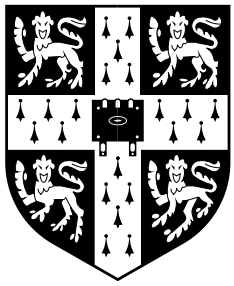
Spatial properties of online social services: measurement, analysis and applications

Salvatore Scellato



Churchill College
University of Cambridge

2012

This dissertation is submitted for
the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation does not exceed the regulation length of 60 000 words, including tables and footnotes.

Spatial properties of online social services: measurement, analysis and applications

Salvatore Scellato

Summary

Online social networking services entice millions of users to spend hours every day interacting with each other. At the same time, thanks to the widespread and growing popularity of mobile devices equipped with location-sensing technology, users are now increasingly sharing details about their geographic location and about the places they visit. This adds a crucial spatial and geographic dimension to online social services, bridging the gap between the online world and physical presence.

These observations motivate the work in this dissertation: our thesis is that the spatial properties of online social networking services offer important insights about users' social behaviour. This thesis is supported by a set of results related to the measurement and the analysis of such spatial properties.

First, we present a comparative study of three online social services: we find that geographic distance constrains social connections, although users exhibit heterogeneous spatial properties. Furthermore, we demonstrate that by considering only social or only spatial factors it is not possible to reproduce the observed properties. Therefore, we investigate how these factors are jointly influencing the evolution of online social services. The resulting observations are then incorporated in a new model of network growth which is able to reproduce the properties of real systems.

Then, we outline two case studies where we exploit our findings in real application scenarios. The first concerns building a link prediction system to find pairs of users likely to connect on online social services. Even though spatial proximity fosters the creation of social ties, the computational challenge is accurately and efficiently to discern when being close in space results in a new social connection. We address this problem with a system that uses, alongside other information, features based on the places that users visit. The second example presents a method to extract geographic information about users sharing online videos to understand whether such videos are going to become locally or globally popular. This information is then harnessed to build caching policies that consider which items should be prioritised in memory, thus improving performance of content delivery networks.

We summarise our findings with a discussion about the implications of our results, debating potential future research trends and practical applications.

Acknowledgments

First of all, I would like to thank Cecilia Mascolo, my supervisor: she has been nurturing and advising me throughout my years at Cambridge and her support and guidance have been fundamental to shape my research and focus my efforts. She was the first to suggest that I apply for a PhD and I cannot overstate how much that suggestion has changed my life. Then, I thank Vito Latora, who patiently taught me the joy of doing research when I was still a very young student, and who encouraged me to pursue bigger goals and study in Cambridge. Third, I want to thank Mirco Musolesi: his constant mentoring in the first months of my PhD helped me to improve my research and technical skills.

I have been lucky enough to enjoy many collaborations and to have many co-authors who deserve my gratitude: Prithwish Basu, Chloë Brown, Andreas Kaltenbrunner, Renaud Lambiotte, Ilias Leontiadis, Liam McNamara, Vincenzo Nicosia, Anastasios Noulas, Bence Pásztor, John Tang, Yana Volkovich, Murtaza Zafer. I would also like to thank Jon Crowcroft for many interesting discussions and great insights; I am also indebted to Richard Gibbens and Timothy Griffin, who have guided my thesis with wisdom and useful suggestions. Also, I would like to thank the friends I have made in the lab, especially Tassos, Ilias, Liam, Haris, Andrius, John, Christos, Kiran, Kharsim, Jisun, Narseo, Bence, Daniele: my life in Cambridge has been enjoyable and fun because of them and I will always cherish all the great moments we spent working in the lab or being somewhere else.

A highlight of the recent years is the time I spent during my internship at the Google office in Zürich, where I learnt how to use my scientific curiosity to build products and services that benefit millions of users all around the world and where, at the same time, I was allowed to test some of the ideas and insights contained in this dissertation. A big thanks goes to the entire YouTube Analytics team and especially to Anders Brodersen, who guided me with precious personal feedback,

and to Mirjam Wattenhofer, who always kept me motivated during my internship, working side-by-side with me.

While all those who worked with me deserve my utmost gratitude, I am certainly more indebted to my parents: their sacrifices, their encouragement, their guidance and their love have supported me throughout my education, enduring the physical distance existing between us. I thank my sister, Francesca, always ready to chat with me and soothe my homesickness when I am away from Sicily and always ready to spend time with me when I am there. I also thank Antonio, Diletta and Isotta, because they have been like family.

I do not even know how to thank my beloved Violetta. She has been through this PhD with me, month after month, spending hours with me on Skype, sitting through unpolished versions of my talks, reading drafts of my papers, visiting me in Cambridge and waiting for me in Sicily. She has always believed in me, even when I did not believe in myself: I know that without her I would not be writing this. I hope we will remember these years forever, together.

Contents

The First Law of Geography is: everything is related to everything else, but near things are more related than distant things.

Waldo R. Tobler

1 Introduction

The popularity of online social networking services has grown at an extraordinary pace over the last years, dramatically altering how hundreds of millions of users spend their time. The numbers are overwhelming: to date, Facebook has more than 900 million active users, with half of them using the service on a daily basis, while Twitter has more than 350 million users; YouTube and LinkedIn have similarly large user bases, while dozens of smaller services boast millions of active users.

The importance of online social networks is growing in parallel with their popularity: as users spend more and more time socialising online, service providers can offer sophisticated features such as targeted advertising, personalised recommendations and content delivery. As a result, it becomes of crucial significance to study and understand the online behaviour of such users and the characteristics of the social connections that bind them. Also, the need to design and build systems and applications that revolve around online social connections has sparked research into understanding the structure of online ties among users.

These attempts have often adopted methods of complex network [RB02, BLM+06] and social network analysis [WF94], combining them with algorithmic techniques drawn from computer science to manage massive social graphs. Thanks to the availability of large-scale data about online social interactions, it is now possible to make sense of billions of friendship connections and to infer notable patterns and useful properties.

Some seminal works have characterised the structure of these online social net-

works, finding common properties which appear across different social services. For instance, heavy-tailed degree distributions, where a non-negligible fraction of users have many social connections, and the presence of locally dense social communities appear as the most significant traits of many different online social networks [MMG+07, AHK+07]. Other work has focussed on studying the evolution of online social networks over time, trying to reproduce their observed properties with generative models that are inspired by mechanisms purportedly driving user behaviour [KNT06, LBKT08].

In general, studying online social networks has allowed researchers to extend the scope of traditional social network analysis, scaling up to millions of individuals and billions of social links. The combination of large-scale data analysis, insights provided by sociological theories and problems arising from system engineering has resulted in a plethora of applications and systems that mine online social interactions to provide suggestions, offer recommendations and filter information. This has impacted the Web in an unprecedented way, as these features are profoundly different from the predominantly static content, lacking any personalisation, that was available to users only a decade ago.

In fact, insights into the properties of online social services can be exploited to design novel applications that provide recommendations about items [Gol08], answer Web search queries [EC08, HK10] and reduce spam [GKF+06], among many other examples. Information related to online social ties can even be used to improve existing distributed systems and applications: for instance, by taking into account how people use online social services to share and consume content items, it becomes possible to optimise delivery and storage of online content [THT+12].

More recently, the widespread adoption of powerful mobile devices has led to a dramatic change in the way the Web is accessed. In particular, every day hundreds of millions of individuals use a smartphone to interact online with their friends. The launch of new operating systems for mobile devices has drastically reshaped which applications and services are available to end users. Both Google's Android and Apple's iOS, the two most successful smartphone platforms to date, offer application stores where developers can publish, and sell, their mobile applications. The abundance of such applications further stimulates the adoption of mobile devices: the overall effect is that this field is developing fast and in many directions.

An important related aspect is the increased access to social networking services through mobile devices: mobile users spend more minutes every day interacting with social applications than desktop users [Mas10] [1]. Simultaneously, mobile Web access has caused a substantial shift in the feasibility of pervasive and ubiquitous

[1] As early as in 2010 mobile users were using Facebook on their mobile devices on average for 45 minutes a day, while desktop users only for 32 minutes.

services.

The deployment of location-based services has been made possible by the location-sensing capabilities of these devices; they are able to generate location-tagged information and enable users to share their physical whereabouts. As a result, online services are increasingly becoming *location-aware*.

1.1 How geography affects online social networking

The combination of the upsurging popularity of online social networking, especially on mobile platforms, and the rise of mobile location-based services allows us to merge together two facets of user behaviour that were previously difficult to connect, **adding a crucial spatial dimension to online social networking services**. For the first time, the wealth of information about online social interactions can be augmented with geographic information. Online social networks were previously studied ignoring their spatial properties, as these were not accessible. Now, instead, users can be considered to be embedded in a data-rich geographic space.

1.1.1 The rise of location-based mobile services

This connection between social and local services has been epitomised by location-based online social services such as Foursquare[2], Brightkite[3] and Gowalla[4], which have attracted millions of users in recent years. These services are targeting a mainly mobile user audience: they are based on the concept of disclosing the presence of the user at a particular venue, broadcasting such notifications to friends. It is crucial to stress that not only the geographic location of each user is revealed to these services, but also a detailed set of additional data related to individual places: for instance, users could disclose that they are in a stadium, visiting a museum or spending time in a cafeteria.

At the same time, reviews, tips or other information related to such places can be generated and shared. Therefore, a vast and detailed user-generated catalogue of venues is continuously growing within each service, compiled by users themselves and providing fine-grained data about where people go. Places, with their simultaneous online and offline presence, represent a new entity that drives and shapes user behaviour, bridging the gap between physical location and online activity [CTH+10].

[2] `http://www.foursquare.com`
[3] `http://www.brightkite.com`
[4] `http://www.gowalla.com`

More generally, all online social services are increasingly becoming location-aware, allowing users to create and access information about their geographic whereabouts. The trend is progressively going from specialised providers offering *location-as-a-service* to a widespread new concept of *location-as-a-feature*, where every online social platform integrates geographic information into their services. For instance, Facebook recently introduced a new feature allowing every single piece of information generated on the service, being it a status update, a photo or a notification from a third-party application, to be tagged with a specific spatial location. Hence, spatial details related to online social activities become progressively more available and exploitable.

1.1.2 The effect of geographic space on online social ties

Among the many interesting research questions sparked by the availability of spatial data on online social services, a fundamental one is whether geographic space affects social interactions taking place on the Web.

Systems where space and distance constrain connections between networked entities have been extensively studied, like in transportation networks [KT06], Internet router connections [YJB02, BGG03], power grids [AAN04] and urban road networks [CSLP06]. In general, metric distance directly influences these systems by imposing higher costs on the connections between distant entities. When there is a cost associated with link length, the appearance, and the persistence, of longer links is usually compensated by some other advantage. As an example, long-distance commercial flights are often directed to well-connected airport hubs.

However, social networks have been largely studied from a purely topological perspective, focussing mainly on the structure of the graph. Some sociologists have studied the effect of geographic distance on social ties before the advent of online social services, with the underlying expectation that most individuals would try to minimise the efforts to maintain a friendship link by interacting more with their spatial neighbours. This would be in accordance with the broad "Principle of Least Effort" theorised and proposed by Zipf to explain multiple facets of human behaviour [Zip49]. Individuals could be less likely to meet people who live further away because overcoming distance needs more time or more money, in other words, more effort.

In fact, as early as 1941 Stewart observed an inverse relationship between distance and the likelihood of friendship between college students [Ste41]. Similar statistical regularities have been later observed in new housing developments [FSB63], residences for the elderly [NL75] and urban interactions [ML76]. Nonetheless, the connection costs imposed by spatial distance may not be important in social systems,

particularly when focussing on online interactions. The Internet and, in general, other communication technologies may potentially lessen the costs associated with social interaction, removing geographic barriers and reducing overhead.

1.1.3 An historical perspective

As McLuhan theorised in 1962 [McL62], years before the inception of the Internet, the enhanced transmission speed of information given by modern mass media would turn the world into a "Global Village". Thirty years later, such a concept became widely popular thanks to the birth of the Web, which fostered the idea that people can communicate with ease and simplicity as a single, planetary community.

It is reasonable to say that, thanks to the Web, people now are connected, and keep in touch, with greater simplicity and proficiency than at any time in the past. As proposed by Cairncross, spatial distance may finally cease to play a rôle because of the increasing availability of affordable long-distance travel and cheap communication channels, resulting in the inevitable "Death of Distance" [Cai01], while other scholars have similarly discussed the "End of Geography" [TL88]. The implied consequence is that in this new scenario the process of friendship formation might easily become completely disentangled from spatial distance [Gra98].

Interestingly, similar arguments had been already put forward when other technological breakthroughs were made. For instance, the introduction of the telegraph in 1844, with an initial 40-mile link between Washington and Baltimore, provided for the first time the effective separation of communication from transportation, freeing the transmission of information from the constraints of geography, as discussed by Carey [Car89]. This idea goes back to Cooley, who wrote in 1894 that "Space – distance – as an obstacle to communication has so nearly been overcome that it is hardly worth considering" [Coo94].

Similar considerations can be made about the reactions sparked by the introduction of the telephone or the radio: common people and academic scholars anticipated a far-reaching revolution, bound drastically to alter how individuals would communicate with each other. Yet, as it became apparent after each individual innovation, a new communication technology hardly cancels out or completely replaces existing systems. Instead, it is easily adopted to maintain and nurture social communication channels that were already in place: face-to-face contacts and the shared experience of spatial locality remain dominant across communication media [HW01].

Similar reasoning might apply to social interactions on the Web: they could reflect social ties and contacts that develop and exist through other communication channels, such as face-to-face encounters or phone calls. The effect of distance on such social ties would then be still important, even if online communication tools

are widely available. In reality, precisely because of the latest technological changes in travel and communication, evidence suggests that social groups have become "glocalised" [WH99], with both extensive short-range links and occasional long-distance relationships. Even more convincingly, some initial results clearly demonstrate that online social connections are more likely to appear at shorter geographic distances [LNNK+05, BSM10].

As the death of distance seems postponed, space and proximity might continue to play a pivotal rôle on the Web, influencing whom individuals connect to and how they interact with others. Thanks to the wealth of geographic information increasingly available, it is possible to understand the effect that space and distance have on online social services. This is likely to provide a more complete picture of social interactions on the Web, with important and far-reaching implications. In addition, as the relative importance of the Web grows, this knowledge might shed more light on social behaviour in a broader sense.

1.2 Potential implications

Augmenting social structure with geographic information adds a new dimension to social network analysis and a large number of theoretical investigations and practical applications can be pursued for online social systems, with many promising outcomes.

From one point of view, spatial information can help to explain social phenomena taking place online, such as the creation of friendship ties or the spreading of information. Even though the structure and the dynamics of social networks have been under scrutiny for many years [CFL09, WF94], only a few works have addressed how geographic distance affects online social ties [LNNK+05, BSM10]. These initial results still leave untouched issues such as how online users establish new social connections over space and whether their online interactions are affected by distance. Similarly, users could be characterised by their preference towards global, long-range interactions rather than towards local, short-distance online ties, in order to classify their behaviour and profile them.

On the other hand, location-sharing on online services opens possibilities for new applications and systems. Details about the type of places where individuals go are increasingly available, providing rich information about user preferences and choices. Applications such as local search, content recommendation and advertising would greatly benefit from such geographic information. Search queries about local content could be targeted to nearby users, while both advertising and recommender systems could better profile users by knowing how their social ties stretch over space, thus improving their accuracy. Moreover, information about social links, content

consumption and geographic location can reveal how tastes and interests disseminate over an online social service. Some potential applications of these ideas include targeted advertisement, more effective content spreading (e.g. shop promotions, local news, job openings) and even local activism and advocacy.

Finally, large-scale systems would greatly profit from a better knowledge of how online users are connected over space and how information spreading over space creates demand for content and services around the planet. In particular, with the recent rising interest in cloud services [Hay08] and content delivery networks [Lei09], it has become extremely important to understand the geographic patterns of traffic requests. A challenging problem is to understand whether it is possible to improve the design of such systems by exploiting the geographic properties of social processes. For instance, popularity of content can be geographically and temporally characterised to devise new strategies for replica placement and caching.

1.3 Thesis and its substantiation

As we have discussed, gaining knowledge about how geographic space influences online social services could be of great importance to understand better many research problems and to improve systems related to these services. The effect of geographic distance seems still to be present in the online world: a more complex and broad research question regards how spatial and social factors simultaneously influence the structure of online social networks and the dynamic processes that take place on them. Closely related to this theme is the problem of exploiting the spatial dimension of online user behaviour to provide better and more useful features in online social networking services and to devise new systems and application.

Consequently, the **thesis of this dissertation** is that *the study of the spatial characteristics of online social interactions is useful to provide a more comprehensive understanding of their structure and to build more efficient and effective systems and applications on top of them.*

We substantiate this statement with two closely related threads of research. First, we aim to expand the understanding of the spatial properties of online social networks, focussing on measuring, analysing and modelling such properties and their connection to social patterns. Second, we plan to demonstrate that such spatial characteristics can be used in the design of new systems and applications related to online social networking services.

1.4 Contributions and chapter outline

This thesis offers three major contributions: firstly, the measurement and the analysis of social and spatial properties of online social services, secondly, the study of models which capture the spatial and social properties of user behaviour on such services, and finally the design and the evaluation of applications and systems that exploit spatial and geographic information in online social networks.

As we have considered, the impact of spatial distance on online social networks seems still to be important, even though the Internet and the Web allow individuals to communicate easily and cheaply. As a consequence, the properties usually observed in online social networks could be influenced by geographic distance in a variety of different ways. Hence, in Chapter 2 we introduce and explore the properties of online social services, discussing the rôle of space in shaping them. We also examine whether location-sharing features, which reveal the spatial patterns of online social interactions, might be changing how users engage with online social platforms. This discussion provides insights into why the spatial properties of these online services are of significant importance to understand better online user behaviour and to build related systems.

The rest of the dissertation presents our novel contributions, which are summarised as follows:

- In Chapter 3 we discuss the effect that spatial factors have on online social platforms through a comparative study of the spatial properties of the social graphs arising among users of popular online services. We exploit location data available on such services to embed users in geographic space, studying the resulting social networks as spatial networks. We define two **randomised null models** of the social graph that take into account either only the spatial properties or the social properties of the original graph: this allows us to discern what characteristics we would observe if only spatial, or social, factors were in place. Using these two null models we discuss the interplay between the spatial and social dimensions, which generates a wide heterogeneity of properties across different users. We also propose two new network measures, **node locality** and the **geographic clustering coefficient**, which help to differentiate users with respect to their preference for short-range or long-distance ties.

- In Chapter 4 we aim to understand the temporal evolution of an online social network and its spatial properties with a longitudinal study of a real service. Our goal is to define basic evolutionary models that can reproduce the social and spatial patterns observed in the real data and the properties discussed in

Chapter 3. We show that social factors and spatial distance simultaneously influence the establishment of new user connections: this can be modelled as a **gravitational attachment** process that mimics the attraction forces between physical bodies influenced by mass and distance. At the same time, we note that triadic closure is also strongly shaping the creation of social links, although this process appears to be driven purely by social factors. These findings allow us to propose a new **gravitational model of network growth**, which is able to reproduce the social and spatial properties observed in real networks. We further discuss how our new model compares to other frameworks previously introduced to study spatial networks.

- In Chapter 5 we explore one practical application that takes advantage of spatial data available on online social networks: **link prediction**. Link prediction systems have been largely adopted to recommend new friends in online social networks using data about social interactions. We propose to exploit an additional source of information: the places people visit. We study the problem of designing a link prediction system for online location-based social networks. We investigate how users create new connections over time and we study the relative link prediction space: we find that about 30% of new links are added between **"place-friends"**, i.e., between users who visit the same places. We show that this prediction space can be made 15 times smaller, while still 66% of future connections can be discovered. Finally, we define new prediction features based on the properties of the places visited by users, which are able to discriminate potential future links among them. Building on these findings, we describe a supervised learning framework which exploits these prediction features to predict new links between friends-of-friends and place-friends, offering high link prediction performance.

- In Chapter 6 we explore a different application that benefits from the constraints imposed by spatial distance on online social connections: **video content delivery on a planetary scale**. More and more, the diffusion of content items happens on online social networks, where social cascades can be observed when users increasingly re-post links they have received from others. We take advantage of the fact that such social cascades can propagate in a geographically limited area to discern whether an item is spreading locally or globally. This informs cache replacement policies used in content delivery networks, which utilise this information to ensure that content is kept close to the users who may be interested in it. We build a proof-of-concept geographic model of a distributed content delivery network and we simulate its performance on real traces; our evaluation shows that we improve cache hits by up to 70% with respect to cache policies without geographic and social information.

To conclude, in Chapter 7 we discuss and summarise the insights offered by this dissertation and we explore their consequences, presenting directions for further research.

1.5 List of publications

During the course of my Ph.D. I have had the following 20 publications. Thanks to many fruitful collaborations, I had a chance to contribute to several different projects: hence, not all the following works contribute to this dissertation. In more detail, this introduction is inspired by [Sce11], Chapter 3 draws from [SMML10a] and [SNLM11], Chapter 4 is based on [ASM12], Chapter 5 is based on [SNM11] and Chapter 6 is inspired by [SMMC11].

Works related to this dissertation

[SMML10a] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, Vito Latora. Distance Matters: Geo-social Metrics for Online Social Networks. *Proceedings of the Third Workshop on Online Social Networks (WOSN 2010)*, co-located with USENIX, (Boston, Massachusetts, USA), June 2010.

[SMMC11] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, Jon Crowcroft. Track Globally, Deliver Locally: Improving Content Delivery Networks by Tracking Geographic Social Cascades. *Proceedings of the 20th World Wide Web Conference (WWW 2011)*, (Hyderabad, India), March 2011.

Named "Publication of the Year 2011" by the Cambridge Computer Lab Ring.

[SNLM11] Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, Cecilia Mascolo. Socio-Spatial Properties of Online Location-Based Social Networks. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, (Barcelona, Spain), July 2011.

[SNM11] Salvatore Scellato, Anastasios Noulas, Cecilia Mascolo. Exploiting place features in link prediction on location-based social networks. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD 2011)*, (San Diego, California, USA), August 2011.

[Sce11] Salvatore Scellato. Beyond the Social Web: The Geo-Social Revolution. *ACM SIGWEB newsletter*, Autumn 2011 issue.

[**ASM12**] Miltiadis Allamanis, Salvatore Scellato, Cecilia Mascolo. Evolution of a Location-based Online Social Network: Analysis and Models. *Proceedings of the 12th ACM International Internet Measurement Conference (IMC 2012)*, (Boston, Massachusetts, USA), November 2012.

Other works

[**DEM+10**] Vladimir Dyo, Stephen A. Ellwood, David W. Macdonald, Andrew Markham, Cecilia Mascolo, Bence Pásztor, Salvatore Scellato, Niki Trigoni, Ricklef Wohlers, Kharsim Yousef. Evolution and Sustainability of a Wildlife Monitoring Sensor Network. *Proceedings of the Eighth ACM Conference on Embedded Networked Sensor Systems (SenSys 2010)*, (Zürich, Switzerland), November 2010.

[**TSM+10**] John Tang, Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora. Small-world behavior in time-varying graphs. *Physical Review E* **81** (2010), no. 5, 055101(R).

[**SM11**] Salvatore Scellato, Cecilia Mascolo. Measuring User Activity on an Online Location-based Social Network. *Proceedings of the Third International Workshop on Network Science for Communication Networks (NetSciCom 2011)*, co-located with INFOCOM 2011, (Shanghai, PRC), April 2011.

[**SLM+11a**] Salvatore Scellato, Ilias Leontiadis, Cecilia Mascolo, Pritwish Basu, Murtaza Zafer, Understanding Robustness of Mobile Networks through Temporal Network Measures, *Proceedings of the 30th IEEE International Conference on Computer Communications (INFOCOM 2011)*, mini-conference track, (Shanghai, PRC), April 2011.

[**SMML10b**] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, Vito Latora. On Nonstationarity of Human Contact Networks. *Proceedings of the Second Workshop on Simplifying Complex Networks for Practitioners (SIMPLEX 2010)*, co-located with ICDCS 2010, (Genoa, Italy), June 2010.

[**SMM+11**] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, Vito Latora, Andrew J. Campbell. NextPlace: A Spatio-Temporal Prediction Framework for Pervasive Systems. *Proceedings of the Ninth International Conference on Pervasive Computing (Pervasive 2011)*, (San Francisco, California, USA), June 2011.

[**NSMP11**] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, Massimiliano Pontil. Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. *Proceedings of the Third Workshop*

on Social Mobile Web (SMW 2011), co-located with ICWSM 2011, (Barcelona, Spain), July 2011.

[SLM+11b] Salvatore Scellato, Ilias Leontiadis, Cecilia Mascolo, Pritwish Basu, Murtaza Zafer. Evaluating Temporal Robustness of Mobile Networks. *IEEE Transactions on Mobile Computing*, 15 November 2011. IEEE Computer Society.

[ASW12] Anders Brodersen, Salvatore Scellato, Mirjam Wattenhofer. YouTube Around the World: Geographic Popularity of Videos. *Proceedings of the 21st World Wide Web Conference (WWW 2012)*, (Lyon, France), April 2012.

[YSL+12] Yana Volkovich, Salvatore Scellato, David Laniado, Cecilia Mascolo, Andreas Kaltenbrunner. The length of bridge ties: structural and geographic properties of online social interactions. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM 2012)*, (Dublin, Ireland), June 2012.

[NSL+12] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, Cecilia Mascolo. A tale of many cities: universal patterns in human urban mobility. To appear in *PLoS ONE*.

[BNS+12] Chloë Brown, Vincenzo Nicosia, Salvatore Scellato, Anastasios Noulas, Cecilia Mascolo. The Importance of Being Placefriends: Discovering Location-focused Online Communities. *Proceedings of the Fourth Workshop on Online Social Networks (WOSN 2012)*, co-located with SIGCOMM 2012, (Helsinki, Finland), August 2012.

[KSV+12] Andreas Kaltenbrunner, Salvatore Scellato, Yana Volkovich, David Laniado, Dave Currie, Erik J. Jutemar, Cecilia Mascolo. Far from the eyes, close on the Web: impact of geographic distance on online social interactions. *Proceedings of the Fourth Workshop on Online Social Networks (WOSN 2012)*, co-located with SIGCOMM 2012, (Helsinki, Finland), August 2012.

[NSLM12] Anastasios Noulas, Salvatore Scellato, Neal Lathia, Cecilia Mascolo, A Random Walk Around the City: New Venue Recommendation in Location-Based Services. *Proceedings of the Fourth IEEE International Conference on Social Computing, (SocialCom 2012)*, Amsterdam, The Netherlands, September 2012.

The social sciences are granted eternal youth because findings must be revisited.

Max Weber

2

Online social services: an overview

Starting from instant chat and bulletin boards systems, and then with email, newsgroups and discussion forums, people have engaged in online social interactions since the inception of the Internet. Nonetheless, at the beginning of the 21st century a new generation of "social networking services" came to prominence, revolutionising, within a few years, the way users access the Web and spend their time online. Nowadays services like Facebook and Twitter boast hundreds of millions of active users; in addition, the amount of time spent by users on such sites is soaring.

These online services are already so popular, and their usage so entrenched in our society, that they have even changed our common language, introducing words such as *unfollow* ("verb: stop tracking a person, group, or organisation on a social networking site") or adding new meanings to existing words such as *friend* ("verb: add someone to a list of friends or contacts on a social networking website")[1].

In this chapter we provide a broad overview of online social networking services, discussing their classification and their main properties. We also review related systems and applications that exploit social information to offer a wide set of functionalities. Our main emphasis is on highlighting how spatial and geographic factors may influence the properties of online social connections, considering their consequential impact on social-based systems and applications.

[1]The Oxford English Dictionary introduced these modifications in 2011 and 2010, respectively.

Chapter outline Section 2.1 introduces a definition of social networking services and then examines their general taxonomy. The most important properties of the social networks arising among online users are discussed in Section 2.2, which reviews a vast set of related works in the literature. In Section 2.3 we describe the new generation of social services mainly based on location sharing and we discuss the connection they offer between online social interactions and real entities in the physical world. Systems and applications related to online social services are discussed in Section 2.4. Finally, Section 2.5 draws lessons from the research results we survey, discussing how this dissertation differs from the growing body of literature, opening new research directions.

2.1 Classification of social networking services

The abundance of online services providing social features or supporting social interactions among users makes it difficult to define properly when a service can be considered an online social networking platform. We discuss a definition recently proposed by scholars and, then, we analyse a classification of these services according to the semantics of the social connections they support and the related structural directionality. We conclude this section with a review of recent results focussing on the characteristics of the most important online social services.

2.1.1 Core features of social networking services

As discussed earlier, social interactions were taking place online before the rise of services such as Facebook and Twitter. For instance, instant messaging applications were already widely used about 15 years ago to keep a list of friends and chat with them in real time. However, social networking sites present a unique set of characteristics that clearly differentiate them from earlier services.

One concise definition relies on the fact that a social networking service is a Web-based service that offers these three features to its users [bE07]:

1. the creation of a public or semi-public profile within a bounded system;
2. the ability to create and maintain a list of other users with whom they share a connection;
3. the ability to view and traverse their list of connections and those made by others within the system.

From this definition we note that what identifies social networking services is not the fact that they enable online social interaction between connected members, but,

more importantly, that users can carefully craft and make publicly visible their social connections, and interactions, on the service. This is the crucial feature that differentiates social networking services from older online communication tools.

A consequence of this visibility of social ties is that they often represent a display of social identity [Db04]. However, there are deep differences in the structural nature and the social semantics of ties supported by different services. These factors are of paramount importance to understand the characteristics of a given social networking service, its target user base and the social interactions it supports and fosters; equally, the applications and systems that each service enables will be different. These variations are also likely to impact the spatial properties of the resulting social ties.

2.1.2 Link semantics

As different online social networking services offer different features, and target different users, the significance of the social connections that their users establish varies widely. A coarse-grained yet effective distinction is between services that *support* social relationships and services that *create* connections between people who share interests. For instance, Facebook users tend to recreate online the set of friendship and acquaintance ties they have in their daily lives, while LinkedIn supports work-related connections, mainly to help job-seekers and foster professional collaboration.

Other social networks try to entice users with particular interests, such as Last.fm for music-lovers, aNobii for individuals passionate about books, Flixster for cinephiles and Flickr for amateur and professional photographers. Finally, social services such as Twitter and Google Plus encourage their users to connect to other users, celebrities, media organisations or product brands, effectively supporting a wide range of link semantics.

Implication: The effect of space and distance differs across these services as the focus of link semantics shifts from supporting existing social relationships to connecting users who share interests. Users connecting because of a common interest in a given topic might be less likely to be constrained by geographic proximity, as their tie is fostered by factors less related to the social sphere.

2.1.3 Link structure

At a structural level, the main property of online social platforms is whether new social connections established by users are unidirectional or bidirectional: in the

latter case, both users must confirm the existence of the social tie that connects them on the service.

Social services which aim to support trusted relationships, such as Facebook and LinkedIn, adopt the bidirectional model: users are required to request, and accept, the creation of friendship connections. In this case there is only an implicit directionality of the resulting link, as one user has to initiate the contact and the other has to react. Instead, services based on shared interests tend to adopt the unidirectional model, where users can connect with freedom without requiring any reciprocation. In this case the act of creating a friendship connection is more oriented towards the consumption of user-generated information: for instance, book reviews, photos, music tracks, videos or Web links.

This consumer-producer paradigm, where a user generates information and a set of "followers" receive it, has been hugely successful. Such unidirectional connection models, more akin to subscriptions than to friendship connections, are epitomised by services such as Twitter and YouTube, where every user profile is seen as a channel of updates which is pushed to the subscribers. Google launched its own social networking service in 2011, Google Plus, with a unidirectional sharing model that allows users to subscribe to any other user's updates; at the same time, each user can specify the preferred audience for every single piece of shared information, with highly granular privacy control. Shortly after, Facebook amended its historical preference for mutual bidirectional connections and allowed users to "subscribe" to each other, providing sharing filters for users to limit the visibility of each post.

Bidirectional ties are more difficult to establish: malicious users or spammers can indiscriminately make unidirectional links to several unrelated users, whereas bidirectional links must be vetted by both parties. Even when unidirectional links are arising because a user has a genuine interest in connecting to another user, there is little indication that the connection has a social nature, unless the tie is reciprocated. On the other hand, unidirectional ties allow a richer social structure to emerge, since there are now four possible relationships between any couple of users. Tie asymmetry can be exploited to infer status or reputation, as the PageRank mechanism used by Google does to rank Web pages [PBMW99]. TunkRank, based on the same principles of PageRank, has been proposed to rank influential Twitter users based solely on link structure [Tun09].

Implication: We expect spatial constraints to be stronger when online connections reflect relationships that are as close as possible to social ties in real life: as a consequence, social services requiring bidirectional ties could be more affected by geographic distance, whereas services allowing unidirectional followers could be less constrained by space.

2.1.4 Examples

Even though only in the last decade online social networking has seen an explosion of interest, social interactions were taking place online much earlier. In fact, Ward Christensen and Randy Suess created the Computerised Bulletin Board System, or CBBS, in 1978 to allow computer enthusiasts to exchange messages in a computerised way, predating modern social forums[2]. While CBBS was used actively by only a handful of users, current online social platforms have a massive user base. For instance, Facebook boasts more than 900 million active users, over 50% of whom access the service on any given day [Fac12]. Similarly, LinkedIn has more than 135 million members over all the planet, establishing professional ties with one another [Lin12].

Modern online social services have offered researchers the opportunity to study online social interactions at an unprecedented scale. In addition, since online services keep track of such interactions in a structured and digital way, they offer computation-friendly means of recording, storing, accessing and manipulating the graph of social relationships. However, the sheer size of modern social networking services has generated extremely large social graphs, which pose several analytical and technical challenges. The social connections between users, together with any additional available information, are often obtained through Application Programming Interfaces (APIs) exposed by each service to provide access to their data in a structured form. Yet, this data acquisition process may be severely time-consuming, as API calls are typically rate-limited and closely monitored by the service providers.

At the same time, analytical techniques, algorithms and computational resources used to study the resulting graphs need to cope with the size of current social platforms. For instance, one of the largest social systems ever analysed consisted of about 30 billion instant messaging conversations exchanged over 30 days among 240 million users, gathering about 4.5 terabytes of text logs for each day [LH08]. An even larger example is a recent study, which adopted a methodology based on probabilistic counters to compute graph distances between Facebook users [BBR+11]: the authors analyse a network with about 720 million users and 70 billion friendship connections, discovering that, on average, Facebook users are only 4 hops apart, with smaller regional networks exhibiting even lower values.

In the following paragraphs we discuss different kinds of online social services that have been studied by researchers, ranging from online communities predating the inception of the Web to modern large-scale social platforms.

[2] A brief yet vividly sketched account of online social networking is The Cartoon History of Social Networking [Lon11]. Other short reports with similar material include [Nic09, Sim09, Bia11].

Earlier online communities

Before the advent of current online social networking services, online services were neither explicitly recording nor storing social ties between users. Yet, personal home pages were linking to each other and users were replying to each other on forums, leaving a digital trace of these social interactions. Thus, social graphs arising in older online communities tend to be inferred based on records of such interactions.

One of the first works on the analysis of online social relationships used email conversations to infer a social graph and study shared interests among people [SW93]. ReferralWeb [KSS97] was a system built to reconstruct a social network among individuals by mining their online presence, sifting through Web links across home pages, co-authorship and citations in academic papers, news archives and organisation charts, with the aim of exploiting the network structure to locate experts across different topics. Similarly, social statistics about online interactions in the LambdaMOO Multi-User Dungeon were collected by an autonomous social software agent, Cobot [IKK+00]. Conversations taking place on instant messaging platforms were used to extract a planetary-scale social network and investigate homophily effects [LH08].

There are other online platforms where social features do not represent the main focus, such as online shopping and product reviews sites. Such services also support social interactions between members, usually involving products in which users share an interest: such social graphs have been exploited to enhance confidence in sellers, by providing explicit feedback, as well as to provide useful product recommendations. Several such social networks have been examined: examples include Epinions [RD02], Amazon [LAH07] and Overstock [SWB+08].

Modern online social services

With the rise of the Web 2.0, loosely defined as the set of online services centred around consumption and sharing of dynamic user-generated content, the digital representation of social connections between users became of pivotal importance, giving birth to the first online services mainly focussed on social features. As these services accumulated millions and millions of users, researchers turned their attention to them.

One of the earliest studies analysed the social networks arising among the members of two services, the photo sharing community Flickr and the social networking site Yahoo!360, finding that each network is segmented into a well-connected core of users and a fringe of disconnected smaller communities and isolated nodes [KNT06]. More studies soon followed, investigating different social services and offering the first evidence that online social networks exhibit certain universal features [MMG+07,

AHK+07]. Other online communities that have attracted the attention of researchers include the hugely popular, albeit now in decline, MySpace [CW08, TRW09] and the social aggregation service FriendFeed [GGM+09].

Some services have received abundant attention due to their sheer popularity and to their importance, with hundreds of millions of users accessing them on a daily basis. We consider a sample of studies for three of the most common platforms on which people interact: Facebook, Twitter and YouTube.

Facebook In addition to the basic topological properties, various social processes have been investigated on Facebook, since it allows users to interact in a rich manner across several different media (statuses, photos, likes). Such user interactions reveal a picture very different from the entire graph of social connections: when social links with lower interaction levels are removed the network shrinks and changes structure, thus raising the challenge to individuate when a tie on Facebook represents a meaningful social bond between two users [WBS+09, VMCG09]. Further, since its inception, user engagement on Facebook is so pervasive that temporal periodicities connected to human rhythms can be extracted from user interactions [GWH07]. The analysis of information diffusion, for instance in terms of users sharing Web links received from their friends, has been another process largely investigated on Facebook [SRML09]. In particular, a recent work by Bakshy et al. [BRMA12] confirms that being exposed to information through online friends increases the likelihood of re-sharing that information.

Twitter Launched in 2006 as an SMS-based service, Twitter has been steadily growing as the main service for publicly sharing status updates: its public nature, together with the unidirectional nature of its social connections, has created a unique social ecosystem that has attracted plentiful of research attention. The motivations that brought early adopters to use Twitter were discussed in [JSFT07], while a classification of Twitter users was proposed in [KGA08]. As a few Twitter users gained celebrity status by accumulating millions of followers, researchers addressed the problem of finding the most influential users, adopting different metrics to quantify the impact of user influence [CHBG10]. A large-scale study of the entire social network of Twitter users found low levels of reciprocity among users, unlike other social networks, but a considerable number of news URLs being shared and forwarded over social connections [KLPM10]. Based on these findings, the hybrid nature of Twitter as both a social network and a news media site has been suggested.

YouTube Even though YouTube is not strictly a social service, since it mainly allows user to publish and share their video creations, social interactions take place

on a massive scale: users subscribe to one another's channels and comment on video items published by others. Video popularity on YouTube has been extensively studied: one of the central findings is that there is a strong heterogeneity across videos, with a few of them rising to extremely high levels of popularity and the vast majority experiencing only a handful of views [CKR+09]. Far from being irrelevant, the huge number of non-popular items still drives user consumption, attracting people interested in a vast quantity of niche topics: this effect has been popularised by Anderson as the "Long Tail" [And06] and greatly impacts systems designed to host and deliver YouTube videos. The temporal dynamics of video popularity has also attracted research efforts, trying to understand the factors that drive success for video items [FBA11]: in particular, both social sharing and external influence affect video popularity growth, giving rise to characteristic patterns in temporal evolution of the number of daily hits [CS08].

2.2 Characteristics of online social networks

The graphs arising among the members of online social services exhibit certain characteristic features, many of which are equally found in other types of social networks. Some of these properties have even entered common knowledge and popular culture. Many of these features are not unique to social networks, but are commonly found in other networked systems.

We present in this section a review of the most important features observed in social graphs, arising in online services or elsewhere. Our intention is to introduce the characteristics that are traditionally associated with social systems and to discuss which properties are likely to be influenced by the spatial constraints imposed by geographic distance.

