the existence of different behavior in link delay as the network growth, i.e. link are faster in the network early stage. The different behavior during the different Facebook macroscopical snapshots can be explained

by considering that in the early phase the information about the establishment of new friendship relation reaches more quickly the nodes, while during the network consolidation it slows down. Although link analysis delay at a macroscopical granularity is able to highlight differences among diverse network snapshots, we can deepen the growth analysis increasing the temporal granularity.

Temporal statistics. The static analysis on the aggregated link delay behavior is able to capture the impact of quick links on the overall network. However, we can obtain more information by taking into account when links are created. This way we obtain the temporal trend of the link delay classes as shown in Fig.6.2(c). According to the relevant link delay distributions we divide links in delay groups and we compute, for each week, the percentage of new links belonging to the different classes. In Facebook dataset we observe different behaviours, in particular the component of links having a delay within a week and representing the 32% of the overall links is characterized by a quite constant trend. In fact the dark green class in Figure 6.2(c) is always between 20% and 40%. A more evident phenomenon happens on March 26, 2008 where we observe a drastic increase in the volume of delayed links. This date corresponds to the introduction of "People You May Know" (PYMK) Facebook functionality. By analysing the delay we can highlight *i)* how this features acts and *ii*) how long its effect lasts. The friend recommendation system highly amplifies the tendency, already acting in the network, at establishing old potential that could be created many time ago. In fact 60% of the links created in the week of the PYMK introduction have a delay greater than 6 months and the 20% could be created an year before. Observing the class trends the weeks after PYMK introduction we note that the initial behavior in preferring delayed links disappears and after the summer of 2008 reaches percentages pre-PYMK. Although the link delay let emerge and capture interesting characteristic in edge creation process, it is not able to capture the reason behind, i.e. which process causes the observed effects or which algorithm has been adopted in the early stage of the PYMK feature. In Section 6.5 we deep into these effects correlating them to the triadic closure process. In order to quantify the instability we measure the standard deviation of the classes time-series. The more the standard deviation is far from 0 the more the time-series is disperse and fluctuates in time. We apply the stability time-series analysis the 10 weeks after March 26 to reduce the PYMK effects. The analysis of the standard deviation shows that a fluctuated and stable components simultaneously act during the Facebook New Orleans growth: 1) unstable quick links and 2) stable more delayed links.

6.4.3 Link speed and link peerness

The link delay observations provided so far are independent of the reciprocal network age of the nodes involved in links. However, the birth date b allows us to verify whether or not the link establishment privileges pairs of nodes having similar network ages. In order to observe this type of attitude, we introduce the notion of *peerness* p of a link:

Definition 6.2. Given $G = (V, E)$ and the birth date function b, the peerness $p : E \to \mathbb{R}$ of a link (u, v) is defined as

$$p((u, v)) = |b(u) - b(v)| \tag{6.2}$$

In Figure 6.2(b) we show the measures of peerness obtained from our dataset. We observe that the probability of having a link between peer nodes is low and nearly 50% of links are established between nodes with different creation times 5 months apart. We argue that this result is the direct consequence of the bursty behaviours in the edge creation process, as previously observed in [51]. In fact old nodes continue to generate or receive (as we consider an undirected graph) links even if they are ageing because 20% of links are respectively created 9 months or one years later a node's birthday

By comparing link delay and peerness distribution, as shown in Figure 6.2(b), we can derive that low link delay has an higher influence on the link population than low link peerness. In fact the impact of low peerness (≤ 7 days) is lower (at most 10%) than the contribution given by low delay links (at most 50%) This proposes the link delay as a relevance feature in the future growth of the networks and on the processes happen on them.

6.4.4 Link delay and edge locality

We can verify the previous hypothesis about how friend's presence in the network propagates comparing two kind of locality of the edge: topological and temporal. The notion of topological locality of a link (u, v) coincides with measuring the number of hops h it spans, i.e. the length of the shortest path from u to u removing the edge (u, v), while the temporal proximity is given by the link delay. We reasonably assume that the information about the node v's presence is proportional to the distance between u and v, i.e. further away a node is the longer the other node has to search for it. The previous assumption could be true if a node exploits only the network structure, but what happen if external effects speed up the propagation of the node presence? We answer comparing the topological distance and the link delay of each edge in the network. In Fig.6.2(d) we study the distribution of the link delay in four

classes of geodesic distance (2 hops,3 hops,4 hops and >4 hops[1]). We observe an unexpected result, in fact the distribution for the > 4-hops class lies above the other classes, supporting the fact that edges that connect nodes more than 4 hops distant establish earlier than closer nodes. The same behavior, to a lesser extent, involves also the 4-hops class. This fact seems an evidence that the information of a node's presence spread through channels different from the network. One could object that the low level of delay is due to the first months when the services has a smaller group of subscribers and as a consequence the phenomenon degrades as the network growth. Fig.6.2(d) shows that the tendency of non-local edges to establish fast links remains quite constant also at the end of sample. As shown in Figure 6.2(d) edges than span four or more hops are more likely a lower delay w.r.t. those ones which span closer nodes. In Figure 6.2(e) we report the absolute volume and the percentage of links covering more than 4 hops during the last year of the dataset divide in the delay classes adopted in the temporal analysis. We choose the last years to have a sample as numerous as studying the delay tendency . Also in this case the lowest components (≤ 7 and $(7-14]$) score the $40-50\%$ of links along the period indicating a stable behavior.

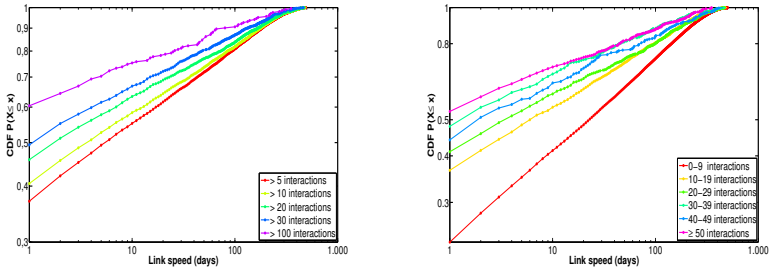
In general a quite surprising behavior involving link delay and edge locality exists. While some works [92], [41] report that closer nodes are more likely to establish a new link, our results suggest that different temporal behaviors occur. Links that span farther nodes establish generally faster. In such case the network connectivity is less important than external information or event in the creation of this kind of friendships. Furthermore this behavior is constant during the network evolution; that suggests a background component not heavily influenced by external shocking events.

6.4.5 Link delay and dyadic interactions

The Facebook dataset also offers a great opportunity to underscore the relation between the temporal properties of the links and the processes occurring on top of them. In particular, we consider the interactions between two users of a dyad. The basic assumption we want to verify is that link delay and interactions are related, as the delay is a eligible marker of the strength of a connection between two nodes. So, the faster a link the more likely becomes a high quantity of interactions between the nodes incident with it.

We examine the interdependence between interactions and link delay by considering different subsets of active users, where a user pair (u, v) is *active* if $n_\iota(u, v) \geq 5$ according to the definition in [155]. In order to compare our results with those in Vismanath *et al.* work, we adopt groups of pairs

[1] For computational constraints distances have been computed by a truncated version of the shortest path Dijkstra's algorithm that terminates 4 hops far the node, so >4 hops class might contain edges that connect components previously disconnected.



(a) Link delay and interactions (b) Intercation classes and link delay

Fig. 6.3 6.3(a) CDFs of the link delay within different levels of total interaction (overlapping classes). 6.3(b) CDFs of link delay within different not overlapping classes of wall interaction.

who exchange more than 5, 10, 20, 30 and 100 wall posts, then we compute the relative link delay. Figure 6.3(a) shows the link delay CDF of each single group. At this point we make two observations. Firstly, as the number of interactions on the wall increases, the more likely it becomes that these interactions occur on links at a low delay. This trend shows up more clearly in Figure6.3(b), where we divide user pairs in non overlapping groups based on numbers of interactions. In this case too, we observe the CDFs of the link delay in each group. We see in the figure that the link delay behaviors varies if we consider the set of links with either zero or few interactions (red line), where the 38% of edges have a delay of one week; or if we look at > 100 group, where we find that 66% of the links took place in the same span of time. So, link delay can be used as a good indicator of the level of interaction between the linked nodes.

Secondly, we can link the previous observation to the features of the pairs of nodes which exchange more than 100 wall posts observed in Vismanath *et al.*[155]. The authors found that pairs with very high levels of activity maintain a quite constant rate of interaction over time with respect to the other groups and keep their relationship alive for a long time. In the light of these facts, link delay can be used not only as a indicator of strength but also as a marker of the persistence of the interactions.

In general we say that link speed as a certain importance both in the heterogeneous evolution of the network and in the prediction of the level of interaction between a user pair. In some way interactive and active nodes try to establish as soon as possible relationships with people with whom they can exchange information or establish some form of useful communication. Meanwhile, less important types of connections can wait. Furthermore the tendency of non-local edges to establish fast links might suggest that external factors, such as friendships external to the network, are acting.

6.5 The Triadic Closure Process

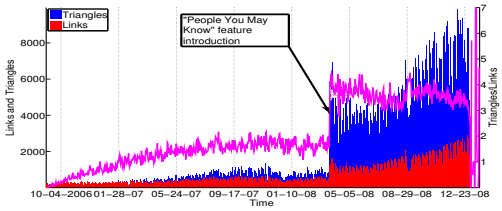


Fig. 6.4 Number of new links (red) and triangles (blue) formed during the growth of Facebook New Orleans, sampled each day. The magenta line represents the ratio between the triangle and the links created in a day (y-scale on the right).

The availability of temporal annotated networks has allowed the study of the network evolution over time and has led to a deep understanding of the mechanisms governing node and link arrival and creation. In this chapter we are dealing with the evolution mechanisms of online social networks, ergo we focus on the basic growth principle underpinning these networks: *triadic closure*.

Observed as one of the most frequent processes of link formation, triadic closure has been widely adopted in different disciplines. For instance, the sociological principle for triadic closure is the transitivity of friendship, which says that two individuals have a high likelihood of reaching to a friendship status if they share a common friend. The transitivity of friendship has been proved to be effective in modeling network evolution and predicting future link formation. Despite its commonly accepted value, the foundational principles governing triadic closure have not yet been analyzed in depth.

In this section we characterize the triadic closure process by delving into its temporal aspects. In particular we consider two perspectives. First we analyze the triangle formation growth by the network point of view by counting the new formed triangles on the overall dataset periods. To reach this goal we propose an algorithm able to extract temporal annotated triangles. Temporal information are used to monitor the number of triangles day-by-day. By analyzing the triangle time-series we are able to map sudden changes of the triangle formation onto events external to the network, such as the introduction of a new feature in the service or seasonal events involving most of nodes.

We move apart from the network perspective to embrace a microscopical viewpoint focusing on the formation of the single triangles. In particular we study the speed of the formation of triangles. Our goal is to shed a light on the dynamical properties of the triadic closure process by introducing the formalism to capture the time a triangle takes to be established. This way we

define the triadic closure delay and we show how this new metric captures different behaviors in the dataset under investigation.

6.5.1 Temporal triadic closure

We believe that the dynamical analysis of network evolution cannot disregard the transitivity closure process, for the literature has shown its importance in the formation of social networks - despite the fact that a temporal analysis of the triadic closure poses both algorithmic and methodology issues. The first concerns the extraction and counting of temporal annotated triads. While many approaches have been proposed in literature, most are suitable for static networks and so cannot be adopted in our microscopical view. Our approach in the study of triadic closure dynamics advocates, rather, an extension of the triangle enumeration methods in order to swallow the temporal information. [2] Our starting point is the observation that time annotation impacts the number of isomorphic triangles. In a simple undirected graph the number triangles isomorphic to $u\hat{v}z$ is 6 (3!), while in the temporal case the ordering induced by the time makes the isomorphism disappear. To extract all the temporal triangles we adopt Algorithm 1. At the end of its execution, it returns set Δ containing all the triangles $u\hat{v}z$ such that $\tau((u,v)) < \tau((v,z)) < \tau(z,u))$ without repetitions. For each triangle $u\hat{v}z$ in Δ we also store the τ values of the links to keep track of the instant of the closure.

Once the triangles have been extracted, we have all the information we need to study the triangle creation process during the network evolution. In Figure 6.4 we show the volume of triangles that are created daily. In Facebook during the three years of observation, we count more than 1.7 million temporal triads. By observing the triangle trend in the two datasets over the overall periods, we note a general inhomogeneity in the triadic closure process.

Obviously, triangle and link volumes are strictly related, as the increasing number of triangles could impact the overall number of new links. This is true even though not all links derive from the triadic closure process; for instance, they could be the consequence of new node arrival or some other effects, namely a preferential attachment process or search for new friends by graph exploration. Mixing the above mechanisms could result in new links and triads trends. For example Viswanath *et al.* [155] showed that the preferential attachment mechanism alone provides an expected number of triangles lower than in real social networks.

[2] Methods for frequency-based pattern and temporal graph matching [13] are not suitable for our purposes because of combinatorial arguments based on integer partition. For each integer i we should extract all the triangles such that the relative times of their links sum to i. In fact we are not interested in temporal pattern shifted in time.

A first hint in the validation of the impact of triadic closure on the creation of new links can be given by the comparison between the arrival of the new edge and the formation of the new triangle. In fact by comparing the link and the triangle time-series in Figure 6.4, we observe the same trend; an increase/decrease in the new link volume corresponds to an increase/decrease in the number of triads, the ratio between the variables. In order to quantify this relation, in the same figure we plot the ratio between the triangle and the link time-series. Thus we can quantify, on average, the number of triangles closed by a link. Obviously, the mean value does not account for the per-link fluctuation, although it gives an idea of the role played by the triadic closure process. Despite the above limitations, by analyzing the link/triangle ratio, a surprising result emerges from the triadic closure counting on the Facebook network. As evident in Fig.6.4, the network shows an abrupt transition after March 26, 2008. This date corresponds to the introduction of "People You May Know" (PYMK) Facebook functionality, which promptly impacts both network and triangles. In fact, prior to the launch of PYMK, the triangles/links ratio rapidly increases and stabilizes at the greater value of 3-4 triangles/link. In this case we note how the PYMK mechanism highly impacts the microscopic characteristic of the network structure. In particular, it supervises the link positioning in such a way as to highly prefer the triadic closure. This strong effect cannot be captured by analyzing only the link creation over time. In fact we observe only a medium increase in the new link volume as shown in Fig.6.4. This observation stresses the importance of adopting different indexes in describing the network evolution; in fact, the number of new links alone would not be enough to let the phase transition

Algorithm 1 Temporal Triangle Extraction

```
1: procedure EXTRACTION(𝔾)
2:     Δ = ∅
3:     for all u ∈ 𝕍 do
4:         for all v ∈ Γ_T(u) do
5:             t_1 = τ(u, v)
6:             for all z ∈ Γ_T(v) do
7:                 t_2 = τ(v, z)
8:                 if t_2 > t1 then
9:                     for all y ∈ Γ_T(z) do
10:                        t_3 = τ(z, y)
11:                        if y = u ∧ t_3 > t_2 then
12:                            Δ.add(uvz)
13:                        end if
14:                    end for
15:                end if
16:            end for
17:        end for
18:    end for
19: end procedure
```

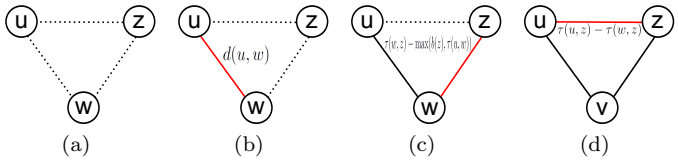


Fig. 6.5 Steps in the formation of the triangle and the definition of its constitutive elements. In 6.5(a) the potential triangle (dotted links) which will form at the end of triadic closure. In 6.5(b) the link (u, w) (red) establishes in $t_{\Delta}(u, w)$ unit time. In 6.5(c) the second link (w, z) forms in $t_{\Delta}(w, z)$ time and finally in 6.5(d) the last link takes $t_{\Delta}(z, u)$ to be created and the process ends.

emerge in the triadic closure process. We see how events external to the network topology can highly influence its dynamical properties. In our case the event is external to the network but internal to the service. This observation implies that the introduction of new features on the service could rapidly and massively modify the structure it manages, in a sort of feedback effect.

We have shown how the triadic closure process is a fundamental mechanism in the growth of online social networks and how it impacts their evolution. Nevertheless, we only consider the result of the process, i.e the triangle which has already been formed, and make no mention of how it got formed. The question which promptly comes to mind is how long we have to wait before observing the triadic closure effects, i.e. how long a triangle takes to be established.