2.2.1 Node properties

A first observation about social graphs is that they tend to be *sparse*: a large fraction of node pairs are not connected, with the result that the number of links is comparable to the number of nodes. The degree of a node, that is, the number of connections it has to other nodes, is of particular interest when analysing complex networks. When each node has the same topological properties, one would expect a homogeneous degree distribution to arise, where every node degree is tightly distributed around a well-defined average value. In these scenarios nodes are almost interchangeable, lacking any distinguishable individual feature. Instead, real networked systems often exhibit a much broader distribution, with a noticeable positive skewness or, in other common terms, with a *heavy tail*. Such distributions have

been found, among other examples, both in the network of Web pages [HA99] and in the graph of connections between Autonomous Systems on the Internet [FFF99].

This broad degree distribution has two important consequences: first, individual nodes are different from a topological perspective, as there exist huge variations between them; second, the co-existence of a few highly connected nodes and a large number of poorly connected elements has important consequences for the overall network structure and for processes such as information diffusion [Rap53], epidemic spreading [PSV01] and attack tolerance [AJB00]. In online services, where connections can easily be established and then accumulated, this heterogeneity can be even more marked [KNT06, MMG+07].

Another remarkable feature of social graphs is that they tend to exhibit positive correlations between node degrees [New02]. In other words, individuals with more connections tend to connect to other individuals with many connections: conversely, individuals with fewer ties connect among themselves. In contrast, other complex networks with a similar degree distribution exhibit negative correlation, with high degree nodes preferentially connecting to low degree ones [YJB02, BBPSV04]. Positive assortativity, together with transitivity (Section 2.2.2), are the two main structural features that differentiate social networks from other types of networks: their presence seems to be intimately connected to the tendency of individuals to cluster in groups or communities (Section 2.2.4) [NP03]. As a result, these structural properties have important effects on several dynamic processes that take place on social networks.

Implication: The heterogeneity observed across nodes in social networks could translate to a similar heterogeneity with respect to spatial properties. In particular, popular users with large numbers of connections might tend to be *less* affected by spatial distance than users with fewer connections. Given the importance of such highly connected individuals for the processes taking place over the social network, the spatial implications would become significant.

2.2.2 Transitivity

Sociologists have often discussed the fact that social connections are *transitive*, in the sense that friends of friends are also likely to be friends [WF94]. Several theories have been put forward to understand what drives this behaviour: Heider was the first to suggest that people tend to seek *structural balance* in their relationships [Hei46]. The idea of balance focusses on the concern that individuals have about how their personal attitudes and opinions coincide with the attitudes and opinions of their friends. Thus, if John is friends with Isaac, and Isaac is friends with Kim, then if

John is not friends with Kim there will be a perceived imbalance[3]. Even though the original theory was proposed to explain the more general case of directed relationships, structural balance has been an inspiration to many triadic closure models that aim to mimic the formation of new social ties [BC96].

A crucial property of social networks is that they exhibit high levels of transitivity: high transitivity is not common in other networked systems such as biological and technological ones. In fact, one would expect a certain number of existing triads in a network arising merely because of a certain level of edge density: indeed, such a value can be calculated in a straightforward way in simple random null models, where probabilities of connections between nodes can be theoretically manipulated [EMB02]. Hence, transitivity is often quantitatively measured by counting the number of triangles, or triads, in the social graph, and checking whether this value is larger than one would expect in a random graph of comparable size and density [NWS02]. While in several non-social systems the amount of transitivity present can be explained by simple random models, thus suggesting the absence of significant mechanisms influencing local network structure, the same is not true for social networks. In general, social graphs exhibit a far higher degree of transitivity than their random counterparts [NP03].

From another point of view, the abundant presence of such triangles might be seen as evidence of tightly connected local neighbourhoods or groups. Each social tie belonging to a triangle is a short-range connection between two nodes that are already close to each other, since they share at least one common neighbour. In complex networks literature, this property has also been referred to as *clustering*: in particular, the clustering coefficient of a single node is defined by taking into account the number of triangles present in its neighbourhood with respect to the total number of possible triangles [WS98]. Again, social networks tend to have higher values of the average clustering coefficient than random graphs of comparable size.

Statistical analysis of the evolution of large-scale online social systems suggests that the likelihood that two individuals will establish a social connection greatly increases as they have more friends in common [LBKT08] or as they are affiliated to more common groups or communities [KW06], confirming the transitive nature of online social ties. This has an important impact on link prediction and recommendation systems, which aim to find pairs of disconnected users who are likely to establish a friendship tie.

Implication: Transitivity appears to be a strong influence behind the creation of social ties and its importance for online services is unquestionable. Since spatial distance affects the likelihood of connection between individuals, one would expect

[3]More precisely, Heider's structural balance theory pertains to the perceptual level, thus the perceived existence of a social relationship is far more important than its real existence.

that triads tend to arise between spatially close individuals. However, whether this process is purely driven by spatial factors, by social ones, or by a combination of them is open to discussion.

2.2.3 The small-world effect and navigability

The rich local structure present in social networks, which implies a large number of short-range connections, would suggest that, on average, the path length between two random nodes can be large. For instance, a regular lattice possesses an ordered local structure, with edges between nodes which are close to each other, but longer paths are needed to connect nodes that are further away. However, social networks tend to exhibit short distances between individuals, as found in random graphs that lack clustering.

This concept has been popularised to the general public by Stanley Milgram's famous experiment about social chains of acquaintances [TM69], being termed as the "small-world effect". Milgram's experiment took place in 1967: he asked 160 randomly selected individuals living in the US town of Omaha, Nebraska, to deliver a package to a specified person living in Boston, Massachusetts, about 1,500 miles away. Participants merely received a limited set of information about the target: name, address and occupation. However, participants were not allowed to mail the package directly; they had to rely, instead, on their own friends and acquaintances, forwarding the package to someone they felt could get it "closer", in any sense, to the final destination. These friends or acquaintances were provided with the same instructions, until the package was eventually delivered to the target in Boston. For each transition a postcard was sent back to Milgram with the details of the receiving party.

Even though the majority of these delivery chains did not reach the target, 64 packages were successfully dispatched to the final destination. Among these, the average path length was close to 6 hops, which led to the popular culture reference about the "six degrees of separation" between any couple of individuals on the planet[4].

The apparent incongruity between clustering at a local level, which could lead to long paths as in regular lattices, and the global property of short connection paths was unravelled by Watts and Strogatz in their groundbreaking work [WS98]. A regular lattice structure can be slightly modified by randomly rewiring a few edges, creating long-range shortcuts; even a negligible number of rewired edges dramatically decreases the average path length, while the local clustering remains almost

[4]Milgram never used this exact phrase in his original works; rather, it was later made popular by the Broadway play and, later, by the movie bearing the same name [Gua90]

unchanged. The conceptual implication of the Watts-Strogatz model is that the small-world effect in real networks represents a middle point between a totally ordered and a completely disordered system. Further, the powerful practical implication is that any information or signal can quickly propagate through small-world networks.

The existence of such short paths on the network does not imply that they can be found through decentralised mechanisms with only limited local knowledge. In fact, participants in Milgram's experiment struggled to locate the target in the majority of cases, breaking the chain before completion. The problem of navigability in a social network was initially discussed by Kleinberg considering a simple graph model where long-range connections are added to a regular lattice with probability proportional to $r^{-\alpha}$, where r is the Manhattan distance between nodes [Kle00]. In this model, a simple decentralised greedy search strategy is adopted to forward messages to a target: such a strategy is only successful, with an expected path length $O\left(\log^2 N\right)$ in the number of nodes N, only if $\alpha = d$, where d is the dimensionality of the lattice. Any other value of α results in asymptotically longer paths. This suggests that the correlation between long-range connections and local network structure provides important cues to navigate the network.

Both in Milgram's experiment and in similar more recent reenactments the path followed by the forwarded message moves geographically closer to the target [DMW03]. Together with information about occupation, geographic cues are predominantly used to select the next step in the chain, particularly in early steps [KB78]. The interplay between friendship and geography with respect to the navigability of real-world social networks has been discussed in a work by Liben-Nowell et al. [LNNK+05]. Through simulation of the message-forwarding experiment on a large-scale online social network in the USA, a purely geographic greedy search strategy is shown to achieve results comparable to the original experiment by Milgram. Yet, the social network exhibits long-range connections spanning geographic distance d with probability proportional to $d^{-\alpha}$, with $\alpha \approx 1$. The apparent contradiction between the navigability of the network, which would require $\alpha \approx 2$ as the Earth's surface is two-dimensional, and the spatial distribution of connections is then resolved by considering the large variance in population density and adopting a novel geographic notion, *rank-distance*, to control for heterogeneous population density and hence reconcile the two results.

Implication: As demonstrated by Liben-Nowell et al. [LNNK+05], and also as suggested by Backstrom et al. [BBR+11], geography plays a huge rôle in forming the global structure of social networks, with important consequences for the navigability of the network itself. However, the relationship between the topological position of social connections and their geographic properties is still unknown.

2.2.4 Community structures

Individuals tend to organise themselves into social groups at different scales: families, working and friendship groups, villages, towns and nations. The properties of these groups have been studied for a long time, as they constitute the building blocks of human society [Col64]. Researchers in the social sciences have often investigated the emergence of structurally cohesive groups of individuals and the tendency of the members of such groups to exhibit high levels of homogeneity [WF94].

From a structural point of view, the presence of such groups causes the social network to exhibit specific patterns. More precisely, the structure of social ties between people presents strong local inhomogeneities, with a higher concentration of connections within social groups and fewer links between them. This feature of networks has been termed *community structure* and has been observed not only in social networks but in numerous other networked systems: for instance, links among Web pages reveal communities which likely correspond to clusters of sites related to the same topic [DGP07].

Yet, even though an abundance of measures have been proposed [WF94, GN02, FLGC02, NG04, RCC+04], no clear consensus has been reached on a quantitative definition of community. Several algorithms have been devised to partition a network into communities [New04, DDADG05, For10]: each method relies on a given assumption about which properties a community should exhibit and most of them are computationally intensive and unable to scale to large networks. A fast and widely adopted algorithm is the Louvain method [BGLL08], based on the optimisation of the modularity measure proposed by Newman and Girvan [NG04] and able to scale to large networks.

Finally, Newman and Park suggested that both transitivity of friendship and assortativity can be captured by a simple model where social connections arise from users' affiliations to one or more groups or communities [NP03]. This suggests that the existence of social groups has a strong impact on the properties of social networks.

Implication: The existence of communities in social networks reflects the tendency of individuals to form groups. These groups tend to be constrained by geography, with smaller communities being spread over a smaller area [OAG+11]. Since people close together are more likely to be connected, dense communities might arise merely because of geographic proximity. Researchers have therefore proposed community detection methods that control for spatial proximity to extract more meaningful communities [EEBL11]. Since the relationship between social and spatial factors could be more complex than the one captured by controlling for the distance-dependent probability of connection, a better understanding of spatial patterns of social connections could greatly benefit our understanding of communities.

2.2.5 Homophily, social influence and information diffusion

Homophily is the tendency of people to establish links with other similar and like-minded individuals. Homophily is a powerful driver of social link formation [MSLC01]; several studies have shown that there are important connections between rising similarity between individuals and potential social connections [CCH+08]. Many systems to predict social link formation are based on the assumption that similar users might want to connect with each other [LNK07]. However, homophily can also lead to segregation. Schelling demonstrated that, even in an uncomplicated model, global patterns of spatial segregations can arise from homophily sought by agents at a local level, even if no individual actively seeks segregation [Sch69]. Local homophily can thus easily partition a social system into spatially segregated components.

While similarity tends to foster new social links, existing ties can cause individuals to become more similar by fostering the spreading of trends and innovations [Rog95]. A few studies have investigated the influence that friends can have on product adoption [SS98]. Such viral adoption, that is, behaviour spreading due to person-to-person recommendation, has been intensively researched, trying to understand its dynamics and control its evolution [LAH07]. However, other results suggest that evidence of viral spreading could be greatly overestimated, since a large fraction of such social pressure could also be explained by the fact that connected users are likely to share interests, precisely because of homophily effects [AMS09].

Intimately related to social influence, information diffusion is another important process taking place on online social services. Pieces of information or content items are thought to face lower barriers than products and behaviours when spreading over social connections, quickly and widely reaching a large portion of the network [GGLNT04]. Such an information dissemination process over the network one hop at a time is often referred to as a *social cascade* [BHW92]. Social cascades have been investigated in sociology, economics and marketing for more than 60 years: an eminent example is the threshold model proposed by Granovetter, which models information propagation as a local process depending on friends' adoption [Gra87].

More recently, researchers have harnessed large-scale datasets to track and study social cascades in online services. Adar and Adamic studied the diffusion of information in blogs by adopting epidemic models of spreading [AA05]. Another large-scale characterisation of information cascades using data from Flickr was presented by Cha et al. [CMG09]: their findings show that information does not spread widely through the network, but remains close to the initial seed. A more recent study on URL diffusion over Twitter also supports this claim [RBC+11], since the authors find that cascades extend only over a few additional hops beyond the initiator.

Implication: Overall, there appears to be a strong connection between the be-

haviour of similar individuals and their social connections. This has important spatial implications when social connections tend to be constrained by geographic distance, as viral spreading can be similarly spatially limited. Equally, other dynamic processes taking place on social networks could have a strong spatial component.

2.2.6 Temporal evolution

Networks are dynamic entities: they evolve over time, as their edges are rearranged and altered, but they equally change when new nodes and connections are added and deleted. One of the first attempts to describe the growth of real networks over time is the *preferential attachment* model, put forward by Barabási and Albert to explain the emergence of a scale-free network with power-law degree distribution between Web pages [BA99]. This model relies on a generic growth mechanism where new nodes with constant degree are continuously added to the network and their new connections preferentially attach to already well-connected nodes. Further analysis of the preferential attachment model revealed that it results in a degree distribution where the probability of nodes having k connections is given by $P(k) \propto k^{-\gamma}$, with $\gamma = 3$ [BRST01]. At the same time, several modified versions of the preferential attachment model have been put forward to describe patterns arising in other complex networks; a thorough review is given by Boccaletti et al. [BLM+06].

The preferential attachment model reproduces the stationary properties of scale-free networks, but it also imposes two important constraints on network growth: the average node degree remains constant over time and the network diameter is a slowly growing function of the number of nodes. However, in an extensive study of several networks spanning different domains, Leskovec et al. [LKF05] found that real networks tend to exhibit increasing average degree and decreasing diameter: graphs appear to be *densifying* and *shrinking* at the same time. Hence, they propose two new families of growth models based on community-driven node connections and on a copying mechanism of neighbours' links through an epidemic spreading process.

While online social networks exhibit scale-free structure, they also present high levels of clustering. This appears in contrast with the preferential attachment model, which predicts vanishing clustering as the system size grows. This discrepancy has been addressed by a tunable model that adds a triad formation step to the original mechanisms [HK02]. The final result is that the scale-free degree distribution is maintained but the level of clustering can be increased up to what is found in existing networks. This modification suggests that a triangle closing mechanism could be as important as the preferential attachment principle in driving network growth: both these processes mimic how social networks could grow, namely by means of users connecting to popular individuals and to friends of friends.

The evolution of online social networks at a microscopic level has been quantitatively studied by Leskovec et al. [LBKT08]. Their findings suggest that the three main processes driving network growth are a preferential attachment mechanism for new nodes joining the network, inter-edge waiting times distributed according to a truncated power-law and a simple triangle closing mechanism to provide clustering.

Implication: The two main processes driving the growth of social networks seem to be a global attachment process that favours popular nodes, and a local clustering mechanism that results in triads, communities and clusters. The interplay between these two forces and geographic factors is yet to be understood fully, in order to devise evolutionary models that describe the spatial properties as well as the social characteristics of social networks.

2.3 Location-based social services

The emergence of online social networking services has revolutionised the Web, driving the transition from a static, one-way information transfer to a dynamic, user-generated and interactive communication. Another important trend we have seen is the growing popularity of powerful mobile devices, making available accurate and cheap location-sensing technologies to mobile users.

The combination of these two powerful trends has recently resulted in an innovative generation of services: location-based social networks. These platforms combine geographic services, such as geocoding and geotagging, with online social interactions, enabling users to generate and share information about the locations they visit. In this section we briefly summarise the timeline of location-based social services; we then focus on the concept of place as an integral part of the mobile experience offered to users by these services. Finally, we discuss the influence that these novel location-sharing features might have on online user behaviour.

2.3.1 A brief history of location-based social services

The first attempts at building location-based services with explicit support for user social interactions were mainly research experimental designs, exploring the implications and consequences of these services in a controlled and monitored environment. Paulos and Goodman explored how "familiar strangers" whom people were meeting repeatedly in public spaces could be approached through a mobile application [PG04]. Similarly, Eagle and Pentland designed and built the Serendipity mobile system, enabling users to detect and identify proximate people [EP05]. The semantic concept of place was explored by Wang and Canny [WC06], studying user

reactions to an idea of location going beyond the raw geographic coordinates often used in previous location-based services.

The first large-scale commercial location-based social service to gain appreciable traction among users was Dodgeball. Created in 2000, Dodgeball was a mobile application to distribute location-based information to friends, in order to facilitate social gatherings within cities [Hum07]. Among the many innovations introduced by the service, Dodgeball pioneered the concept of a "check-in" as used by today's location-based social services, in the form of a SMS text message sent to the central server with the indication of the user's current location; this information was then sent to the user's friends, again via SMS messages.

Dodgeball was bought by Google in 2005, which later discontinued the service to make way for their own location-based system, Google Latitude. However, other location-based social services soon followed. Brightkite was founded in 2007 as a social networking website that allowed users to share their location with their friends, providing a crowdsourced database of places that users could access and modify. Similar features were offered by Gowalla, another location-based social network created in 2009; this again supported check-ins, which were shared with friends on the service. Also in 2009, the original creators of Dodgeball launched Foursquare, an innovative location-based social network that added game mechanics on top of traditional place check-ins; the user with the highest number of check-ins in the last 60 days is deemed the *mayor* of a place, encouraging location sharing as users compete to win such "mayorships". Foursquare has since overshadowed the other services, accruing millions of users and becoming the most popular location-based social service available. It now allows users to leave tips about places, create lists of places to visit; it also features a sophisticated place recommendation system.

The landscape of location-based social services includes many other examples. The most popular online social services have launched location-based features, enabling users to specify a particular location when they share an item, a status or a photo. For instance, Twitter offers the option of tagging each status update with a geographic location. Facebook has launched a similar feature, allowing its users to specify a location or a place when posting updates.

These developments are likely to bring location-based features to the vast majority of users, expanding the initial nucleus of eager early adopters. Other services mainly used on mobile devices, such as the popular photo sharing application Instagram or the local businesses directory Yelp, take advantage of a large catalogue of venues to facilitate the creation and retrieval of information related to physical places. Overall, the main technological trend seems to be the convergence of social, mobile and local applications towards a unique user experience, with the potential to bridge the gap between online information and the physical world.

2.3.2 Features of location-based social networks

There are many location-based services available to mobile users, with purposes ranging from searching for local businesses and locating the user on a map, to recommending interesting places to tourists. As online social networks become increasingly location-aware, the line between purely location-based services and more general online social platforms is blurring.

This dissertation focusses mainly on the interplay between spatial factors and online social services; hence, we concentrate on location-based services that feature both a strong social component and an engaging experience revolving around physical places. Our aim is not a complete classification of location-based social services, and we focus our interest on services based on these key concepts:

- users establish social connections among them and can interact with each other, publicly sharing their friend lists;
- users are able to access, search and modify a database of *places* and their related information;
- users mainly interact with the service through a mobile device, which is able to personalise the service based on the user's current location;
- users can voluntarily disclose to the service the exact place where they are, through an action referred to as a *"check-in"*: such information can then be made public or shared only with friends.

As individuals use these location-based social services they leave behind them digital traces of both their social interactions and their spatial movements. These are enriched with data about the nature of the places visited by each user, adding an additional layer of information with rich semantic implications.

Location-based social networks represent the ideal systems to investigate the spatial properties of online social services. First of all, they uniquely and simultaneously provide data about both social connections and geographic movements, making detailed spatial analysis possible. In addition, user location information in these services is often more accurate than text-based descriptions available in other online systems [HHSC11], since it is directly acquired through location-sensing mobile devices. Finally, to this date, these services have already accumulated hundreds of thousands, and sometimes millions, of users, thus enabling large-scale studies that can uncover general properties and trends.

2.3.3 The importance of places as online entities

The main innovation introduced by location-based services is that user activity takes place in the physical world, and not in intangible Web space. Thus, in order to be represented in the online setting, real locations have to be mapped to a new virtual counterpart, the *place*.

These places, also called *venues* in the context of such services, represent references to geographically placed physical entities. The concept of place on location-based services is strongly similar to the concept of *Point Of Interest* (POI), widely used in cartography to signal that a set of coordinates on a map is relevant or interesting with respect to a given context. In our scenario, a place is a human-defined POI; the difference between a location and a place is subtle but important. Whereas a location is a geographical construct immutable over time, like a fixed point on the Earth's surface, a place is a POI loosely coupled with a location. In other words, in theory a place can move to another location and still be the same place, maintaining its semantic implications for users [Gal10].

The importance of representing places on the Web is a theme that goes beyond location-based platforms. In fact, there is a potentially vast set of additional data that can be added to the virtual representation of a place, such as an address, a name, a description, category or type information, contact details, and so on. As online services start offering the ability to create and search for information related to physical places, new applications and systems will be designed and implemented to take advantage of these new layers of information. To achieve this, new standards and methods will be required to harmonise how places are referred to and represented on the Web; the importance of this aspect is confirmed by the existence of the W3C Points Of Interest Working Group, launched in 2010 with a "mission to develop technical specifications for representation of POI information on the Web"[5]. As more and more places are represented on the Web, and as providers exploit this vast catalogue to build location-based services, the connection between the online and the physical worlds will become stronger and more interesting.

2.3.4 The impact of location sharing

The landscape of location-based social services appears exciting and highly dynamic: given their infancy, there are some important factors that are influencing user adoption and, as a consequence, the characteristics of user activity.

Since these platforms are still in a relatively early stage, they tend to attract mainly enthusiastic early adopters; as discussed by Rogers, these users are eager to try new

[5]More information is available online: `http://www.w3.org/2010/POI/`.

innovations and tend to be young, to have high social status, to be highly educated and to be socially open [Rog95]. This means that the user audience of these services is far from being representative of the bulk of Web users, and even more different from the overall composition of the society.

The act of sharing one's location raises important privacy concerns; users are increasingly aware that information about their whereabouts can be highly sensitive and easily misused. For instance, in 2010 a Web service called "Please rob me" [6] was set up to extract automatically Foursquare check-ins publicly shared on Twitter, raising users' awareness about giving away the type of information a burglar would love to have. In addition, privacy groups argue that the privacy policies of online companies collecting location data are "uneven at best and inadequate at worst" [Mor10]. Overall, when users are concerned about sharing their location they could react by selectively avoiding to disclose when they are at certain places or by entirely refusing to join a location-based service.

Finally, these services need to be accessed via applications available only on mobile devices, because they take advantage of location-sensing technology to help the user navigate nearby venues. While smartphones are quickly widening their user base, the considerations about the particular demographics of early adopters equally apply in this case. At the same time, affluent people living in cities are known to be more eager to try online social networking services[7]. This bias could be of greater importance for location-based platforms, as users located in cities with a high density of places are more likely to be enticed to join than users living in less populated and sparse areas that offer less spatial variety. In addition, the potential interest in discovering new venues is higher in dense urban environments than in smaller settings, where users may already be familiar with most of their surroundings.

2.4 Social-based systems and applications

The study of the properties of online social services, which has largely drawn from related findings in physics and social sciences, has several practical applications. In particular, as users spend more and more time using online social platforms, a better understanding of the structural properties of the resulting social graph is needed to design architectures and services appropriately.

In this section we discuss some examples of systems related to online social services, emphasising how spatial information could be exploited to improve existing design

[6] `http://pleaserobme.com/`

[7] This was true as early as 2009 for large-scale services, as presented by a Nielsen report (available online at `http://blog.nielsen.com/nielsenwire/online_mobile/the-more-affluent-and-more-urban-are-more-likely-to-use-social-networks/`).

choices or to create new applications. The mechanisms and factors that drive the temporal evolution and growth of the social network are of paramount importance to design link prediction systems. At the same time, the diffusion of information and trends across social ties and the importance of homophily greatly impact the design of recommender engines. Finally, since online content consumption is increasingly driven by social sharing, a wide range of storage and delivery infrastructure solutions could largely benefit from information extracted from the social connections among users.

2.4.1 Link prediction systems

Social networks are highly dynamic, since they grow and change over time with the addition of new edges as individuals engage in new interactions or lose contact with old acquaintances. Online social services equally exhibit temporal evolution, primarily because new users join and establish new connections from scratch, but also because existing members form new relationships. By understanding the mechanisms by which social networks evolve it becomes feasible to predict which social relationships are likely to appear in the future.

More generally, the problem of *link prediction* is an important task in network science, with important applications in every field that makes use of network models to represent real systems. Despite this, the link prediction problem was initially formulated explicitly for social networks by Liben-Nowell and Kleinberg [LNK07] as the task of accurately predicting the edges that will be added to the network during a future temporal interval, using a snapshot of the network at current time. In this formulation there is emphasis on predictive power rather than on capturing global structural features, such as the degree distribution or the level of clustering. Initial results demonstrated that different proximity measures, based on information entirely contained in the social network itself, can be effectively exploited to algorithmically predict which new ties will occur.

Link prediction methods enjoy increasing popularity thanks to their importance for online social services. Such services strive to entice and retain their users by offering them a pleasant experience, which often involves a rich set of social interactions. Since users with more friendship connections benefit more from the service themselves, systems to recommend and suggest the creation of new online ties are put in place. Such recommending engines largely draw from models that predict which social connections are more likely to develop. Since Facebook launched the "People You May Know" feature [Rat08], devoting vast engineering efforts to finding other members that users might want to add as Facebook friends, it has become customary to deploy such systems on social platforms.

It is interesting to note that these approaches only focus on finding suitable recommendations in a subset of the prediction space: namely, Facebook considers only pairs of users who already share at least one friend. This is due to the fact that in a large graph the total number of node pairs grows quadratically with the number of nodes. Facebook, with its 900 million users, would face a prohibitively large task if it searched the entire prediction space for useful predictions.

More recently, researchers have advocated supervised approaches to link prediction, given the possibility of modelling the task as a binary classification problem. In particular, Lichtenwalter et al. [LLC10] have presented a detailed analysis of challenges in link prediction systems, discussing the extreme imbalance inherent in link prediction tasks, where the number of positive instances is overwhelmed by the number of negative cases. They propose to mitigate these problems by treating prediction separately for different classes of potential friends.

Implication: Information about spatial movements of users across places can reveal a great deal about their preferences and interests, thus improving models of link prediction based on user similarities. In addition, spatial distance could be used as a filtering mechanism to select promising prediction candidates on which more complex models could be run.

2.4.2 Recommender engines

As soon as the Web allowed users to access unprecedented amounts of information, recommender systems emerged to help individuals navigate such large collections of data according to their personal interests. These systems are designed and built around the principle of *collaborative filtering*, the central hypothesis of which is that content items should be ranked "based on the premise that people looking for information should be able to make use of what others have already found and evaluated" [ME95]. The key insight here is that the collective set of user preferences can be used to help each individual. Similarity between users is computed according to their item preferences. Then, user preferences that are unknown for certain items can be predicted by considering ratings by similar users.

Social connections have mainly been introduced in recommender engines as trust relationships; in this context, trust is considered to be the level of belief established between two individuals [JIB07]. In a much broader sense, social trust can be described as the willingness to take some action as the result of receiving information from a given producer [Gol08]. Hence, the main challenge is to estimate the level of trust between users. Trust-based connections can be implicit, when inferred from user preferences over the set of items, and explicit, when based on declared social links between users. In the latter case, connections between users can be established

by requiring individuals to rate other users explicitly and quantify whether they agree with their item preferences, or by asking them to import or otherwise reveal their friendship ties.

While such connections may merely be adopted as an alternative to item-based similarity measures, recommendations also have a powerful social aspect that goes beyond user similarity. In fact, individuals turn to their friends for recommendations in their daily lives, seeking advice about products or content items. Furthermore, users' perception of recommendation is strongly influenced by social aspects; experiments have shown that users overwhelmingly favour recommendations from familiar rather than from similar individuals [BSH07]. Hence, in order to improve recommender systems and to provide more personalised recommendation results, researchers have proposed to incorporate the large amount of social network information available on online services into recommendation engines [MZL+11]. These first attempts demonstrate that social ties might greatly extend the scope and improve the performance of recommendation systems.

Implication: A large fraction of information tends to have a spatial locality of interest, such as news, politics, sports, shopping items and restaurants. Moreover, as the focus of recommender systems shifts from online content to physical entities such as places, the spatial factors shaping social user behaviour are bound to become increasingly more relevant.

2.4.3 Social-inspired system design

A natural way of exploiting the properties of the social connections between members of an online service is to guide and suggest design choices regarding its supporting infrastructure and its software implementation. Given the sheer size of their user audiences and their planetary scope, successful online social services face the problem of replicating as much of their data as possible across the globe, in order to distribute the load on their infrastructure and reduce service latency. Numerous features offered by such services correspond to a many-to-many paradigm, with users simultaneously producing and consuming content through their social connections. This creates many complex inter-dependencies between data items, complicating the design of the service infrastructure and of the distributed storage architecture.

However, by exploiting the complex structure of the social network formed by users' connections, design choices can be made to optimise the services offered to users. Wittie et al. showed that Facebook interactions take place mainly between friends in the same geographic region [WPD+10]; hence, they propose to ease the load experienced by a centralised server infrastructure by introducing distributed regional proxies, reducing latency experienced by users at the same time. The inherent

community structure arising from social connections has been exploited to partition service users into communities and, thus, optimise data distribution across storage locations, as done in a company's email network [KGNR10] and on Twitter [PES+10].

Since social connections between users drive content consumption on online services, other attempts have tried to improve how such content is delivered and served. Silberstein et al. [STCR10] sought to optimise personalised news feeds by differentiating users based on their content production rate. By querying in real-time content produced by high-rate users and caching content created by low-rate users, they reduce both latency and system load. The viral diffusion of content across online social connections has been exploited to rank popularity of content items dynamically based on the fraction of social-generated requests they experience, and to optimise content storage across disks in order to reduce energy consumption [SC10].

Finally, by observing network properties of individual users it becomes possible to find outliers that substantially deviate from expected behaviour; these individuals could be malicious users or could signal that a particular user account has been compromised. For instance, a security mechanism devised by Backstrom et al. [BSM10] is based on a probabilistic framework to predict the geographic whereabouts of Facebook users by inspecting where their friends are located; when a user logs in from an unexpected location, additional security mechanisms could be set in place to prevent fraudulent activity.

Implication: The geographic properties of online social services have a huge potential to inspire the design of systems and infrastructure; in fact, some research attempts have already explored ways of exploiting some characteristics such as spatial locality of access patterns. However, a better understanding of the spatial characteristics of dynamic social processes would greatly expand the scope of social-inspired system design.

2.5 Present dissertation and future outlook

This chapter has reviewed the different types of online social services available to users and the most important characteristics of the social networks arising among their members, introducing also the new generation of location-based social services and their particular characteristics. We then discussed systems exploiting the characteristics of online social services.

Since social ties are constrained by spatial distance, our discussion mentioned several times its effect on all the main characteristics usually exhibited by online social networks. Furthermore, the emergence of the mobile Web makes location-based services

a pervasive reality, allowing the introduction of a novel entity in the online realm: the place. This innovation enables users to interact not only among themselves, but also with spatial entities, revealing their whereabouts and their usual geographic movements. In summary, the spatial properties of online social services are increasingly more accessible and, at the same time, more important to understand their structure fully and to design effective systems and applications.

This dissertation is a step in this direction. Given the impact that geographic space has on online social ties, as discussed in Section 1.1, we plan to study and understand the relationship between social and spatial factors. To this end, we present a comparative study of the spatial properties of different online social services and we discuss new measures that can be used to capture social and geographic characteristics of online users simultaneously (Chapter 3). These findings are then extended to define a temporal model of network growth that reproduces social and spatial patterns seen in real data (Chapter 4).

We also present two practical applications that stem from the availability of spatial data about online social networks: a link prediction engine based on the places visited by users (Chapter 5), and a family of caching policies for distributed content delivery networks that exploit content diffusion over the social graph to discover geographic patterns of item popularity (Chapter 6).

Given the results and findings about online social networks discussed in this chapter, many research directions that consider space and geography could be further explored. We will consider them at the end of this dissertation (Chapter 7).

Science is what we understand well enough to explain to a computer. Art is everything else we do.

Donald Knuth

3

Measurement and structure

As discussed in the previous chapter, location-aware capabilities are gradually being offered as a feature by many online services. People appear more willing to share information about their geographic position with friends, while companies can customise their services by taking into account where the user is located. Therefore, online platforms are able to gather data not only on social interactions, but also on users' physical movements. Such information is available for the first time at unprecedented scales, posing new questions about what spatial properties of user behaviour it might reveal.

In this chapter our aim is to exploit this simultaneous availability of social and spatial data given by online services to study the spatial properties exhibited by the social connections arising among users. By embedding the social graphs in physical space, we will focus on understanding whether, and how, spatial and social characteristics are related to each other. Our findings suggest that geographic space influences users in a heterogeneous way; hence, we will present new social network measures that simultaneously capture social and spatial aspects. These measures allow users to be distinguished quantitatively according to their geographic and social characteristics, enabling comparative analysis of different spatial social graphs.

Chapter outline In Section 3.1 we present and describe our data collection methodology, which has allowed us to acquire extensive traces about the geographic and social properties of three popular online location-based social services. These

digital records about the interplay between online friendship ties and user locations make it possible to pursue a set of further investigations.

In Section 3.2 we take advantage of user check-ins to assign a geographic position to each individual, embedding social connections in space. This allows us to study the spatial properties of online friendship ties, with emphasis on the effect that geographic distance might have on online relationships. Our analysis focuses on whether space homogeneously affects users or if, instead, different individuals tend to have online ties spanning different geographic distances, leading to a more heterogeneous system. By adopting randomised null models we are able to investigate the statistical significance of the empirical properties found in these networks. Furthermore, we discuss how social and spatial factors jointly influence the structure of connections between users, resulting in heterogeneity between users.

Since our analysis suggests that space influences the social connections established by online users, we then aim to include the effect of spatial distance in standard network measures. Thus, in Section 3.3 we define two novel geosocial measures that combine standard network properties with spatial distance. We demonstrate the effectiveness of these measures in capturing the impact of geographic distance on online social ties with a case study; we compare location-based social services to other online platforms where location-sharing features are not dominant, discussing observed similarities and differences. Using our measures we discover that spatial constraints do not uniformly affect different categories of online social services; users of location-based services exhibit a stronger preference for short-range social connections than users of content-sharing platforms. We discuss the implications of our findings in Section 3.4, while Section 3.5 reviews related results and Section 3.6 concludes the chapter.

3.1 Data collection methodology

Our first goal has been to collect extensive traces about the spatial and geographic properties of online social services. We have acquired traces from three different online social platforms. Each service offers location-sharing features, such as check-ins, or status updates where places are tagged, as well as social networking features.

As we detail in this section, for each service we gathered data about both social ties and user check-ins. This allows us to extract the social network arising among users and to assign a geographic *home location* for each user, effectively embedding the social graph in geographic space. When users do not explicitly specify their location, we assign to each user the geographic location of the place where he/she has made the greatest number of his/her check-ins. We discard users who have no check-ins.

3.1.1 Brightkite

Brightkite was founded in 2007 as a social networking website that allowed users to share their location, to post notes and to upload photos through different interfaces. It was initially based on the idea of performing check-ins at places, where users can see who is nearby and who has been there before. Brightkite users could establish bidirectional friendship links and send public and private messages to each other. Brightkite has recently transformed its service, offering a group-messaging mobile application.

Our measurement took place when the service was still mainly based on mobile location-sharing. Brightkite offered a public API which provided geographic coordinates of user home locations and lists of friends. We fetched data from their API in September 2009; we seeded a crawler with 1,000 users randomly selected from the public timeline offered by Brightkite and then we exhaustively retrieved friends and followed social links until we collected all users connected to the seed users. The resulting dataset contains information about 54,190 users; it represents a complete snapshot of a location-based social platform in its initial evolution phase, including the entirety of its social graph.

3.1.2 Foursquare

Foursquare is a location-based social networking service launched in 2009 which engages its users in a competition. Users check in at venues in order to be awarded points that contribute to their position on a leaderboard. In addition, users can publish and share tips and suggestions related to the places they visit. The service also enables the creation of bidirectional friendship links.

Foursquare does not provide public access to user check-ins. However, many Foursquare users choose to push their check-in messages to Twitter automatically, which provides a public API to search and download these messages. By using this API, we have recorded approximately 4 million tweets, each one containing a check-in made by a Foursquare user during June 2010. These messages come from about 250,000 different users and cover about 1.5 million locations on the planet. We estimate that our sample contains approximately 20% to 25% of the entire Foursquare user base at collection time, or between 40% and 50% of all active Foursquare users [1].

Since Foursquare does not provide unauthorised access to users' friend lists, we have acquired friendship ties that Foursquare users have between them on Twitter, where they are publicly available, to extract a social network. Our assumption is that a friendship connection between two Foursquare users is likely to be present also on

[1] Foursquare reached 1 million users in April 2010 [Sie10].

Twitter, if the two users are Twitter users as well. While this resulting social graph is not expected to be identical to the original Foursquare graph, we will show that it conveys meaningful information, providing results comparable to the other datasets.

3.1.3 Gowalla

Gowalla is a location-based social networking service created in 2009. Users can check in places through a dedicated mobile application; such check-ins are then pushed via notifications to other friends on the service and, by linking accounts, to Twitter and Facebook. The friendship relationship is mutual, requiring each user to accept friendship requests to allow location sharing. Gowalla was discontinued at the end of 2011 as the company was acquired by Facebook.

We acquired a complete snapshot of Gowalla in August 2010. Gowalla provided a public API offering access to information about user profiles, friend lists, user check-ins and place details. For every user we have gathered the user profile, the friends list and the list of all the places where the user has checked in. The API did not provide unauthorised access to fine-grained temporal information about user check-ins. However, the API presented the timestamps of the earliest and latest check-ins for each place where a user had checked in. Since users were identified by consecutive numeric IDs, we were able to exhaustively query all user accounts and download all the aforementioned information.

3.2 The spatial structure of online social networks

Our exploration now focuses on understanding the spatial structure of the social graphs in these online services. Specifically, we want to focus at first on global properties, studying the effect of geographic distance on online ties and the likelihood of connection between individuals at different distances from each other. Then, given the heterogeneity shown by individuals in social networks, exhibiting a wide variability in their number of connections, we expect their spatial properties to be similarly heterogeneous. In other words, space could homogeneously affect users or, instead, some individuals may exhibit a preference for long-distance connections.

In this section we address these issues with a comparative study of data collected from the three services described previously. At first, we focus on the global spatial properties of the social graphs: we study whether distance affects social ties by looking at the likelihood of connection between users as a function of their geographic distance. We then shift the focus of our analysis to individual users, showing that both their connections and the triads they belong to are influenced by spatial proximity.