6.5.2 Triadic Closure Delay

The triadic closure process has never been deeply analyzed in temporal networks evolution despite the aforementioned relevance in shaping the structure. The total amount of time a triangle takes to be closed and the temporal relation among the constitutive links still await investigation.

6.5.2.1 Definitions

In analogy with the physical definition of speed, we need to define the respective concepts of time and space as applied to the triangle case. Quite trivially, we can observe that the number of edges of the triangle (3) plays the role of physical space. Ergo, space represents a constant. Consequently, triadic closure delay is linearly proportional to time. So, what we really need is the definition of time necessary for the formation of a triangle.

The time definition is tricker since we are considering a dynamic process where the components could appear at different times. For example, from figure 6.1 we can see that the creation of the red edge depends on the presence of nodes u and v and that this same fact applies to the remaining links. To take into account these arguments, we employ the *birth date* $b(u)$, which denotes the time of the first appearance of node u into the networks.

Once the birth date has been defined, we can normalize the triangle creation time swallowing the temporal gap of node appearance, thereby capturing the real feasibility of a triangle. To attain a global definition of triadic closure time, we focus on the definition of its constitutive elements. We indicate the normalized time of the link (u, w) in a triangle $\Delta = u\hat{w}z$ as $t_\Delta(u, w)$.

To give a general definition of the triangle delay, we follow the steps that characterize the triadic closure process shown in Figure 6.5. In 6.5(a) we show the potential triangle with no links among nodes. The first element to be created is the red link (u, w) in Figure 6.5(b) and we have to measure how long it takes to be established. It corresponds to the delay of the link (u, w), so

$$t_\Delta(u, w) = d(u, w)$$

.

The triadic process is still in node w as the link (w, z) has not been created yet. Two possible situations could arise just before the creation of the red link in Figure 6.5(c): 1) node z is already in the network, so $\tau(u, w) > b(z)$ and $t_\Delta(w, z) = \tau(w, z) - \tau(u, w)$; 2) node z is absent, so the closure has to wait for its appearance. In the latter case we have $b(z) > \tau(u, w)$, so we discount the waiting time of the process in the node w, $\tau(u, w) - b(z)$, obtaining $t_\Delta(w, z) = \tau(w, z) - b(z)$. Putting together the conditions we obtain the general definition for $t_\Delta(w, z)$:

$$t_\Delta(w, z) = \tau(w, z) - \max(b(z), \tau(u, w))$$

The last step involves the creation of the link w, z as depicted in Figure 6.5(d). By definition of b and the ordering of the time values of the links, at the creation of the link (w, z), nodes w and z are already participating, so

$$t_\Delta(u, z) = \tau(u, z) - \tau(w, z)$$

.

Once the delay of each constitutive element has been defined, we can introduce the definition of the triadic closure delay.

Definition 6.3. Let $\mathbb{G}$ be a temporal undirected graph and $u\hat{w}z$ a temporal admissible triangle, i.e. $\tau((u, w)) < \tau((w, z)) < \tau(z, u))$, the *triadic closure delay* of $u\hat{w}z$, $d(u\hat{w}z)$ is defined as

$$d(u\hat{w}z) = d(u, w) - \max(b(z), \tau(u, w)) + \tau(u, z)$$

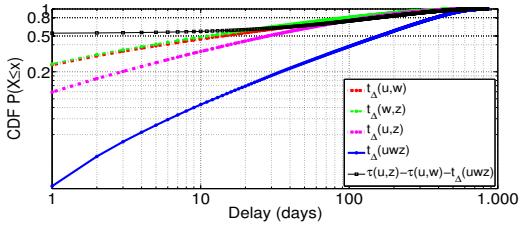


Fig. 6.6 CCDF of the triadic closure delay and its constitutive elements measured on Facebook. The black CDF represents the effect of the normalization of the triadic delay definition w.r.t. the simple definition that does not consider the node arrival process.

From the above definition we must observe that the triadic closure delay does not depend on the creation time of the middle link (w, z), while the normalization given by the birth date is quite important since it involves all the nodes. In Fig.6.6 we quantify the effects of the normalization through the distribution of $\tau(u, z) - \tau(u, w) - d(u\hat{w}z)$, that represents the difference between the triadic delay and the delay not normalized, *i.e.* $\tau(u, z) - \tau(u, w)$. We can observe that the normalization impacts on 50% of triangles, in particular 40% of triads are involved in a delay normalization of more than a month.

6.5.2.2 Triadic closure delay properties

We analyze the triadic closure delay and the t_Δ of each link in a triangle. The temporal information not only allows us to measure the triangle delay but also to temporally place them. This enabled us to verify weather or not the fast triangle trend is stable during the network growth and to see if external mechanisms, e.g. Facebook PYMK, have modified this trend. As for the properties of triadic closure delay, in Fig.6.6 we report the triadic closure delay CDF for Facebook. We observe that most triangles have a high speed growth, given that half of the triangles close in five months at most. This fact stresses the importance of the study of triangle formation dynamics. In effect, triadic closure impacts the network structure both significantly and suddenly. The Facebook delay behaviour shows some features. First, $t_\Delta(u, w)$ and $t_\Delta(w, z)$ have the same distribution and secondly, they slightly differ from the $t_\Delta(u, z)$ distribution.

To obtain an insight into the triadic closure delay, we wonder if by knowing only a subset of the constitutive element delays we can infer the general delay of the triangle, i.e. all the element are necessary to predict the triadic closure delay. This corresponds to verifying weather or not certain relationships occur among the different elements. For example, if $t_\Delta(u, w)$ and $t_\Delta(w, z)$ are fast, what can we say about the delay of (u, z). Will it be fast too? To stress possible relationships among the elements we adopt two approaches. First we

randomize the t_Δs to delete any relation between the triangle elements, then we compare the resulting delay distribution with the real one. The randomization is obtained by shuffling the elements in each column of the matrix $T_\Delta = [t_\Delta(u,w), t_\Delta(z,w), t_\Delta(u,z)]$. As reported in Fig.6.7 the delay and the shuffled delay distributions are quite similar. This observation suggests a lack of a particular relation between the delay elements of a triangle. Furthermore we confirm the above result by computing the correlation matrix among $t_\Delta(u,w)$,$t_\Delta(z,w)$ and $t_\Delta(u,z)$. More specifically we find correlation coefficients close to 0 (from 0.03 to 0.06) for each pair of variables.

Generally, these observations suggest that the single delay of the constitutive elements are not sufficient for explaining the total delay of triangles. In other words, the triadic closure delay cannot be predicted by simply observing a single element.

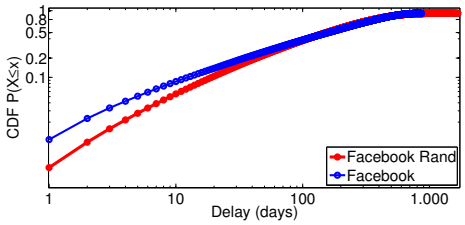
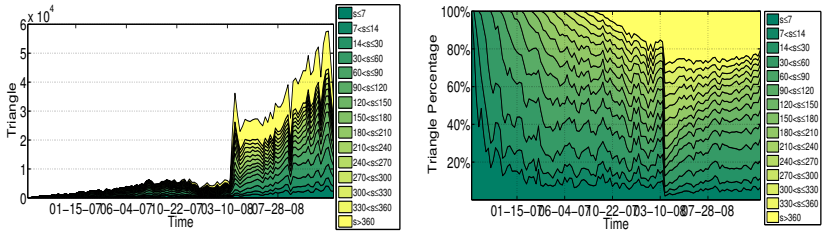


Fig. 6.7 Comparison between the triadic closure delay CDFs in real dataset and the relative randomization.

Delay dynamics in the network growth. As shown in Figure 6.4 the triangle formation trend is not regular, although, as in the link delay, we are to able to capture what kind of triangles (low or high delay) contributes to the observed irregularity. To deepen into the temporal mechanisms in Figure 6.8 we shows the different contributions of the delay classes of triangles in the growth of the network. In the left figure we divide the new triangles created in each week into 15 classes according to their delay. In the right figure we maintain the same classes but we normalize the contribution of each class with respect to total number of triangles formed during the week. This way we quantify the absolute and the relative contribution of each class to the triangle formation dynamics. For example a class could undergo a rapid increase (absolute volume) but have an overall low impact (relative) just because a general boost of formation activity of the triads. As for the absolute volume, we observe that the eight first delay classes keep slowly increasing. That accounts for a component of fast triangles which is independent from the stage of the network and that involves a quite similar number of triangles. The other classes, characterized by a higher delay, manifest after the PYMK service and maintain quite constant till the end of the sampling period. As concerns the relative volume trends, we observe that the PYMK service primarily acts on the 'old' triangles. This implies that the suggestion mechanism,

merely based on the common-neighbors algorithm for friend recommendation, awakes latent links, remained asleep until the service introduction. Moreover we can also quantity the long period effects of the mechanism. Specifically, we observe that after a brief period from the PYMK introduction, the relative volumes stabilize, with the exception of the [30-90] classes which slightly increase. In general by analyzing the triadic closure delay during the temporal evolution of Facebook we were able to quantify the effects and the impact of the PYMK feature.



(a) Facebook New Orleans (b) Facebook New Orleans

Fig. 6.8 6.8(a): The number of triangle created in each week and divided in different delay classes. 6.8(b): The volume trends (percentage) of the triangle delay classes during the growth of Facebook. Class green intensity is inversely proportional to triangle delay (upper and lower bound of the classes are in days).

6.6 Conclusion

In this chapter we have shown the temporal dynamic of an online social network, Facebook, adopting a microscopical approach based on the physical time. In this setting we provide the definition of two new metrics: the link delay and the triangle delay. These quantities allow us to measure how quickly a possible link between two nodes establish and how fast the triangle formation is, respectively. In general we find that a slower component combines with a fast one in a non-stationary way. In particular, by correlating the non-stationary periods with events external to the network topologies, we are able to measure the effects of the last ones on the evolution of the social media. The most representative results is the introduction of the friend suggestion mechanism on Facebook, which establishes many latent links and close a great amount of triangles. These results hihglight that the study of the growth of online social network can not disregard the other surrounding systems as underlined by the contagious network paradigm. Preliminary results on an important Chinese online social network not only confirm but also stress the above features discussed for Facebook.

Our work could be extended and deepen in various directions. The interesting results about the correlation between the volume of interactions and the link delay indicates that this quantity might represent a fundamental feature for the interaction prediction. In fact, following the method proposed by Steurer and Trattner [147], interactions between users in online social networks can be predicted by exploiting features obtained from the static topology and temporal data. Also link prediction can benefit from the temporal characteristic. More precisely we can combine the classical topological features like the common neighbours with the temporal information as the delay of the links connecting common neighbours. This advocates for a factor analysis on the features which determine the triangle delay or machine learning techniques able to predict if two hops nearby nodes will quickly close the triangle. As for the link prediction problem [95] we might evaluate the accuracy and recall performance of the various algorithms taking into account the non-stationary periods. In fact the experimental evaluation of the performances assume during the time period interleaving the snapshot on which we measure the features and the snapshot where we evaluate the performances, the mechanisms driving the network growth do not change. As concerns the modelling we plan to add the temporal information about the triadic closure process to the model proposed by Shin *et al.* [144]. Finally, if the datasets will be made publicly available , we plan to extend our analysis to other online social media as Google+ or to other information networks as Wikipedia.

Chapter 7
Conclusion

In this thesis we have addressed the close interplay among mobility, offline relationships and online interactions and the related human networks at different dimensional scales and temporal granularities. By generally adopting a data-driven approach we move from small datasets about physical interactions mediated by human-carried devices, describing small social realities, to large-scale graphs that evolve over time (*e.g.* Renren and Facebook), as well as from human mobility trajectories to face-to-face contacts occurring in different geographical contexts. Recalling the main topics presented in the introduction, the main contributions of the works presented in this thesis can be categorized as follows:

Human Mobility and Social Interaction: Patterns and Models. We explored in depth the relation between human mobility and the social structure induced by the overlapping of different people's trajectories on publicly available GPS traces collected in urban and metropolitan areas. The main questions driving our approach are: *What is the network structure of the contacts induced by the human mobility? What are their characteristics? Can we synthetically generate such kind of human networks?* In order to answer to these questions we have defined the notion of geo-location and geo-community which are operational in describing in a unique framework both spatial and social aspects of human behavior. The first step in our approach has been the development of a time efficient procedure for their extraction from GPS traces. The most immediate results concern the statistical properties of the human movements at the finest temporal and spatial granularities. The obtained results on mobility features considering both geo-locations and the geo-communities viewpoint confirm most of the observations reported in literature. In addition to the classic properties we investigated how distance influences the choice of the next destination and we introduced a classification of the geo-locations which results in a categorization of the users in two mobility profile. The contact network obtained from the superposition of many human mobility traces, show the classical properties of human network as an high modular structure and small-world properties. In fact the shared social

physical foci, *i.e.* geo-communities, result in group of interconnected people while mobility favors the creation of ties among the different groups, letting small-worldness appear. Through the concept of geo-community we can answer to the previous questions by modeling the human mobility adopting a bipartite graph. Thanks to this graph representation we can generate a social structure that is plausible w.r.t. the real interactions, by adopting some graph projections based on classical random graph models. The obtained random projections have been evaluated by comparing the typical features of complex network analysis with real-world ones. The evaluation process results in a parameter estimation method which maximizes the network similarity between the real and the synthetic networks. In general the modeling approach has the merit for reporting the mobility in a graph-theoretic framework making the study of the interplay mobility/sociality more affordable and intuitive.

Our modeling approach also results in a mobility model, Geo-CoMM, which lies on and exploits the idea of geo-community. Essentially the model follows the location-based paradigm, where users are assigned to a subset of places. The users assignment essentially relies on an affiliation network (bipartite graph) on top of which we run a finite time homogeneous Markov chain where the states are the geo-communities linked to the moving node and where the transition probability is a function of the rank on their distances. The Markov chain drives the movement among geo-communities, while within a geo-community we adopt a particular variety of non-uniform random model similar to a Lévy walk. The model represents a particular instance of a general framework we provide. A framework where the social structure behind the preferred-location based mobility models emerges. The framework provides a temporal extension of the user/geo-community affiliation network, in which the users' mobility impacts the activation of the connection user/geo-community. Reporting Geo-CoMM in this framework we can highlight its social characteristics, in fact the social structure emerges from the model without imposing any social graph or social overlay. Finally we validate Geo-CoMM on spatial, temporal and pairwise connectivity features showing that it reproduces the main statistical properties observed in real traces. We also focus on the social graph induced by the contacts. In particular we provide the characteristics of contact graph of some traces and we show that the model can properly reproduce them.

Our work could be extended in many ways. As concerns geo-community extraction we could implement different density-based clustering algorithms. The best choice should be hierarchical algorithm as OPTICS [6] or HDB-SCAN [30], as to reduce the threshold problems which affect the classical density-based algorithms and to simplify the temporal and spatial scale analysis. Moreover the choice mechanism could be further deepened by introducing semantic labels to the extracted points of interest. This way we can quantify when and how social, work and agenda contexts heavily act on the choice of the next destination. As regards the human mobility modeling, the time-varying bipartite graph approach seems to be promising. A further extension

can be represented by the introduction of synchronization mechanisms which aim at reproducing the periodicity patterns. Usually the synchronization has been introduced by ad-hoc activation functions applied on the nodes which correspond to the physical places, or by an additional context layer. These techniques further increase the complexity of the model. A better approach should let periodicity emerge as a consequence of the synchronization among nodes. Finally the main problem in both fields is the 'quest' of mobility data. In the light of some results about LBSN data [166], the collection of GPS data is now the best solution.

Multidimensional Networks for Offline/Online Social Interaction. The work in Chapter 4 represents a first effort to provide a complete overview of the close connection between online and offline sociality. The understanding of the interplay between the aforementioned levels have not only social implications, *e.g.* how people interact, who the closest friends are, how they exploit different media to maintain or establish new relationships; but also could be exploited in the organization of the enterprises or in the business process. The completeness of the gathered dataset, albeit limited to a small scenario, and the proposed multidimensional complex network allowed us to deeply understand how the characteristics of users in the distinct networks impact each other. The first and main questions to be answered are: *Do offline and online social networks capture the same interactions and relationships? Do the role and the importance of a user maintain among the dimensions?* Answering to these questions is fundamental. A positive answer should mean that we could use Facebook or other online social services to monitor our sociality expressed externally to these services. On the contrary, our work shows how offline and Facebook friends are different. This way we confirm and worsen the general intuition that online social networks have shifted away from their original goal to mirror the offline sociality of individuals. As for the role and the social importance, it becomes apparent that social features such as user popularity or community structure do not transfer along social dimensions, as confirmed by our correlation analysis of the network layers and by the comparison among the communities. In the light of the previous results we ask whether offline information could modify the online connectivity. We show that by introducing external information the users' role and their groups may change. In general, together with Magnani *et al.* [101], this work represents the first effort in understanding the relation between offline and online relationships reporting them to multidimensional network framework.