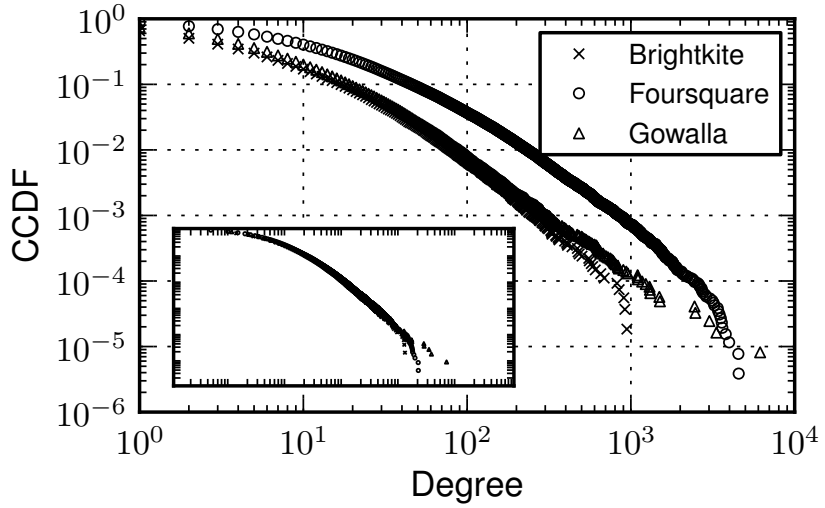


Figure 3.1: Empirical Complementary Cumulative Distribution Function (CCDF) of the number of friends for each user in Brightkite, Foursquare and Gowalla. The inset shows the same distributions rescaled by dividing by the average number of friends in each network; the three datasets fall on the same curve.

Service	N	K	N_{GC}	$\langle k \rangle$	$\langle C \rangle$	H_{EFF}	$\langle D \rangle$	$\langle l \rangle$
Brightkite	54,190	213,668	50,896	7.88	0.181	5.73	5,683	2,041
Foursquare	258,706	2,854,957	254,532	22.07	0.191	5.90	8,494	1,442
Gowalla	122,030	577,014	116,910	9.28	0.254	5.43	5,663	1,792

Table 3.1: Properties of the traces: number of nodes N and edges K in the social network, number of nodes in the largest connected component N_{GC}, average node degree $\langle k \rangle$, average clustering coefficient $\langle C \rangle$, 90th percentile of shortest path length distribution H_{EFF}, average geographic distance between nodes $\langle D \rangle$ [km], average link length $\langle l \rangle$ [km].

We assess the statistical significance of the observed spatial properties by designing two randomised null network models; the comparison of the real graphs with the null graphs sheds light on whether spatial and social factors are influencing the structure of the network. Overall, we observe robust and universal features across the three services; this suggests that there might be social processes that are not specific to one particular online tool adopted by users.

3.2.1 Global spatial properties

We study the spatial properties of the social networks by representing them as *spatial graphs*, where nodes are positioned in a space equipped with a metric. In our case, online users are located over the 2-dimensional surface of the Earth and we adopt the great-circle distance as our metric; the distance D_{ij} between any two nodes i and

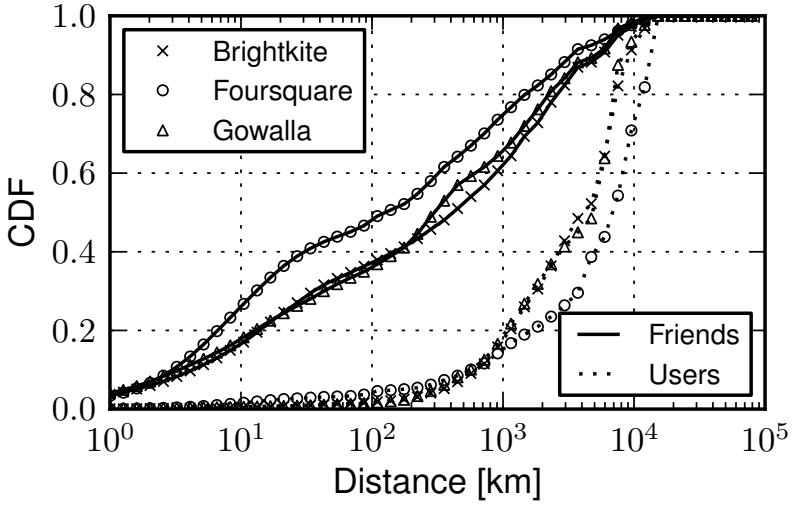


Figure 3.2: Empirical Cumulative Distribution Function (CDF) of the geographic distance between all users (dotted line) and between connected friends (solid line) for the three datasets. Logarithmic binning has been adopted to estimate the number of samples in each range of values.

j is then computed given their geographic coordinates. Then, the social network can be represented as an undirected graph G with N nodes and K links, with users as nodes and where a link exists for each social tie (i.e., a person lists another user as one of his/her friends). We assign a length l_{ij} to each social link so that $l_{ij} = D_{ij}$. Each social link may be undirected or directed; in the latter case, the existence of a link from node i to node j does not imply the existence of the reverse link from j to i. Unless explicitly specified, we always consider undirected connections in our graphs.

The general properties of these three datasets are reported in Table 3.1. The social networks are heterogeneous in size, ranging from 54,190 nodes in Brightkite to 258,706 in Foursquare; the average degree is lower in Brightkite and Gowalla, respectively 7.88 and 9.28, than in Foursquare, where users have on average 22.07 friends. In all the networks, the largest connected component, also known as the giant component, contains the vast majority of the users. The degree distributions for the three networks are reported in Figure 3.1; they all show a heavy tail, with some users having thousands of friends. Rescaling the degree distributions by dividing by their average values results in a common trend, as shown in the inset. These networks also exhibit high values of the average clustering coefficient, between 0.18 and 0.26, and short topological distances between their nodes, with 90% of all pairs being fewer than 6 hops away from each other. These properties confirm the small-world nature of these networks, as found in many other online social systems.

The average geographic distance between users $\langle D \rangle$ is consistently larger than the average distance between friends $\langle l \rangle$ across all the datasets; while the first value

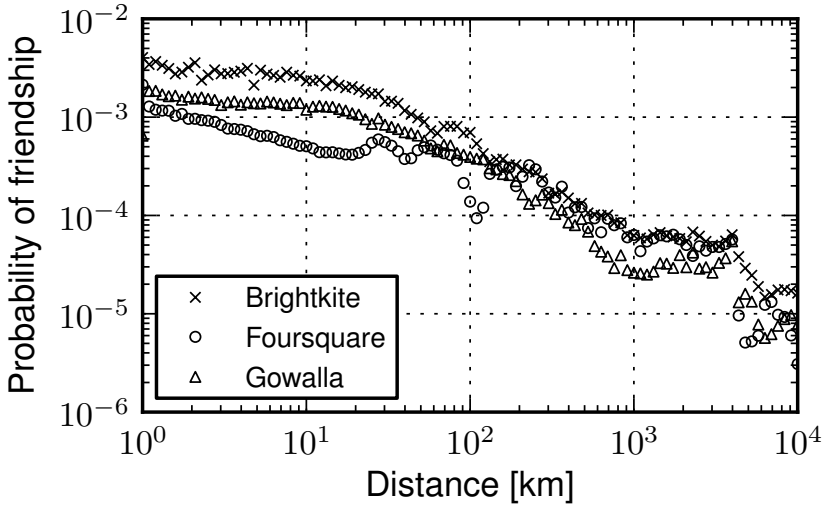


Figure 3.3: Probability of friendship between two users as a function of their geographic distance, for the three datasets under analysis. Logarithmic binning has been adopted to estimate the probability in each range of values.

ranges between 5,600 and 8,500 km, the latter has much smaller values, between 1,400 and 2,000 km. This already provides evidence that the probability of a social link between two users decreases with distance; we will further investigate this relationship later. The distribution of social link length is comparable across the three datasets, as shown in Figure 3.2: about 40%-50% of all couples of friends are within 100 km, with more than 3% of all links being shorter than 1 km. The distribution of distances between users, also depicted in Figure 3.2, is different; about 50% of users are at distances larger than 4,000 km, across all the networks.

3.2.2 The effect of distance on online friendship

To further investigate whether social links are more likely to exist between geographically close rather than distant users, we estimate the probability of friendship $P(d)$ as a function of distance d by counting L_d, the number of social links with length d, and by estimating N_d, the number of pairs of users at distance d. This gives us $P(d) = L(d)/N(d)$. As discussed before, this relationship has been found in other studies to be close to a law $P(d) \sim d^{-\alpha}$, with values of α ranging between 1 and 2.

As reported in Figure 3.3, our datasets present noisy patterns: we notice an almost flat probability in the range 1-10 km, then all curves decrease as distance grows, reaching another steady probability between 1,000 and 4,000 km. This final plateau might denote a background probability that affects ties spanning more than 1,000 km. Similar constant trends at short and long geographic ranges have also been found in other online systems [LNNK+05, BSM10]. The appearance of social ties longer than 4,000 km is constrained by the fact that both Europe and North America,

where a large proportion of users are based, are not large enough to allow such long-range connections and the distance between these two regions is about 6,000 km.

We find that our traces are closer to a decay $d^{-\alpha}$ with $\alpha = 0.5$, whereas larger exponents have been found in other similar studies; it seems that in the location-based services under analysis long-range social ties have a relatively higher probability of occurrence than in other social systems. A potential explanation of this behaviour is that these platforms are relatively new, so they have mainly attracted early adopters. These users tend to be tech-savvy, with many existing long-distance online friendship ties which they bring to these services. This might not happen in other types of social networks, such as those extracted from mobile phone communications or Facebook interactions, which are already mature.

Indeed, mobile phone connections exhibit a larger exponent α than online social networks: phone conversations are much more constrained by geographic distance than interactions on Facebook. This might be due to the fact that mobile phones represent a mature technology, adopted by the vast majority of the population. Also, mobile phone calls could often be exchanged to arrange face-to-face meetings, which take place between spatially close individuals, as discussed by Calabrese et al. in a large-scale study of mobile phone call records [CSBR11]. As location-based services become more mainstream their user audience may broaden and include individuals who are affected by distance in a stronger way.

Decoupling social and spatial factors: network randomisation

After these initial investigations, we assess the statistical significance of the empirical spatial properties of these networks using two *randomised models*, which capture either the geographic or the social properties of the original social networks and randomise everything else:

- **Geo**: this null model keeps the user locations unmodified and then creates a social link between two users at distance d according to the relative probability of friendship $P(d)$ (as reported in Figure 3.3).
- **Social**: this null model keeps the social connections as they are, shuffling at random all user locations.

The overall properties of these models are a direct consequence of their definition. Both models result in a network with exactly the same number of nodes and, on average, the same number of edges. The Social model has the social properties of the original network, including degree distribution, clustering coefficients and topological network distances, but link geographic lengths are distributed as the

pairwise user spatial distances: as a result, the average link length becomes higher than in the original network, with $\langle l_{SOCIAL} \rangle = \langle D \rangle$. On the other hand, the Geo model has the same distribution of link geographic lengths as the original network, so that $\langle l_{GEO} \rangle = \langle l \rangle$, but the social properties are now lost: the degree distribution is peaked and has no heavy tail, while the average clustering coefficient is much lower, since there are fewer social triads. However, the two network models present similar distributions of topological distances, with about 90% of all pairs of nodes always within 6 hops.

We will exploit these two null models by comparing their properties to those of the real networks, in order to understand whether the observed socio-spatial characteristics might be explained in terms of simple geographic or social factors. Every analysis performed on a randomised null model will be averaged over 100 different random realisations.

3.2.3 User spatial properties

We now focus on individual users, studying the extent to which their social ties stretch across space. We define

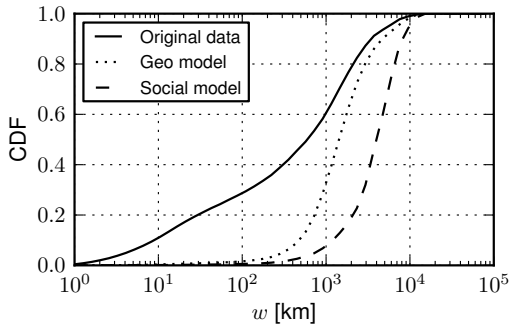
$$w_i = \frac{1}{k_i} \sum_{j \in \Gamma_i} l_{ij} \tag{3.1}$$

to be the *friend distance* of user i, where Γ_i is the set of neighbours of node i and $k_i = |\Gamma_i|$ is its degree. The overall distribution of w is reported in Figure 3.4 for the original social network and for the two randomised versions. The existence of values over all geographic scales is due to the existence of users with different characteristic lengths of interaction. For instance, about 10% of users have connections with an average length shorter than 10 km, whereas around 20% of users have values of friend distance above 2,000 km. Links with different geographic lengths do not appear homogeneously across all users; instead, there is heterogeneity between users, with some having only short-range connections and others with long-distance ties. These correlations are stronger than one would expect by chance; in fact, the two randomised models suggest that values of w should be more peaked around the average, not spread over a large range of magnitudes.

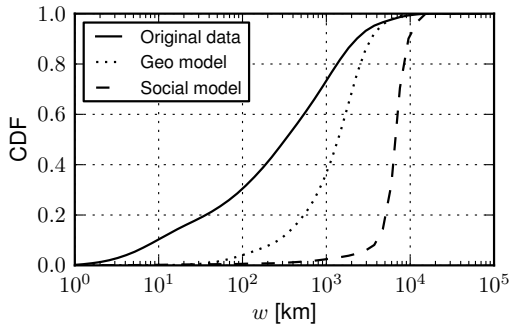
Another interesting result is obtained by studying the correlation between the friend distance w_i and the degree k_i. We study the user *distance strength* [BBPSV04], defined as

$$s_i = \sum_{j \in \Gamma_i} l_{ij} = k_i w_i \tag{3.2}$$

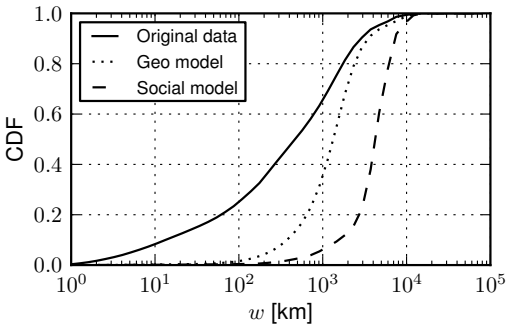
and then we compute the average distance strength $s(k)$ for all users with degree k. In the absence of any correlation, this measure should be linearly correlated with the



(a) Brightkite



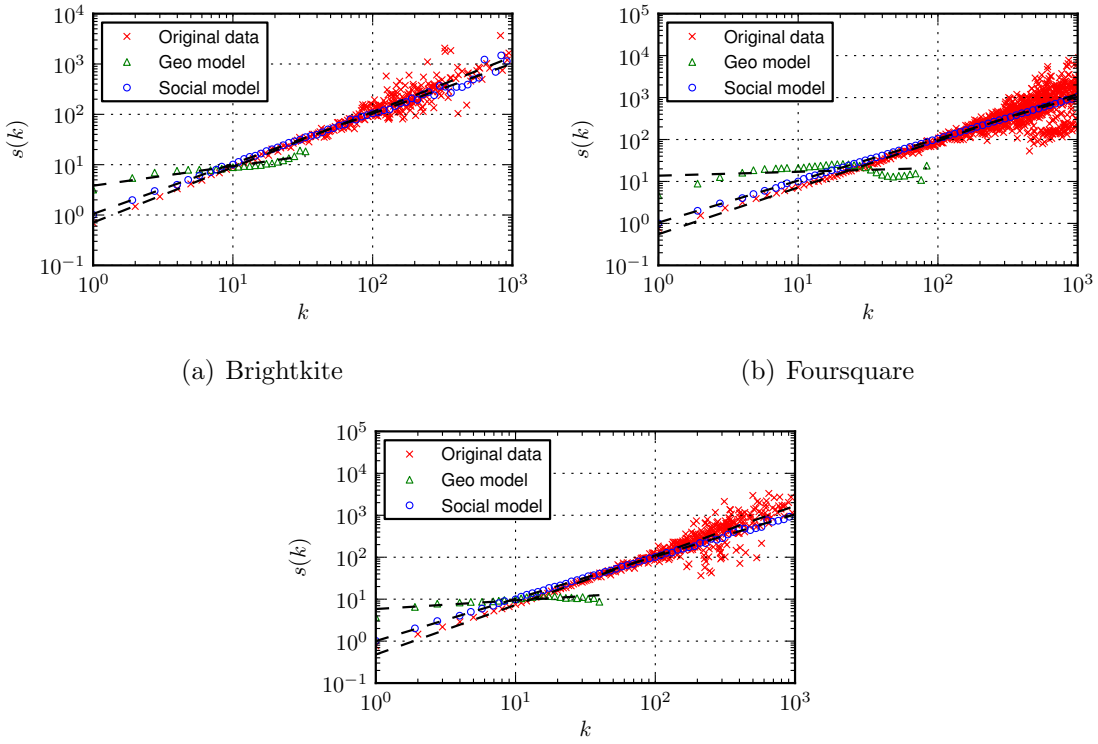
(b) Foursquare



(c) Gowalla

Figure 3.4: Empirical Cumulative Distribution Function (CDF) of the friend distance w for each user in the social network, together with the distributions obtained in the randomised models.

degree $s(k) \sim k\langle l \rangle$, while a relation of the form $s(k) = Ak^{\beta}$ with $\beta \neq 1$ or $A \neq \langle l \rangle$ would imply correlation between the distance strength and the degree. In particular, $\beta > 1$ signals that users with more friends tend to have longer connections than users with fewer friends. This relationship is reported in Figure 3.5 for the three datasets under analysis: we obtain values of β in the range 1.10-1.18 across the different networks, showing weak positive correlation. Real data reveal a pattern much closer to the Social model, which has $s(k) \sim k$, with $\beta = 1$, than to the Geo model, which has much lower values of β in the range 0.2-0.4, showing negative correlation between node degree and friend distance. The case of the Geo model suggests that if only spatial factors were shaping social connections, then users would accumulate several links only when these links are predominantly covering short geographic distances. In reality, as users add more and more friends, on average their link length slightly increases. This contrasts with what is found in the null models, providing evidence that users with more connections tend to have friends further away.



(a) Brightkite

(b) Foursquare

(c) Gowalla

Figure 3.5: Average distance strength $s(k)$ as a function of node degree k for the original network and for its two randomised versions. Each trend is fitted by a law $s(k) \sim k^{\beta}$.

3.2.4 Spatial properties of triads

We now shift our attention to understanding the geographic properties of social triangles. Social networks usually present many triads, resulting in high values of clustering coefficient. Our networks exhibit similar patterns, with clustering values between 0.18 and 0.26. We extract 377,438 triangles in the Brightkite social networks, 18,764,129 in Foursquare and 1,327,559 in Gowalla. Between 70% and 86% of all links in each social network belong to at least one triangle, given the highly clustered structure of these social networks. We find that social triangles arise at a wide range of geographic lengths. Investigating the probability that a link belongs to a triangle as a function of its length provides a surprising result: this probability is largely unaffected by distance, as shown in Figure 3.6. A link is equally likely to belong to a social triangle regardless of its length. A related result was found by Lambiotte et al. [LBD+08]: many spatially local clusters of people tend to appear in mobile phone communication networks, with social links below 40 km more likely to participate in social triads, but then this likelihood reaches a constant value for longer links. As we have already seen, online behaviour appears less sensitive to distance than mobile phone communication. Overall, the trend that longer links

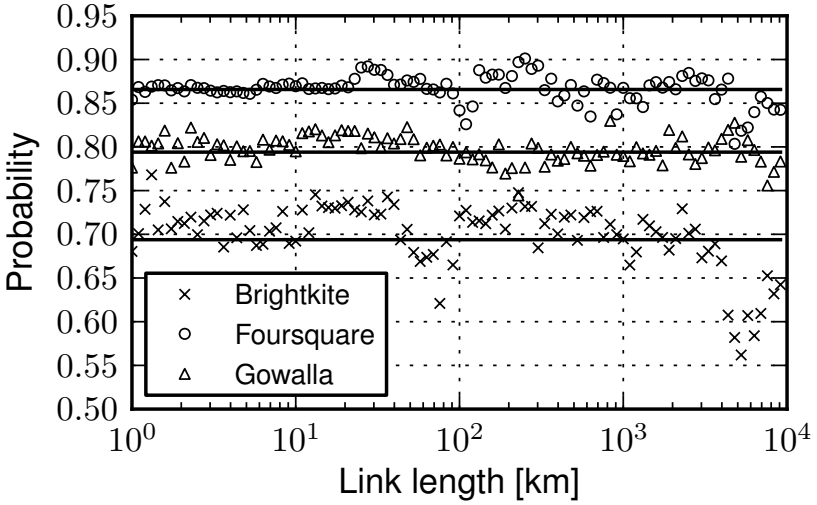


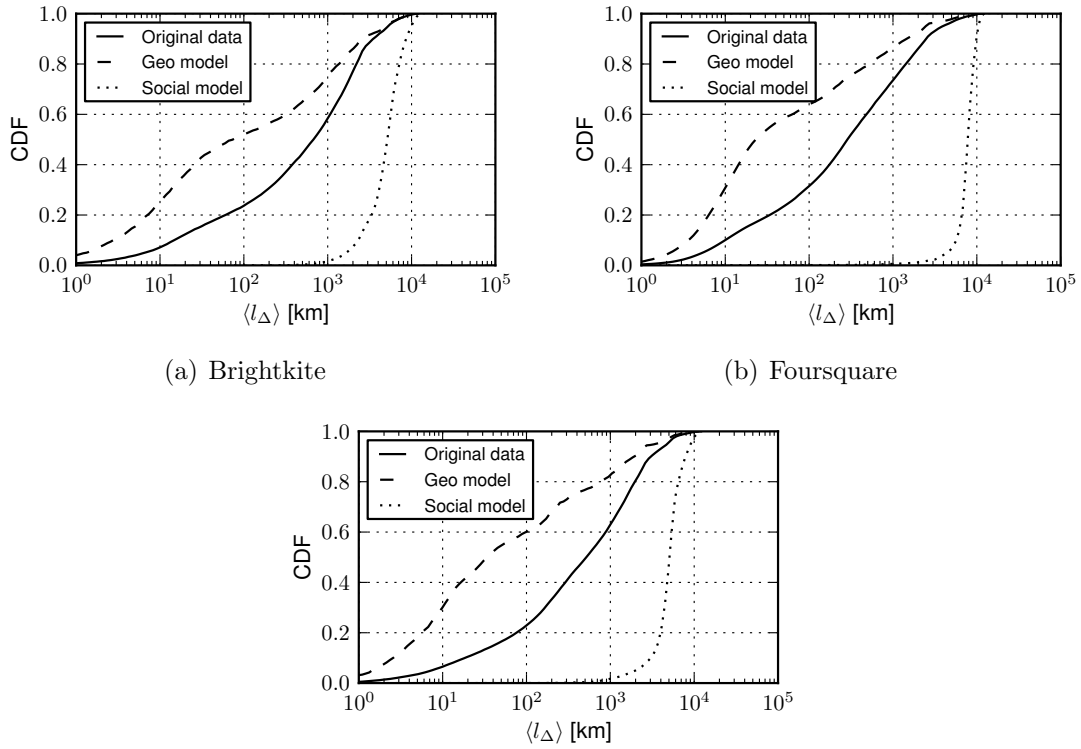
Figure 3.6: Probability that a social link belongs to a triangle as a function of its geographic length for the three datasets under analysis. The solid lines show the average probability that a link belongs to a triangle for each network.

equally participate in social triangles holds also in our datasets.

To assess user heterogeneity, for each social triangle we compute the *geographic triangle length* l_Δ. That is, we compute the arithmetic average of the spatial length of the three links that constitute the triangle. Then, we compute $\langle l_\Delta \rangle_i$ for each user i, which averages l_Δ over all the triangles he/she belongs to. This value does not take into account how many triangles a user might belong to, as the clustering coefficient does so by normalising with respect to the number of potential triangles. Instead, we aim to assess merely the geographic span of a user's social triangles, however many there are. In Figure 3.7 we show the distribution of $\langle l_\Delta \rangle$ over all users: triangles with different geographic span do not arise equally among all users, but instead there are users with smaller triads and users with wider ones.

For example, there are at least 20% of users with an average triangle length below 100 km, while the top 20% have values above 2,000 km. This heterogeneity is much higher than one would expect if space did not matter, as the Social model mainly exhibits values in the range 1,000-10,000 km. However, if social mechanisms were not a factor at all, then social triads should be smaller, as the Geo model shows. The existence of both local, short-range triads and global, long-distance ones needs to be related to the influence both of geographic distance and of social processes such as homophily, triadic closure and focus constraint [MSLC01, Gra73, Fel81].

We further study this heterogeneity arising among users by computing the average $\langle l_\Delta \rangle$ for users with a given degree, as a function of the degree. In these social networks $\langle l_\Delta \rangle$ increases with the number of friends, as shown in Figure 3.8. This effect is not present in the randomised networks: the Social model shows no correlation at all,



(a) Brightkite

(b) Foursquare

(c) Gowalla

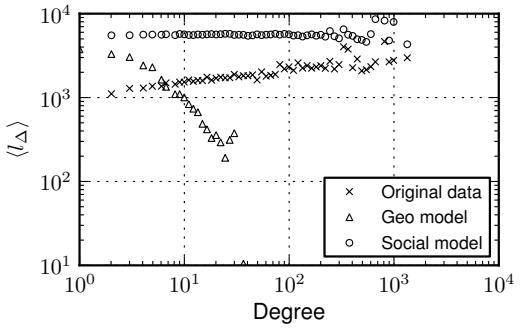
Figure 3.7: Empirical Cumulative Distribution Function (CDF) of average triangle link length $\langle l_\Delta \rangle$ for the original network and for the two randomised models.

while the Geo model exhibits the opposite trend, with smaller triangles appearing among users with higher degrees. Apparently, there are both social and geographic factors influencing social triangles, since having only one type of factor does not reproduce the empirical data.

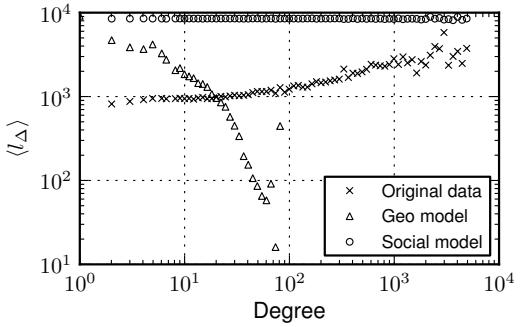
These results signal that users with fewer friends tend to generate social triangles on a smaller geographic scale, while users with more friends belong to triangles with longer links. This confirms the strong connection between the number of connections of a given user and the geographic distance of these friendship connections.

3.3 Geosocial network measures

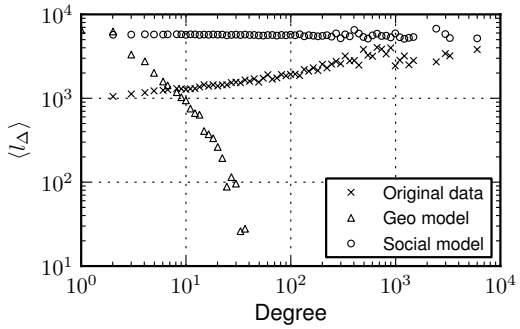
Online social networks are affected by geographic distance but, as discussed in the previous section, users exhibit large variations in the spatial characteristics of their social ties. Since network measures have been used extensively to differentiate the social properties of users, our aim moves now towards defining measures that combine network properties and spatial distance, allowing us to identify also geosocial differences. Measures that augment social structure with geographic information would add a new dimension to social network analysis and could enable a large



(a) Brightkite



(b) Foursquare



(c) Gowalla

Figure 3.8: Average geographic triangle link length $\langle l_\Delta \rangle$ for all users with degree k, as a function of degree k. Results for the original network and for the two randomised models.

number of practical applications related to social systems. One example will be presented in Chapter 6.

In this section we present two novel geosocial measures: the *node locality*, which quantifies how much an individual is connected to a local rather than global set of friends, and the *geographic clustering coefficient*, which extends the standard notion of clustering coefficient by taking into account the extent to which clusters of people are connected by short-range ties. These measures offer a way to compare individual users inside the same network but also to compare different spatial social networks, regardless of the particular geographic space they span.

In order to explore the benefits offered by our measures, we apply them to four different online social networks which provide location information for their users. We compare social graphs with different spatial properties by choosing two location-sharing social services, one blogging community and a social micro-blogging platform. The range of functionalities offered by these services and the differences arising with respect to their semantics will enable us to use our measures to understand the effect of spatial distance across different online platforms.

In particular, using our new measures we show that the four services present con-

trasting characteristics, which may be explained by different attitudes of their users towards the social and geographic aspects of online friendship. Location-based platforms exhibit users with higher preference for short-range social connections than sharing-based services, where social ties appear less constrained by spatial distance.

3.3.1 Augmenting network measures with spatial distance

Our main motivation to introduce these measures is given by the heterogeneity we have observed across users in Section 3.2: in online social services users with social ties, and social triangles, spanning only short distances coexist with users having long-distance connections. Even though we have observed correlations of spatial properties with the number of connections users have, individual users could still deviate from the average behaviour.

When defining our new geosocial network measures our chief objective is to individuate users who predominantly exhibit short-range connections: this might be useful for a wide set of purposes, such as geographically caching data related to online interactions [WPD+10] and distributing content items by sending them where traffic requests are anticipated to arise [THT+12]. Our aim is to create measures that take into account the fact that users have different numbers of connections; we achieve this by introducing normalisation by node degree. Our measures should capture spatial properties regardless of the geographic scale of the social system. In other words, each measure should be related to the spatial size of the social graph, so that one can compare social connections within a city to connections between individuals across a large country. We will highlight in our discussion how our measures meet these requirements. The definitions of our new network measures are based on the spatial representation of a social graph already introduced in Section 3.2.1.

Node locality The first measure quantifies the geographic closeness (i.e., the locality) of the neighbours of a certain node to the node itself. Let us consider an undirected geographic social network, a node i with a particular geographic position and the set Γ_i of its neighbours. The node degree k_i is the number of these neighbours, that is $k_i = |\Gamma_i|$. Then, the *node locality* of i can be defined as a measure of how geographically close its neighbours are and it is computed as follows:

$$NL_i = \frac{1}{k_i} \sum_{j \in \Gamma_i} e^{-l_{ij}/\beta} \tag{3.3}$$

where β is a scaling factor to avoid extremely small values of node locality when links have large lengths. This definition fits the three main requirements that we discussed earlier. By definition, NL_i is always normalised to be between 0 and 1, where the latter is achieved only when all friends of a user are in the same location

as the user herself. The value of β can be chosen so that networks with different geographic size can still be compared to each other, as we will discuss more in detail later. We adopt an exponential decay to highlight social ties that span over short geographic distances and to penalise longer ties. Finally, normalisation by node degree is needed to take into account the huge heterogeneity observed in the degree distribution of such graphs.

In a similar way, in the case of directed graphs the *node in-locality* can be defined considering only the incoming connections of a node

$$L_i^{IN} = \frac{1}{k_i^{IN}} \sum_{j \in \Gamma_i^{IN}} e^{-l_{ji}/\beta} \tag{3.4}$$

where Γ_i^{IN} is the set of its incoming neighbours and $k_i^{IN} = |\Gamma_i^{IN}|$ is the in-degree of the node. The *node out-locality* is defined in a similar manner considering only outgoing links:

$$L_i^{OUT} = \frac{1}{k_i^{OUT}} \sum_{j \in \Gamma_i^{OUT}} e^{-l_{ij}/\beta} \tag{3.5}$$

Geographic clustering coefficient While node locality captures how close the neighbours of a node are, another measure is needed to quantify how connected the neighbourhood of a node is over space. The *geographic clustering coefficient* can be defined as an extension of the clustering coefficient used for complex networks. The clustering coefficient measures the fraction of existing triangles among the neighbours of a given node, compared to the number of possible triangles. This geographic adaptation gives more weight to triangles when they are formed by nodes that are close to each other than when nodes are at a greater distance. The geographic clustering coefficient of node i is thus defined in the same way as the clustering coefficient, but each existing triangle between nodes i, j and k is assigned a weight w_{ijk} defined as:

$$w_{ijk} = e^{-\frac{\Delta_{ijk}}{\beta}} \tag{3.6}$$

where Δ_{ijk} is the maximum length among the three links, that is $\Delta_{ijk} = max(l_{ij}, l_{ik}, l_{jk})$. We define $w_{ijk} = 0$ if there is no link between j and k.

Since this measure uses the maximum weight among all the links of a triangle, it focusses on nodes that are all close to each other: when just one of the three nodes is not close to the other two, the weight will immediately decrease. This emphasises social triangles where the three users are close to each other. Again, the parameter β is used to scale the values of the measure with respect to the geographic span of the entire network.

In the case of directed graphs, as for the standard clustering coefficient, we consider triangles containing undirected links joining node i to its neighbours and directed

links for the remaining side. If we consider Γ_i as the set of all the neighbours of node i (considering both incoming and outgoing links), with $k_i = |\Gamma_i|$, the geographic clustering coefficient is defined as:

$$GC_i = \frac{1}{k_i(k_i - 1)} \sum_{j,k \in \Gamma_i} w_{ijk} \tag{3.7}$$

where the sum is extended only to existing triangles. Since there are exactly $k_i(k_i-1)$ different ordered couples of neighbours in Γ_i, GC_i is normalised to be between 0 and 1 by definition. This definition also fits the same requirements met by node locality.

Choice of scaling factor When using these geosocial measures on different spatial networks, one should be able to compare results across them regardless of the specific geographic span. These measures should scale with the geographic size of a system, so that when the system is enlarged, or made smaller, the measures do not change. For instance, one could compare a spatial social graph arising between individuals in an urban area, with a maximum span of about 50 km, and a graph formed from connections across a large country, with distances up to 1,000 km. In other words, these measures should capture whether nodes preferentially exhibit short-range rather than long-distance ties with respect to the expected spatial dimension of the system.

This is accomplished by choosing an appropriate value for the scaling factor β used in Equations 3.3 and 3.6. Using the same value for every network, the graph whose nodes are at shorter distances from each other might have higher values of geosocial measures than the other. Instead, we want to be able to compare the spatial structure of two networks even if it arises at different geographic scales, e.g., city-wide or nation-wide. A reasonable choice for each network is to adopt a scaling factor β equal to the average spatial distance between all its nodes. This choice is dependent only on the positions of the nodes of the social graph, not on their links. It is worth noting that when considering a single social network the scaling factor becomes less important and could be ignored by setting $\beta = 1$, for instance.

3.3.2 Case-study: assessing the impact of location-sharing

We explore the effectiveness of our measures by applying them to a set of four different social services with location information about their users.

Traces

We use traces from four different social services, created with different goals and offering different features to their users. Two of them, Brightkite and Foursquare,

Dataset	N	K	$\langle k \rangle$	$\langle C \rangle$	$\langle D_{ij} \rangle$	$\langle l_{ij} \rangle$	$\langle NL \rangle$	$\langle GC \rangle$
Brightkite	54,190	213,668	7.88	0.181	5,683	2,041	0.82	0.165
Foursquare	258,706	2,854,957	22.07	0.191	8,494	1,442	0.90	0.173
LiveJournal	992,886	29,645,952	29.85	0.185	6,142	2,727	0.73/0.71	0.146
Twitter	409,093	182,986,353	447.29	0.207	6,087	5,117	0.57/0.49	0.108

Table 3.2: Properties of the datasets: number of nodes N and edges K, average node degree $\langle k \rangle$, average clustering coefficient $\langle C \rangle$, average distance between nodes $\langle D_{ij} \rangle$ [km], average link length $\langle l_{ij} \rangle$ [km], average node locality $\langle NL \rangle$ (in/out), average geographic clustering coefficient $\langle GC \rangle$.

are location-sharing services: the other two are LiveJournal, a blogging network, and Twitter. They all provide static geographic information about their users, in explicit or implicit form (e.g., geographic coordinates or a city name). The traces used for Brightkite and Foursquare are the same as those described in Section 3.1, while for LiveJournal and Twitter we describe our collection methodology here.

LiveJournal LiveJournal is a community of bloggers with over 10 million active users as of the end of 2010. Users can keep a blog or a journal and establish friendship connections to other users. Each user provides a personal profile, which often includes home location, personal interests and a list of other bloggers considered as friends. Friendship links are not always reciprocal. The data collection process involved both crawling the social network links through the API and downloading the user profile pages. The duration of the collection was 9 days, from November 2 to November 9, 2009, obtaining a sample of 1,502,684 users. Given the 1,226,412 users who provided location information, we successfully obtained a meaningful geographic location for 992,886 users.

Twitter Our crawling process was seeded collecting 1,000 seed users from the public timeline, which shows a list of the 20 most recent tweets posted by users with unrestricted privacy settings to the entire service. The duration of the data crawling was 6 days from December 3 to December 8, 2009, gathering information about profiles and follower lists for 814,902 different users. Of these, 535,653 reported some information about their home location. We have successfully geocoded 409,093 users, translating their location information into a point on the Earth.

To extract and collect information from LiveJournal and Twitter we crawled a sample of users employing snowball sampling: the data extraction starts from a set of seed users and expands the extraction by following the outgoing links of these users to reach new users, and so on.

General properties

In Table 3.2 we compare some basic properties of the traces under analysis. The graphs extracted from these services are different in size: Brightkite and Foursquare have average node degrees $\langle k \rangle$ of 7.8 and 22.0 respectively, LiveJournal has an average degree of about 30 and Twitter shows a larger value of 447. The case of Twitter is peculiar: this social network encourages users to follow a large number of other users and, since no reciprocation in link creation is needed, it is easier for a user to accumulate a large number of social connections. Also, Twitter users may accumulate many connections as sources of updates, as the service is used as a news feed by many [KLPM10]. Our samples show a giant component containing almost all the nodes in each sample.

Differences between the social graphs are present also with respect to the average clustering coefficient $\langle C \rangle$: Twitter and Foursquare have higher coefficients of 0.207 and 0.191 respectively, while LiveJournal scores 0.185 and Brightkite 0.181. In addition, since LiveJournal and Twitter are represented as directed graphs, we report their value of reciprocity ρ [GL04], which measures how likely each link is to be present in both directions and spans from $\rho = 1$ for perfect reciprocity to $\rho = -1$ if each link is present only in one direction. We have $\rho = 0.69$ for LiveJournal, while Twitter has $\rho = 0.79$. Hence, both networks exhibit high values of reciprocity, although Twitter appears more symmetric; this property might be related to the fact that it encourages more reciprocal interactions than LiveJournal.

Geographic properties

After investigating the social structure of the services, we now analyse their geographic properties. One of the most important characteristics is the geographic distance that social connections span; even if a link between two users denotes some sort of social relationship, it is also important to take into account how it stretches across space. First of all, the networks under analysis present different values of the average distance $\langle D_{ij} \rangle$ between users: Foursquare users exhibit an average distance of about 8,500 km, while in Brightkite this value goes over 5,600 km and in LiveJournal and Twitter it is above 6,000 km. The higher value in Foursquare reflects the wider global audience of that service, popular in the USA, in Europe and in Asia.

In Figure 3.9 we compare the cumulative probability distribution of edge length for the four different social networks. The distributions for Foursquare and Brightkite have already been presented in Figure 3.2 and they are reported here again to aid comparison with the other two services. Foursquare has the smallest average, only 1,442 km, yet only about 4% of its links are shorter than 1 km. Similarly, only about

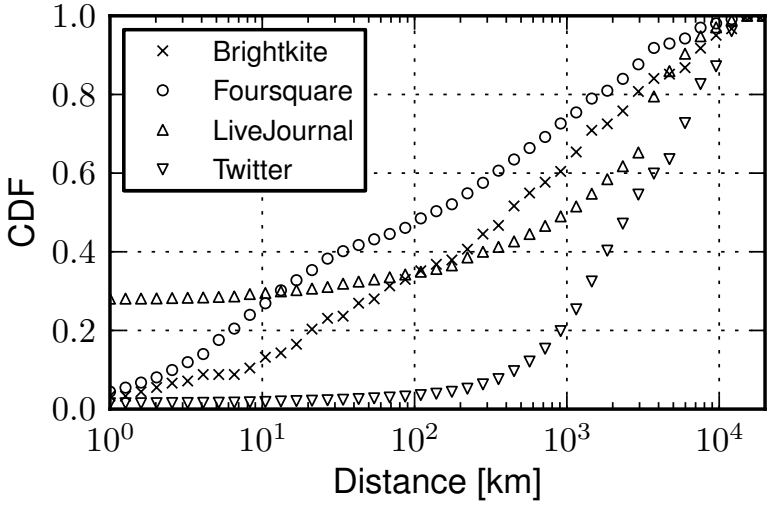


Figure 3.9: Empirical Cumulative Distribution Function (CDF) of the geographic link length for the four datasets. Logarithmic binning has been adopted to estimate the number of samples in each range of values.