The study of the interplay between offline and online social networks is still in its infancy, mainly due to few datasets gathering offline/online data. This way an extension of this work is direct towards an intensive and widely data collection. A mobile phone application seems to be the right solution as the widespread diffusion of mobile devices. To this aim the Funf [1] platform, released by the MediaLab at MIT, represent the most mature solution. On

[1] www.funf.org

the other side, as reported by Magnani e Rossi [100] online social dimensions are different one another. For instance Twitter is more an information and news media than a social network [88]. For these reasons the understanding of their superposition plays a fundamental role in the diffusion of contents. Also in this scenario mobile phone application could be exploited to gather social network data, as mobile phone are becoming the main interface towards the social services.

Growth and Temporal Evolution in Online Social Networks. Online social networks (OSNs) have been extensively studied in the last decade, with most efforts focusing on static properties computed on single snapshots of the network. Gradually, attention of the community has shifted towards temporal properties of OSNs, with the goals of understanding patterns and mechanisms underlying their growth. However, most temporal studies are limited to analyses of dynamics using logical clocks [92], despite the growing recognition that a full understanding of graph dynamics requires analyses using physical clocks. In Chapter 5 and 6 we analyze the microscopical characteristic from a physical time perspective, *i.e.* considering the graph evolution as a graph time-series and not as a function of the network basic properties (number of nodes or links). As for the physical time in a user-centric viewpoint, we investigated the bursty nature of the link creation process in online social network. We prove not only that it is a highly inhomogeneous process, but also identify patterns of burstiness common to all nodes. In terms of edge creation, users are inactive for most of their lifetimes, and concentrate their link activity in a number of short regular time periods. These observation have been confirmed in successive work by Kikas *et al.* [84] on the Skype social network. With respect to that work based on time-series clustering, we are also able to characterize node activity.In fact we developed a new methodology based on the acceleration of degree growth, which allows us to highlight the internal structure of link creation bursts.

We believe using acceleration as a general metric to characterize network dynamics prompts future work in studying link generation mechanisms. In particular, defining different phases of edge creation hints at the possibility of characterizing users into distinctive activity levels that correlate with their likelihood of adding social links. Some preliminary results confirm this intuition: when nodes (users) first join the network, they create links based on the preferential attachment mechanism; while in later bursts, nodes seem to explore (acceleration phase) and densify (deceleration) in far regions of the graph that could correlate to spatial areas. These results open the door for new generative models [140] that consider different phases of the node activity.

In Chapter 6 we focus on the dynamic formation of two fundamental network building components: *dyads* and *triads*. We propose two new metrics to aid the temporal analysis on physical time: link creation delay and triangle closure delay. *Link delay* is the time period required before two users of a network create a link between them, *i.e.* the delay between when a friendship

is possible and when a friendship link actually forms. On the other hand, triangle closings capture the transitivity of friendships, which has proven to be effective in modeling network evolution and link prediction. *Triadic closure delay* captures the time gap between the formation of two connected edges (an open triangle) and the closure of the triangle with the third and final edge. These two metrics enable us to study the dynamic creation of dyads and triads, and to highlight network behavior that would otherwise remain hidden. In our analysis, we find that link delays are generally very low in absolute time, meaning if two people want to become OSN friends, they do so very shortly after both have joined the network. Link establishment is especially fast in early stages of social networks. In addition, link delay results are largely independent of the dates people join the network. To highlight the social nature of this metric, we introduce the term *peerness* to quantify how well linked users overlap in lifetimes. Finally, we study if link delays correlate to distance in the social network, and find that links spanning more distant nodes generally form faster. Our *triadic closure delay* takes into account how long a temporal triangle takes to form. We first introduce an algorithm to extract of temporal triangle which enables us to monitor the triangle formation process, and to detect sudden changes in the triangle formation behavior, possibly related to external events. In particular, we show that the introduction of "People You May Know" Facebook functionality had a disruptive impact on the triangle creation process in the network.

Our work could be extended and deepen in various directions. The results about the correlation between interactions and the link delay could be exploited in the interaction prediction task. Also link prediction can benefit from the temporal characteristics. More precisely we can combine the classical topological features with the introduced temporal information. We plan to make a factor analysis on the features which determine the triangle delay or adopt machine learning techniques able to predict if two hops nearby nodes will quickly close the triangle. As for the link prediction problem [95] we might evaluate the accuracy and recall performance of the various algorithms taking into account the non-stationary setting, a field in machine learning theory. As concerns the modeling we plan to add the temporal information about the triadic closure process to the model proposed by Shin *et al.* [144].

References

1. Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25:211–230, 2001.
2. Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive Mobile Computing*, 7(6):643–659, 2011.
3. Hirotsugu Akaike. Information measure and model selection. *Bulletin of the International Statistical Institute*, 1983.
4. Leman Akoglu and Christos Faloutsos. Rtg: a recursive realistic graph generator using random typing. *Data Mininig Knowledge Discovery*, 19(2):194–209, 2009.
5. Frédéric Amblard, Arnaud Casteigts, Paola Flocchini, Walter Quattrociocchi, and Nicola Santoro. On the temporal analysis of scientific network evolution. In *Proceedings of the International Conference on Computational Aspects of Social Networks*, CASoN '11. IEEE, 2011.
6. Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: ordering points to identify the clustering structure. *ACM SIGMOD Record*, 28(2):49–60, 1999.
7. Nils Aschenbruck, Aarti Munjal, and Tracy Camp. Trace-based mobility modeling for multi-hop wireless networks. *Computer Communications*, 34(6):704–714, 2011.
8. Daniel Ashbrook and Thad Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal Ubiquitous Computing*, 7(5):275–286, 2003.
9. Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06. ACM, 2006.
10. Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.
11. Albert-Lszl Barabsi and Rka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
12. Alain Barrat, Ciro Cattuto, Martin Szomszor, Wouter Van den Broeck, and Harith Alani. Social dynamics in conferences: Analysis of data from the live social semantics application. In *Proceedings of the 9th International Semantic Web Conference*, ISWC '10. IEEE/ACM, 2010.
13. Michele Berlingerio, Francesco Bonchi, Björn Bringmann, and Aristides Gionis. Mining graph evolution rules. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, ECML-PKDD '09, 2009.

14. Michele Berlingerio, Michele Coscia, and Fosca Giannotti. Finding redundant and complementary communities in multidimensional networks. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11. ACM, 2011.
15. Michele Berlingerio, Michele Coscia, Fosca Giannotti, Anna Monreale, and Dino Pedreschi. As time goes by: discovering eras in evolving social networks. In *Proceedings of the 14th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining*, PAKDD '10. ACM, 2010.
16. Michele Berlingerio, Michele Coscia, Fosca Giannotti, Anna Monreale, and Dino Pedreschi. Foundations of multidimensional network analysis. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '11. IEEE/ACM, 2011.
17. Michele Berlingerio, Michele Coscia, Fosca Giannotti, Anna Monreale, and Dino Pedreschi. Multidimensional networks: foundations of structural analysis. *World Wide Web*, pages 1–27, 2012.
18. Michele Berlingerio, Fabio Pinelli, and Francesco Calabrese. Abacus: frequent pattern mining-based community discovery in multidimensional networks. *Data Mining Knowledge Discovery*, 27(3):294–320, 2013.
19. Greg Bigwood, Devan Rehunathan, Martin Bateman, Tristan Henderson, and Saleem Bhatti. Exploiting self-reported social networks for routing in ubiquitous computing environments. In *Proceedings of the 2008 IEEE International Conference on Wireless & Mobile Computing, Networking & Communication*, WIMOB '08. IEEE, 2008.
20. Vincent D. Blondel, Anahí Gajardo, Maureen Heymans, Pierre Senellart, and Paul Van Dooren. A measure of similarity between graph vertices: Applications to synonym extraction and web searching. *SIAM Review*, 46(4):647–666, 2004.
21. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), 2008.
22. Chiara Boldrini, Marco Conti, and Andrea Passarella. The sociable traveller: human travelling patterns in social-based mobility. In *Proceedings of the 7th ACM International Symposium on Mobility Management and Wireless Access*, MobiWAC '09. ACM, 2009.
23. Phillip Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, pages 1170–1182, 1987.
24. Vincent Borrel, Franck Legendre, Marcelo Dias De Amorim, and Serge Fdida. Simps: using sociology for personal mobility. *IEEE/ACM Transaction on Networking*, 17(3):831–842, 2009.
25. Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.
26. Piotr Bródka, Przemysław Kazienko, Katarzyna Musiał, and Krzysztof Skibicki. Analysis of neighbourhoods in multi-layered dynamic social networks. *International Journal of Computational Intelligence Systems*, 5(3):582–596, 2012.
27. Piotr Bródka, Pawel Stawiak, and Przemyslaw Kazienko. Shortest path discovery in the multi-layered social network. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '11. IEEE/ACM, 2011.
28. Kenneth P. Burnham and David R. Anderson. Multimodel inference. *Sociological Methods & Research*, 2004.
29. Tracy Camp, Jeff Boleng, and Vanessa Davies. A survey of mobility models for ad hoc network research. *Wireless Communications and Mobile Computing*, 2(5):483–502, 2002.
30. RicardoJ.G.B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, VincentS. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery*

and Data Mining, volume 7819 of *Lecture Notes in Computer Science*, pages 160–172. Springer Berlin Heidelberg, 2013.

31. Augustin Chaintreau, Pan Hui, Jon Crowcroft, Christophe Diot, Richard Gass, and James Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, 6(6):606–620, 2007.
32. M. Chiappalone, A. Novellino, I. Vajda, A. Vato, S. Martinoia, and J. van Pelt. Burst detection algorithms for the analysis of spatio-temporal patterns in cortical networks of neurons. *Neurocomputing*, 65-66:653–662, 2005.
33. Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11. ACM, ACM, 2011.
34. Hyunwoo Chun, Haewoon Kwak, Young-Ho Eom, Yong-Yeol Ahn, Sue Moon, and Hawoong Jeong. Comparison of online social relations in volume vs interaction: a case study of cyworld. In *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement*, IMC '08. ACM, 2008.
35. Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6), 2004.
36. Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4), 2009.
37. Marco Conti, Silvia Giordano, Martin May, and Andrea Passarella. From opportunistic networks to opportunistic computing. *IEEE Communications Magazine*, 48(9):126–139, 2010.
38. David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.
39. Mark E. Crovella and Azer Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transaction on Networking*, 5(6):835–846, 1997.
40. Elizabeth M. Daly and Mads Haahr. Social network analysis for routing in disconnected delay-tolerant manets. In *Proceedings of the 8th ACM International Symposium on Mobile ad hoc Networking and Computing*, MobiHoc '07. ACM, 2007.
41. Easley David and Kleinberg Jon. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.
42. Sergei N. Dorogovtsev and Jose F.F. Mendes. Accelerated growth of networks. *Handbook of Graphs and Networks*, page 318, 2003.
43. Nathan Eagle, Alex Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of The National Academy of Sciences*, 106, 2009.
44. Nathan Eagle and Alex (Sandy) Pentland. Reality mining: sensing complex social systems. *Personal Ubiquitous Computing*, 10(4):255–268, 2006.
45. Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, KDD '96. ACM, 1996.
46. Anna Förster, Kamini Garg, Hoang Anh Nguyen, and Silvia Giordano. On context awareness and social distance in human mobility traces. In *Proceedings of the Third ACM International Workshop on Mobile Opportunistic Networks*, MobiOpp '12. ACM, 2012.
47. Santo Fortunato. Community detection in graphs. *Physics Reports*, 486:75 – 174, 2010.
48. Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

49. Sabrina Gaito, Elena Pagani, and Gian Paolo Rossi. Opportunistic forwarding in workplaces. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, WOSN '09. ACM, 2009.
50. Sabrina Gaito, Elena Pagani, and Gian Paolo Rossi. Strangers help friends to communicate in opportunistic networks. *Computer Network*, 55(2):374–385, 2011.
51. Sabrina Gaito, Matteo Zignani, Gian Paolo Rossi, Alessandra Sala, Xiaohan Zhao, Haitao Zheng, and Ben Y. Zhao. On the bursty evolution of online social networks. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, HotSocial '12. ACM, 2012.
52. Wei Gao, Qinghua Li, Bo Zhao, and Guohong Cao. Multicasting in delay tolerant networks: a social network perspective. In *Proceedings of the 10th ACM Interational Symposium on Mobile Ad Hoc Networking and Computing*, MobiHoc'09. ACM, 2009.
53. Sanchit Garg, Trinabh Gupta, Niklas Carlsson, and Anirban Mahanti. Evolution of an online social aggregation network: an empirical study. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, IMC '09. ACM, 2009.
54. Daniel Gayo-Avello, Panagiotis T. Metaxas, and Eni Mustafaraj. Limits of Electoral Predictions Using Twitter. In *International AAAI Conference on Weblogs and Social Media*, SocialCom '11, 2011.
55. Yong Ge, Hui Xiong, Zhi-hua Zhou, Hasan Ozdemir, Jannite Yu, and K. C. Lee. Top-eye: top-k evolving trajectory outlier detection. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10. ACM, 2010.
56. Joy Ghosh, Sumesh J. Philip, and Chunming Qiao. Sociological orbit aware location approximation and routing (solar) in manet. *Ad Hoc Network*, 5(2):189–209, 2007.
57. Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07. ACM, 2007.
58. Sergio Gómez, Albert Díaz-Guilera, Jesús Gómez-Gardeñes, Conrad J Perez-Vicente, Yamir Moreno, and Alex Arenas. Diffusion dynamics on multiplex networks. *Physical Review Letters*, 110(2):028701, 2013.
59. Neil Zhenqiang Gong, Wenchang Xu, Ling Huang, Prateek Mittal, Emil Stefanov, Vyas Sekar, and Dawn Song. Evolution of social-attribute networks: measurements, modeling, and implications using google+. In *Proceedings of the 2012 ACM Conference on Internet Measurement*, IMC'12. ACM, 2012.
60. Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 2008.
61. Roberto Gonzalez, Ruben Cuevas, Reza Motamedi, Reza Rejaie, and Angel Cuevas. Google+ or google-?: dissecting the evolution of the new osn in its first year. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13. ACM, 2013.
62. Mark Granovetter. The strength of weak ties: A network theory revisited. *Sociological Theory*, 1(1):201–233, 1983.
63. Junjun Hao, Shuiming Cai, Qinbin He, and Zengrong Liu. The interaction between multiplex community networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(1):016104–016104, 2011.
64. Ramaswamy Hariharan and Kentaro Toyama. Project lachesis: Parsing and modeling location histories. In *Geographic Information Science*, volume 3234 of *Lecture Notes in Computer Science*, pages 106–124. Springer Berlin Heidelberg, 2004.
65. Tristan Henderson, David Kotz, and Ilya Abyzov. The changing usage of a mature campus-wide wireless network. In *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking*, MobiCom '04. ACM, 2004.
66. Theus Hossmann, Theus Hossmann, George Nomikos, George Nomikos, Thrasyvoulos Spyropoulos, Thrasyvoulos Spyropoulos, Franck Legendre, and Franck Legendre. Collection and analysis of multi-dimensional network data for opportunistic networking research. *Computer Communications*, 35(13):1613–1625, 2012.