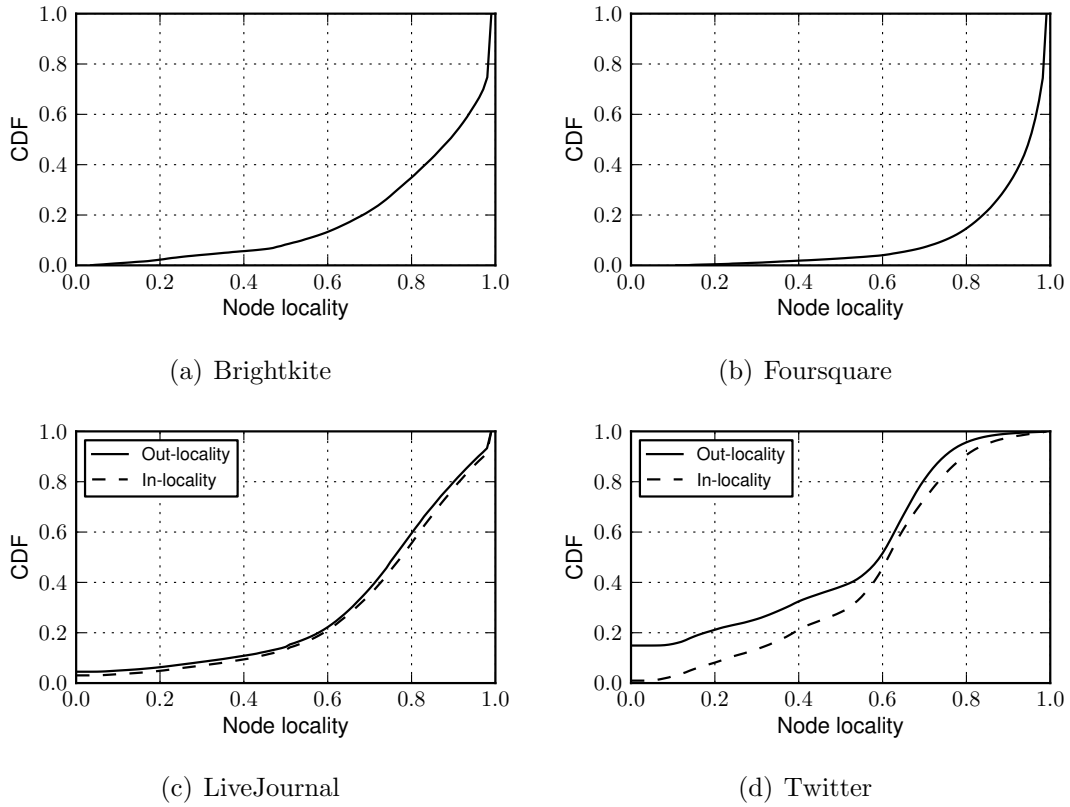
4% of the links in Brightkite are extremely short, with a global average length of 2,041 km. However, between 60% and 70% of links are shorter than 1,000 km in these two networks. An opposite trend appears in LiveJournal, with around 30% of links being shorter than 1 km, but an average link length of 2,727 km. Finally, Twitter links have an average length of 5,117 km: below 5% of these links are shorter than 100 km, while more than 80% are longer than 1,000 km. This is a clear indication that Twitter users are likely to be engaged with a global audience of followers, even though there are also short-range social connections.

3.3.3 The effectiveness of geosocial measures

After discussing the social and geographic properties of these services, we apply our two novel measures, *node locality* and *geographic clustering coefficient*, which blend social and spatial factors together.

Node locality

The probability distributions of node locality for the four datasets are shown in Figure 3.10. The main observation is that in Brightkite, Foursquare and LiveJournal there is a non-negligible fraction of users with node locality close to 1. Hence, there are some users who have social connections only with other individuals within a close geographic distance. In the Brightkite network about 40% of users have a node locality higher than 0.90, and an even higher proportion is seen in the Foursquare



(a) Brightkite

(b) Foursquare

(c) LiveJournal

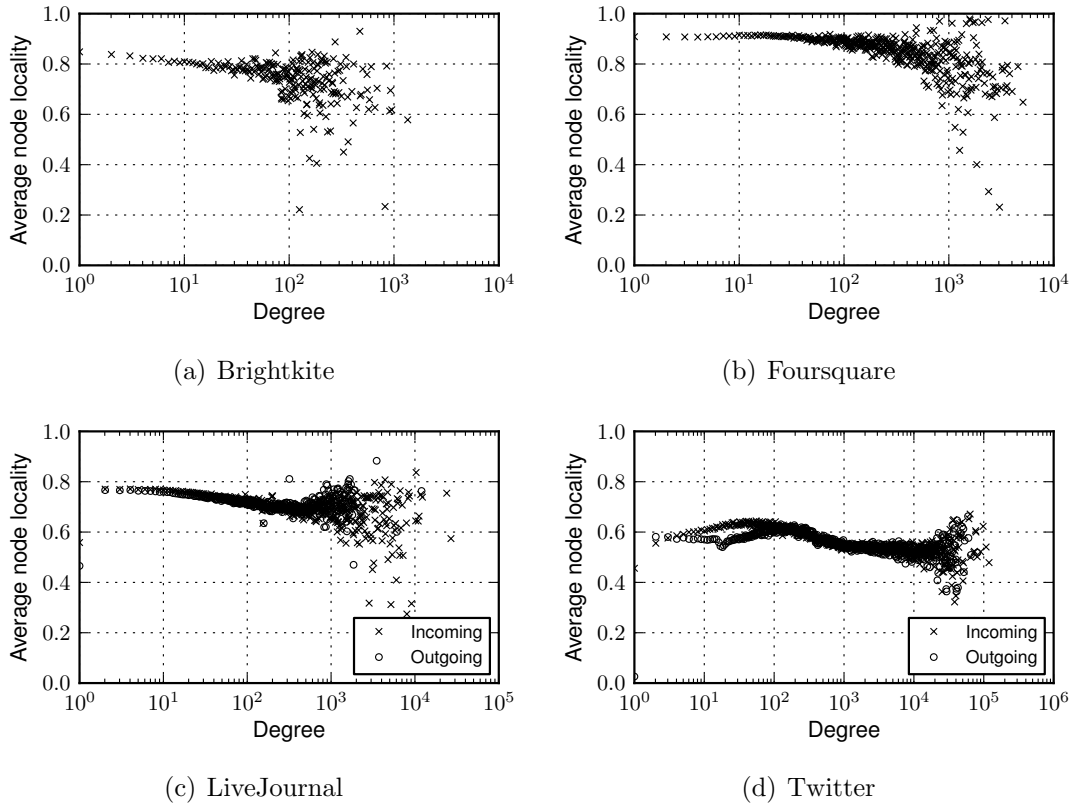
(d) Twitter

Figure 3.10: Empirical Cumulative Distribution Function (CDF) of node locality for each dataset.

dataset. Users of these location-based services exhibit high average node locality: Brightkite has an average value of 0.82, while in Foursquare this value is 0.90.

In the LiveJournal network only 20% of users have a node locality higher than 0.90 and the mean values are 0.73 for in-locality and 0.71 for out-locality. The node locality distribution appears similar both for in- and out-locality. In Twitter the distribution of node locality shows fewer nodes with high values. This may provide evidence that Twitter users are more likely to engage with a geographically spread set of individuals rather than only with users at closer distances. Moreover, in- and out-locality exhibit different patterns, since there are more than 15% of nodes with an out-locality of 0, probably nodes without outgoing connections. The average values are lower than in the other networks: 0.57 for in-locality and 0.49 for out-locality.

These results show that location-based services such as Brightkite and Foursquare are characterised by short-range friendship links between users, resulting in a vast proportion of them having high values of node locality. Thus focussing merely on user location, rather than on what users share and post, may give more opportunities to discover potential friends who live nearby. In contrast, these patterns are not present in social networks that are less centred on user location; in LiveJournal



(a) Brightkite

(b) Foursquare

(c) LiveJournal

(d) Twitter

Figure 3.11: Average node locality as a function of node degree for each dataset. For directed networks the relationship is shown both for incoming and outgoing links.

users have connections with heterogeneous length and this effect is even greater in Twitter. Their users may be more interested in becoming friends with individuals who post and share interesting content rather than simply with people at close distance.

Node locality and node degree

We now analyse the correlation between node degree and node locality to understand the geosocial properties of users with different numbers of connections. The average node locality as a function of node degree for Brightkite and LiveJournal is shown in Figure 3.11: node locality is slowly decreasing with node degree and only users with many connections have lower values of node locality. Since LiveJournal and Twitter are modelled as directed graphs we investigate the correlation in both directions. In LiveJournal the decreasing trend is evident for both in- and out-locality. Twitter users reach a maximum value of out-locality as their number of outgoing links grows larger than 100, whereas in-locality shows a maximum just before 100 incoming connections. Both relationships then decrease until they reach a plateau.

While it is expected that nodes with larger degrees exhibit smaller locality values, since it is statistically more likely that they are connected to distant users, this

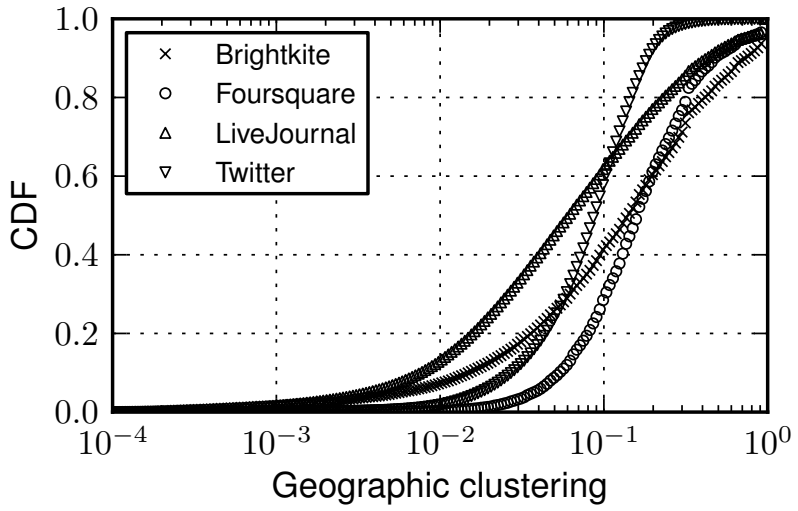


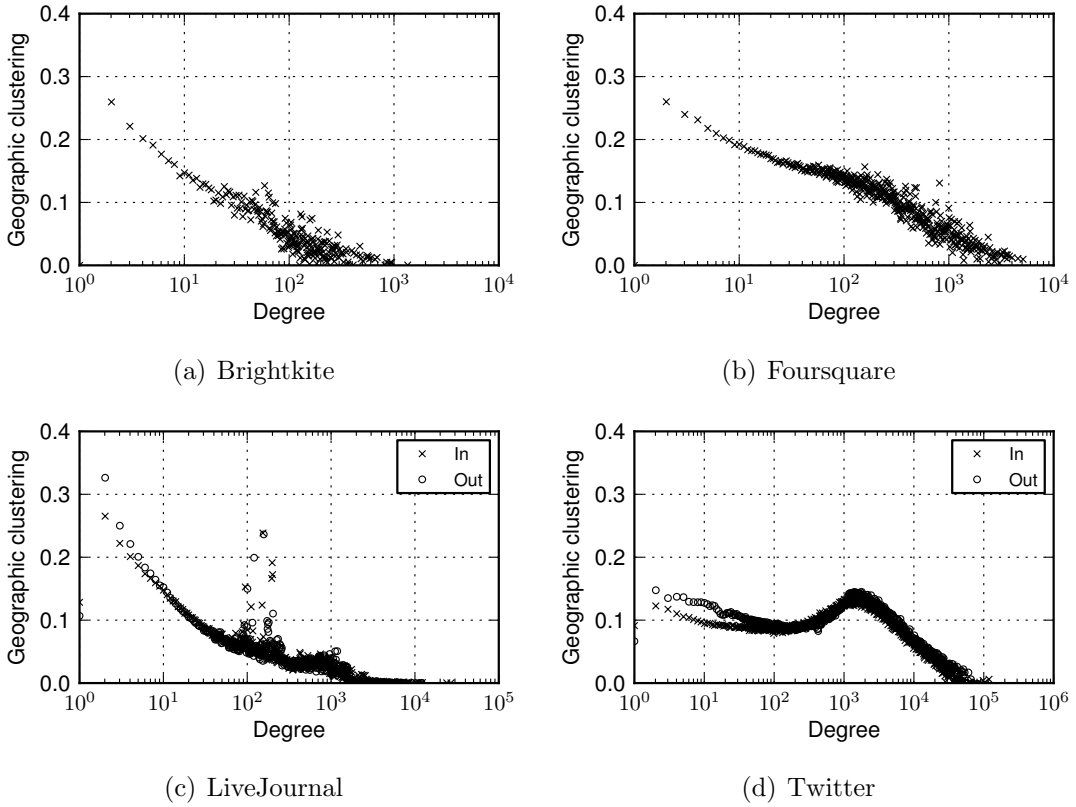
Figure 3.12: Empirical Cumulative Distribution Function (CDF) of geographic clustering coefficient for each dataset. Logarithmic binning has been adopted to estimate the number of samples in each range of values.

behaviour is not observed in Twitter: users with about 10 outgoing connections have lower values of out-locality, but as the out-degree grows there is a maximum at 100. One possible explanation is that users with a small number of links are probably mainly connected to popular accounts, e.g., institutions, media and commercial entities, which are usually not geographically close to them.

Note that during the data collection, when users joined Twitter they were presented with a list of 20 popular users to follow; these celebrities are unlikely to be located close to the joining user. As a consequence, people who just added those suggested connections when they joined the service may have ended up with a small number of connections which are not close from a geographic point of view.

Geographic clustering coefficient

The other geosocial measure that we have studied is the geographic clustering coefficient. Since social networks are widely known to be characterised by the presence of triangles, the aim of this measure is to understand whether triplets of mutually connected users are more likely to be geographically close or, instead, distant from each other. A user with high geographic clustering coefficient has neighbours who are tightly interconnected and close to the user themselves and to each other. The four datasets exhibit different values of geographic clustering coefficient; while Brightkite has an average value of 0.165 and Foursquare of 0.173, the average for LiveJournal is 0.146 and Twitter scores 0.108. Also, the first two datasets exhibit a geographic clustering coefficient close to their standard clustering coefficient, while LiveJournal and Twitter present lower values when geographic distance is taken into



(a) Brightkite

(b) Foursquare

(c) LiveJournal

(d) Twitter

Figure 3.13: Average geographic clustering coefficient as a function of node degree. For directed networks the relationship is shown both for in- and out-degree.

account. Thus, in the former two networks clusters form at shorter distances than in the latter.

The probability distributions of the geographic clustering coefficient are shown in Figure 3.12. The higher mean values for Brightkite and Foursquare are explained by the fact that a non-negligible portion of users have a coefficient of 1.0: about 10% in Brightkite and about 5% in Foursquare. On the other hand, in Twitter higher values are less likely to be observed and there is no discernible proportion of users with a coefficient of 1. These results show that location-based platforms tend to have more geographically confined triangles than social networks more focussed on content production and sharing such as Twitter.

Geographic clustering coefficient and node degree

We now investigate the relationship between geographic clustering coefficient and node degree. As reported in Figure 3.13, in Brightkite, Foursquare and LiveJournal the geographic clustering coefficient steadily decreases as the number of neighbours grows: thus, if a user has only few friends they are more likely to create connections with people nearby. On the other hand, Twitter shows a different behaviour: the geographic clustering coefficient is slowly decreasing as the degree increases, but

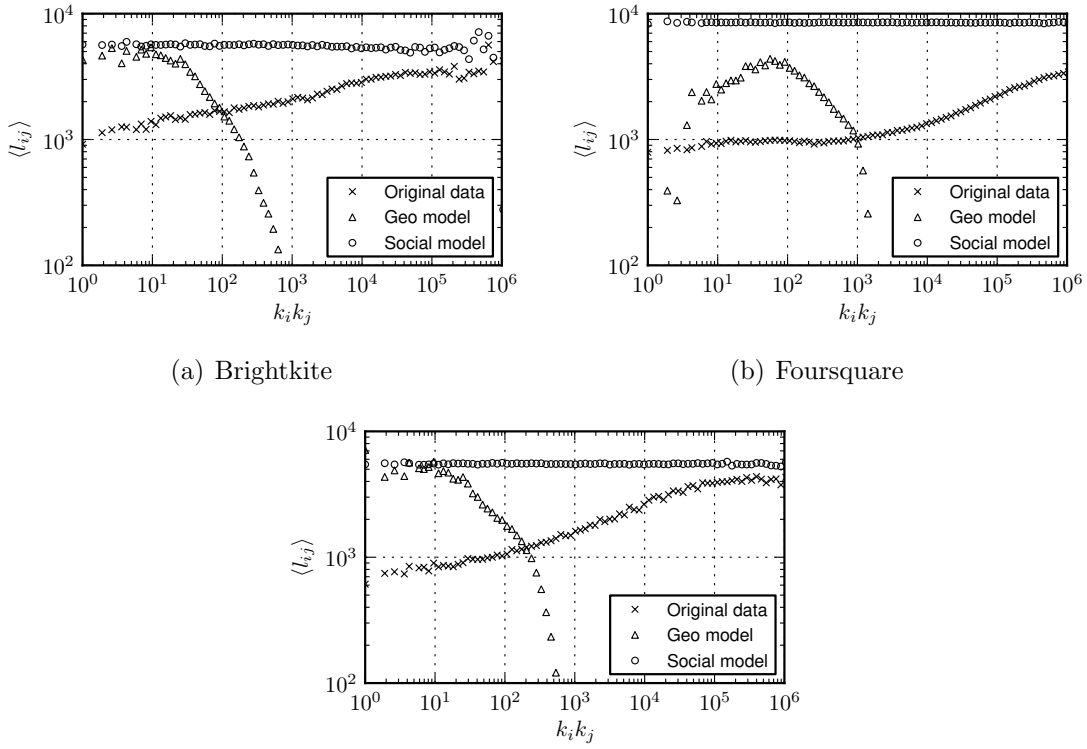
then it grows again until reaching a local maximum around the value of 1,000, while it decreases again for larger degrees.

This particular property of the Twitter network may be explained by the existence of users which are popular only in a particular region: they have both incoming and outgoing links with a large audience which has, however, several interconnections on a confined scale. Indeed, a user which is locally popular might have lower values of node locality because of his/her large audience, as shown in Figure 3.11, but users which are following him/her are also likely to share the same interests (since they follow the same popular user) and to become connected with each other. Instead, when a user reaches a wider popularity, his/her followers will be both more geographically spread and less interconnected.

3.4 Discussion and implications

In this chapter we have seen that spatial distance affects the social structure of online services and that users exhibit different geographic and social characteristics, with weak positive correlation between the number of friends and their average distance. Also, a similar heterogeneity appears with respect to social triads, with users participating in geographically wider triangles as their degree increases. Our findings appear robust across the three traces under analysis, as they arise regardless of the particular service we consider, the data collection methodology, the time elapsed since the creation of the service or the number of users in the social graph. However, the properties we observe in the real systems do not appear in the two randomised versions of these networks; therefore, *their socio-spatial structure cannot be explained by taking into account only geographic factors or social mechanisms.*

Indeed, this claim can be further supported by considering the average length of a link l_{ij} as a function of the *product of the degrees* $k_i k_j$. As observed in Figure 3.14, longer links tend to arise between users with more friends, while links connecting users with fewer friends tend to be much shorter. This effect signals significant correlation between users' social properties and their spatial behaviour. In fact, it is not seen at all in the Social model; this suggests that there might be an underlying spatial process taking place that results in this correlation, since social ties are not equally likely to appear regardless of their geographic length. On the other hand, the Geo model exhibits the opposite trend, with shorter links appearing mainly between well-connected users. Hence, distance is not the only factor affecting the link formation process; in other words, when only mechanisms that depend on geographic distance are in place, a user accumulates many friends only where there are many potential friends living nearby, i.e., if he/she is located in an area with high density of users. Furthermore, this geographic model cannot reproduce how some users



(a) Brightkite

(b) Foursquare

(c) Gowalla

Figure 3.14: Average link length $\langle l_{ij} \rangle$ as a function of the product of the node degrees $k_i k_j$ for the original network and for its two randomised versions.

accumulate thousands of friends, creating a heavy-tailed degree distribution.

Since both social factors and geographic space need to be considered when studying these systems, we have proposed two new geosocial measures, node locality and the geographic clustering coefficient, that take spatial distance into account. Using these novel network measures we have been able to study four different online services from a social and a spatial perspective at the same time. Our measures have highlighted that purely location-based social networking services, which mainly focus on the geographic dimension of social interaction, tend to have users with stronger preference for spatially short social ties. In contrast, services based more on the idea of sharing information and content have users with lower node locality and geographic clustering coefficient values, showing that space has a weaker effect on these services.

All these findings support the claim that accurate modelling of spatial social networks requires the incorporation of processes combining social and spatial factors. We have discussed how the effect of node degree on network evolution is captured well by the preferential attachment model, where the probability of connection between nodes i and j, P_{ij}, is proportional to the degree of node j, $P_{ij} \propto k_i$. The effect of geographic distance can be included in this attachment probability, $P_{ij} \propto k_i k_j f(D_{ij})$, where f is a decreasing *deterrence function* of the geographic

distance D_{ij} between the nodes. Thus, long distances tend to be covered only to connect to important hubs, while nodes with fewer connections become attractive when they can be reached over a short distance. When the deterrence function has a simple functional form such as $f(d) \sim d^{-\alpha}$, then the probability of a connection between two nodes becomes similar to the gravitational attraction between celestial bodies, $P_{ij} \propto \frac{k_i k_j}{D_{ij}^{\alpha}}$. Hence, these attachment models are known as *gravity models* [Car56, ES90].

A gravity model balances the effect of spatial distance with other node properties; the underlying assumption is that longer (and more expensive) ties will appear mainly between important entities, while a node will connect to an unimportant one only if they are close to each other. These models have long been used to model connections in spatial networks such as trade flows across countries [BMS+08], traffic flows in highway networks [JWS08] and mobile phone calls between cities [KCRB09].

Gravity models are only a first tentative step to reproduce the patterns we have observed. In particular, gravity models only focus on pairs of nodes, without taking into account social effects such as triadic closure and focus constraint [Gra73, Fel81]. Furthermore, even though the degree of a node represents a reasonable choice as a "mass" variable for a social gravity model, any notion of "importance" in a social network might avoid quantification, thus making the definition of a sound social gravity model hard to specify. Such individual importance may be an exogenous variable which affects the socio-spatial structure, such as being a well-known celebrity or any other type of individual popularity or social influence measure. It is likely that any social gravity model will need to take into account this type of heterogeneity across individuals. We will explore these issues in more depth in Chapter 4.

3.5 Related work

In this chapter we have presented our findings concerning the spatial properties of online services, studying the spatial structure of the social graph arising among individual users. Hence, the main thread of research related to these results involves studies of the spatial properties of complex networks and social systems.

The effect of geography on complex networks has been studied mainly in systems such as transportation networks and infrastructure networks, that is, structures able to convey energy, matter or information at different scales and in different scenarios. Some examples include Internet router connections [YJB02, BGG03], airline flights between airports [GN06], subway networks [LM02], electrical power grids [SRCCMV08], urban road networks [CLP06, CSLP06], maritime cargo shipments [HZ09] and other systems where nodes are embedded in a metric space. A

more complete review discussing many of these examples in depth has been compiled by Barthélemy [Bar11].

This abundance of work on spatial networks does not extend to social systems: mainly because geographic data about individuals have been difficult to obtain, especially for large systems, social network analysis has often neglected the spatial perspective. Thanks to some dedicated and relatively small-scale data collection efforts, some sociologists have studied the effect of distance on social ties. For instance, through a longitudinal study spanning different decades, Mok and Wellman discuss how spatial proximity influenced social interactions among resident of a neighbourhood in Toronto before the advent of the Web [MW07], but also when online communication tools became widely available [MW09]. Their results suggest that the effect of distance remained strong over 20 years, even though communication was made easier and more effective over the entire range of geographic distances.

Some studies have explored a fundamental spatial property of social networks: the probability $P(d)$ of having a social connection between two individuals as a function of their distance d. Even though there seems to be agreement that $P(d)$ decreases with distance, the exact relationship between these two variables is still unclear. Lambiotte et al.[LBD+08] found that it decays as $P(d) \sim d^{-2}$ in a mobile phone communication network, while Liben-Nowell et al. [LNNK+05] found a different relationship $P(d) \sim d^{-1} + \epsilon$ among online bloggers on LiveJournal in the USA, ϵ being a constant probability which acts on online communities regardless of distance. In another study, Backstrom et al. [BSM10] have similarly found spatial scaling $P(d) \sim 1/d$ of Facebook connections; they show that this association appears so strong and important that it can be safely exploited to infer where users are located only from the location of their friends.

It has also been proposed that the spatial structure of social networks might be scale-invariant, with a universal distribution $P(d) \propto d^{-1}$ [HWL+11]. Butts suggests that this relationship between physical distance and connection probability in social networks is so important that it can be used to explain entropy and predictability of the social structure, provided that an upper bound can be defined for the likelihood that distant individuals are connected [But03]. Our study goes beyond the investigation of this basic relationship between ties and distance, analysing user heterogeneity and addressing the interplay between spatial and social factors.

Finally, a few other studies have investigated the structural properties of a location-based social network and how social and geographic distance influences the creation of new connections between its users [LC09a, LC09b]. Such investigations fail to address the heterogeneity we have observed across users and do not discuss whether the global structure of the network is influenced by social and spatial factors. In doing this, our findings represent some important initial steps towards a better

and more comprehensive understanding of the spatial properties of online social networks.

3.6 Summary

Location-sharing features offered by online social services allow users to engage in a new way with the physical world around them and with their friends. At the same time, as individuals create and share location-tagged information, they reveal their geographic position and their spatial movements. In order to understand the effect of space and distance on online social connections we need to measure and interpret traces extracted from such services.

In this chapter we have addressed these issues, offering a series of results. We have presented a large-scale study based on traces collected from mobile location-based social services. We have used check-in data to assign geographic positions to users and to study their social graph as a spatial network. Through a comparative study of three different services, and by adopting randomised null models to disentangle social from spatial factors, we have discovered that the spatial properties of users are heterogeneous. These findings will be revisited in Chapter 4 when we discuss the temporal evolution of a spatial social network.

Finally, we have devised two new network measures that combine social and spatial characteristics: node locality and the geographic clustering coefficient. We have explored their potential by using them to assess the effect of spatial distance on different online social services. We will also see in Chapter 6 that these measures can be successfully used when designing content delivery distribution systems.

Essentially, all models are wrong, but some of them are useful.

George E. P. Box

4

Modelling social network growth over space

Connections established between users of online social networks are often influenced by basic social mechanisms such as preferential attachment, which captures how popular users attract more and more new links, and triadic closure, which models how shared friends are powerful indicators of future friendships of a user. However, we have seen that geographic distance is also a factor which impacts online social links; our results indicate that spatial proximity fosters the creation of online social ties. Yet, the underlying spatial processes that might shape the creation of new social links are still largely unknown.

Different evolutionary models have been proposed and tested to explain the growth of online social networks; in many cases they describe the behaviour of individual nodes that results in some global structural properties observed in real-world systems, such as power-law degree distributions and high clustering coefficients [LBKT08]. The fundamental importance of such models is due to the fact that they often highlight universal characteristics of user behaviour; for instance, mechanisms such as preferential attachment and triadic closure are thought to mimic the actions of individuals creating their social connections, thus offering practical insights to predict future links [LNK07, LLC10]. However, these growth models often neglect factors that are not inherently connected to the structure of the social network itself; in particular, they neglect spatial distance.

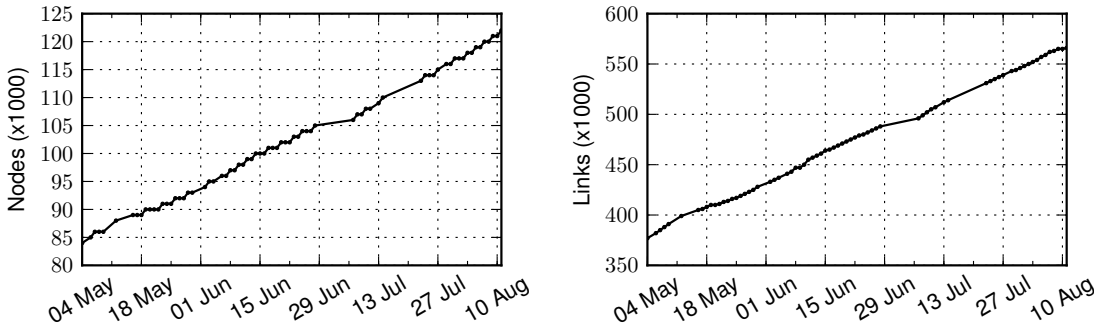
Among the many interesting questions arising from the inclusion of spatial information in social network analysis, we aim to understand how distance is affecting the establishment of new online social ties. In the previous chapter we have seen that by considering only social or only spatial factors one cannot reproduce the overall properties of real spatial social networks. The open question is therefore how to merge space and social influences successfully to reproduce what is observed in real graphs. In this chapter we answer this question: we present a detailed study of the temporal evolution of a social network by quantifying the impact of spatial and social factors on the growth of a popular location-based social service. We exploit our findings to propose a new model of network growth, which is able to describe the social and spatial properties observed in real networks.

Chapter outline In this chapter we study the temporal evolution of a social network using daily snapshots of a popular location-based social service, with information about users' location and social connections. We discuss how the network grows over time in Section 4.1. Using this fine-grained temporal information about network evolution, in Section 4.2 we test whether different edge attachment and triadic closure models can explain the observed data, adopting an approach based on *likelihood estimation*. This methodology allows us to compare quantitatively different evolution models according to their statistical ability to reproduce real events. Section 4.3 studies the temporal patterns of individual user behaviour, namely the lifetime of a node, that is the amount of time a user is actively creating new edges, and the inter-edge waiting time, which governs the amount of time elapsed before a node will create a new edge.

Based on these findings, in Section 4.4 we describe a new model of network growth which is able to reproduce both the social and spatial properties observed in the real data. Our results show that geographic constraints should be considered when dealing with online social networks and suggest that a gravitational attachment model is able to capture the effect of geographic distance on users. We consider the implications of our study in Section 4.5, reviewing related work in Section 4.6. Section 4.7 closes the chapter.

4.1 Measuring network growth

In this section we describe the temporal evolution over 4 months of a popular location-based social service, Gowalla. We have acquired traces which include fine-grained temporal information about when individual social connections between users were created; this enables us to explore the temporal evolution of the Gowalla social network.



(a) Nodes

(b) Links

Figure 4.1: Temporal evolution of the number of nodes (a) and links (b) in the spatial social network of Gowalla users, as captured by our periodic measurements.

4.1.1 Data collection

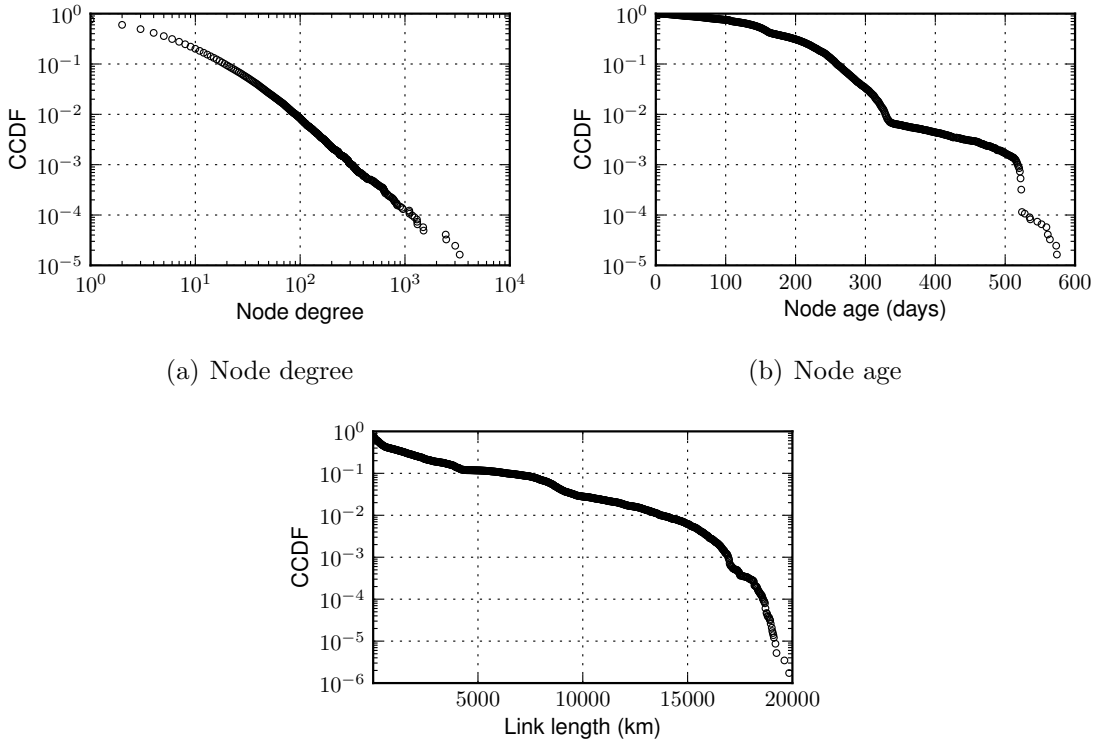
We have downloaded daily snapshots of Gowalla data between May and August 2010 by accessing their public API, obtaining for each snapshot the data outlined in Section 3.1. This dataset represents a sequence of *complete* snapshots of a large-scale location-based service, allowing us to study how the social network grows over time and over space. In particular, we have temporal information about all the social links created during our measurement process; this enables us to study the social and spatial factors that may influence the creation of social links.

The dataset contains about 400,000 registered users at the end of our measurement period, but only a fraction of them are actively using Gowalla, while many of these accounts do not show any type of activity: no social connections and no check-ins. There are 183,709 users with at least one check-in and 162,239 with at least one friend. We focus our analysis on 122,030 active users who have both friends and check-ins.

Notation

Formally, we represent the spatial social network of Gowalla users as an undirected graph. We denote by N and K the total numbers of nodes and edges, while $G_t = (N_t, K_t)$ is the graph composed of the earliest t edges $(e_1, \ldots, e_t)$, with G_T being the final network at the end of the measurement process. The time when edge e was created is $t(e)$ and $t(u)$ is the time when node u joined the service, that is, the time when it created its first connection or made its first check-in. The age of node u at time t is denoted as $a_u(t) = t - t(u)$. The degree of node u at time t is $k_u(t)$, while the number of nodes with degree k at time t is denoted as $n_k(t)$.

As in the previous chapter, every node of this network is embedded in a metric space: we assign each user to the geographic location of the place where he/she



(a) Node degree

(b) Node age

(c) Link length

Figure 4.2: Complementary Cumulative Distribution Function (CCDF) of node degree (a), node age (b), and link geographic length (c) at the end of the measurement period.

has made the greatest number of his/her check-ins at the end of the measurement period. Since we do not change user locations over time, link lengths and distances between nodes do not change either.

4.1.2 Basic properties

The number of nodes and the number of links grow approximately linearly over time, with the number of links growing at a faster pace. On average, the network gains about 375 new nodes and about 1,900 new edges per day, as shown in Figure 4.1. Note that the graph at the end of our measurement period corresponds to the Gowalla dataset already studied in Chapter 3.

In Figure 4.2(a) we present again the degree distribution, which exhibits a heavy tail. In contrast, both the distributions of node age and link geographic distance, in Figures 4.2(b)-4.2(c), do not exhibit heavy tails but instead an almost exponential decay (notice the linear x-axis). There is a large fraction of short-range geographic connections: about 50% of social links span less than 200 km, with only a small fraction being longer than 4,000 km. At the same time, the distribution of node

age shows that nodes have joined the network with some irregular temporal spikes; overall, about 99% of all nodes have joined the service in the last 300 days.

4.2 Modelling network growth

In this section we study how the creation of individual social links is influenced by both global and local properties of the network, adopting a methodology based on the Maximum Likelihood Principle to evaluate and compare how a set of models describe the empirical data.

In more detail, we analyse two core facets of temporal network evolution:

- **how edges are created**: we test different attachment models that select the target of a new connection given the social and spatial properties of network nodes;
- **how social triangles are created**: we test a family of triadic closure mechanisms based on node properties and spatial distance.

Our results demonstrate that node degree and spatial distance are simultaneously influencing edge creation, suggesting that *a gravitational attachment model describes real network evolution better than purely social or spatial models*; we also find that *social factors are more important than spatial constraints when an edge closes a triangle*. These findings will be revisited and exploited in Section 4.4, where we will define a new model of network growth.

4.2.1 Maximum likelihood estimation

When assessing whether a model reproduces empirical properties observed in the data, a common approach is to test if global properties are equally found in the real data and in the output of the model. Instead, we take advantage of the fine-grained temporal information present in our traces and we adopt a quantitative approach to compare how different models describe the empirical traces. We directly compute the likelihood that a model has of generating the events observed in our sequence of traces. The Maximum Likelihood Principle can then be applied: historically, this principle has been used to compare numerically a family of models and, as a result, pick the "best" model (and parameters) to explain the data [Sti02].

Studying networks with likelihood methods requires a probabilistic model describing the evolution of the graph itself. In other words, the network is considered the result of an evolutionary stochastic process which has driven its growth, both in terms of

new nodes and new edges [WBHS06]. For instance, the preferential attachment model discussed in Section 2.2.6 describes the evolution of a network in terms of the probability of connection between new nodes and existing nodes. Given real data about the evolution of a network, one can test the extent to which the assumptions of preferential attachment model are supported by the data.

In our case, estimating the likelihood of a model M involves considering each individual edge $e_t = (i, j)$ created during our measurement period and computing the likelihood $P_M(e_t)$ that the source i selects the actual destination j according to the model M. Thus, the likelihood $P_M(G)$ that model M reproduces graph G is given by the product of the individual likelihoods according to model M:

$$P_M(G) = \prod_t P_M(e_t) \tag{4.1}$$

We use log-likelihood for better numerical accuracy, obtaining

$$\log(P_M(G)) = \log(\prod_t P_M(e_t)) = \sum_t \log(P_M(e_t)) \tag{4.2}$$

Equation (4.2) suggests a simple algorithm to compute the log-likelihood of a given model M: for each new edge created during the graph evolution, we compute the probability that it would be created according to model M, we take the logarithm of this probability and we sum all the values obtained for each edge. When this procedure is repeated for several models, we can choose the model with the highest likelihood to explain the data.

Since every edge is undirected and we do not have information about which user initiated the social contact, we consider every new edge $e_t = (i, j)$ in both directions in the rest of our analysis, in order to avoid any bias. This methodology can be extended easily to handle directed graphs.

4.2.2 Modeling edge attachment

We investigate the effect of three global characteristics on the creation of individual social links between users: node degree, node age and spatial distance between nodes.

Attachment by node degree According to the preferential attachment model [BA99], the probability of creating a new connection to a node is proportional to the number of its existing connections. The cumulative advantage held by high-degree nodes results in a degree distribution with a heavy tail, as some nodes accumulate a large number of connections. We test whether a similar mechanism is compatible with

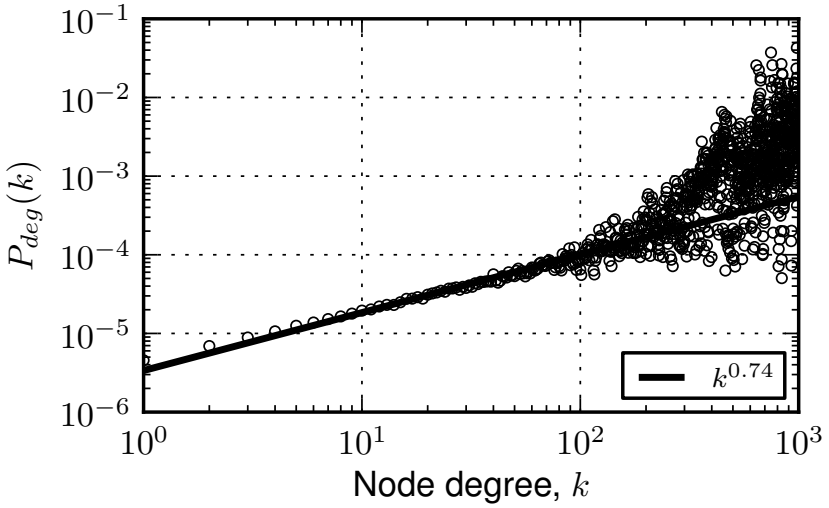


Figure 4.3: Probability of creating a new social link as a function of node degree.

our data by computing the probability $P_{deg}(k)$ that a new link will be created with a node with degree k:

$$P_{deg}(k) = \frac{|\{e_t : e_t = (i,j) \wedge k_j(t-1) = k\}|}{\sum_t n_k(t-1)} \tag{4.3}$$

where the numerator counts how many edges have been created connecting a node with degree k and the normalisation factor considers how many nodes with degree k were present when the edge was created. If preferential attachment is not governing the growth, then $P_{deg}(k)$ should not depend on k. However, we see in Figure 4.3 that $P_{deg}(k) \propto k^{0.74}$, showing that nodes with higher degree are more likely to attract new edges than nodes with fewer connections. Although the trend is not exactly linear as in the original preferential attachment model, node degree is related to the creation of new edges.

Attachment by node age The period of time a node has been part of the service could also be a factor affecting the creation of edges. Older nodes might have more visibility or more authority in the network; at the same time, when new users join the network they might experience intense activity as they search the network for potential connections. We compute $E(a)$, the number of edges created by nodes of age a normalised by dividing by the total number of nodes that ever achieved at least age a:

$$E(a) = \frac{|\{e_t : e_t = (i,j) \wedge t(e) - t(i) = a\}|}{|\{n : T - t(n) \geq a\}|} \tag{4.4}$$

where T is the time when the last node joined the network during the measurement period. As depicted in Figure 4.4, there is a spike at age 0: this represents nodes

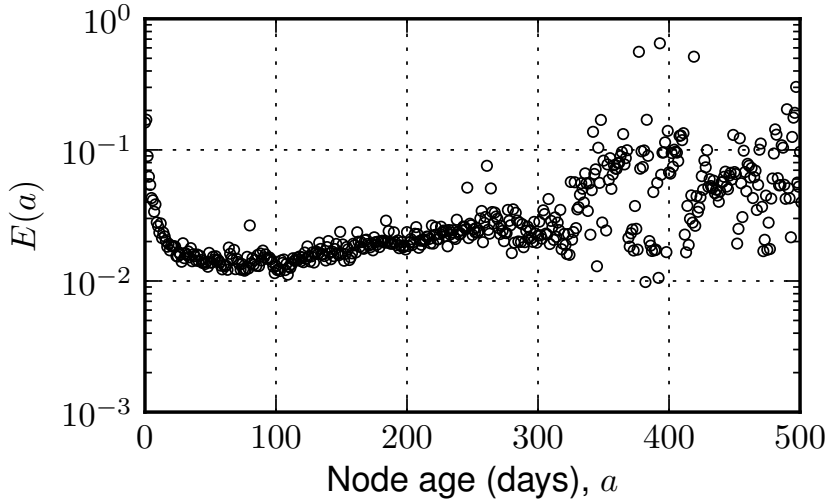


Figure 4.4: Probability of creating a new social link as a function of node age.

that join the network, create some links and then never come back. The number of created edges then quickly goes down with age a but grows again for higher values of a. Many links are created when a node joins the network, followed by lower levels of edge creation; older nodes then tend to establish further links.

Attachment by spatial distance Another important factor for edge attachment may be geographic distance. We compute the probability $P_{geo}(d)$ that a new edge spans geographic distance d, normalised by dividing by the number of nodes at distance d from the source:

$$P_{geo}(d) = \frac{|\{e_t : e_t = (i,j) \wedge D_{ij} = d\}|}{\sum_t |\{n : D_{in} = d\}|} \tag{4.5}$$

Our data show how $P_{geo}(d)$ decreases with distance d, as reported in Figure 4.5, even though the trend appears noisy. In particular, the data roughly follow a trend $P_{geo}(d) \approx d^{-\alpha}$ with $\alpha = 0.6$. While a similar functional form has been found in other spatial social networks, but with different exponents α, in this case it is measured at the level of individual edge creation events. Geographic distance affects the edge creation process in a clear way: as already discussed in Section 3.2.2 in the static scenario, even when considering individual edge creation events, longer links have a lower probability of appearance than short-range ones.

Evidence of gravity effects in network growth

We discussed in Section 3.4 how gravity models are a suitable choice to combine social and spatial properties. Our aim is now to uncover evidence to support the

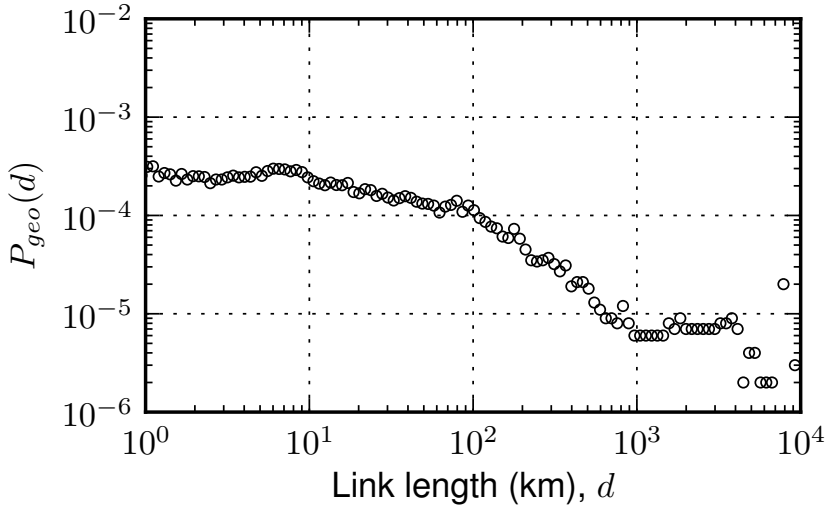


Figure 4.5: Probability of creating a new social link as a function of the geographic distance of the node. Logarithmic binning has been adopted to estimate the number of samples in each range of values.

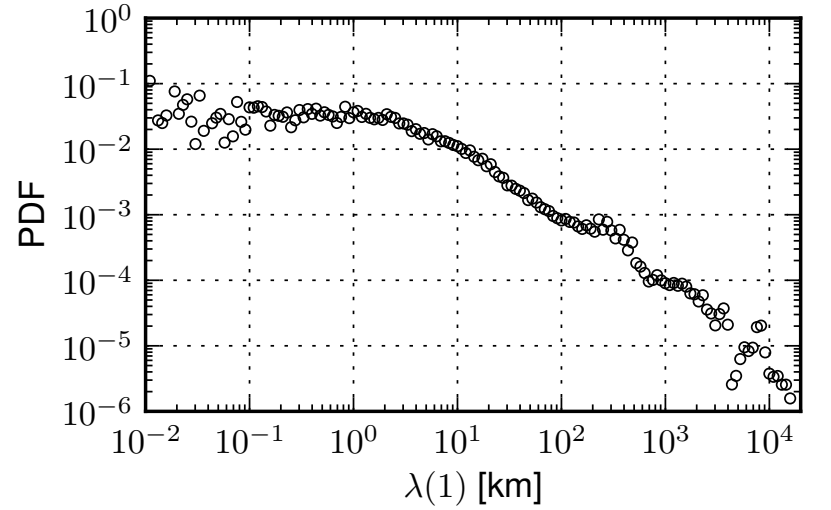


Figure 4.6: Probability Distribution Function of $\lambda(1)$, the geographic span of the first social link created by a user. Logarithmic binning has been adopted to estimate the number of samples in each range of values.

hypothesis that a similar mechanism can reproduce patterns observed in the real data.

A consequence of the gravity model is that nodes with higher degree tend to attract longer links. We therefore define $\lambda_i(k)$ as the geographic length of the k-th edge created by user i and we study how the probability distribution of $\lambda(k)$ changes for different values of k. The probability distribution of $\lambda(1)$ is reported in Figure 4.6. The distribution can be roughly divided into two regions: social connections shorter than 5 km, a threshold compatible with the size of many urban areas, exhibit a

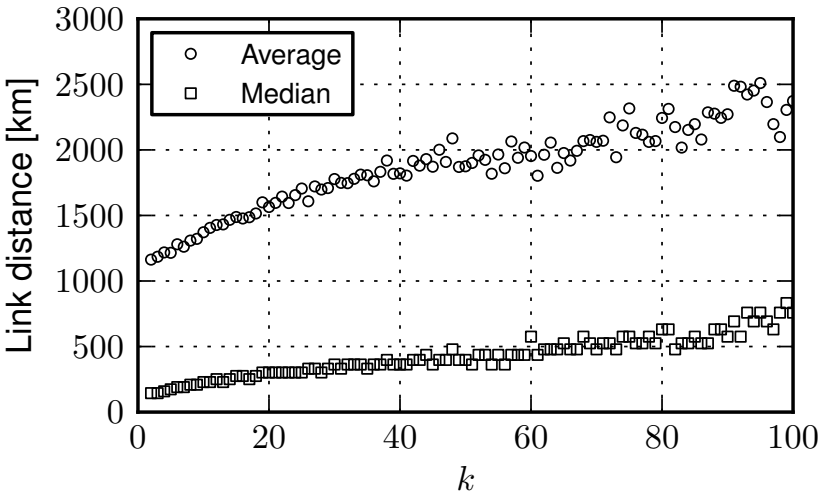


Figure 4.7: Average and median geographic span gap of the k-th edge created by a node as a function of k.

constant probability, whereas the probability of creating longer ties quickly decays.

This preference for short-range ties confirms our previous analysis of edge attachment mechanisms. The influence of degree k on the geographic properties of social links appears strong; as shown by Figure 4.7, both the average and the median value of the geographic length $\langle\lambda(k)\rangle$ of the k-th edge increase with k. While the average length of the first edge is about 1,100 km, the 100th edge is about 2,400 km. The median value shifts accordingly with k, increasing from 150 km to more than 900 km for higher degrees. These findings are compatible with a gravity model where node degree and geographic distance simultaneously influence social connections created over space.

Evaluation of attachment models

We have discovered that individual node properties and geographic distance affect edge creation. Our aim is now to understand what type of edge attachment models better approximate network temporal evolution.

We deliberately choose simple models, since our goal is not to reproduce exactly the temporal evolution of the network but, rather, to understand which factors mainly drive its growth. We consider four different edge attachment models, each one with a single parameter α:

- D: the probability of creating an edge with node n is proportional to a power α of its degree: $k_n(t)^\alpha$;
- A: the probability of creating an edge with node n is proportional to a power α of its age: $a_n(t)^\alpha$;

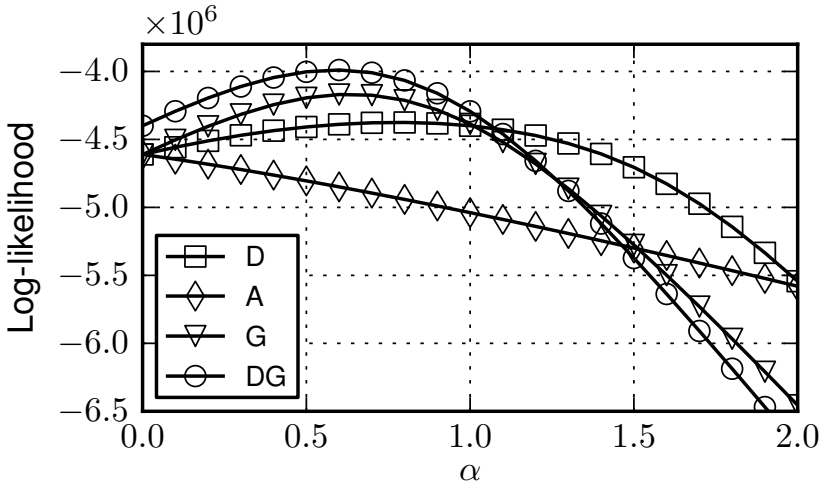


Figure 4.8: Log-likelihood of each edge attachment model as a function of their parameter α. The gravity model DG outperforms all the others.

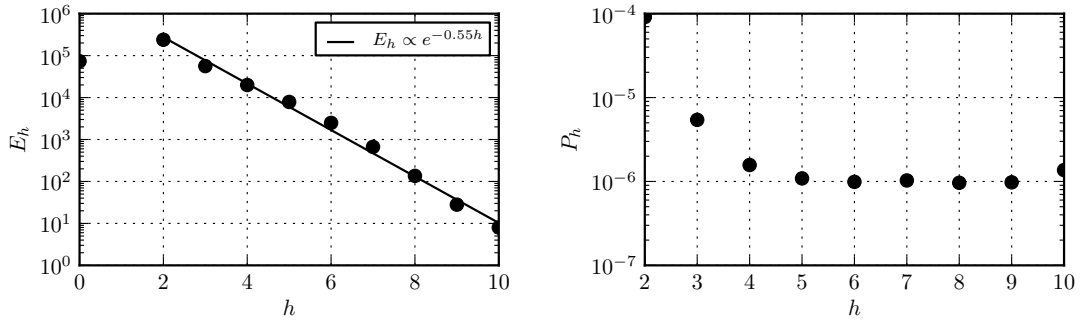
G: the probability of creating an edge with node n is inversely proportional to a power α of its geographic distance from i: $D_{in}^{-\alpha}$;

DG: the probability of creating an edge with node n is proportional to its degree and inversely proportional to a power α of its geographic distance from i: $k_n(t)D_{in}^{-\alpha}$

We will make use of the Maximum Likelihood Principle presented in Section 4.2.1 to compare and evaluate which model and which parameters better reproduce the real evolution. Figure 4.8 displays the log-likelihood values obtained by each model as a function of the parameter α. First, we note that the models G and DG, which incorporate geographic distance, have higher log-likelihood than the other two models D and A, with the overall maximum log-likelihood achieved by DG. The maximum log-likelihood for DG is achieved for $\alpha \approx 0.6$, which is in agreement with the results obtained measuring $P_{geo}(d)$. Node age does not seem a key factor for edge attachment, as the model A shows decreasing values of log-likelihood for values of α between 0 and 2. Model D reaches its highest log-likelihood for $\alpha = -0.8$ and fails to outperform G and DG. Hence, it seems that *the main driving factors in edge attachment are node degree and geographic distance* and that a gravity model which combines them is the most suitable option.

4.2.3 Modelling triadic closure

The edge attachment models previously proposed only take into account the influence of global network properties on new edge creation. However, local network properties can be equally or more important; for instance, new links often tend to



(a) (b)

Figure 4.9: Number of new links E_h created between nodes h hops away (a) and probability P_h that a new link connects nodes h hops away (b). The single E_h value at $h = 0$ denotes the number of edges connecting nodes previously in separate disconnected components.

connect users who already share friends, creating social triangles that are extremely common in social networks [LNK07]. Hence, in this section we study how new links result in new social triangles and whether different triangle-closing models can reproduce the patterns observed in the data.

The importance of triangle-closing links

Social connections tend to link together individuals who are already at close social distance: the vast majority of new links tend to be between nodes that already share at least one friend, thus only 2 hops away from each other, with larger social distances exponentially less likely [LBKT08]. We observe a similar pattern in our data: Figure 4.9(a) shows that the number of edges E_h that connect nodes h hops away exponentially decays with h. A few edges also connect nodes that were not in the same connected component, as when a new node joins the network and creates its first link.

A better understanding of this process can be achieved by considering not only how many new links connect nodes h hops away, but also considering the number of nodes at that social distance. In fact, since E_h exponentially decreases with h and the number of available nodes increases with h, the probability P_h that a new link spans h hops must be decreasing much faster than exponentially. More precisely, we compute P_h as

$$P_h = \frac{|\{e_t : e_t = (i,j) \wedge h_{t-1}(i,j) = h\}|}{\sum_t |\{n : h_{t-1}(i,n) = h\}|} \tag{4.6}$$

where $h_t(i,j)$ is the number of hops between nodes i and j at time t. Figure 4.9(b) plots P_h as a function of h: the probability decays quickly and finally reaches a

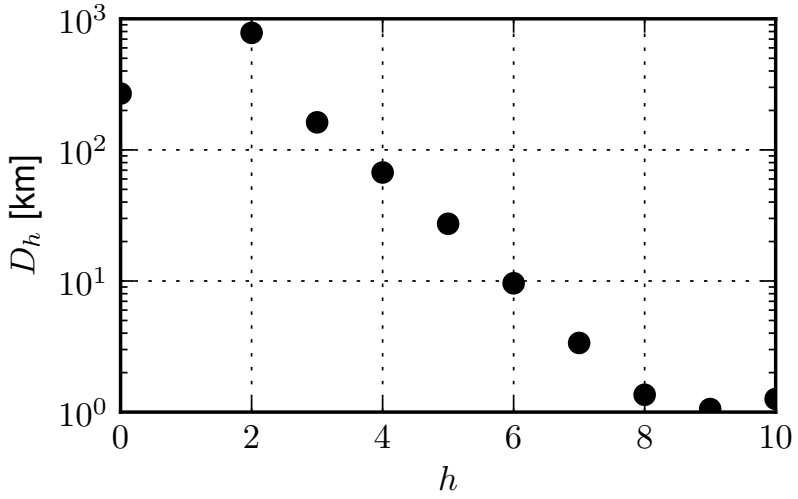


Figure 4.10: Geometric average geographic distance D_h of new links created between nodes h hops away. The single D_h value at $h = 0$ denotes the average geographic distance of links connecting nodes previously in separate disconnected components.

constant value. Thus, triadic closure seems to be the predominant factor shaping network growth over time: new edges are most likely to connect people who already share at least one friend, closing social triangles.

Given the importance of triadic closure, we focus on how this process is affected by spatial properties. We want to understand whether there is any interplay between the geographic distance and the social distance that a new link spans. A first indication is given by the geometric average geographic distance D_h of all new edges that connect nodes previously h hops away, shown in Figure 4.10. There is an evident trend: *social connections at shorter social distance tend to have higher geographic distances, while links spanning more hops have lower spatial distance.* A potential explanation is that both social and geographic distance tend to affect the edge creation process: a new link is created either between users sharing friends, even if they are far from each other, or between spatially close users, even if they have no friends in common. In particular, it appears that *geographic proximity is as important as social closeness*: both factors are shaping the network, but in different ways.

In summary, our analysis of triadic closure confirms that two users sharing at least one friend are much more likely to create direct connections than two users without friends in common. At the same time, geographic distance appears again as a driving force, even if in a different way, influencing the creation of edges between users without friends in common.

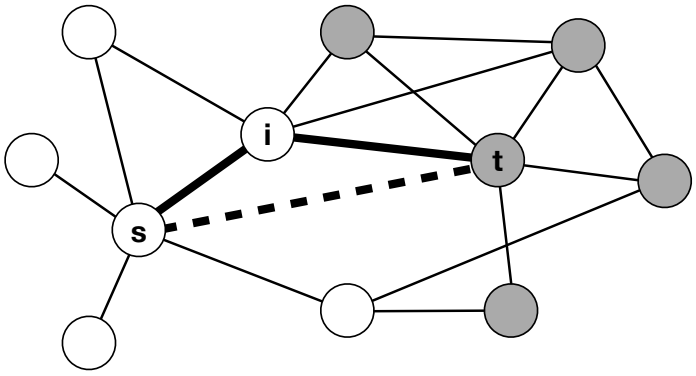


Figure 4.11: In a triangle-closing model node s creates an edge by selecting first an intermediate node i, which then selects target node t: the edge (s, t) is thus created. Different strategies to select a node's neighbour can be combined. All candidate nodes two hops away are shaded in grey; the baseline model picks at random among these candidates.

Triangle-closing models

Since about 60% of new edges close social triangles, such triangle-closing edges represent an important aspect of network growth. Hence, our aim is now to understand what factors influence which node to choose when a edge is closing a triangle. Again, we make use of the Maximum Likelihood Principle to test whether different triangle-closing models would be able to generate the triangles created during the real network evolution.

We consider the case when a source node s chooses another target node t located two hops away to create a new link, as illustrated in Figure 4.11. A simple model would be for node s to choose t uniformly at random from all the nodes at a distance of 2 hops, which will be our baseline model. We then take into account more complex models where node s first chooses according to a given strategy an intermediate node i among its neighbours and then picks a target t among i's neighbours with a potentially different strategy. The edge (s, t) is then created, closing the triangle (s, i, t).

Since every strategy involves only choosing a node i among the neighbours of a given node n, we consider five different strategies to choose i:

1. random: uniformly at random among node n's neighbours;

	random	shared	degree	distance	gravity
random	**12.34**	9.48	-3.47	-28.17	-35.26
shared	**14.54**	11.47	-0.95	-24.74	-34.46
degree	7.33	5.16	-6.79	-25.17	-41.98
distance	-0.92	-3.70	-16.94	-39.32	-41.53
gravity	2.71	0.25	-12.11	-33.01	-43.18

Table 4.1: Performance of different triangle closing models: rows show the model used to pick the intermediate node, and columns show the model used to then pick the target node. The value in each cell gives the percentage improvement of model log-likelihood over the baseline model, which chooses a random node two hops away from the source.

2. shared: proportional to the number of shared friends between i and n;

3. degree: proportional to the degree of the neighbour i;

4. distance: inversely proportional to the geographic distance between i and n;

5. gravity: proportional to the degree of the neighbour i and inversely proportional to the geographic distance between i and n.

Since there are 5 different triangle-closing models, there is a total of 25 different combinations. We compute the log-likelihood of each combination and we measure the percentage improvement over the log-likelihood of the baseline model. The results are presented in Table 4.1: the general trend is that random and shared offer the largest improvements over the baseline, with a maximum improvement of 14.54% in the combination shared-random and 12.34% for random-random. Models based on degree or on distance have performance much lower than the baseline, with degradation of up to 40% when the gravity model is adopted. In particular, the random-random model works surprisingly well, as it favours connections between nodes that have multiple 2-hop paths between them and that have higher degrees, while being extremely simple and computationally fast.

These results show that *triadic closure is mainly driven by social processes, while geographic distance is not an important factor*. Nonetheless, triangle-closing mechanisms only model some aspects of network evolution, as non-local edges are still needed to globally connect the network and create the small-world effect. In summary, social processes at a local level seem complementary to spatial factors that are shaping the network at a global level. As we will see, we need both to model real network evolution successfully.

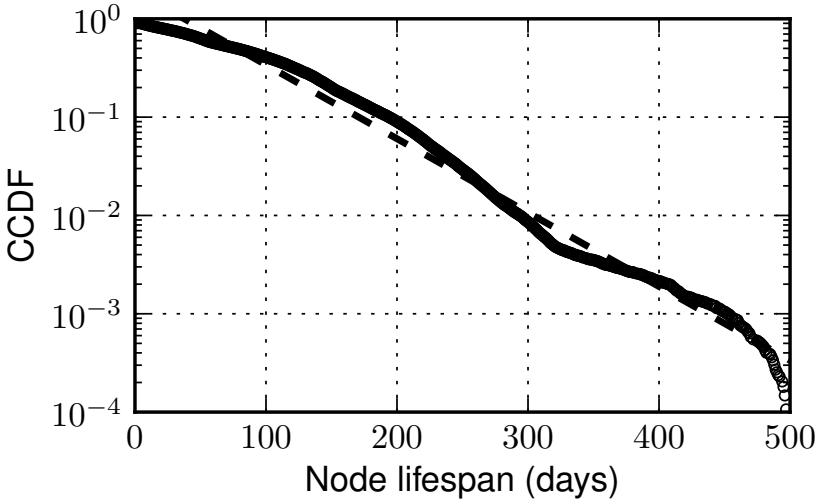


Figure 4.12: Complementary Cumulative Distribution Function (CCDF) of node lifespan and exponential fit.

4.3 Temporal aspects of network growth

After considering the edge attachment and the triangle-closing processes, we now shift our attention to the temporal properties of the network evolution. In this section we study how users create new connections as they spend more time on the network. Our aim is to understand these temporal patterns in order to capture them and reproduce them in a network growth model, rather than discussing their statistical description.

At first, we consider the amount of time users are active on the network, their lifespan, and then we investigate whether nodes with different degrees tend to establish new edges at a different pace and at different geographic distances. We consider only nodes that have joined the network after our measurement process started, in order to avoid any bias in the estimation of their temporal properties: in this way we are observing their entire temporal evolution from the very first moment they appear in the system.

4.3.1 Node lifespan

We define the *lifespan* of a node to be the temporal difference between the time the node created its last edge and the time when the node joined the service. The lifespan of a node is likely to affect the network evolution, since nodes that cease to be active stop creating new edges, affecting the global properties of the whole network.

Figure 4.12 shows the distribution of lifespan for all users: the distribution shows

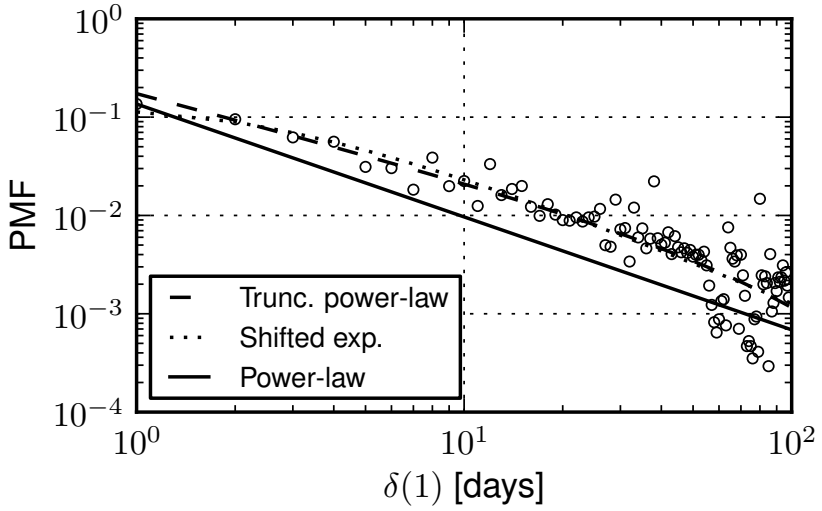


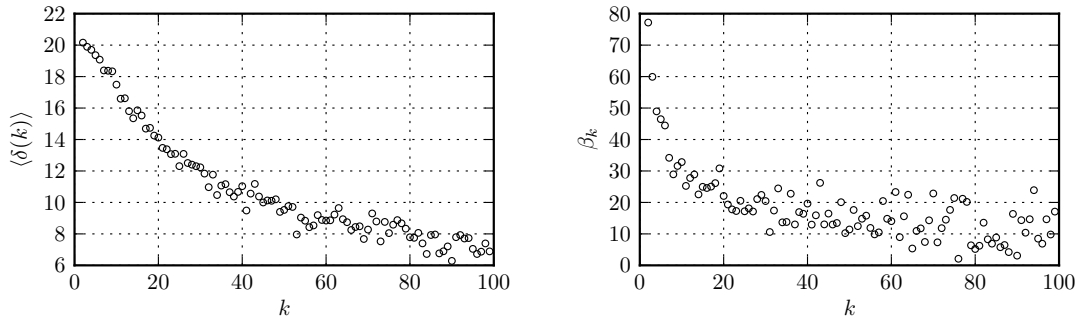
Figure 4.13: Probability Mass Function (PMF) of $\delta(1)$, the temporal gap elapsing between the times when the first and the second edges are created by a user. The fits are a truncated power law $\delta(1)^{-\alpha}exp(-\delta(1)/\beta)$, a shifted exponential $\delta(1)^{\beta-1}exp(-\lambda\delta(1)^{\beta})$ and a power law $\delta(1)^{-\alpha}$.

approximately exponential behaviour, with a deviation only at longer lifespans for few users who were early adopters and started using the service from the very first days. The fit is reasonably good for a wide range of lifespan values and it can be used to capture and reproduce how long nodes stay active in the network.

4.3.2 Inter-edge temporal gap

Different users can show significant differences in the pace at which they add new edges: some users can be faster and more active than others. In addition, users who have been active for a longer period on the service may attract new friends at a faster rate. We define $\delta_i(k)$ as the temporal gap between the k-th and $k+1$-th edges of user i and we study the distribution of $\delta(k)$ across all users for different values of k.

Figure 4.13 displays the probability distribution of $\delta(1)$, the amount of time between the first and the second edges created by a user. Even though many users add their second edge after a few days, some users wait several weeks. Hence, there is a wide range of variability in how quickly nodes start adding new edges after they join the network. The distribution can be captured by different functional forms: we find that an exponentially truncated power law $p(\delta(1)) \propto \delta(1)^{-\alpha_1}exp(-\delta(1)/\beta_1)$ yields a slightly higher log-likelihood than a pure power-law, a shifted exponential and an exponential, even though the average log-likelihood improvement over the exponential fit is below 5%. For our modelling purposes we will use the exponentially



(a) (b)

Figure 4.14: Arithmetic average value $\langle\delta(k)\rangle$ (a) and exponential cut-off β_k (b) of the truncated power law $p(\delta(k)) \propto \delta(k)^{-\alpha_k} exp(-\delta(k)/\beta_k)$, which approximates the probability distribution of the temporal gap $\delta(k)$ between the k-th and $k+1$-th edges as a function of node degree k.

truncated power law, which provides the highest likelihood even for different values of k.

Then, we study the effect of the current degree k on the temporal behaviour of the user: in particular, we are interested in seeing whether the probability distribution of $\delta(k)$ changes with k. A first indication is given in Figure 4.14(a), which plots the average temporal gap $\langle\delta(k)\rangle$ between the k-th and $k+1$-th edge for different values of k: users with higher degrees tend to wait, on average, for a shorter amount of time before adding a new edge. In fact, users wait on average 20 days before adding their second edge but only 7 days when they have about 100 friends.

It is not surprising that nodes with higher degree add links at a faster pace: given a fixed temporal period, as in our measurement, higher degree nodes add more links than lower degree ones, so their activity has to be greater in the same temporal period. Nonetheless, we are still interested in capturing this heterogeneous temporal behaviour, as this fosters heterogeneity in the degree distribution as well [LBKT08].

We find that the same truncated power law $p(\delta(k)) \propto \delta(k)^{-\alpha_k} exp(-\delta(k)/\beta_k)$ holds for different values of k, always offering the fit with the highest log-likelihood. While we find that α_k tends to be unrelated to k, the exponential cut-off β_k becomes smaller as k grows larger, as seen in Figure 4.14(b). The effect of this trend is that nodes with higher degrees are less likely to wait for a longer time span, as the truncated tail of the power law $P(\delta(k))$ increasingly constrains higher gap values.

4.4 A new spatial model of network growth

In this section we build upon the set of results found in this chapter and we propose a new model of network growth with spatial information. Our model combines a *gravitational attachment* process with a triangle-closing model to algorithmically grow a spatial network edge-by-edge. We demonstrate that the resulting synthetic network exhibit social and spatial properties of the true network, whereas a similar model without the effect of geographic distance does not.

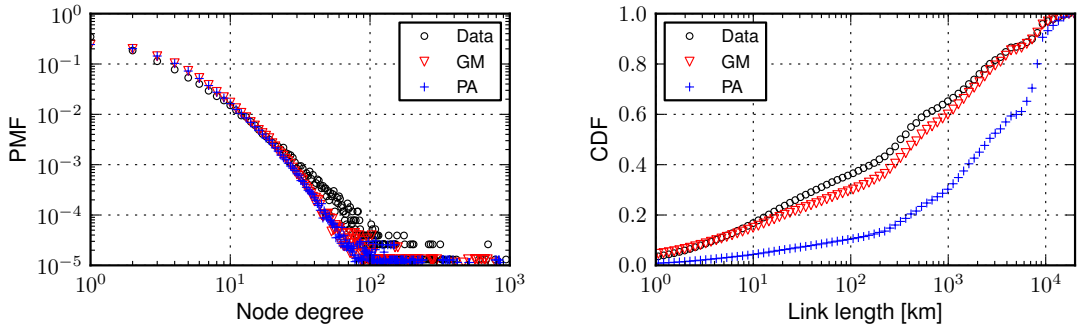
We have seen that a gravity-based mechanism describes how new edges are created (in Section 4.2.2), while we have discussed how triadic closure is mainly shaped by social factors rather than geographic ones (in Section 4.2.3). These two mechanisms seem to be complementary; while the former is responsible for edges connecting together different parts of the network, the latter seems involved in the creation of local edges between nodes that already share a friend. We analysed how nodes tend to create new edges faster and faster as they acquire more connections in Section 4.3. Building on all these results our aim is now to define a network growth model that is able to capture and reproduce the spatial and social properties of the real network.

4.4.1 A new gravitational attachment model

Following the methodology presented in [LBKT08], we describe our model as an algorithm to grow a network one node, and one edge, at a time:

1. A new node u joins the network and positions itself in physical space.
2. Node u samples its lifetime from an exponential distribution;
3. Node u adds its first edge to some node v according to a gravity model, with probability directly proportional to the degree of v and inversely proportional to their geographic distance;
4. Node u with degree k samples a time gap δ from a truncated power-law probability distribution, dependent on degree k, and then goes to sleep for δ time steps;
5. When node u wakes up, if its lifetime has not expired yet it creates a two-hop new edge using the $\mathsf{random} - \mathsf{random}$ triangle-closing model and then repeats step 4.

The different processes included in this model are meant to reproduce the most important properties observed in spatial social networks. The combination of exponential node lifetimes and degree-dependent inter-edge waiting times allows few



(a) (b)

Figure 4.15: Comparison of our gravity-based model (GM), preferential attachment model (PA) and real network: Probability Mass Function (PMF) of node degree (a) and Cumulative Distribution Function (CDF) of link geographic length (b).

nodes to accumulate a higher number of connections, resulting in a heavy-tailed degree distribution. The gravitational attachment favours geographically shorter links and introduces correlations between the number of connections and the spatial properties of these connections, replicating the heterogeneity observed across users. Finally, the purely social triangle-closing mechanism mimics how social triads are not affected by spatial constraints.

4.4.2 Evaluation

In order to test our model we take the Gowalla network at the beginning of our measurement period, denoted with T_1 and we simulate its growth by adding the missing nodes, with their real geographic locations, according to when they joined the real network. However, once they join the network they add new edges according to our algorithmic model. We stop the evolution of the network when it reaches the same number of edges as the real graph at the end of the measurement period, or when all node lifetimes have expired.

To assess better the performance of our model, we compare it to a similar model where in Step 3 a node creates an edge according to the preferential attachment model, thus ignoring geographic distance and considering only node degree in the attachment probability. We refer to our new gravitational attachment model as GM and to the preferential attachment model, which ignores spatial properties, as PA. Both models include the same triadic closure and inter-edge time gaps. The two models are run 100 times with different random seeds and then their properties are averaged over all these realisations. We compute and compare the properties of the networks by only considering edges added after T_1, both in the real network and in the simulated models, to avoid the properties of the initial graph G_{T_1} influencing

the final result.

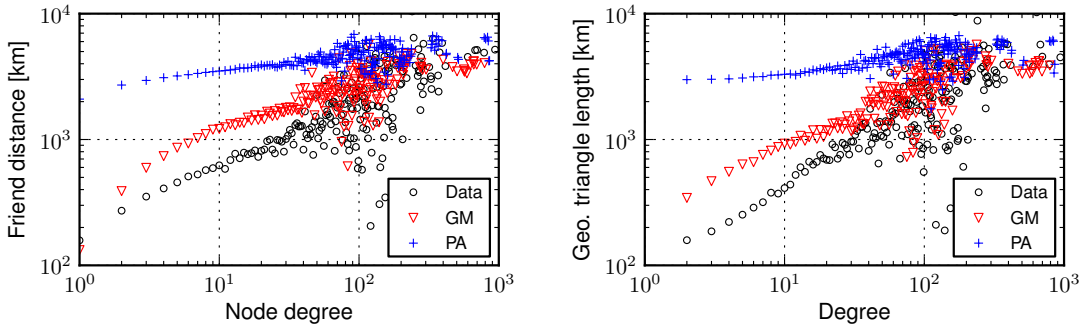
In our comparison we consider four different characteristics. First, we compute the degree distribution and the probability distribution of geographic link length, in order to assess whether these basic social and spatial properties are correctly reproduced. The degree distributions observed in the real network and in the two models are shown in Figure 4.15(a): both models are able to reproduce the distribution, replicating the social properties of the real network. As shown in Figure 4.15(b), the probability distribution of link geographic length is better approximated by the GM model, while the PA model results in social links with longer geographic length. As expected, the PA model fails to create those short-range links found in the real network.

Then, we focus on users and on their heterogeneity using two measures already described in Section 3.2, briefly summarised here. For every user we compute the *friend distance* and we plot the geometric mean value of this measure for all users with k connections, as a function of k. Similarly, for every user we compute the *geographic triangle length* and then we plot the geometric mean value of this measure for all users with k connections, as a function of k. We adopt the geometric mean to combine the values of users with the same degree because such values span several orders of magnitude and we aim to emphasise smaller values that correspond to short-range distance. These two user measures will shed light on whether the two models capture the correlations between user degree and socio-spatial properties that we discussed in Chapter 3.

These two measures highlight a large difference between the GM and PA models. When considering the average friend distance of a user as a function of the user degree, as seen in Figure 4.16(a), both the GM and the PA models show an increasing trend, as the original data. However, the PM model results in values systematically higher than the GM model. Similarly, when considering the average geographic triangle length in Figure 4.16(b), we find that the GM model reproduces the increasing trend observed in the real graph more closely than the PA model, which fails to capture how users with fewer than 10 connections exhibit low values of geographic triangle length.

4.5 Discussion and implications

Our results show that the effect of geographic distance cannot be neglected when online social networks are studied and modelled: preferential attachment mechanisms need to be modified into gravity-based mechanisms, which are able to correctly balance the effects of node attractiveness and the connection costs imposed by spatial



(a) (b)

Figure 4.16: Comparison of our gravity-based model (GM), preferential attachment model (PA) and real network: average friend distance for all nodes with a given degree (a) and average geographic triangle length for all nodes with a given degree (b), both as a function of node degree.

distance. This finding gives rise to interesting implications about processes that may be driving the actions of individuals.

In reality, preferential attachment and triadic closure together are already able to reproduce the global social properties observed in real social networks, namely the degree distribution and the level of clustering. However, neglecting spatial information about where users are located fails to account for the effect of distance: while in real systems users preferentially connect to other users at closer distances, resulting in a considerable fraction of short-range social ties, models without spatial information give rise to an unlikely majority of long-range connections. This is at odds with plentiful empirical evidence, both in offline and online social systems.

Our findings support the idea that distance has a simple effect on the creation of social ties: the probability of connection between two individuals decreases as a negative power of the spatial distance between them. Yet, this effect must be combined with a process based on "popularity" or "visibility" that introduces heterogeneity across users, such as attachment to the best connected nodes, in order to fully recreate the self-reinforcing mechanisms that lead to the scale-free degree distributions observed in social graphs.

Gravity models, already widely and successfully used to understand several types of spatial systems, provide an elegant and insightful way of combining the effect of distance and the influence of popularity. The main implication of the gravity mechanism is that one user may be interested in another because the other user is hugely popular, regardless of their spatial distance, or because the other user is spatially close, *regardless of popularity and importance*. The underlying message, which goes beyond the specific scenario of network growth, is that the powerful First Law of Geography is still at work: near things are more related than distant

things, as stated by Tobler in 1970 [Tob70].