67. Theus Hossmann, Franck Legendre, George Nomikos, and Thrasyvoulos Spyropoulos. Stumbl: Using facebook to collect rich datasets for opportunistic networking research. In *Proceedings of the Fifth IEEE WoWMoM Workshop on Autonomic and Opportunistic Communications*, AOC '11. IEEE, 2011.
68. Theus Hossmann, Thrasyvoulos Spyropoulos, and Franck Legendre. A complex network analysis of human mobility. In *Third International Workshop on Network Science for Communication Networks*, NetSciCom'11. IEEE, 2011.
69. Theus Hossmann, Thrasyvoulos Spyropoulos, and Franck Legendre. Putting contacts into context: mobility modeling beyond inter-contact times. In *Proceedings of the Twelfth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, MobiHoc '11, 2011.
70. W. Hsu and A. Helmy. IMPACT: Investigation of Mobile-user Patterns Across University Campuses using WLAN Trace Analysis. *Arxiv preprint cs.NI/0508009*, 2005.
71. Wei-Jen Hsu, Thrasyvoulos Spyropoulos, Konstantinos Psounis, and Ahmed Helmy. Modeling spatial and temporal dependencies of user mobility in wireless mobile networks. *IEEE/ACM Transaction on Networking*, 17(5):1564–1577, 2009.
72. Pan Hui, Augustin Chaintreau, James Scott, Richard Gass, Jon Crowcroft, and Christophe Diot. Pocket switched networks and human mobility in conference environments. In *Proceedings of the 2005 ACM SIGCOMM Workshop on Delay-tolerant Networking*, WDTN '05. ACM, 2005.
73. Pan Hui, Jon Crowcroft, and Eiko Yoneki. Bubble rap: social-based forwarding in delay tolerant networks. In *Proceedings of the 9th ACM Interational Symposium on Mobile Ad Hoc Networking and Computing*, MobiHoc'08. ACM, 2008.
74. B. Aban Inmaculada, Meerschaert Mark M., and Panorska Anna K. Parameter estimation for the truncated pareto distribution. *Journal of the American Statistical Assocation*, 2006.
75. Jing Jiang, Christo Wilson, Xiao Wang, Peng Huang, Wenpeng Sha, Yafei Dai, and Ben Y. Zhao. Understanding latent interactions in online social networks. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10. ACM, 2010.
76. Hang-Hyun Jo, Márton Karsai, János Kertész, and Kimmo Kaski. Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics*, 14(1):013055, 2012.
77. Hang-Hyun Jo, Raj Kumar Pan, and Kimmo Kaski. Emergence of bursts and communities in evolving weighted networks. *PloS one*, 6(8):e22687, 2011.
78. Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuos Univariate Distribution*. Wiley Series in Probability and Statistics, 1994.
79. Y. Kagan and F. Schoenberg. Estimation of the upper cutoff parameter for the tapered pareto distribution. *Journal of the Applied Probability*, 2001.
80. Jong Hee Kang, William Welbourne, Benjamin Stewart, and Gaetano Boriello. Extracting places from traces of locations. In *Proceedings of the 2nd ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots*, WMASH '04. ACM, 2004.
81. Thomas Karagiannis, Jean-Yves Le Boudec, and Milan Vojnović. Power law and exponential decay of inter contact times between mobile devices. In *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking*, MobiCom '07. ACM, 2007.
82. Marton Karsai, Kimmo Kaski, Albert-László Barabási, and János Kertész. Universal features of correlated bursty behaviour. *Scientific Reports*, 2, 2012.
83. Maurice Kendall. *Rank correlation methods*. Griffin, London, 1970.
84. Minkyong Kim and David Kotz. Extracting a mobility model from real user traces. In *Proceedings of the 25th Annual Joint Conference of the IEEE Computer and Communications Societies*, INFOCOM '06. IEEE, 2006.
85. Jon Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

86. Gueorgi Kossinets and Duncan J Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.
87. Hans-Peter Kriegel, Peer Krger, Jrg Sander, and Arthur Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240, 2011.
88. Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In ACM, editor, *Proceedings of the 19th international conference on World wide web*, WWW '10, 2010.
89. Kyunghan Lee, Seongik Hong, Seong Joon Kim, Injong Rhee, and Song Chong. Slaw: self-similar least-action human walk. *IEEE/ACM Transaction on Networking*, 20(2):515–529, 2012.
90. Vincent Lenders, Jörg Wagner, and Martin May. Measurements from an 802.11b mobile ad hoc network. In *Proceedings of the IEEE WoWMoM Workshop on Advanced Experimental Activities on Wireless Networks and Systems*, EXPONWIRELESS '06. IEEE, 2006.
91. Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09. ACM, 2009.
92. Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08. ACM, 2008.
93. Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05. ACM, 2005.
94. Wen-Hwa Liao, Jang-Ping Sheu, and Yu-Chee Tseng. Grid: A fully location-aware routing protocol for mobile ad hoc networks. *Telecommunication Systems*, 18(1-3):37–60, 2001.
95. David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03. ACM, 2003.
96. Miao Lin, Wen-Jing Hsu, and Zhuo Qi Lee. Predictability of individuals' mobility with high-resolution positioning data. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12. ACM, 2012.
97. Miao Lin, Wen-Jing Hsu, and Zhuo Qi Lee. Detecting modes of transport from unlabelled positioning sensor data. *Journal of Location Based Services*, 0(0):1–19, 2013.
98. Hao Ma, Dengyong Zhou, Chao Liu, Michael R. Lyu, and Irwin King. Recommender systems with social regularization. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, WSDM '11. ACM, 2011.
99. Anmol Madan, Manule Cebrian, Sai Moturu, Katayoun Farrahi, and Alex Pentland. Sensing the health state of a community. *Pervasive Computing, IEEE*, 11(4):36–45, 2012.
100. M. Magnani and L. Rossi. The ml-model for multi-layer social networks. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '11. IEEE/ACM, 2011.
101. Matteo Magnani, Barbora Micenkova, and Luca Rossi. Combinatorial analysis of multiple networks. *arXiv preprint arXiv:1303.4986*, 2013.
102. Matteo Magnani and Luca Rossi. Formation of multiple networks. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 257–264. Springer, 2013.
103. Matteo Magnani and Luca Rossi. Pareto distance for multi-layer network analysis. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 249–256. Springer, 2013.

104. Mary McGlohon, Leman Akoglu, and Christos Faloutsos. Weighted graphs and disconnected components: patterns and a generator. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08. ACM, 2008.
105. Mary McGlohon, Jure Leskovec, Christos Faloutsos, Matthew Hurst, and Natalie S. Glance. Finding patterns in blog shapes and blog evolution. In *Proceedings of the First International Conference on Weblogs and Social Media*, ICWSM '07. AAAI, 2007.
106. Marvin McNett and Geoffrey M. Voelker. Access and mobility of wireless pda users. *SIGMOBILE Mobile Computing Communication Review*, 9(2):40–55, April 2005.
107. Jose F.F. Mendes. Effect of accelerated growth on networks dynamics. In Romualdo Pastor-Satorras, Miguel Rubi, and Albert Diaz-Guilera, editors, *Statistical Mechanics of Complex Networks*, volume 625 of *Lecture Notes in Physics*, pages 88–113. Springer Berlin Heidelberg, 2003.
108. Alan Mislove, Hema Swetha Koppula, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Growth of the flickr social network. In *Proceedings of the First workshop on Online Social Networks*, WOSN '08. ACM, 2008.
109. Abderrahmen Mtibaa, Augustin Chaintreau, Jason LeBrun, Earl Oliver, Anna-Kaisa Pietilainen, and Christophe Diot. Are you moved by your social network application? In *Proceedings of the First Workshop on Online Social Networks*, WOSN '08. ACM, 2008.
110. Peter J. Mucha, Thomas Richardson, Kevin Macon, Mason A. Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.
111. Aarti Munjal, Tracy Camp, and Nils Aschenbruck. Changing trends in modeling mobility. *Journal of Electrical and Computer Engineering*, 2012, 2012.
112. Mirco Musolesi and Cecilia Mascolo. A community based mobility model for ad hoc network research. In *Proceedings of the 2nd international Workshop on Multi-hop ad hoc Networks*, REALMAN '06. ACM, 2006.
113. Seth A. Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12. ACM, 2012.
114. Anirudh Natarajan, Mehul Motani, and Vikram Srinivasan. Understanding urban interactions from bluetooth phone contact traces. In *Proceedings of the 8th international Conference on Passive and Active Network Measurement*, PAM '07, 2007.
115. Atif Nazir, Saqib Raza, Dhruv Gupta, Chen-Nee Chuah, and Balachander Krishnamurthy. Network level footprints of facebook applications. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, IMC '09. ACM, 2009.
116. M.E.J Newman. *Networks: An introduction*. Oxford Press, 2010.
117. Tom Nicolai, Eiko Yoneki, Nils Behrens, and Holger Kenn. Exploring social context with the wireless rope. In *On the Move to Meaningful Internet Systems*, OTM Workshops. Springer, 2006.
118. Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. A tale of many cities: Universal patterns in human urban mobility. *PLoS ONE*, 7, 05 2012.
119. Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, ICWSM '10. AAAI, 2010.
120. Jukka-Pekka Onnela, Samuel Arbesman, Marta C. Gonzlez, Albert-Lszl Barabsi, and Nicholas A. Christakis. Geographic constraints on social network groups. *PLoS ONE*, 6(4), 04 2011.
121. Gergely Palla, Albert-László Barabási, and Tamás Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