Surprisingly, the influence of distance on the formation of local network structure appears negligible. Triadic closure does not seem affected by geographic proximity between individuals; this suggests that network transitivity is chiefly ruled by social factors that seem blind to geographic constraints. The overall picture is that proximity both over space and over the social fabric greatly fosters the creation of new social links; the result is that the likelihood of a new connection increases when two individuals share many other connections or when two individuals are close to each other. We also point out that no friend recommendation mechanism was in place on the online service under analysis during the measurement period.

This dual rôle of distance in the social and in the spatial dimension has promising applications in a wide range of systems. In particular, while the predictive power of social proximity has already been harnessed by a plethora of friend recommendation systems, spatial closeness has largely been ignored in such scenarios. However, geographic proximity is a powerful but simplistic indicator of social ties that are likely to be created. In fact, our model is able to reproduce the global social and spatial properties observed in the real traces, but it could be unable to accurately detect whether two users are going to connect to each other. This lack of precision is due to a lack of information: the spatial distance between the locations of two users is only one variable. In reality, there is more information about user spatial movements, since we can tap into a rich data source: the places that users visit. Thus, geographic distance can be used together with data about user check-ins to provide a more comprehensive picture of user behaviour, likely to offer higher accuracy in modelling how new social ties are established. In Chapter 5 we will discuss how friend prediction systems can be designed based on this consideration.

4.6 Related work

The temporal patterns of network evolution have been the focus of many studies and several models been put forward to describe the basic mechanisms that may drive network growth.

One of the first models of temporal network growth was the simple yet powerful Barabási-Albert (BA) model [BA99], based on two key ingredients: growth and preferential attachment. Inspired by how new Web pages tend to link to already popular pages, this model reproduces the scale-free degree distribution observed in several network systems across different scenarios. The rationale behind the BA model is to focus on the network as a dynamic entity under continuous evolution: hence, by mimicking the dynamic mechanisms that assemble the network over time,

one will be able to reproduce the topological properties of the system at the current time.

The properties of the networks generated by the BA model have been extensively studied and discussed; in particular, we note that BA graphs tend to show a vanishing level of clustering as the system grows in size, and also tend to exhibit negative degree-degree correlations. Hence, the BA model has attracted a considerable amount of attention in the literature from authors who have tried to modify its basic mechanisms to introduce such characteristics that are often found in other networks, such as in social graphs.

A set of works studied the temporal evolution of online social networks, discussing global properties such as densification and diameter reduction observed during the growth of the graph [FMNW03, KNT06]. Even though online social graphs tend to have an heterogeneous degree distribution, in agreement with the preferential attachment principle, these findings highlighted that, in social networks, different mechanisms seem to be in place. Leskovec et al. [LKF05] propose a "forest-fire" copying process: when a new node joins the network and connects to a first neighbour, it starts recursively creating new links among the neighbours of the first neighbour, effectively copying the connections already in place. This process mixes preferential attachment, as more connected nodes are more likely to be selected, and transitivity, which fosters new connections between nodes in social proximity. This confirms that triadic closure is an essential ingredient in the evolution of social graphs, to generate transitivity and community structure.

Several other works have focussed on the importance of triadic closure for social network evolution: Simmel noted that people sharing many friends might be more likely to become connected [Sim08]. This effect was then measured in real social networks [LNK07, KH06] and included in growth models. With respect to these results, our work explores, for the first time, the effect of spatial distance on network evolution. Specifically, we study how distance influences growth mechanisms such as preferential attachment and triadic closure.

Another large body of work has focussed on general models for spatial networks. One of the earliest examples is the Waxman model, where nodes are distributed at random over space and then connected with probability exponentially decreasing with distance [Wax88]. The Waxman model has also been modified as a growth model, where new nodes join the network and connect using a similar rule [KH04]. Barthélemy proposed to combine the preferential attachment rule with spatial distance, studying how the resulting graphs move away from being scale-free as the effect of spatial distance is increased [Bar03], however this case only considered an exponential decay of the effect of distance as in the original Waxman model. Barrat et al. [BBV05] also considered a similar model for weighted networks where prefer-

ential attachment is driven by the weight of the existing connections and hampered by greater spatial distance.

While these works contain the initial ideas related to modifying preferential attachment with spatial influence, they were based on spatial systems such as transportation networks that lacked social properties. Hence, they tend to focus on an exponential decay of the probability of connection as a function of distance, different to what is observed in social graphs, and they ignore properties arising from triadic closure. Our contribution builds on these findings and brings together several different insights in order to obtain a suitable model for spatial social graphs.

Finally, we adopt the maximum likelihood methodology, and we base our growth model on results presented in [LBKT08], where the evolution of four different online social networks was discussed. Again, our work differs as it addresses the effect of geographic distance on the temporal mechanisms that govern network evolution, providing a more complete understanding of the factors driving social behaviour. Furthermore, we describe a model of network growth that successfully reproduces both social and spatial properties observed in real social graphs.

4.7 Summary

In Chapter 3 we put forward the idea that spatial distance still impacts online social services. Users tend to exhibit heterogeneous socio-spatial properties correlated with their number of friends: as they have more and more connections, these connections tend to span longer geographic distances. We suggested that, as found in many other spatial networks, preferential attachment tends to be mitigated by mechanisms akin to gravitational attraction: nodes still tend to connect to high-degree hubs even when these are far away, but less connected nodes can still attract short-range connections from nodes nearby.

In this chapter we have extended our results concerning the structural properties of the spatial social graph by investigating the effect of geographic distance on the temporal evolution of a social network. Based on our findings, we have defined and tested a gravitational attachment growth model that reproduces the structural properties observed in the real spatial social network. This new model highlights basic factors driving network evolution which could greatly impact a vast range of research efforts and practical applications devoted to spatial social networks.

Our model relies on triangle-closing mechanisms to create new edges and grow the network. However, this fails to reproduce how new links can be created even between users that do not share any friend, despite these links being a minority of all new connections. In addition, our model only considers spatial distance between users

when modelling their behaviour, but in reality users might exhibit more complex patterns that require more detailed data about where they are located and where they go. To overcome these limitations, we need to exploit a different source of information to connect users who might not share any friend, but who may geographically close. In the next chapter we will investigate how the physical places that people visit can not only bridge this gap, but also help to obtain precise and accurate predictions about future online social connections.

What I cannot create, I do not understand.

Richard Feynman

5

Link prediction in location-based services

Online social services greatly benefit from recommending new friends to their users, since as users add more and more friends their engagement with the service increases. Hence, link prediction systems have been widely deployed to find which users should be recommended. However, the prediction space faced by these systems is huge and highly imbalanced: given a user, the overwhelming majority of other users are not likely to be suitable friend recommendations. Real recommendation systems merely focus on finding friends in the 2-hop social neighbourhood, i.e., friends-of-friends of a user. For instance, a popular Facebook feature is "People You May Know": launched in 2008, it suggests friends-of-friends that are likely to be suitable for new social connections [Rat08]. As this example suggests, extending prediction efforts to the 3-hop neighbourhood, or even further, may not be worth the effort.

The predominance of new links between users sharing at least a common friend was also confirmed in Chapter 4: a large fraction of all new connections arising in online social networks tend to arise between users exactly two hops away from each other. However, we also discussed how geographic proximity is a factor that impacts new connections. Specifically, ties arising between users that are several hops away seem to be established between spatially close individuals. In other words, being close in space could be as important as being close in the social graph, in order to create a new social tie. Nevertheless, geographic proximity alone could be a simplistic and imprecise indicator of potential future connections, as it lacks the richness of properties that can be exploited when considering the social connections between

two friends-of-friends. This implies that the model of network evolution presented in Chapter 4 could offer insights about what factors mainly drive user behaviour, but would only reproduce the average behaviour when directly applied to predict which individual social ties are likely to be established. A link recommendation system should offer greater accuracy: this requires more information about user actions and more complex models.

In location-based social services there is an unprecedented source of useful additional information about future connections: the places visited by each user. These places also offer a set of promising candidates. Given a certain user, data about venues and check-ins can be exploited at first to select a subset of users as prediction candidates: these "place-friends" represent pairs of users that share at least one common place among their check-ins. Then, the same information can be exploited to identify the candidates that are more likely to become actual social connections.

By exploiting the properties of the venues visited by users it is possible to provide additional predictive power to augment purely social approaches, considering how and where different users check in. In this chapter we build upon these initial considerations with a practical goal in sight: to design a link prediction system for new social connections that exploits data about user check-ins.

Chapter outline To investigate the practical feasibility of our proposal, we study longitudinal data about an online location-based service, Gowalla, with information about friendship connections and check-ins. In Section 5.1 we analyse the link prediction space by investigating how new friendship connections are created over time: we discover that *about 30% of all new links appear among users who check in at the same places*. Thus, these "place-friends" represent disconnected users that can become direct connections.

Hence, we argue that effective link prediction on location-based services can greatly benefit from focussing only on the friends-of-friends and on the place-friends of a user. The challenge is how to exploit the information given by the check-ins of two users, who do not share any friends but who visit the same places, to predict whether they will become direct connections. Towards this goal, in Section 5.2 we define prediction features which quantify how likely users are to become friends considering the places they visit and the properties of these places.

In Section 5.3 we present the proposed prediction system: prediction features based on visited places, combined with other measures, are exploited in a supervised learning framework to predict future links. Our evaluation in Section 5.4 shows the effectiveness of our design choices; the inclusion of information about places and related user activity offers high link prediction performance. These results open new directions for real-world link recommendation systems on location-based social networks,

t	Users	Active users	Places	Check-ins	N	K
1	252,020	148,234	958,823	7,475,401	109,045	476,409
2	291,812	168,925	1,104,771	9,073,157	124,190	559,901
3	325,025	189,512	1,226,847	10,537,516	138,387	630,045
4	382,750	216,734	1,421,262	12,846,151	159,391	736,778

Table 5.1: Properties of our Gowalla dataset across the different temporal snapshots: total number of registered users and active users, total number of different places, total number of check-ins, number of nodes N and edges K in the social graph.

as we discuss in Section 5.5. We present an overview of related results in Section 5.6 and conclude in Section 5.7.

5.1 The importance of place-friends

In this section we analyse how new social ties are created by users of a location-based social service. Our aim is to understand what challenges a prediction system would face and how to overcome them.

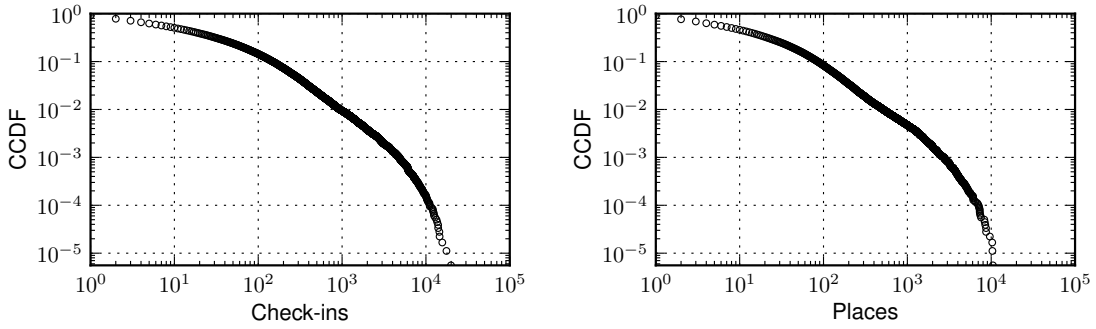
5.1.1 Snapshot properties

We have extracted four monthly snapshots of Gowalla data from the temporal dataset presented in Section 4.1.1. Each snapshot contains the social connections between users at that time and it includes all the check-ins made by users up until that time.

In the 4 consecutive monthly snapshots Gowalla increased its total number of registered users from about 250 thousand to about 380 thousand, as shown in Table 5.1. At each snapshot, about 56% of users are active, that is, have at least one friend or one check-in. Each snapshot of our dataset results in a social graph; as the average number of friends per user grows from 8.73 to 9.24, the social network remains sparse, making link prediction challenging because of the scarcity of social ties.

User check-in activity also presents a heavy-tailed distribution: 90% of users with check-ins have made fewer than 110 check-ins and have visited fewer than 95 different venues, as detailed by Figure 5.1(a) and Figure 5.1(b) Even though users might visit only a few places, users who visit the same places are still more likely, on average, to become friends than would be expected, as we will see later.

Finally, we note that while many users might have social connections and no check-ins, there are also many accounts with check-ins but no friends at all. On average, *only 57% of active users have both some friends and some check-ins, while 26% have*



(a) (b)

Figure 5.1: Complementary Cumulative Distribution Function (CCDF) of the number of check-ins (a) and of the number of places (b) per user for the last snapshot of the dataset (Month 4). The probability distributions do not change significantly across different snapshots.

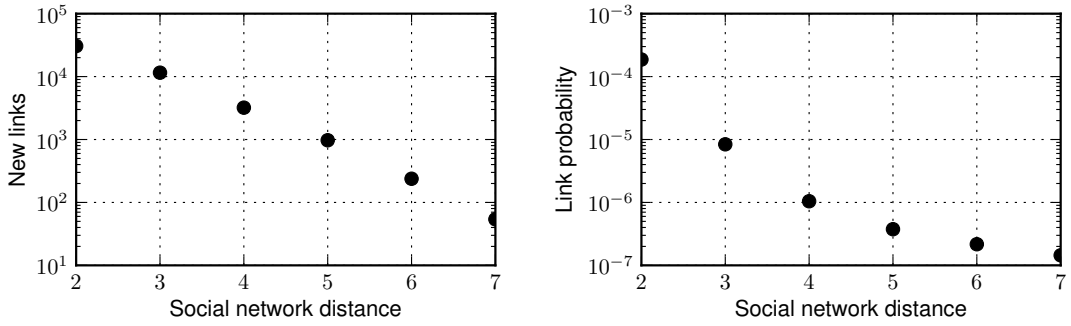
no friends and 17% have no check-ins. This is approximately constant across the temporal snapshots.

5.1.2 Definitions and notation

Formally, we represent each snapshot of our dataset as an undirected graph $G_t = (V_t, E_t)$ for $t = 1, 2, 3, 4$, where t indicates the different snapshots. The set of nodes $V_t = \{u_1, u_2, \ldots, u_{N_t}\}$ is composed of N_t users and the set of edges E_t is composed of pairs of users that are present in each other's friend lists in snapshot t. We define Γ_i^t to be the set of users connected to user u_i in graph G_t, so that $k_i^t = |\Gamma_i^t|$ is the number of friends of u_i in snapshot t. In addition, there are L_t different places $M_t = \{m_1, m_2, \ldots, m_{L_t}\}$ where users have checked in and c_{ij}^t represents the number of check-ins that user u_i has ever made at place m_j until time t. All the check-ins of user u_i until time t can also be represented as a vector $\vec{c_i^t} = (c_{i1}^t, c_{i2}^t, \ldots c_{iL_t}^t)$. Then, Φ_j^t is the set of all users who have checked in place m_j and Θ_i^t is the set of all places where user u_i has checked in, both until snapshot t. Finally, $A_t = \bigcup_{j=1}^{L_t} \Phi_j^t$ is the set of all users with at least one check-in at snapshot t, while $U_t = V_t \cup A_t$ is the set of all users present at snapshot t with at least one friend or one check-in.

5.1.3 Dividing the prediction space

Users adding friendship connections tend to prefer other users "close" to them, either in a social sense or along other dimensions such as geographic proximity or topic interest [LNK07, AA03, EPL09, QC09]. As also discussed in Section 4.2.3, many



(a) Number of new links

(b) Probability of a new link

Figure 5.2: Number of new links appearing among pairs of nodes at different social distance (a) and their relative probability of appearance (b). Pairs of users at closer distance are both generating a larger fraction of all social links and more likely to generate them.

new links appear between individuals at closer social distance to each other, with the 2-hop neighbourhood of single nodes being the largest source of new ties [LLC10].

This holds also for the Gowalla snapshots: as shown in Figure 5.2(a), the number of new links appearing between users who are d hops away exponentially decreases with d. The likelihood that a pair of users at network distance d will have a link in the next snapshot of our dataset decreases sharply with d, as shown by Figure 5.2(b): the probability that two users with at least one friend in common, thus being at social distance $d = 2$, will become friends is above 10^{-4}, but this value quickly drops below 10^{-5} and to 10^{-6} at distance $d = 3$ and $d = 4$, respectively. Hence, pairs of users at larger distances give a weaker contribution to link formation, both in terms of absolute number of new links and likelihood of a new social tie.

Nonetheless, in a location-based social network the social dimension is not the only one to be exploited and investigated. Instead, in our context there is an additional source of information about social ties: the places where users check in. In particular, users may add a new connection not because of a shared friend but because of a shared place.

In order to quantify how users seek and add new friends, for each snapshot and for each user u_i we define two sets of potential friend pairs:

Friends-of-friends

$$S_i^t = \{(u_i, u) : u \in \Big(\bigcup_{u_k \in \Gamma_i^t} \Gamma_k^t \Big) \setminus \Gamma_i^t\} \tag{5.1}$$

Place-friends

$$P_i^t = \{(u_i, u) : u \in \Big(\bigcup_{m_k \in \Theta_i^t} \Phi_k^t \Big) \setminus \Gamma_i^t\} \tag{5.2}$$

Snapshot t	1	2	3
U_t	148,234	168,925	189,512
E_t^{NEW}	43,182 (100.00%)	40,643 (100.00%)	58,238 (100.00%)
S_t^{NEW}	24,174 (56.41%)	21,118 (51.96%)	30,581 (51.51%)
P_t^{NEW}	13,150 (30.01%)	12,572 (30.93%)	20,107 (34.52%)
$S_t^{NEW} \cap P_t^{NEW}$	7,677 (17.52%)	7,131 (17.54%)	10,935 (18.78%)
$S_t^{NEW} \cup P_t^{NEW}$	30,187 (68.90%)	26,559 (65.35%)	39,753 (68.26%)

Table 5.2: Link formation: for each monthly network snapshot we report the total number of active users U_t, the total number of new links appearing among them in the next snapshot E_t^{NEW} and the breakdown of this quantity into new links appearing among friends-of-friends S_t^{NEW} and among place-friends P_t^{NEW}, including the intersection and union of these two latter sets. Percentages are computed with respect to the total number of new links.

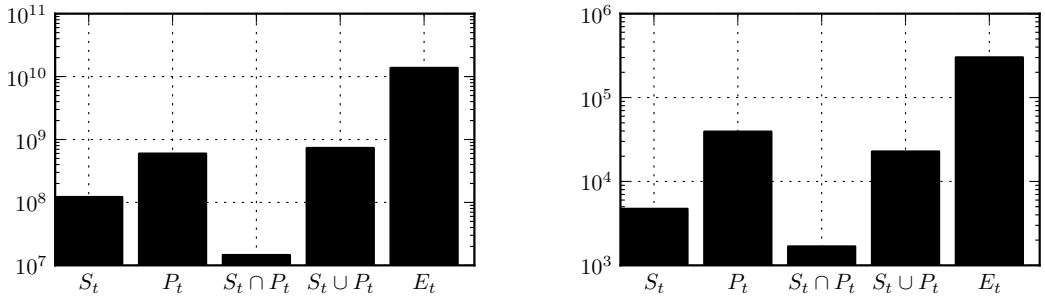
While friends-of-friends are all those users who share at least one friend without being directly connected, place-friends are all those users with check-ins in at least one common place but who are not connected to each other. These two sets may not be disjoint for a given user u_i. Finally, we define two sets containing all the pairs of nodes that are either friends-of-friends or place-friends in a given snapshot: $S_t = \bigcup_{u_i} S_i^t$ and $P_t = \bigcup_{u_i} P_i^t$.

The monthly snapshots of our dataset make it possible to quantify how many new social links appear within these two sets. For every network snapshot $G_t = (V_t, E_t)$ we define $E_t^{NEW} = E_{t+1} \cap ((U_t \times U_t) \setminus E_t)$ as the set of all new links appearing in the next network snapshot $t+1$ between all users already present at snapshot t. In Table 5.2 new links appearing between temporal snapshots are classified according to their origin: $S_t^{NEW} = E_t^{NEW} \cap S_t$ and $P_t^{NEW} = E_t^{NEW} \cap P_t$ are, respectively, the set of new links between friends-of-friends and the set of new links among place-friends.

About two-thirds of all new links appear within $S_t \cup P_t$. In particular, while about 50% of new links appear between friends-of-friends, more than 30% of new links are added between place-friends who check in at the same venues. Finally, about 13% of new links appear between users without any friends in common but who are place-friends.

5.1.4 Reducing the prediction space

In addition to the absolute number of new links appearing between friends-of-friends and place-friends, it is also important to study how link prediction feasibility can vary across these prediction spaces. In a prediction space there are both pairs of users who will become connected and pairs who will not: the performance of prediction



(a) Prediction space size

(b) Imbalance ratio

Figure 5.3: Number of potential friends (a) and imbalance ratio (b) for each class of potential new links: for social potential neighbours S_t, for place potential neighbours P_t, for their intersection, for their union, and for the entire set of users E_t. Results averaged over the temporal snapshots.

approaches depends on the total number of these potential pairs and on the relative proportion of these two classes. Exhaustive approaches would scale with the total number of potential links, which can become prohibitively large for real-world online social networks with millions of users. Also, the two classes can present an extremely skewed distribution, with new links being greatly outnumbered by pairs of users who will never create a social tie. This problem is worsened by the fact that new links are actually the occurrences of greater interest, as prediction systems obtain much more value when predicting that two users will connect than when correctly predicting that they will not.

In Figure 5.3(a) we report the prediction space size for the friends-of-friends set S_t and the place-friends set P_t, including also their intersection and union, along with the size of the overall prediction space for the entire dataset. While there are more than 11 billion pairs of users, there are about 700 million place-friends (P_t) and about 100 million friends-of-friends (S_t), with their intersection reducing the prediction space to about 20 million entities. Thus, by focussing prediction efforts only on place-friends or friend-of-friends the prediction space can be reduced by about 15 times, while still covering two-thirds of all new links.

Then, we study the *imbalance ratio* of a prediction set, which is the ratio between the total number of prediction candidates in the set and the actual number of new links that will appear within it. Imbalance ratios are key indicators of link prediction systems' performance: they express how many real instances should be considered and analysed, on average, before a prediction can be successfully made. Place-friends and friends-of-friends offer lower imbalance ratios than the overall prediction space, as shown in Figure 5.3(b): hence, not only do they offer a smaller prediction space, but the likelihood that new links will be found is also about 20 times higher than

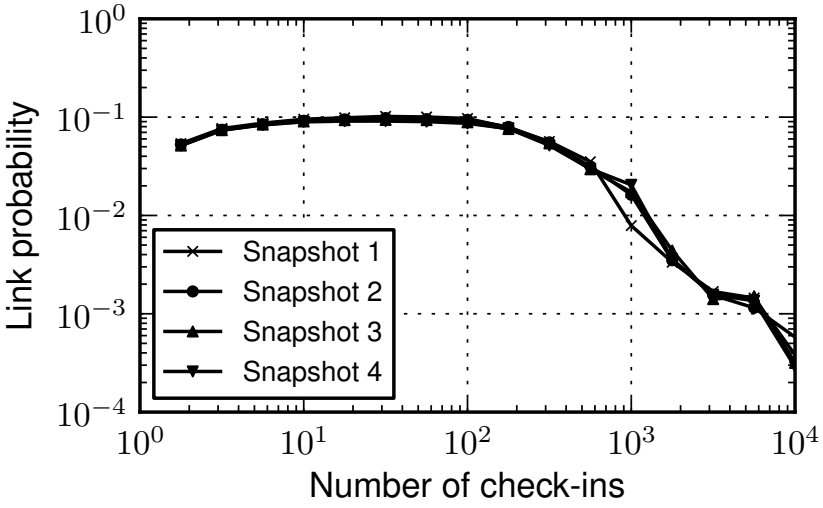


Figure 5.4: Average probability that two users who have checked in a at place are friends, as a function of the number of check-ins in that place.

the average.

However, discovering new ties between users who check in to the same places appears challenging. Not all places have the same importance for different users and, thus, not all places are equally likely to foster new social ties between individuals who visit them. The key idea is then to take advantage of the properties of a place to predict new links.

5.2 Building prediction features

In this section we will describe how the properties of the places visited by users can be exploited in link prediction systems. More broadly, we will also introduce a family of prediction features we will later adopt in our proposed design.

5.2.1 The social properties of places

Places can be characterised by taking into account users' check-ins: in fact, the average probability that two users who have checked in at the same place are friends exhibits a decreasing trend as the place has more check-ins, as shown in Figure 5.4. However, there is not much difference when a place has fewer than 100 check-ins.

A place where only a small number of users regularly check in is likely to be a place with significant importance to them, such as private houses, gyms, or offices. Conversely, a place with a similar total number of check-ins but where these check-ins have been made by many users is likely to be a public place without considerable

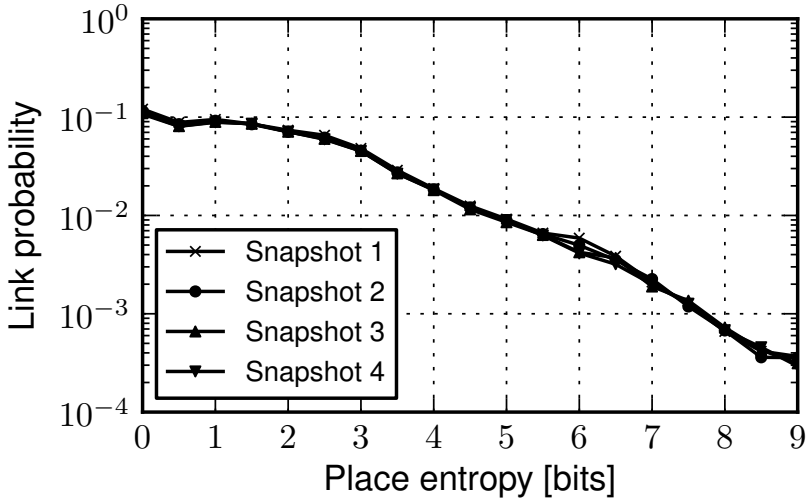


Figure 5.5: Average probability that two users who have checked in at a place are friends, as a function of place entropy.

significance to its visitors, such as touristic places, airports, train stations and so on.

Hence, a more suitable measure of how much a venue promotes social connections among its visitors should take into account both the number of users that check in and their number of check-ins. A feasible combination is to exploit information theory and define an entropy-based measure to assess the importance of a place for social link creation. *Place entropy* has been used in ecology to measure place biodiversity [CTH+10]: the underlying assumption is that a uniform distribution of species in a given physical environment is much more diverse than a skewed distribution, where only a few species are overwhelmingly present.

Let C_k^P be the total number of check-ins made by all users at place m_k and $q_{ik} = c_{ik}/C_k^P$ the fraction of check-ins that user u_i has made at location m_k with respect to the total number of check-ins at place m_k. Therefore $\{q_{1k}, \ldots, q_{Nk}\}$ is a discrete probability distribution that describes how likely it is that a check-in at m_k was made by a certain user. Then, we define H_k as the entropy of place m_k:

$$H_k = -\sum_{u_i \in \Phi_k} q_{ik} \log q_{ik} \tag{5.3}$$

Venues visited by several casual users are less likely to foster the creation of social links between them. Hence, places with higher entropy might result in fewer social links among their visitors than venues with lower values. This is confirmed by Figure 5.5: the average probability that two users who have checked in at the same place are friends decreases as the entropy of the place itself increases. Place entropy seems to have strong discriminative power; as we will see, it is a successful indicator of whether a certain place is likely to result in social ties between its visitors.

Place features	
common_p	$\lvert\Theta_i \cap \Theta_j\rvert$
overlap_p	$\frac{\lvert\Theta_i \cap \Theta_j\rvert}{\lvert\Theta_i \cup \Theta_j\rvert}$
w_common_p	$\vec{c_i}\vec{c_j}$
w_overlap_p	$\vec{c_i}\vec{c_j}/\sqrt{\vec{c_i}^2\vec{c_j}^2}$
aa_ent	$\sum_{m_k \in \Theta_i \cap \Theta_j} \frac{1}{H_k}$
min_ent	$\min(H_k : m_k \in \Theta_i \cap \Theta_j)$
aa_p	$\sum_{m_k \in \Theta_i \cap \Theta_j} \frac{1}{\log C_k^P}$
min_p	$\min(C_k^P : m_k \in \Theta_i \cap \Theta_j)$
Social features	
common_n	$\lvert\Gamma_i \cap \Gamma_j\rvert$
overlap_n	$\frac{\lvert\Gamma_i \cap \Gamma_j\rvert}{\lvert\Gamma_i \cup \Gamma_j\rvert}$
aa_n	$\sum_{z \in \Gamma_i \cap \Gamma_j} \frac{1}{\log(\lvert\Gamma_z\rvert)}$
Global features	
geodist	$\mathrm{dist}(m_{l_i}, m_{l_j})$
w_geodist	$\mathrm{dist}(m_{l_i}, m_{l_j})/c_{il_i}c_{jl_j}$
pa	$\lvert\Gamma_i\rvert\lvert\Gamma_j\rvert$
pp	$\lvert\Theta_i\rvert\lvert\Theta_j\rvert$

Table 5.3: Formal definition of prediction features: Γ_i is the set of users connected to user u_i, c_{ik} is the number of check-ins made by user u_i at place m_k, the vector $\vec{c_i}$ contains all check-ins of user u_i, m_{l_i} is the home location of user u_i, Θ_i is the set of all places where user u_i has checked in, C_k^P is the total number of user check-ins at place m_k and H_k is the entropy of place m_k.

5.2.2 Feature definition

Link prediction methods are based on numeric scores computed for pairs of users. These values tend to capture proximity of two users across different dimensions, with the underlying assumption that pairs of users who are similar or close are likely to develop a social connection between them.

We will consider *social features*, which can be computed for friends-of-friends, *place features*, which can be computed for place-friends, and *global features*, which can be computed for any pair of users, even if they do not share any friend or place. All features are described in Table 5.3 and discussed in the following paragraphs.

Place features

When two users check in to the same places they might have many chances to be in contact with each other and, therefore, to create a new connection. The two features `common_p` and `overlap_p` denote respectively the number and the fraction of common places between two users, while `w_common_p` takes into account the number of check-ins of both users and `w_overlap_p` is given by the cosine similarity of the two check-in vectors.

Then, we define two features based on the entropy of the places that two users share: `min_ent`, the minimum place entropy across all the shared venues, and `aa_ent`, the sum of the inverse of each place entropy value, a measure inspired by the Adamic-Adar similarity score [AA03]. Similarly, we define corresponding features considering the number of check-ins, `aa_p` and `min_p`: in this case the relevance of a shared place is higher if it has only a few check-ins.

Social features

Several link prediction features are based on the assumption that two users who share many common neighbours are more likely to create a direct connection. Thus, given two users we define `common_n` as their number of common neighbours and `overlap_n` as their Jaccard coefficient [Sal83]. In addition, `aa_n` is their Adamic-Adar measure based on the degrees of the shared neighbours [AA03].

Global features

Finally, we define measures that can be adopted for any pair of users, as they are based on their individual properties.

We define m_{l_i} as the "home-location" where user u_i has made the greatest number of their check-ins: this location might not be the place where a user lives, but it gives a reasonable estimation of the place a user seems most attached to. Then, given two users, we compute `geodist` as the geographic distance between their home locations. At the same time, `w_geodist` is the same distance divided by the product of the number of check-ins each user has made at their home location.

Another method to define global features is to consider how many friends users have added or how many places they have visited. We define `pa` as the preferential attachment score of two users, while `pp`, or *place-product*, is given by the product of the numbers of places that each user has visited. These two features tend to capture more active users who tend to visit many places or add many friends.

5.3 System design

In this section we describe our link prediction framework. Our proposal builds on two key choices:

- reducing the prediction space by focussing only on friends-of-friends and place-friends;
- exploiting prediction features based on the places visited by users.

We propose a *supervised learning approach* to link prediction, modelling it as a binary classification problem which adopts the prediction features previously described.

5.3.1 Prediction candidates

Let us consider a dataset snapshot, with U_t being the set of all users and $G_t = (V_t, E_t)$ the relative social network. The link prediction problem can be formulated as follows: given the dataset snapshot at time t as input, compute and return a set of pairs of users $E_t^{PRED} \subset (U_t \times U_t) \setminus E_t$ who are predicted to appear as friends in E_{t+1}.

The entire prediction space $(U_t \times U_t) \setminus E_t$ contains all the potential pairs between users that are not yet connected by a link. Recall that S_t represents the friends-of-friends prediction set and P_t denotes the place-friends prediction set, as defined in Section 5.1.3. Exploiting the findings of our previous analysis of this prediction space in Gowalla, we select three disjoint prediction sets:

1. **Social**: links appearing between users that are friends-of-friends but not place-friends (the set $S_t \setminus P_t$);
2. **Place**: links appearing between users that are place-friends but not friends-of-friends (the set $P_t \setminus S_t$);
3. **Place-social**: links appearing between users that are both friends-of-friends and place-friends (the set $S_t \cap P_t$).

Our choice is motivated by the fact that combining these three prediction sets results in a set of candidates about 15 times smaller than the entire prediction space while still allowing us to predict two-thirds of new social ties, as discussed in Section 5.1.4.

5.3.2 Prediction algorithms

We adopt a supervised learning approach: for every snapshot t, we compute features at time t for pairs of disconnected users and we assign a positive label to each pair if they become connected in snapshot $t + 1$, and a negative label otherwise. Thus, training and test sets are built so that features from a given time interval are mapped to class labels in a future time interval. Hence, given our 4 snapshots, we can create 3 learning sets, each one with labels drawn from the next snapshot.

Classifiers can then be trained to build models and recognise positive and negative items from their features. Motivated by recent results [LLC10], the choice of a supervised learning formulation to address the link prediction problem stems from the heavily skewed distribution of class labels. Unlike unsupervised methods, class distributions are learned by supervised algorithms, allowing more effective discovery of inter-class boundaries and hence better classification performance.

5.4 Evaluation

We now present the experimental evaluation of our link prediction system; this section includes an investigation of the predictive power of each similarity feature and then an analysis of different supervised classifiers that use these features. Our results show how link prediction systems based on our proposal may be feasibly deployed in similar services with high accuracy.

5.4.1 Evaluation strategy

For each snapshot t and for each prediction set we sample disjoint training and test datasets; these datasets are always sampled to maintain the original unbalanced distribution of positive and negative items in the real data. For every item we compute all available prediction features; the only limitations are that in the Social prediction set *place features* are not defined and in the Place prediction set *social features* are not defined. All our evaluation tests have been performed with the WEKA framework, which implements several machine learning algorithms, using default parameters (unless otherwise specified) [WF05].

We adopt Receiver-Operating-Characteristic (ROC) curves as the main tool to evaluate prediction performance [PF01]. ROC curves describe how the fraction of true positives over all the positive cases changes as a function of the fraction of false positive over all the negative cases when the decision threshold varies. A ROC plot is a monotonic non-decreasing plot of true positive rate as a function of false positive rate. A random classifier will result, on average, in the curve $y = x$, while

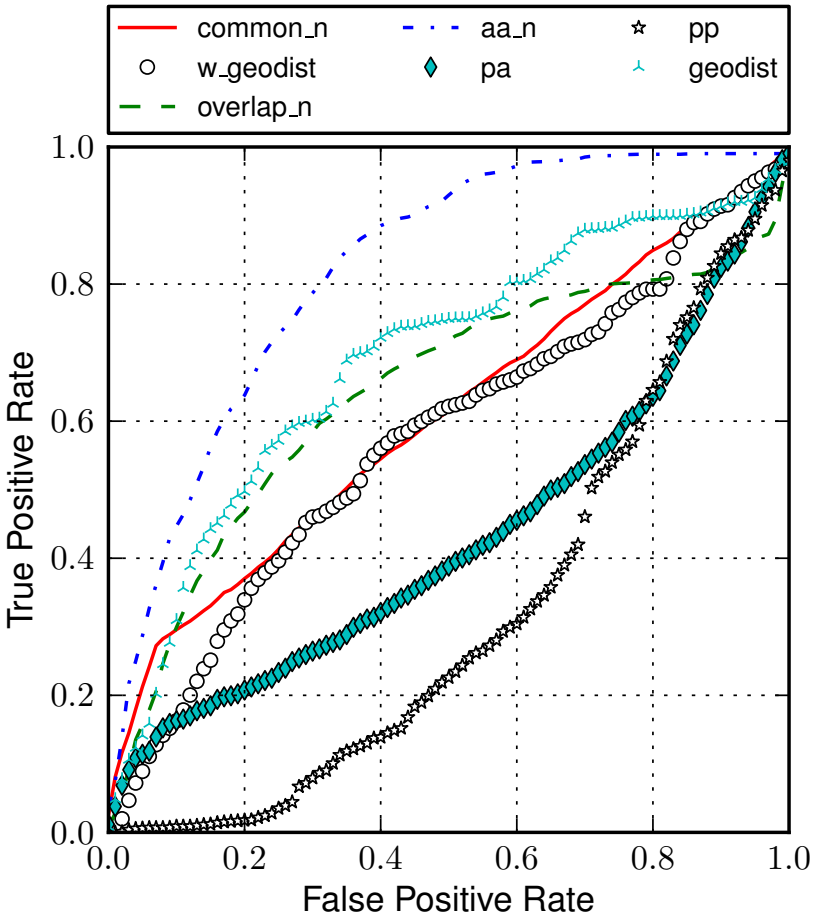


Figure 5.6: ROC curves for individual features used as unsupervised prediction methods on the Social prediction set.

better classifiers will result in curves closer to the upper left corner. ROC curves are particularly able to assess classification performance for highly imbalanced datasets, as in our case. The area under the ROC curve (AUC) is often adopted as a scalar measure of the overall performance.

5.4.2 Individual features evaluation

We first study the predictive power of each individual feature; we compute predictive scores for every pair of disconnected users in the test set and then we numerically rank these candidates according to their score. Given a decision threshold, new links are predicted for all the candidates with scores higher (or lower, depending on the directionality) than the threshold. As we vary the decision threshold we get true and false positives, generating a ROC curve; these curves are presented in

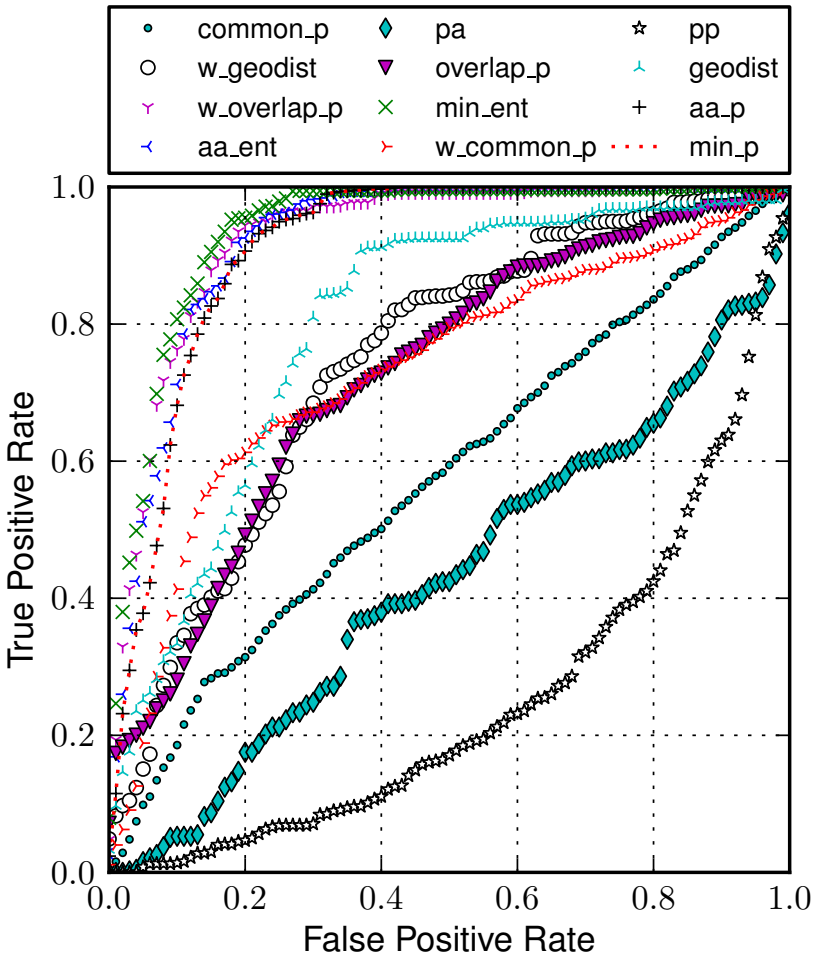


Figure 5.7: ROC curves for individual features used as unsupervised prediction methods on the Place prediction set.