122. Gergely Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 2005.
123. Andr Panisson, Alain Barrat, Ciro Cattuto, Wouter Wouter Van den Broeck, Giancarlo Ruffo, and Rossano Schifanella. On the dynamics of human proximity for data diffusion in ad-hoc networks. *Ad Hoc Networks*, 10:1532–1543, 2012.
124. Michela Papandrea and Silvia Giordano. Location prediction and mobility modelling for enhanced localization solution. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–17, 2013.
125. Nish Parikh and Neel Sundaresan. Scalable and near real-time burst detection from ecommerce queries. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08. ACM, 2008.
126. Erds Paul and Rényi Alfred. On the evolution of random graphs. In *Publication of the Mathematical Institue of the Hungarian Acadamey of Sciences*, pages 17–61, 1960.
127. Alex Pentland. How big data can transform society for the better. *Scientific American Magazine*, 309(4), 2013.
128. Nicola Perra and Santo Fortunato. Spectral centrality measures in complex networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 78(3), 2008.
129. Anna-Kaisa Pietiläinen, Earl Oliver, Jason LeBrun, George Varghese, and Christophe Diot. Mobiclique: middleware for mobile social networking. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, WOSN '09. ACM, 2009.
130. Mikko Pitkänen, Teemu Kärkkäinen, Jörg Ott, Marco Conti, Andrea Passarella, Silvia Giordano, Daniele Puccinelli, Franck Legendre, Sacha Trifunovic, Karin Anna Hummel, Martin May, Nidhi Hegde, and Thrasyvoulos Spyropoulos. Scampi: service platform for social aware mobile and pervasive computing. *Computer Communication Review*, 42(4):503–508, 2012.
131. Ana-Maria Popescu and Marco Pennacchiotti. "dancing with the stars, " nba games, politics: An exploration of twitter users' response to events. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, ICWSM '11. AAAI, 2011.
132. Microsoft GeoLife Project.
133. Josep M. Pujol, Vijay Erramilli, Georgos Siganos, Xiaoyuan Yang, Nikos Laoutaris, Parminder Chhabra, and Pablo Rodriguez. The little engine(s) that could: scaling online social networks. *SIGCOMM Comput. Commun. Rev.*, 40(4):375–386, 2010.
134. Troy Raeder and NiteshV. Chawla. Market basket analysis with networks. *Social Network Analysis and Mining*, 1(2):97–113, 2011.
135. Anatol Rapoport. Spread of information through a population with socio-structural bias: I. assumption of transitivity. *The Bulletin of Mathematical Biophysics*, 15(4):523–533, 1953.
136. William J. Reed and Murray Jorgensen. The double pareto-lognormal distribution: A new parametric model for size distributions. *Communications in Statistics - Theory and Methods*, 2004.
137. CRAWDAD repository.
138. Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, Seong Joon Kim, and Song Chong. On the levy-walk nature of human mobility. *IEEE/ACM Transaction on Networking*, 19(3):630–643, 2011.
139. Giulio Rossetti, Michele Berlingerio, and Fosca Giannotti. Scalable link prediction on multidimensional networks. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ICDMW '11. IEEE, 2011.
140. Alessandra Sala, Lili Cao, Christo Wilson, Robert Zablit, Haitao Zheng, and Ben Y. Zhao. Measurement-calibrated graph models for social network experiments. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10. ACM, 2010.

141. Marcel Salath, Maria Kazandjieva, Jung Woo Lee, Philip Levis, Marcus W. Feldman, and James H. Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 2010.
142. Fabian Schneider, Anja Feldmann, Balachander Krishnamurthy, and Walter Willinger. Understanding online social network usage from a network perspective. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, IMC '09. ACM, 2009.
143. Grant Schoenebeck. Potential networks, contagious communities, and understanding social network structure. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13. ACM, 2013.
144. Xiaolin Shi, Lada A Adamic, and Martin J Strauss. Networks of strong ties. *Physica A: Statistical Mechanics and its Applications*, 378(1):33–47, 2007.
145. Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.
146. Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-Lszl Barabasi. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
147. Michael Steurer and Christoph Trattner. Predicting interactions in online social networks: an experiment in second life. In *Proceedings of the 4th International Workshop on Modeling Social Media*, MSM '13. ACM, 2013.
148. Jing Su, Alvin Chin, Anna Popivanova, Ashvin Goel, and Eyal de Lara. User mobility for opportunistic ad-hoc networking. In *Proceedings of the Sixth Workshop on Mobile Computing Systems and Applications*, WMCSA '04. IEEE, 2004.
149. Michael Szell, Renaud Lambiotte, and Stefan Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31), 2010.
150. Martin Szomszor, Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, and Harith Alani. Semantics, sensors, and the social web: The live social semantics experiments. In *The Semantic Web: Research and Applications*, volume 6089 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2010.
151. Duc A Tran, Khanh Nguyen, and Cuong Pham. S-clone: Socially-aware data replication for social networks. *Computer Networks*, 56(7):2001–2013, 2012.
152. Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*, ICWSM '10. AAAI, 2010.
153. Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.
154. Alexei Vázquez, João Gama Oliveira, Zoltán Dezsö, Kwang-Il Goh, Imre Kondor, and Albert-László Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):036127, 2006.
155. Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, WOSN '09. ACM, 2009.
156. Stanley Wasserman and Katrine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.
157. Duncan James Watts and Steven Henry Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):409–10, 1998.
158. Hsu Weijen, Dutta Debojyoti, and Helmy Ahmed. Profile-cast: Behavior-aware mobile networking. In *Proceedings of IEEE Wireless Communications and Networking Conference*, WCNC '08. IEEE, 2008.
159. Christo Wilson, Alessandra Sala, Krishna P. N. Puttaswamy, and Ben Y. Zhao. Beyond social graphs: User interactions in online social networks and their implications. *ACM Transactions on the Web*, 6(4), 2012.

160. Jierui Xie, Stephen Kelley, and Boleslaw K. Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Survey*, 45(4):1–43, 2013.
161. Osman Yağan and Virgil Gligor. Analysis of complex contagions in random multiplex networks. *Physical Review E*, 86(3):036103, 2012.
162. Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12. ACM, 2012.
163. Shengqi Yang, Xifeng Yan, Bo Zong, and Arijit Khan. Towards effective partition management for large graphs. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGCOMM '12. ACM, 2012.
164. Shusen Yang, Xinyu Yang, Chao Zhang, and Evangelos Spyrou. Using social network theory for modeling human mobility. *IEEE Network: The Magazine of Global Internetworking*, 24(5):6–13, 2010.
165. Giovanni Zappella, Alexandros Karatzoglou, and Linas Baltrunas. Games of friends: A game-theoretical approach for link prediction in online social networks. In *Proceedings of the AAAI 2013 Workshop on Intelligent Techniques For Web Personalization and Recommender Systems*, ITWP '13. AAAI, 2013.
166. Zengbin Zhang, Lin Zhou, Xiaohan Zhao, Gang Wang, Yu Su, Miriam Metzger, Haitao Zheng, and Ben Y. Zhao. On the validity of geosocial mobility traces. In *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*, HotNets-XII, 2013.
167. Xiaohan Zhao, Alessandra Sala, Christo Wilson, Xiao Wang, Sabrina Gaito, Haitao Zheng, and Ben Y. Zhao. Multi-scale dynamics in a massive online social network. In *Proceedings of the 2012 ACM Conference on Internet Measurement*, IMC '12. ACM, 2012.
168. Elena Zheleva, Hossam Sharara, and Lise Getoor. Co-evolution of social and affiliation networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09. ACM, 2009.
169. Yu Zheng, Like Liu, Longhao Wang, and Xing Xie. Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08. ACM, 2008.
170. Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09. ACM, 2009.
171. Changqing Zhou, Nupur Bhatnagar, Shashi Shekhar, and Loren Terveen. Mining personally important places from gps tracks. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop*, ICDEW '07. IEEE, 2007.
172. Matteo Zignani and Sabrina Gaito. Extracting human mobility patterns from gps-based traces. In *Proceedings of Wireless Days 2010*, Wireless Days '10. IFIP, 2010.
173. Matteo Zignani, Sabrina Gaito, and Gianpaolo Rossi. Extracting human mobility and social behavior from location-aware traces. *Wireless Communications and Mobile Computing*, 13(3):313–327, 2013.