Figures 5.6- 5.8 for each prediction set.

In the Social prediction space, as shown in Figure 5.6, the best feature is `aa_n`. Interestingly, we observe how the global features `pa` and `pp` perform worse than a random predictor. This indicates that in the social neighbourhood of a given user global indicators are not as useful as measures based on common friends: this may be a consequence of users having no access to a global view of the network. Instead, global features `geodist` and `w_geodist` perform better, with the former being more accurate than the latter. Overall, `aa_n`, `overlap_n` and `geodist` give the best performance, with AUC values between 0.73 and 0.82.

In the Place prediction space, as reported in Figure 5.7, `min_ent`, `w_overlap_p` and `min_p` show the best results, followed by `aa_ent` and `aa_p`. Sharing places with low entropy values or with a few check-ins seems an important indicator of potential friendship, as well as having a large overlap of visited places. These features achieve

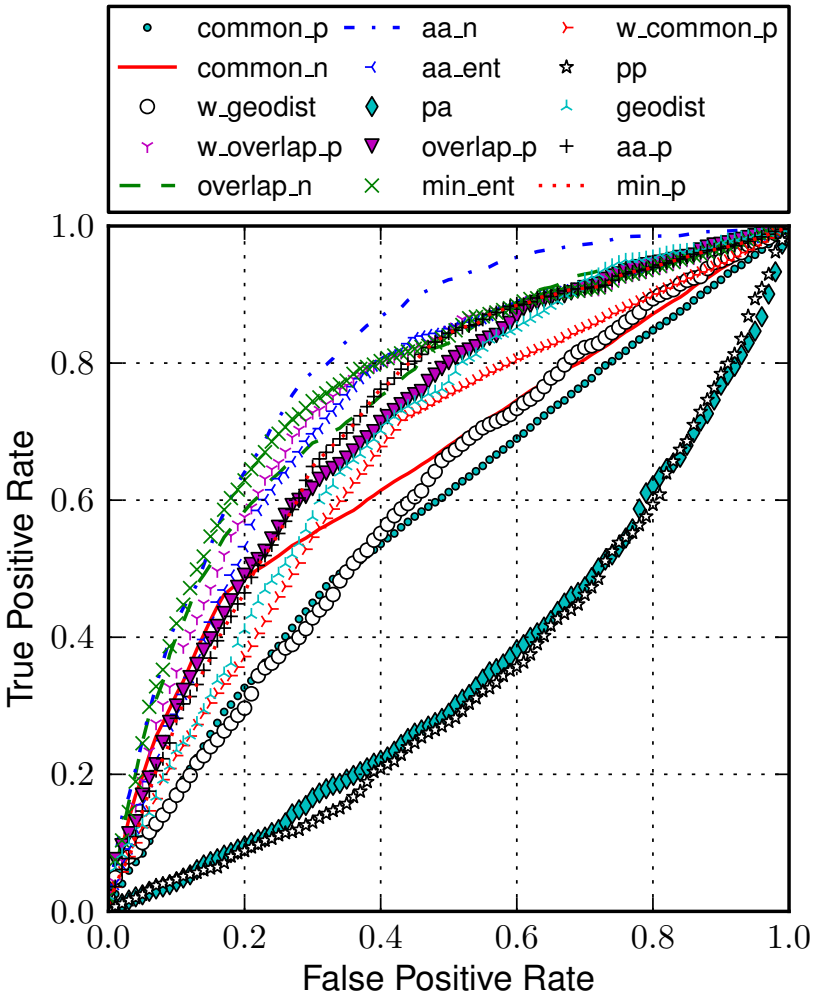


Figure 5.8: ROC curves for individual features used as unsupervised prediction methods on the Place-social prediction set.

high AUC values between 0.87 and 0.90. The other features perform slightly worse, with `geodist` doing better than the others. Global features `pa` and `pp` again show inverted performance, as in the Social case.

Finally, in the Place-social prediction space, as shown in Figure 5.8, all prediction features can be evaluated. Just as `aa_n` dominates in Social and `min_ent` dominates in Place, they also achieve the best results in this case, with the former having a larger AUC (0.80 against 0.76).

In general, prediction performance is higher in the Place set, while prediction within the other two sets achieves lower AUC values. It seems easier to predict links between place-friends than between friends-of-friends: this may be due to the fact that more information is available when two users share visited places than when they share friends. However, the prediction space is much larger in the Place set than in the other two sets, offering an interesting trade-off between prediction effectiveness and

Algorithm	Set	Precision	Recall	AUC
Model trees	S	0.79 ± 0.04	0.28 ± 0.05	**0.91 ± 0.02**
	P	0.87 ± 0.06	0.34 ± 0.06	**0.93 ± 0.01**
	PS	0.92 ± 0.03	0.62 ± 0.07	**0.96 ± 0.01**
Random forests	S	0.92 ± 0.05	0.39 ± 0.05	**0.91 ± 0.02**
	P	0.95 ± 0.04	0.72 ± 0.08	**0.94 ± 0.03**
	PS	0.98 ± 0.04	0.84 ± 0.09	**0.95 ± 0.01**
J48	S	0.63 ± 0.05	0.04 ± 0.01	0.62 ± 0.08
	P	0.86 ± 0.06	0.34 ± 0.04	0.90 ± 0.04
	PS	0.90 ± 0.03	0.64 ± 0.08	0.91 ± 0.02
Naïve Bayes	S	0.01 ± 0.00	0.16 ± 0.02	0.74 ± 0.06
	P	0.01 ± 0.01	0.36 ± 0.04	0.92 ± 0.04
	PS	0.04 ± 0.01	0.22 ± 0.05	0.82 ± 0.06

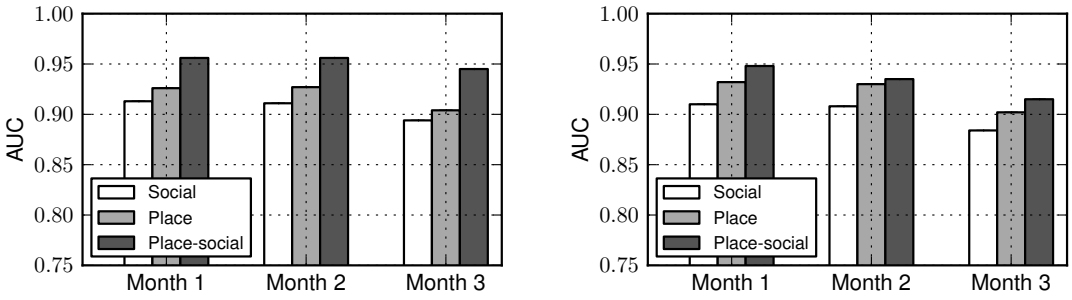
Table 5.4: Precision and recall for the positive items, and overall AUC for different supervised classifiers on the three different prediction sets Social (S), Place (P) and Place-social (PS). Results obtained through 10-fold cross validation and averaged over 20 different random training sets from snapshot $t = 1$.

search complexity.

In essence, the Social set provides good candidates for new links, given its lower imbalance ratio, but then it is difficult to discriminate between them because there is no other information except global features and shared friends. Even if the Place set has higher imbalance ratios, the properties of the places where users check in provide useful information to discover new friendship connections. Finally, the Place-social set, which provides the lowest imbalance ratio across the three sets, is still a source of good candidates like the Social set but since more information is available with respect to location-based user activity, prediction performance is better.

5.4.3 Combining features: supervised learning

We assess whether our prediction features can be combined to characterise a model of link formation across the three prediction sets. Our aim is to achieve at least the same predictive power as the best individual features with a supervised algorithm. We compare the performance of the following classifiers: J48 decision trees (equivalent to decision trees built using the C4.5 algorithm, based on information entropy [Qui93]), Naïve Bayes [Zha04], model trees (decision trees with linear regression models on the leaves [FWI+98]), and random forests (10 decision trees, 4 random features each) [Bre01]. We run 10-fold cross validation over 20 different training set sampled over each prediction dataset and we consider the AUC value as



(a) Model trees

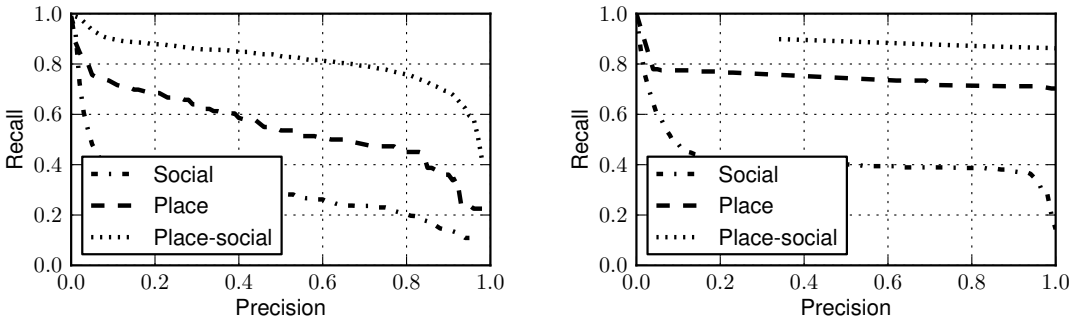
(b) Random forests

Figure 5.9: Prediction performance in terms of AUC of model trees (a) and random forests (b) on the three separate Social, Place and Place-social prediction sets, in each temporal snapshot. Results averaged over 20 random datasets.

an overall performance metric [Qui93]. In addition, we also consider two additional metrics computed over positive items: the average *precision*, that is, the fraction of positive predictions that are correct, and the average *recall*, that is, the fraction of real links that are correctly predicted.

We present our results in Table 5.4. There is variability across different classifiers: the best performance in terms of AUC is given by random forests and model trees, which are the only two methods that outperform individual features across the three prediction sets (the only exception being random forests underperforming on the Place set). Random forests present higher values of precision and recall than model trees.

As random forests and model trees outperform the other methods, we choose these two classifiers for the next part of this evaluation, where we consider prediction performance across consecutive temporal snapshots of Gowalla. In this case, for every snapshot and for each prediction set we sample disjoint training and test sets of equal size and we compute predictions, averaging results over 20 randomly sampled datasets. As seen in Figure 5.9, model trees achieve better AUC values for the three prediction sets and across temporal snapshots. Overall, the two algorithms have lower performance in the Social prediction set, with AUC values between 0.88 and 0.91, whereas Place and Place-social present higher values. Model trees offer slightly better performance than random forests: in particular, the latter algorithm performs worse than individual features on the Place prediction set. A potential explanation for this behaviour is that random forests tend to perform poorly when faced with a large heterogeneous set of features, since randomly chosen features are more likely to include less relevant information [GGCM08]. This may be the case for the Place set, while this is not the case for the Social set, where there are fewer features, nor for the Place-social set, where there are more features but their



(a) Model trees

(b) Random forests

Figure 5.10: Precision-recall curves for model trees (a) and random forests (b) obtained for the three separate prediction sets, averaged across the three temporal snapshots.

prediction performance is more homogeneous. However, investigating the precision-recall trade-off offers a different insight into the prediction performance. Given the same level of precision, random forests consistently achieve higher values of recall than model trees, as described in Figure 5.10. In summary, our prediction framework offers high effectiveness with both methods, since they are able to leverage the information contained in our prediction features.

Finally, to understand the extent to which different feature classes contribute to prediction performance, we focus only on the Place-social prediction set, where all features are used to build the prediction model, and we test the prediction performance that can be achieved by using only one feature class, as compared to the full model. As described in Table 5.5, social features alone show the worst performance, while both place and global features achieve AUC values closer to the full model. Hence, these two latter classes are mainly contributing to the overall performance, as they exploit information about place check-ins (Place features) and geographic distance between users (Global features). Again, this provides evidence that including data coming from location-based activity in the prediction model leads to better performance than purely social-based methods.

5.5 Discussion and implications

Our results arise from two main important design choices: focussing link prediction only on a reduced set of candidate pairs of users and exploiting location-based user activity to define successful prediction features. These two simple ideas are able to improve overall performance of link prediction systems; as a consequence, real-world systems can be deployed, making use of predicted links to suggest friends to users and engage them more with the service. In addition, recommending to a user others

Classifier	Full model	Social	Place	Global
Model trees	0.953 ± 0.015	0.907 ± 0.021	0.927 ± 0.011	0.925 ± 0.012
Random forests	0.932 ± 0.013	0.881 ± 0.019	0.923 ± 0.009	0.928 ± 0.012

Table 5.5: Average AUC and standard deviation for model trees and random forests for the Place-social prediction set when the full set of prediction features is used, and when only a single set of prediction features is used. Results averaged over the three snapshots and over 20 different random training and test sets for each snapshot.

who check in to the same places may be more important in location-based services, since users can directly interact with them when checking in to these places.

Our framework enables the prediction of new social ties even for users who do not yet have any friendship connection, provided that they visit and check in at places. Standard link prediction methods based on social features are of no use in this scenario, since it is impossible to compute prediction features for these isolated users [LCB10]. In some sense, this is a scenario that represents new users of the service: they have signed up, they have checked in at some places but they are not engaging with other users. Thus, predicting their future links might be extremely important to make them more active participants.

5.6 Related work

The link prediction problem in social networks has been under scrutiny for many years. The seminal work by Liben-Nowell and Kleinberg addresses the problem from an algorithmic point of view, investigating how different proximity features can be exploited to predict the occurrence of new ties in a social network [LNK07]. They adopt an unsupervised approach, where scores are computed for all potential candidates and then ranked to obtain the most likely predictions.

More recently, researchers have advocated supervised approaches to link prediction, given the possibility of modelling the task as a binary classification problem. In particular, Lichtenwalter et al. have presented a detailed analysis of challenges in link prediction systems, discussing imbalance problems and proposing to treat prediction separately for different classes of potential friends [LLC10]. While we also adopt a supervised approach, we additionally consider how link prediction can be performed when information not arising from social ties is available.

A related approach to finding online social ties between mobile users has been presented by Cranshaw et al. [CTH+10]: they track a small number of mobile users in the physical world to discover their connections on online social networks. While also focussing on information-based measures, our approach considers a much larger set

of users and studies their activity on a location-based service. Eagle et al. have considered how interactions between people over mobile phones can accurately predict relations between them [EPL09]. Conversely, we consider neither direct interaction nor communication between users to predict social links.

Another work by Crandall et al. [CBC+10] shows how temporal and spatial co-occurrences between people help to infer social ties among them; while their main goal is to put forward a generative model that explains empirical data, our study has a different aim, that is, designing a link prediction system to be used on real-world location-based services. Furthermore, our work deals with a different type of data: since we exploit check-ins at well-defined venues, we can infer that two individuals visited exactly the same place without dealing with generic geographic coordinates. As a consequence, our prediction system achieves higher precision while being more practical for a real-world deployment.

5.7 Summary

In Section 2.3 we discussed how the availability of online location-sharing services provides a window on the spatial properties of social behaviour; in addition, and maybe with more important consequences, such services also highlight the rising importance that physical places have for the Web. As users can seamlessly generate and consume information related to the venues they visit, they leave behind them a trail of digital traces that offers unprecedented opportunities to understand and model their behaviour and to build and design related systems.

Then, in Chapter 4 we found that geographic proximity appears to be a driving factor when users establish new social connections; in particular, when users do not belong to the same social communities, spatial proximity can bring together otherwise disconnected individuals. The challenge appeared to be how to accurately build upon spatial proximity to offer precise predictions about potential future social ties. In this chapter we have shown how the properties of the places that people visit can solve this problem, accurately predicting when two users that visit the same places will become connected.

Specifically, we have focussed on one important application that largely benefits from additional information about where users go: the prediction of new social ties. We have described and evaluated a link prediction model based on properties of the places visited by users of a location-based social network. By focussing only on friends-of-friends and place-friends, and by adopting prediction features based both on social properties and on the features of the places visited by users, link prediction systems can achieve high precision in a smaller prediction space than with exhaustive approaches.

In this chapter we have seen one practical application scenario where the spatial properties of online social services can be successfully exploited. In the next chapter we will present another practical case where such spatial characteristics offer tangible benefit: understanding where requests for online content arise on a planetary scale, to optimise the delivery of such content items.

Here we are now, entertain us!

Kurt Cobain

6

Improving content delivery networks using geosocial measures

The amount of Internet traffic generated every day by online multimedia streaming providers has reached unprecedented levels. For instance, there are more than 4 billion videos viewed everyday on YouTube, which has more than 70% of its traffic coming from outside the USA[1]. These providers often rely on Content Delivery Networks (CDNs) to distribute their content from storage servers to multiple locations over the planet. CDN servers exchange content in a cooperative way to maximise the overall efficiency.

Nowadays content diffusion is fostered by weblinks shared on online social networks, which may often generate floods of requests to the provider through cascading across a user's social links. This type of "word-of-mouth spreading" occurring in these services is already driving many of the daily requests to content providers. In fact, the proportion of traffic generated by social spreading is high and increasing. Broxton et al. [BIVW10] discussed how about 25% of YouTube views are generated via person-to-person sharing, with a much higher fraction during the first days after a video is uploaded. A more recent study by Brodersen et al. [BSW12] presents data about individual videos having, on average, about 37% of views socially generated. Thus, social sharing represents a crucial source of traffic for content providers. Given the increasing size of online social platforms, with hundreds of millions of users, they

[1]YouTube official statistics are available at `http://www.youtube.com/t/press_statistics`

generate millions of accesses to YouTube, accounting for a consistent fraction of the total number of daily requests.

In this chapter we show that geographic information extracted from social cascades taking place over online social platforms can be exploited to improve the design of large-scale systems, such as CDNs. We rely on this novel finding: *social cascades are likely to spread over geographically local distances.* Users tend to share content over short-distance social connections, despite the presence of several long-range links; although many users have these long-distance connections, we have found that about 40% of steps in social cascades involve users that are, on average, less than 1,000 km away from each other.

Our key idea is that *content should be kept on servers that are close to interested users, minimising the impact on network traffic.* In other words, since content servers act as caches of items, we aim to exploit the social and spatial characteristics of the users that are sharing the content in the design of large-scale systems such as CDNs, so that we can serve more requests immediately from the closest server rather than waiting for the content to be transferred to the server from somewhere else.

In order to validate our approach, we analyse a dataset from Twitter containing geographic location, follower lists and tweets. We have tracked the spreading of about one million YouTube videos over this social network, analysing a corpus with more than 334 million messages and extracting about 3 million single messages with a video link. Finally, we have designed a proof-of-concept model of a planetary content delivery network using the geographic properties of the commercial platform once used by YouTube. We show that new cache replacement policies, driven by one of the geosocial measures presented in Section 3.3, improve the overall system performance.

Chapter outline In Section 6.1 we describe content delivery networks and their current problems, discussing the issues that motivate our study. Section 6.2 presents an analysis of social cascades of YouTube links over Twitter; by taking into account user geographic information we are able to investigate the extent of these social cascades and to characterise them over space and time.

In Section 6.3 we describe a model of a content delivery network that exploits the spatial properties of social cascades to characterise individual spreading items, prioritising their presence across different distributed caches, while Section 6.4 reports the results of our evaluation, driven by our cascade dataset. Section 6.5 discusses the implications of our results and Section 6.6 offers a review of related work. Section 6.7 closes the chapter.

6.1 Content delivery networks

A Content Delivery Network (CDN) is a system of networked servers holding copies of data items, placed at different geographic locations. The aim of a CDN is to deliver content efficiently to clients; each request is served by a geographically close server, while content is moved between servers to optimise the quality of service perceived by users. Modern commercial CDNs deploy numerous servers all over the world, often over multiple backbones and ISPs, and offer their services to other companies that want to deliver content to users on a planetary scale, such as dynamic Web pages, software updates, multimedia content, live streams and so on.

6.1.1 Factors impacting performance

CDNs have become progressively more important; the number of users with broadband Internet connections is constantly increasing and along with faster connectivity come greater expectations for better content delivery. Even exploiting additional resources provided by CDNs, this demand puts considerable pressure on the entire Internet. This issue becomes even more important if we consider future trends: as the size of distributed content keeps growing, the distance between server and client becomes more critical to the overall performance, since longer distances increase the likelihood of network congestion and packet loss, which result in longer transfer times [Lei09].

In addition, the geography of the requests influences the performance of CDNs; it would be extremely useful to understand whether an item becomes popular on a planetary scale or just in a particular geographic area. Recent research on YouTube videos confirms that the majority of individual videos tend to exhibit highly localised geographic patterns of popularity, with views mainly arising from a small set of regional areas [BSW12]. This has crucial consequences for CDN performance. A globally popular content item should be replicated at every location, since it experiences many requests from all around the world. On the other hand, when content is only locally popular, it should be cached only in the locations that will experience most requests. The key to such a strategy is being able to predict quickly whether a piece of content will become locally popular, in order to optimise its placement over the CDN before it undergoes the popularity surge.

6.1.2 Improving performance through social cascades

The popularity of content over the Web can be driven by public media coverage or through word-of-mouth spreading [CMG09]. The former takes place when content

is advertised by large information sources, such as search engines, news, and social aggregator websites (e.g., SlashDot, Reddit, etc.). This type of phenomenon often results in globally popular items, which should be widely replicated throughout a CDN, since they are likely to experience requests from all over the world. On the other hand, content may become popular because people share it and talk about it, leading to some sort of viral spreading along social connections. These connections may be real-life contacts or interactions on online social networks, with the latter becoming increasingly common.

As a result, content may easily spread from a small set of users to a vast audience through *social cascades*. The number of content requests generated by these social cascades is hard to estimate. However, recent findings confirm that about 30% of Twitter messages contain URL links, and that YouTube is one of the most popular services present in those URLs [RBC+11]. This suggests that a potentially large number of content requests might be generated by a cascade. The combined effect of the popularity of several online services may cause millions of such requests.

Social cascades can be tracked and analysed; online social services can provide all the information, including user location, to track items shared by users and to understand the properties of a social cascade while it evolves over time. To exploit these aspects to improve CDN performance, we need to understand the key geographic properties of social networks. For instance, we need to study and characterise how social cascades are unrolling over space and analyse whether geography affects the spreading process. In particular, *is it possible to estimate whether cascades will spread globally or locally just from the geosocial properties of the users participating in the cascade?*

In the rest of this chapter we will answer this question, characterising how social cascades evolve over space, and then we will exploit our findings to improve CDN performance.

6.2 Geographic online social cascades

In this section our aim is to extract and study social cascades over a geographic social network. Since online social services represent a popular way of sharing information, a piece of information can quickly spread from one user to another as a virus in an epidemic: somebody shares some new content with their friends, who might share it again, and so on. Here we define two measures that quantify the spatial spreading of a social cascade and the extent of its propagation; we then use them to analyse information spreading over Twitter, focussing on traceable pieces of information: Web links to YouTube videos contained in tweets.

6.2.1 Extracting geographic social cascades

A cascade over a social network begins when the first user shares some content and becomes the initiator of the cascade. After this event, some of his contacts will share the same content again, with the result that the cascade will recursively spread over the social links.

In order to estimate the influence of the social network on the information dissemination process, we combine the information about social connections with the temporal information about the posts of each user.

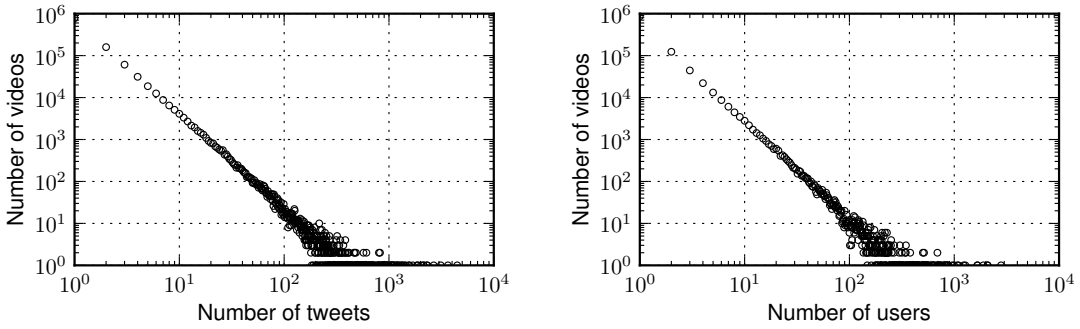
More formally, we say that user B was reached by a social cascade about content c if and only if:

- there is another user A who posted content c *and*
- user A posted content c before B posted it *and*
- there was a social connection from user A to user B when A posted c.

While this does not guarantee the dependency of the posts, in most cases we conjecture that there is a correlation between the two events. If more than one user among the social connections of B posted c, we say that B was reached by the cascade only through the user who posted it last. Therefore, we always have only one previous user in the cascade process. This arbitrary choice will only affect the shape of the cascade, not its size or its overall geographic properties.

In order to describe these social cascades we exploit the spatial social network model defined in Section 3.2. This cascade can be represented as a tree over the spatial network, with the initiator node as the root of the tree. For the same item there might be more than a single cascade. Moreover, the same user may publish the same content at different times. In order to take into account these details, we need to annotate the cascade links with temporal information. Each social cascade is represented as a tree, where a link from user A to user B indicates that user B has received some content as a result of a social cascade from user A. A link between A and B is annotated with the time instants t_1 and t_2: t_2 is the time instant when B posted content c for the first time and t_1 is the time instant of the last time user A posted content c before B did, so that $t_2 \geq t_1$. We define such a cascade step by using a time threshold: consecutive steps in a cascade must be within 48 hours of each other.

In order to investigate the geographic properties of a social cascade we define two measures:



(a) Tweets (b) Users

Figure 6.1: Number of tweets containing a given video link (a) and number of users tweeting a given video link (b).

- the **geodiversity** of a cascade is the geometric mean of the geographic distances between all the pairs of users in the cascade tree;
- the **georange** of a cascade is the geometric mean of the geographic distances between the users and the root user of the cascade tree.

We adopt the geometric mean since geographic distances span several orders of magnitude in our dataset. For a given social cascade these two quantities are correlated, however they can be used to emphasise different properties of the cascade. The geodiversity is computed between all the pairs of users in a cascade, regardless of whether they are connected or not, while the georange is only related to the cascade initiator. On the other hand, the georange allows us to understand how close the initiator of a cascade is to the other people involved in it.

6.2.2 Cascades of YouTube links on Twitter

In this chapter we use the social network extracted from the Twitter dataset already described in Section 3.3. We extract a directed graph from our traces where each node represents a user with a geographic location and a link from user A to user B means that user A follows user B. We recall that this graph has $N = 409,093$ nodes and $K = 182,986,353$ directed links.

Through the Twitter API used to collect this dataset we had also access to the 3,200 most recent tweets for each user in our geographic Twitter graph. We downloaded these tweets for all the users in the graph, obtaining 334,407,185 tweets. The duration of the data crawling was 12 days, from February 1 to February 11, 2010. For each tweet we have crawled the author, the time when it was sent and the actual content of the message. From these tweets we have isolated 570,617 messages containing a direct link to a YouTube video. We have extracted all the messages

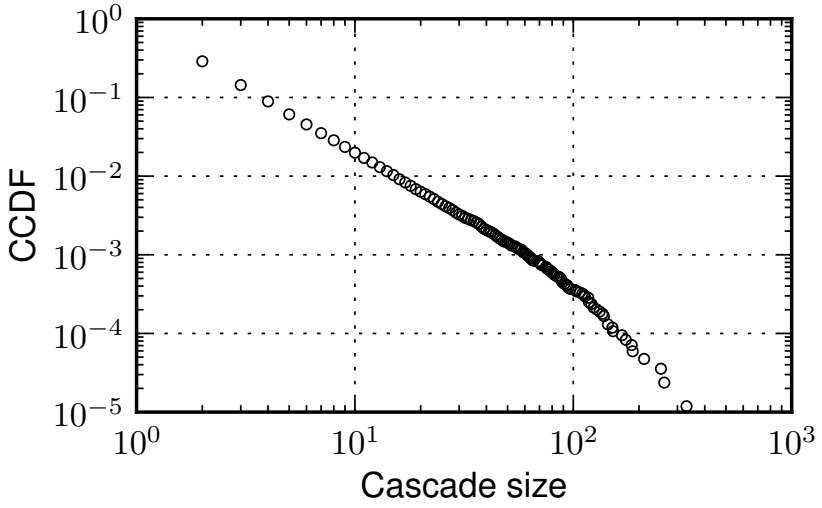


Figure 6.2: Complementary Cumulative Distribution Function (CCDF) of the size of social cascades.

containing a URL shortened with URL shortener services, obtaining an additional 2,332,390 messages with a YouTube link.

Thus, after removing invalid YouTube links, we extract a total of 2,903,007 tweets containing a valid direct link to a YouTube video. These links point to 1,111,586 different YouTube videos, hence some videos are contained in more than a single tweet. The average number of tweets per YouTube video is 2.61. In Figure 6.1 we present the popularity distribution for the video links we have extracted: both the distribution of the number of tweets containing a given video link and the number of different users tweeting a given video link have heavy tails. Thus, there is a very small amount of videos that are tweeted more than 4,000 times or by thousands of different users, while the majority are tweeted only 1 or 2 times by a few users. Such a popularity distribution can greatly affect content delivery, since popular items can easily dominate in terms of number of requests. Furthermore, every tweet is potentially spawning many more actual video requests, since all followers of the author can view the link and follow it. While difficult to estimate, this portion of additional traffic might constitute a large fraction of social-driven web traffic.

Then, we use the cascade definition presented earlier to analyse the tweets and extract 84,337 social cascades for 63,798 different videos. Each cascade involves the initiator and at least one another user. Unfortunately, we have no information about when a user started to follow another one, so we assume that all the social relationships that we have in our social graph were in place when the tweet was sent.

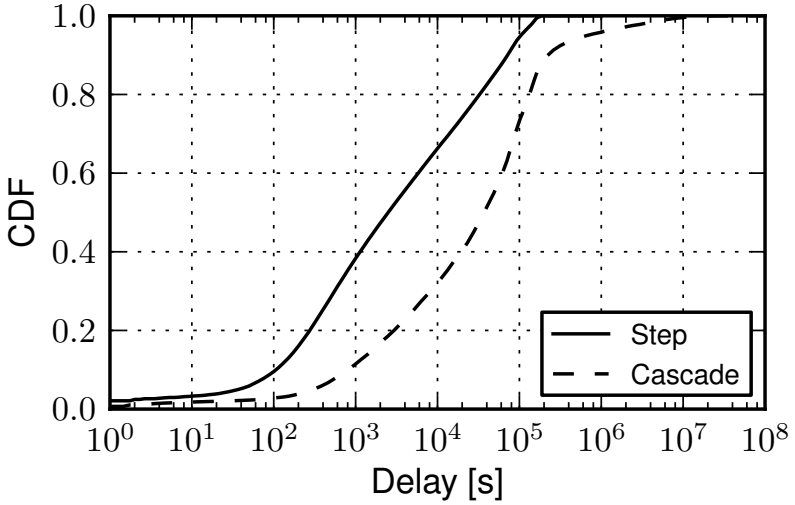


Figure 6.3: Cumulative Distribution Function (CDF) of time delay between two consecutive tweets and total cascade duration (from the first to the last tweet). Cascade duration is shown only for cascades with at least two users in addition to the initiator.

6.2.3 Analysis of social cascades

We define the *size* of the cascade as the number of users involved in it, including the initiator. In Figure 6.2 we report the distribution of the cascade size: we notice again a long tail, with more than 60,000 cascades involving only two nodes and a few cascades reaching up to hundreds of users. This measure of popularity demonstrates that it is rare to have large cascades, but when they do take place they can become extremely large. Again, it is worth noting that social cascades include only users who have tweeted a certain video link; however, each tweet can be viewed by all the followers of the author, thus the potential audience that a YouTube video may reach by means of a social cascade is much larger, even if only a few users are involved.

In Figure 6.3 we illustrate the distribution of the *time delay* between two consecutive tweets in a cascade. About 40% of the tweets in cascades have a delay of about 15 minutes from the previous message, with around 10% having a delay of around 2 minutes. This result shows that YouTube links can spread on Twitter on a time scale of some minutes, even though further spreading does happen even after some hours. This indicates that links to videos can quickly spread over the social network, potentially leading to many views in a short period of time. In Figure 6.3 we show the distribution of cascade duration from the first tweet to the last tweet for each cascade with at least 2 users in addition to the initiator: about 80% of the cascades end within 24 hours, with 40% ending in under 3 hours.

In Figure 6.4 we show the distribution of the *geographic distance* between authors of two consecutive tweets in a social cascade. Around 10% of cascade steps are less

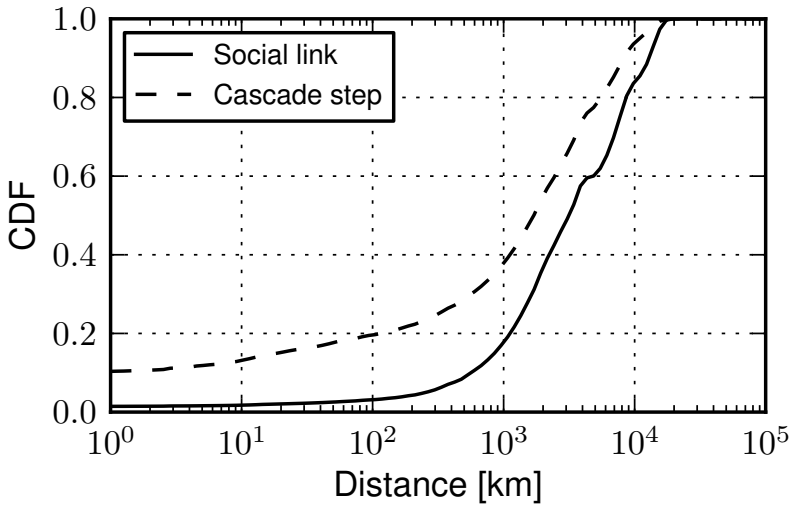


Figure 6.4: Cumulative Distribution Function (CDF) of cascade step distance and of social connection distance: social cascades take place over short-range social connections. Logarithmic binning has been adopted to estimate the number of samples in each range of values.

than 1 km, with 20% of them shorter than 100 km and more than 30% shorter than 1,000 km. This result is in slight contrast with the distribution of link lengths of the Twitter network, already presented in Figure 3.9 and again shown in Figure 6.4 to aid comparison: even if fewer than 5% of the social connections are shorter than 100 km, within cascade steps this fraction increases up to 20%. Content spreading through social cascades is more likely than expected to travel over geographically short-range social connections rather than over the more numerous long-distance links.

In Figure 6.5 we show the distribution of *geodiversity* and *georange* for all the social cascades that involve at least two users in addition to the initiator. About 40% of these cascades have geodiversity lower than 1,000 km, with around 20% of geodiversity values lower than 300 km. Thus, even though many cascades reach a broad audience, some of them remain geographically limited. On the other hand, about 90% of georange values are smaller than 1,000 km, with about 30% of values smaller than 100 km. This is an indication that a cascade may take place in a broad region but with each user still close to the initiator.

6.2.4 Geosocial measures and social cascades

Finally, we are interested in properties of a social cascade that may help us predict its geographic spreading from the very first messages that are tracked. Towards this aim, we use one of the two geosocial measures introduced in Section 3.3: node locality. Recall that this measure indicates whether a user has social connections

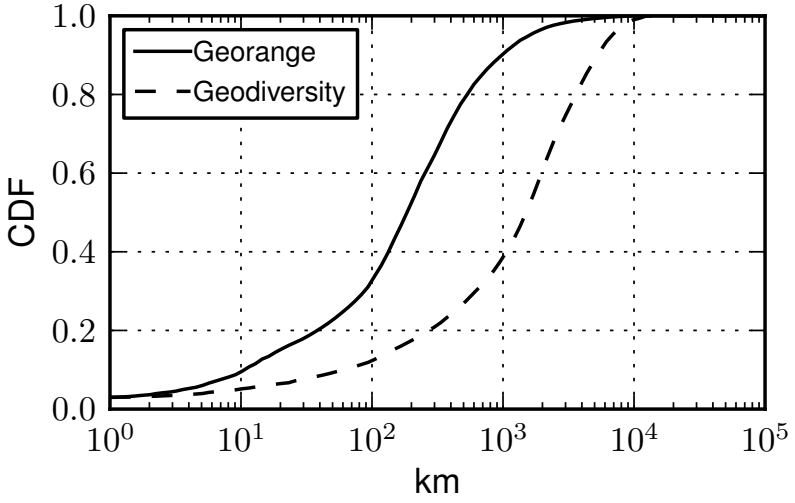
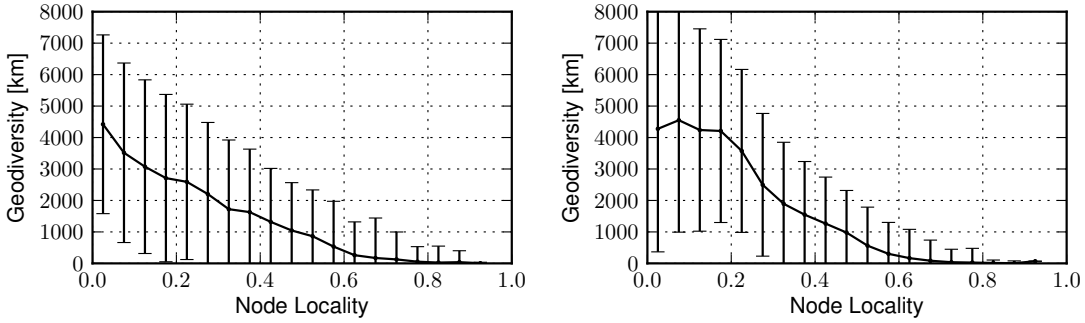


Figure 6.5: Cumulative Distribution Function (CDF) of geodiversity and georange for social cascades with at least 2 users after the initiator. Logarithmic binning has been adopted to estimate the number of samples in each range of values.

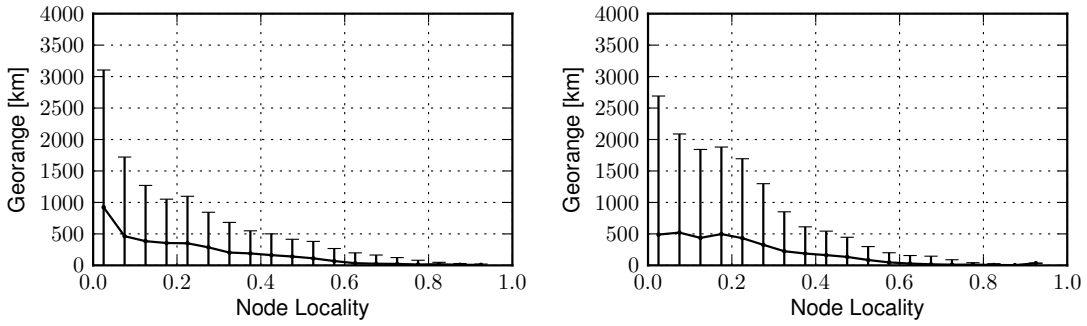


(a) (b)

Figure 6.6: Average geodiversity of a social cascade as a function of the average locality of the first nodes in the cascade: locality of the first node (a) and of the first two nodes (b). Error bars show standard deviation around the average.

mainly over short-range distances, with a node locality close to 1, or over longer distances, with a value closer to 0. The node locality of a user offers an indication about the potential geographic spread of the information passing through the user.

We investigate whether the node locality of the first users who participate in a social cascade is related to the final geodiversity and georange values. We report in Figure 6.6 the average cascade geodiversity as a function of the average locality of the first users involved in the cascade. We observe that even the initial locality of the first user is correlated with the geographic spreading of the cascade and with a reduction in the variance of this spreading. Moreover, by including the locality of the second user we get a stronger relationship. A similar result can be seen in Figure 6.7 for the georange: in this case the correlation is clearer, with less variance



(a) (b)

Figure 6.7: Average georange of a social cascade as a function of the average locality of the first nodes in the cascade: locality of the first node (a) and of the first two nodes (b). Error bars show standard deviation around the average.

especially for high locality values. It is important to consider both the correlation with the average value and with the reduction of the variance, denoting a more indicative estimation for higher values of node locality.

Thus, the final properties of a cascade can be estimated even from the users involved in the initial stages. Also, even the geographic and social properties of the initiator are sufficient to understand whether a cascade will spread locally or globally, and by taking into account a few more steps we are able to give a more accurate estimate of the final outcome.

Given the importance of social cascades and their geographic properties, being able to correlate their properties with the geographic range they will reach makes it possible to exploit these findings to improve the design of cache replacement strategies for CDNs.

6.3 Distribution of content using geosocial measures

We have described the geographic properties of social cascades. In this section we exploit these findings in the design of a proof-of-concept CDN that adopts geosocial measures to improve caching performance.

6.3.1 Assumptions and model

We envisage a single entity able to access information about content shared by users on social networks and control the CDN which delivers the content that users are sharing. This can be mapped to reality in various ways: i) assuming that CDNs will

have access to information from online social services about the cascade dynamics, which is reasonable as they are providing the content sharing service or ii) assuming that, plausibly, in future online social networks and content providers might merge into single entities or cooperate (e.g., companies like Facebook and Google already offer social networking features and serve online content).

We model our system as a collection of server clusters placed around the planet. Each cluster contains a certain number of servers: we assume that all servers within the CDN have identical properties. The only difference between clusters is the number of deployed servers. We assume that there is a central catalogue of content items: clients from all over the world request content items from the CDN and they are redirected to the geographically closest server. If the server already contains the requested item, it is immediately served. Otherwise, the item is retrieved from another portion of the CDN and served.

We assume that, as observed in real systems [Lei09], different clusters are interconnected by a dedicated network. Then, we assume that it is faster to move content between servers to bring an item as close as possible to the client, than to redirect the request to another server further away that already holds a copy. This seems plausible, even if geographic distance may not always be the only factor influencing performance.

Server clusters act as caches: they keep copies of already requested items for future requests, but they have finite storage. A cache replacement strategy is used to remove an item from the cache when it is full. We also assume that the servers within a cluster coordinate to act as a single large cache. Therefore, every server can host up to k items and if there are N servers in a cluster, that cluster is equivalent to a single cache able to host kN different items. This simplifies the definition of the model but still captures the heterogeneity of cluster sizes around the planet. We do not model file size: we assume that the size of a file does not vary much across the items, as we have observed in our specific dataset of YouTube videos.

6.3.2 Model parameters

In order to ground our model in reality we have parametrised it with the real properties of Limelight, the commercial CDN once used by YouTube to deliver content to users, as measured by Huang et al. [HWLR08][2]. Limelight has clusters of servers deployed at 19 different locations around the world and each cluster has a different number of servers. In Table 6.1 we report the details of each server cluster: Limelight deploys 2,830 out of its 4,147 servers in the United States, where there are 10

[2]This paper has been withdrawn by Microsoft due to some criticisms about the system performance results presented. However, we only use information about server locations from this work.

Location	Country	Servers	Location	Country	Servers
Washington	USA	552	Frankfurt	Germany	314
Los Angeles	USA	523	London	UK	300
New York	USA	438	Amsterdam	Netherlands	199
Chicago	USA	374	Tokyo	Japan	126
San Jose	USA	372	Toronto	Canada	121
Dallas	USA	195	Paris	France	120
Seattle	USA	151	Hong Kong	Hong Kong	83
Atlanta	USA	111	Changi	Singapore	53
Miami	USA	111	Sydney	Australia	1
Phoenix	USA	3			

Table 6.1: Geographic distribution of the server clusters in the Limelight network.

clusters. Europe and Asia are served only by seven clusters in total and Australia only by one, while the rest of the world does not contain any cluster.

In our model, cache size should be interpreted with respect to the total number of items present in the system and not as an absolute number, since we do not have access to the whole YouTube item catalogue. Hence, we will also express cache size as a percentage of the total data catalogue. As an example, since we have about 1 million videos in our dataset, a cache size of 100 items is comparable to a cache that can host about 0.01% of a real catalogue; in the case of YouTube, with hundreds of thousands of videos added every day, there are more than 100 million videos, hence this would represent a cache size with more than 10,000 different videos.

6.3.3 Content caching policies

We now define the caching policies adopted by our model to store and replace content within the servers. A server cluster adopts a *cache replacement strategy* to remove an item when the cache is full and a new request arrives. Each strategy assigns priorities to the items in memory and, when a deletion is needed because the cache is full, the item with the lowest priority is removed. The priority of an item might be updated whenever a request for that item is issued.

Our approach is to use standard caching policies and then augment them with geosocial information. Each policy assigns a priority $P(v)$ to a video v and, when a video has to be removed, that with the lowest priority is chosen for deletion. A random choice is made when more than one video has the lowest priority. We adopt three different caching policies: *Least-Recently-Used (LRU)*, *Least-Frequently-Used (LFU)* and *Mixed*.

In LRU the priority of a video v is given by $P(v) = clock$, where *clock* is an internal counter incremented by one whenever a new item is requested. This policy provides a simple aging effect: when an item is not requested for a long time, it is eventually removed. However, it does not take into account item popularity. In LFU the priority of a video v is given by $P(v) = Freq(v)$, where $Freq(v)$ is the number of times video v has been requested since it was stored in the cache for the last time. LFU favours popular content: if an item receives a large number of requests it will stay in the cache for a long time. However, LFU is less flexible: an item that was popular in the past tends to stay in the cache even if it is not requested anymore. The Mixed policy combines both LRU and LFU features and the priority of video v is given by $P(v) = clock + Freq(v)$, in order to balance both temporal and popularity effects [Che98]. In this case *clock* starts at 0 and it is updated for each replacement with the priority value of the removed file. Thus, a video increases its priority when it is requested many times, but, if there are no more requests, it will eventually be removed from the cache.

Then, we define two *priority weights* for each video v, based on the geosocial characteristics of users participating in the social cascades involving this video, measured using node locality:

- **Geosocial**: the weight of video v is given by the sum of the node locality values of all the users who have posted a message about it, even if they are not involved in a social cascade;

- **Geocascade**: the weight of video v is given by the sum of the node locality values of all the users participating in the item's social cascade (or cascades, if an item happens to be posted on more than one cascade).

These weights are used to capture the idea that if a video is tweeted many times by users with high node locality values, then it is likely that it is spreading in a local region, thus future requests will hit the same content server. While the first weight takes into account all the messages regarding a particular content item, the second one only uses the messages caused by a social cascade. By using two different weights based on geosocial information, we want to investigate the contribution of social cascades with respect to using only geographic information of social ties.

The weight of every tweet with a link to a video is updated according to whether that tweet is or is not in a cascade. For every request, content servers get also the video weight and multiply it with the priority of the underlying cache replacement policy. Hence, we have three versions of every cache replacement policy: with no weight, with a Geosocial weight and with a Geocascade weight.

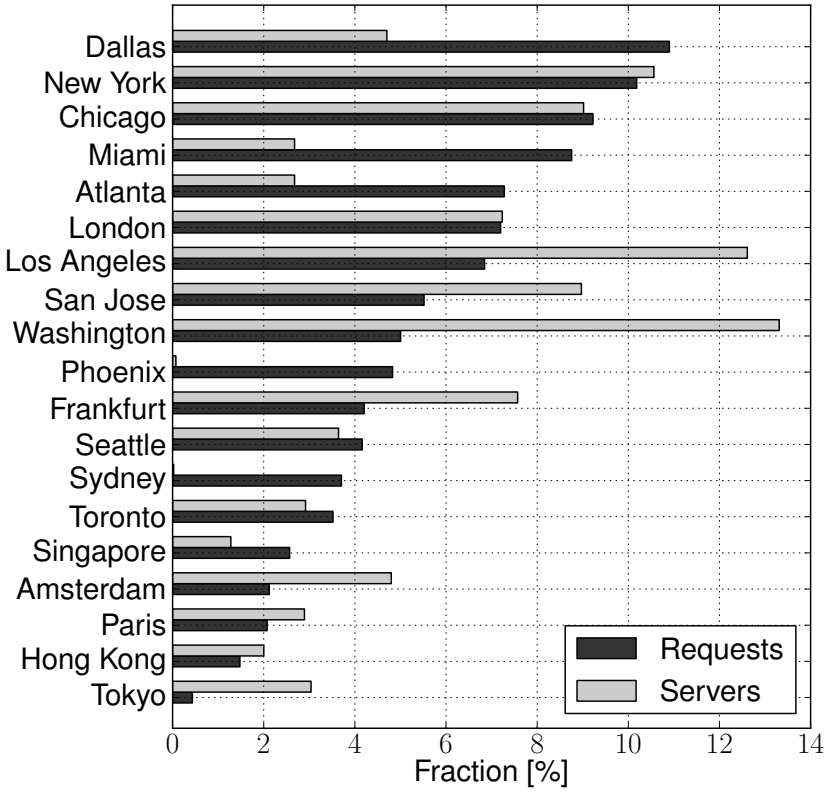


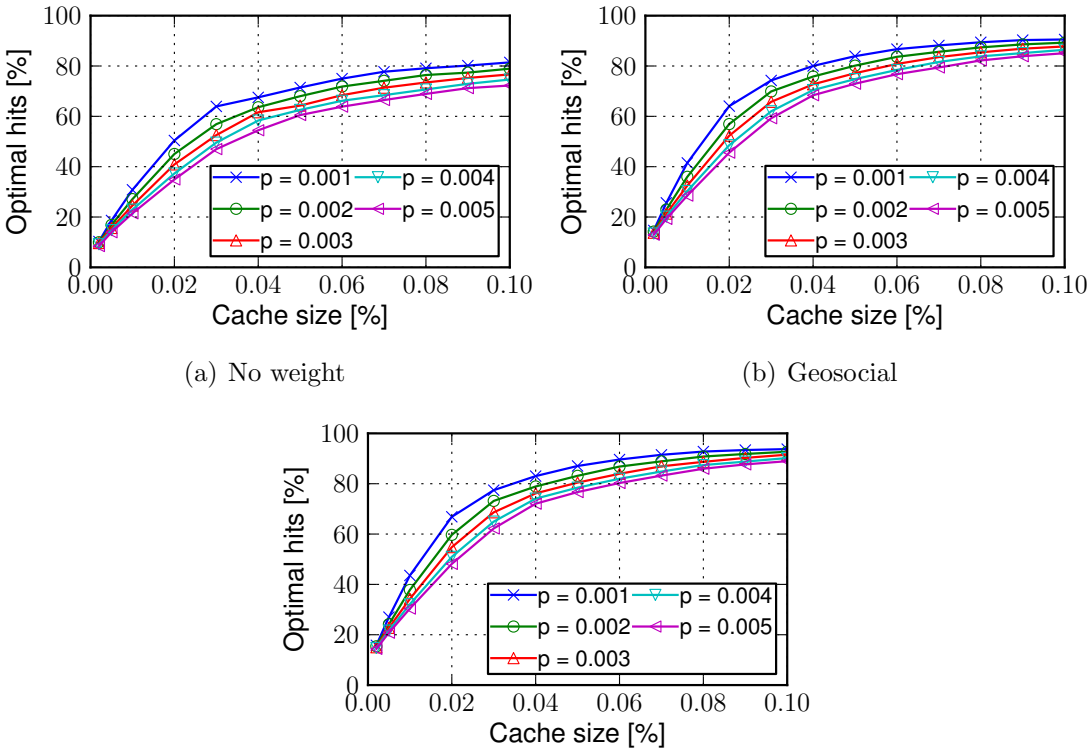
Figure 6.8: Fraction of video requests handled by each cluster and fraction of servers contained in each cluster. Different workloads do not significantly change the distributions of requests.

6.4 Evaluation

In this section we test our idea that information extracted from geographic social cascades can effectively be exploited to improve the performance of CDNs. We have investigated through simulation how different cache replacement policies impact the performance of the system. Our results show that global system performance can be improved with respect to standard policies, which means potentially avoiding millions of video file transfers per day.

6.4.1 Simulation strategy

In our simulation we create a sequence of content requests to the CDN directly from the Twitter messages within our dataset. We assume that every video contained in a Twitter message is requested by each follower of the author with a certain probability p and with a random temporal delay modelled with the same distribution of delay between cascade steps. This assumption is simple and can be far from reality, as the real load is likely to be a function of the particular user and the particular content



(a) No weight

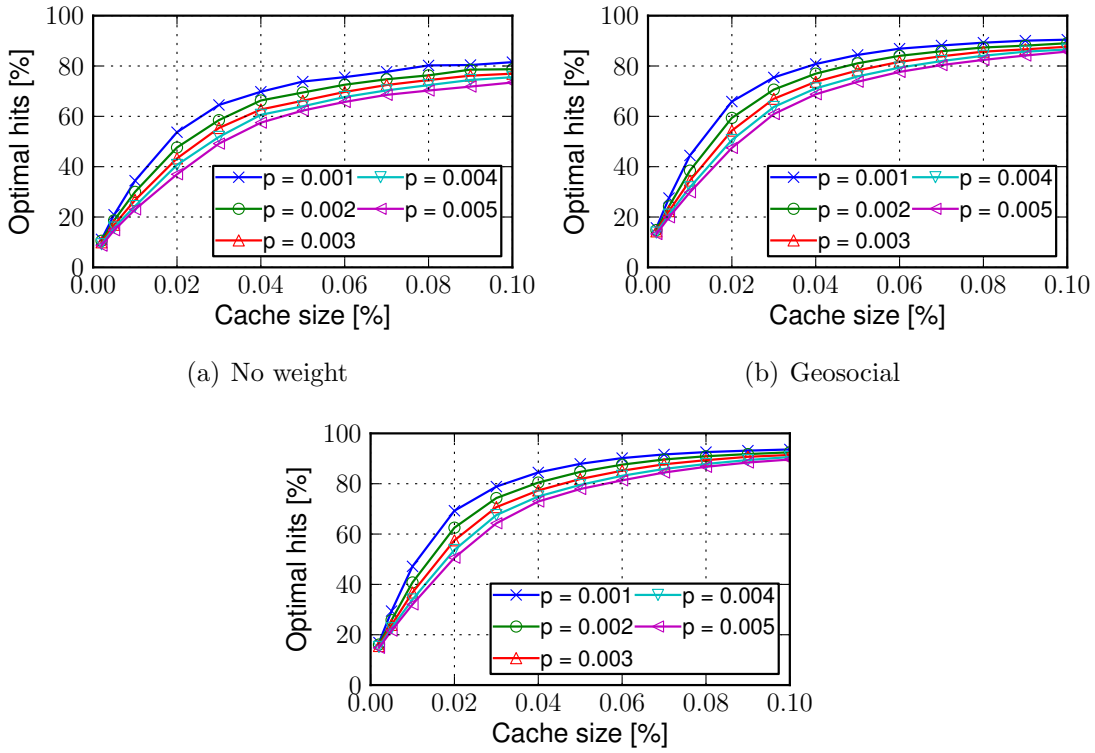
(b) Geosocial

(c) Geoscade

Figure 6.9: Percentage of total hits with respect to the infinite cache case as a function of cache size for the **LRU** cache polity and different weights: no weight (a), Geosocial weight (b) and Geocascade weight (c). Cache size is expressed as a fraction of the entire data catalogue. Every simulation is run 20 times with randomly generated workloads and the average is presented (standard deviation is negligible and not shown).

item: nonetheless, we do not have precise information about real traffic requests spawned by Twitter messages. However, our simulation results show performance improvements for any value of p we adopted. We generate 5 different workloads, corresponding to the values of $p = 0.001, 0.002, \ldots, 0.005$, and we run every workload 20 times, averaging the results.

We always route a request to the server cluster closest to the user. However, the geographic distribution of the requests does not change for p, since it is only influenced by the geographic distribution of Twitter users, which does not change for different workloads. As shown in Fig. 6.8, some servers receive much more traffic than others: as an example, the cluster in Dallas accounts for more than 11% of global requests. Additionally, some locations receive a large fraction of traffic even though they contain only a small number of servers. These properties may impact the performance of the cache replacement strategies for different locations.



(a) No weight

(b) Geosocial

(c) Geoscade

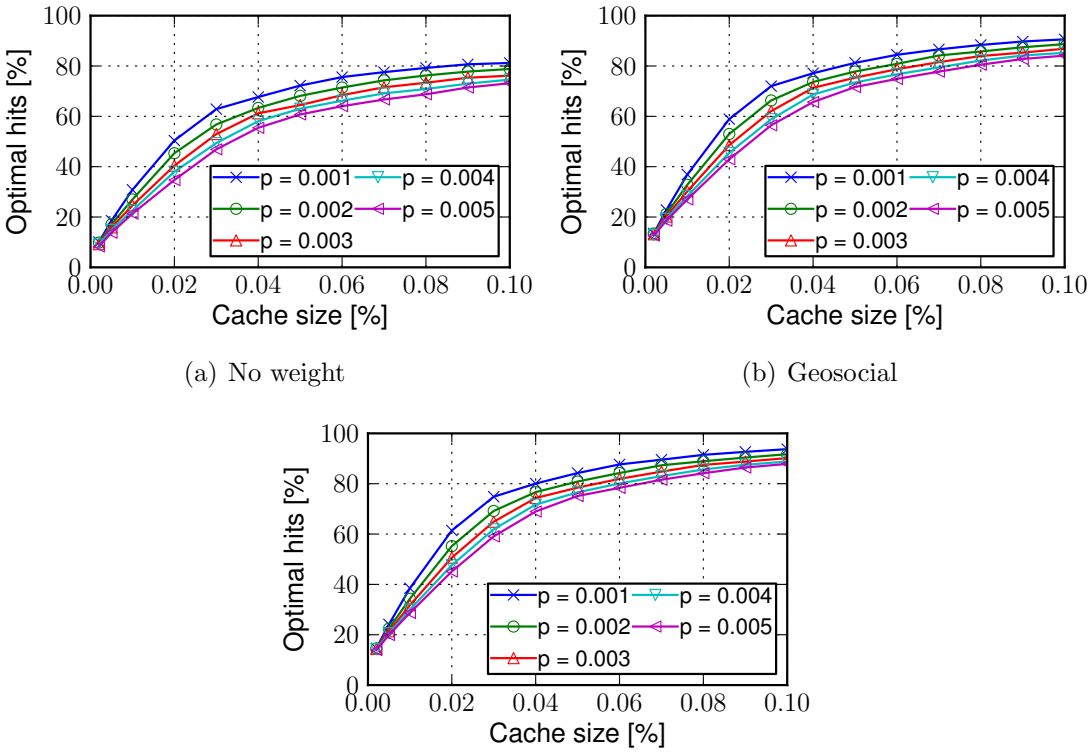
Figure 6.10: Percentage of total hits with respect to the infinite cache case as a function of cache size for the **LFU** cache polity and different weights: no weight (a), Geosocial weight (b) and Geocascade weight (c). Cache size is expressed as a fraction of the entire data catalogue. Every simulation is run 20 times with randomly generated workloads and the average is presented (standard deviation is negligible and not shown).

6.4.2 Global performance

First we investigate the performance of different policies with respect to the case of infinite cache size, i.e., in conditions where no item is ever removed from the cache. The number of hits in this case is the maximum achievable, both on each cluster and globally.

As a global performance metric for our system we consider all the hits on all the clusters; every request is directed to the closest server and there it may result in a hit or a miss. For each cache replacement strategy and for each different workload, we compute the total number of hits obtained and we take the ratio between this value and the performance with infinite cache. This metric shows how different policies react when some parameters of the system are changed, but it does not emphasise differences in their performance.

In Figures 6.9-6.11 we show the change in system performance as a function of cache



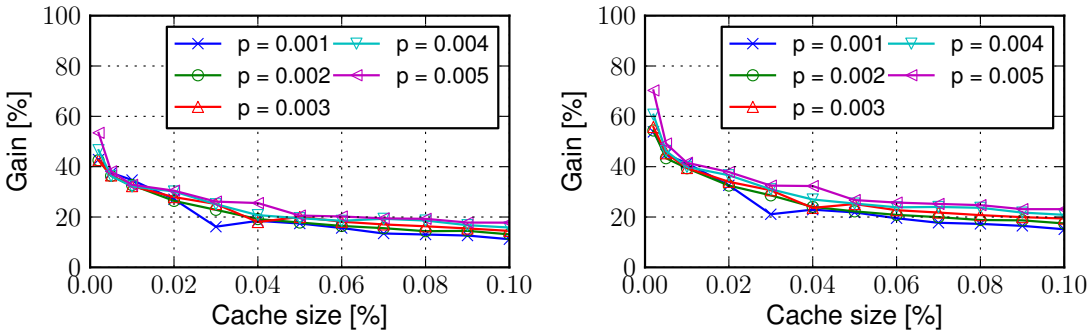
(a) No weight

(b) Geosocial

(c) Geoscade

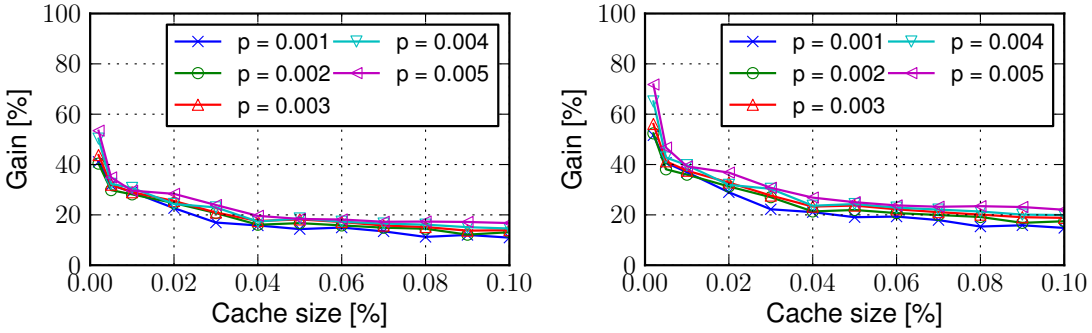
Figure 6.11: Percentage of total hits with respect to the infinite cache case as a function of cache size for the **Mixed** cache polity and different weights: no weight (a), Geosocial weight (b) and Geocascade weight (c). Cache size is expressed as a fraction of the entire data catalogue. Every simulation is run 20 times with randomly generated workloads and the average is presented (standard deviation is negligible and not shown).

size and for different workloads; when the size increases, every policy steadily improves its performance. Larger workloads have worse performance, but differences between them disappear at larger cache sizes. Moreover, as the cache size grows larger, all workloads reach a plateau, since increasing the cache size beyond a certain limit provides a diminishing performance increment. This is due to the fact that there is a portion of content that is requested only a few times and for which caching policies can hardly offer advantages. In addition, we observe that while using no weights results in the lowest hit ratio, by adopting instead the Geosocial and Geocascade weights we achieve noticeable improvements, because the servers are now able to identify geographically popular items and keep them in memory for future local requests. However, we need a direct comparison to appreciate the difference in performance achieved by using these weights.



(a) Geosocial (b) Geocascade

Figure 6.12: Increment (%) of average performance with respect to the case without weight as a function of the cache size and for different workloads when the **LRU** strategy is used with Geosocial weight (a) and with Geocascade weight (b).



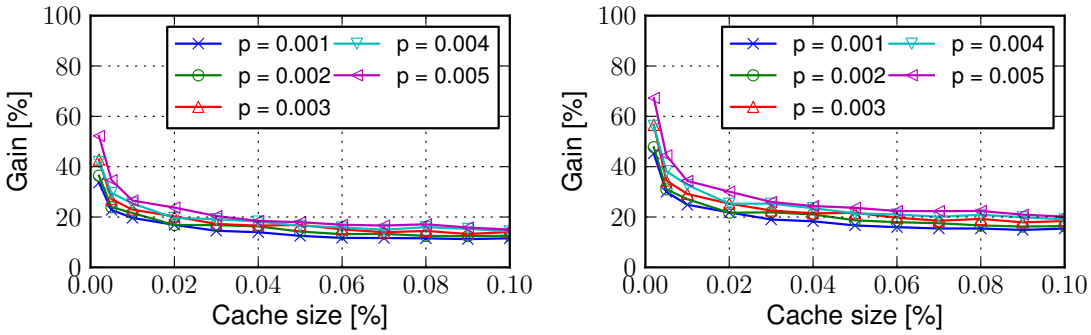
(a) Geosocial (b) Geocascade

Figure 6.13: Increment (%) of average performance with respect to the case without weight as a function of the cache size and for different workloads when the **LFU** strategy is used with Geosocial weight (a) and with Geocascade weight (b).

6.4.3 Policy comparison

In order to understand which policy provides better results, we evaluate the relative performance improvements between the weighted policies and the other strategies.

We illustrate in Fig. 6.12 the performance increment when we augment the LRU strategy with geosocial information. Geosocial-LRU reaches a maximum 55% performance increment, while increasing the cache size results in a smaller increment. Instead, Geocascade-LRU achieves more than a 70% increment over LRU for smaller cache sizes, while the benefit decreases as cache size increases. In Fig. 6.13 we investigate how the use of priority weights improves LFU. Geosocial-LFU achieves a top increment of about 50% against LRU with small cache sizes, with the increment going down as the size increases. However, the improvement is larger in the case of the Geocascade weight, with a maximum increment of 70% and a smaller decrement with cache size. Finally, in Fig. 6.14 we investigate the difference between the Geo-



(a) Geosocial (b) Geocascade

Figure 6.14: Increment (%) of average performance with respect to the case without weight as a function of the cache size and for different workloads when the **Mixed** strategy is used with Geosocial weight (a) and with Geocascade weight (b).

cascade and the Geosocial weight for the Mixed cache policy. Again, the Geosocial weight gives a maximum improvement of 50%, while the Geocascade one improves the baseline performance by up to 65%.

Both weights improve cache performance, since they recognise content that is more likely to become popular only locally and to result in many requests to the same local servers. Indeed, items that are popular on a global scale may be requested from different servers around the planet and may not trigger cache prioritisation in single CDN clusters. Furthermore, including information about the spreading of social cascades appears to be a better predictor of local popularity, since the Geocascade weight gives higher performance than Geosocial. It is also important to note that the performance improvement is smaller when the cache becomes larger. Indeed, with a cache so large that it can host 0.1% of the entire data corpus, it becomes easier to accommodate more items and performance easily reaches a saturation point, as seen in Figures 6.9-6.11. For a given cache size larger workloads have a larger relative improvement, since their absolute performance is lower.

6.5 Discussion and implications

The main result presented in this chapter is that locality information from social cascades can be extracted and used to improve large-scale system design. We see a great potential in exploiting geographic properties of online user communication. Geographic locality of online interactions can be exploited to do pre-fetching of Web content, caching of normal HTTP traffic, datacentre design and placement and even to devise security mechanisms [WPD+10, BSM10, THT+12].

In addition, our approach can be generalised to be used on a number of different social platforms. The information needed can be efficiently exposed by an anonymised

API, which could provide only the aggregated geosocial measures corresponding to a given cascade of a certain shared item. Moreover, information coming through public Twitter feeds, private Facebook posts and emails can be anonymised and exposed in order to classify items according to their geographic popularity and feed this information into CDNs. As with any other anonymisation procedure, this approach would anyway present associated risks.

In the specific example we have discussed, improvement largely depends on cache size: when it is possible to cache a considerable portion of the whole item catalogue, cache policies matter less and the improvement obtained by social information is smaller. However, if cache size is not sufficient to store that portion, because it is too small with respect to item size or because the catalogue contains too many items, geosocial properties can make a difference. Moreover, if in the future social cascades can be tracked on a larger scale, the advantage given by geosocial measures may impact not only CDN caching policies but other large-scale systems in general.

As already mentioned, our results are obtained using a sample from a single service. Although it is generally unknown which portion of the traffic directed to CDNs is coming from online social networks, it is not inconceivable that this traffic may become considerable; the fraction of messages containing content in our dataset is already appreciable and the number of users of social services is still increasing.

An improvement in the number of cache hits for requests coming from these services, as observed in our simulation, would mean that millions of video daily requests could be served locally instead of being transferred over the network. In addition, videos are getting larger, with higher quality demanded by users, meaning bulkier files. Caches need to grow larger and larger to cope with this trend or, alternatively, need to cache fewer items. This is impacting (and will impact increasingly more) on the running costs of modern CDNs. For instance, Limelight runs a global private fibre-optic network that avoids sending files over costly public Internet connections. As a result, any reduction in the number of files sent across the network would reduce the investments needed in network infrastructure, which account for a considerable part of the total expenditure of a CDN [QWB+09].

6.6 Related work

Two research areas are related to this discussion: the analysis of online social cascades and the design of large-scale CDNs.

Social Cascades Social cascades have been studied in sociology, economics and marketing for more than 60 years; an eminent example is the threshold model proposed by Granovetter [Gra87]. Recently, thanks to the availability of large datasets,

many other studies have been presented. In [AA05] Adar and Adamic analyse the diffusion of information in blogs by applying epidemic models of information spreading. Similarly, a characterisation of cascades using data from Flickr, a photo-sharing website, is illustrated in [CMG09].

Finding ways of harnessing the potential of information constantly generated by users is a key and promising research area for the networking community and it is still largely unexplored. An initial proposal was presented by Sastry et al. [SYC09]: their suggestion is to place replicas of items already posted by a user closer to the location of friends, anticipating future requests. Our proposal is a different example that uses information extracted from social cascades to effectively improve the performance of large-scale networked systems, and, more specifically, of CDNs. In addition, we present a large-scale study of geographic social cascades that supports our claims.

Content Distribution Networks Given the success and economic importance of CDNs, many solutions to improve the performance of this class of systems have been proposed with respect to the location-aware selection of servers. Key examples of experimental systems in this area are Meridian [WSS05], a node selection mechanism based on network locality, and OASIS [FLM06], an overlay anycast service infrastructure. WhyHigh is a system to redirect queries based on the measurements of the latency of the Google's CDN [KMS+09]. This system is not only based on geographic proximity but also on measurements of client latencies across all CDN nodes, in order to identify the prefixes with inflated latencies.

While these systems have used some knowledge of the geographic properties of traffic load to improve performance, we have also taken advantage of information from online user interaction to enhance the content placement decision process.

6.7 Summary

Taking into account how online social services are affected by spatial distance could improve system design, as we argued in Section 2.4.3. Already in Chapter 5 we have demonstrated that the additional layer of spatial information about user behaviour can greatly benefit applications based on data mining. Furthermore, spatial properties of online platforms become important when services are deployed on distributed architectures that span and serve the entire planet. Since content storage and content delivery must happen on a global scale, because online platforms serve hundreds of millions of users all around the world, spatial constraints affecting user interactions are of vital importance to improve resource usage.

In this chapter we have shown how geosocial properties of users participating in online social cascades can be exploited to improve the efficiency of caching in CDNs.

We have studied cascades on Twitter, finding that users preferentially share content over short-range links, despite the significant presence of long-distance connections. Using one new geosocial measure introduced in Section 3.3, we have taken advantage of these findings to design content caching policies that prioritise content that experiences geographically local popularity, validating our design through model simulation. While our study is limited in scope by the choice of online social network and dataset, our results are more generally applicable and the impact of the approach could be high for large-scale systems whose traffic is driven by online social services.

The feeling is less like an ending than just another starting point.

Chuck Palahniuk

7

Reflections and outlook

The general public as well as academic scholars believed that the Internet, as many other historical breakthroughs in human communication technology, would turn the entire planet into a "Global Village", where space is irrelevant and geographic distances do no longer matter. This thesis is inspired by an important and contrasting idea: individuals are affected by spatial proximity and geographic factors in their online social interactions.

This dissertation has supported this thesis with a body of work largely made possible by two increasingly important trends: the advent of the mobile Web and the popularity of online social networks. As users access online services through location-sensing devices, service providers gather data about where individuals are located and where they go, together with information about their social interactions. This exposes the spatial properties of the social connections arising on the Web, generating possibilities for study and analysis; also, this opens the door for a wide new range of systems and application.

This dissertation has explored both these threads: we first focussed on studying and understanding how spatial and social properties of online social services are related to each other. We then proposed new ways of taking advantage of spatial data to provide, respectively, better link prediction engines and better caching of content in planetary delivery networks.

7.1 Summary of contributions

When considering the effect that geographic constraints could have on online social services, the properties of their social graphs need to be reconsidered taking into account the metric space where individuals are embedded. Thus, in Chapter 3 we adopted a methodology that treats the social graph as a spatial network, associating a geographic distance to every social connection. We observed that users with more connections appear less constrained by geographic distance and, using two different randomised null models, we assessed how the observed properties could not arise from social or spatial factors alone: both dimensions ought to be jointly considered.

These findings were revisited in Chapter 4, where we analysed the temporal growth of an online social network with respect to its spatial properties. We found evidence that the creation of new links can be reproduced by a gravitational attachment model, where new connections are created with nodes that are either already well connected or spatially close. This mechanism combines together a purely social property, the number of connections a user already has, with a purely spatial measure, the geographic distance between users. To our surprise, we found that only social factors seem to drive triadic closure, which appears largely unaffected by distance. We combined these observations to define a new model of network growth that reproduces the spatial and social properties observed in real data.

The study of the spatial properties of online social services has offered new insights about user behaviour, inspiring new systems and applications. Hence, we have covered and discussed two practical cases where spatial properties of online social networks are explicitly used, respectively, to enhance friend suggestion engines and to improve caching policies used in distributed content delivery networks.

In Chapter 5 we described how friend prediction systems can rely on the places that users visit to find suitable candidate for predictions, reducing the overall prediction space and still covering a large fraction of future connections. We have shown that the properties of the places that two users share can be used to build prediction features; we have proposed and evaluated a supervised learning approach which achieves high performance.

We discussed a different application in Chapter 6: adopting caching policies in content delivery networks to serve items over the planet. Our key idea is that content consumption is fostered by users sharing items over online social networks: as their social connections are constrained by space, we can understand which items are popular on a geographically local scale by tracking their spreading by means of social cascades. Again, we have shown how exploiting the spatial properties of online social interactions can effectively improve delivery performance, using a trace-driven simulation of global content requests.

Returning to our initial thesis, we feel that geography and space affect online social interaction in a strong and straightforward way: users will want to connect to other individuals who are "popular", regardless of where those users are located, or to other individuals who are "close", even though they might be relatively unheard of. In other words, the mere fact of being spatially close to someone else greatly increases how interesting, or socially attractive, that individual is. At the same time, other factors such as homophily and transitivity appear less affected by space; instead, they might be influenced by properties such as similarity, user preferences and other measures of like-mindedness. Overall, it seems that being in proximity, either in the geographic sense or by sharing common interests, is what brings new social connections to life.

In summary, we believe that the three main spatial properties that online social services exhibit are the likelihood of friendship connection that decreases as an inverse power of spatial distance, strong correlations between spatial properties and node degree, and the lack of spatial constraints on social triads. These properties may hold for other social systems embedded in space and are likely to strongly influence other characteristics commonly observed in social graphs, such as community structure and the properties of navigable network paths.

7.2 Future directions

From these considerations, geographic proximity arises as yet another factor that brings individuals to connect to each other. This simple observation also suggests many possibilities for future work.

First, a key question is whether the effect of spatial distance can be introduced beyond the gravitational attachment process. Even though our new model reproduces some social and spatial properties observed in the real graphs, it may fail to replicate many other characteristics such as community structure, transitivity and small-world behaviour that are observed in online services. Thus, two different but related threads of work need to be carried out: the analysis of the spatial properties of these phenomena, in order to understand how geographic distance influences different facets of social networks, and the extension of the network growth model to take into account these new findings.

In fact, another property of online social networks likely to be strongly influenced by distance is community structure. The existence of social groups is as important as the effect of spatial distance to fully understand how individuals establish social ties. Since geographic proximity fosters connections, large and dense communities might be more likely to arise between individuals close to each other rather than far apart. Recent results on mobile phone interactions confirm that communities are

constrained by geography and only groups with more than 30 members gradually lose these spatial limitations, spanning wider areas [OAG+11]. In Chapter 4 we jointly considered social and spatial factors to reproduce the properties observed in real online social networks: this study could be further extended to understand whether our model reproduces the communities present in real scenarios and, if it fails, what modifications would be needed to capture how real social communities are created over space.

Then, on a broader scope, the close relationship between how users create new friendship ties and how users move across places has to be explored. As discussed in Chapter 5, the places visited by users can reveal which new social connections are likely to arise in the future. This relationship can be reversed with equally promising results, as users can be influenced by their friends when choosing which new places to visit. By simultaneously considering these two influences, from places to social connections and from social connections to places, there is the potential to expand our understanding of each of these two processes. More importantly, this might allow us to define joint models that describe how mobile users behave from both a social and spatial point of view.

The close inter-dependence between user mobility and social connections becomes even more compelling when considering long-range social connections, spanning large distances even across continents. One could posit that such ties were created when users were in direct spatial proximity at some point in the past; then, one of the two parties migrated somewhere else, effectively creating a long-range social connection. This would imply that a more meaningful explanation of the geographic patterns of social interaction could lie in individual migration patterns, spanning short and long distances and ranging from daily movements to long-term relocation shifts. A recent study by Levy discusses how such migration patterns exhibit statistical regularities that could explain the observed effect of geographic distance on social connections [Lev10]. More data and more studies in this direction could help shed more light on this initial insight.

7.3 Outlook

The importance that spatial proximity holds in connecting people together has been recently exploited by a new generation of mobile applications that use location-sensing technology available in modern devices to continuously acquire data about other users nearby. Mobile applications such as Highlight[1], Banjo[2], Sonar[3] and

[1] `http://highlig.ht`
[2] `http://ban.jo`
[3] `http://www.sonar.me`

Glancee[4] help users discover new friends by matching geographically close users according to their shared interests and personal profiles. These new services seem to pave the way for a broader trend: location data will be increasingly available on online services and ingrained in the features they offer. As powerful mobile devices become mainstream, the potential audience of location-based services could easily grow as online social services did over the last years, changing the Web yet again.

Overall, this dissertation has made a step in addressing how the spatial properties of online social networks can be used effectively to understand and model their structure and to design and deploy related systems. As location-based data will be increasingly more available, our findings and results open the door to a vast range of future possibilities. Our results are relevant both to researchers and to practitioners: the former would benefit from these insights when studying online social services, while the latter could be aware of these additional possibilities when building systems and applications related to online social platforms.

Despite the clear effect that space and geography have on online users, we imagine that taking advantage of the spatial properties of social services in real scenarios would still present interesting design challenges. Facing these challenges requires inventive thinking and technical knowledge that stem from the particular domain of interest: our hope is that this work and its results can facilitate and inspire such creative process, by offering a new perspective on online social services.

[4] `http://www.glancee.com`

Bibliography

[AA03] L. A. Adamic and E. Adar, *Friends and neighbors on the Web*, Social Networks **25** (2003), no. 3, 211–230.

[AA05] E. Adar and L. A. Adamic, *Tracking information epidemics in blogspace*, Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005) (Compiègne, France), IEEE Computer Society, 2005, pp. 207–214.

[AAN04] R. Albert, I. Albert, and G. L. Nakarado, *Structural vulnerability of the north american power grid*, Physical Review E **69** (2004), no. 2, 025103.

[AHK+07] Y. Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, *Analysis of topological characteristics of huge online social networking services*, Proceedings of the 16th World Wide Web Conference (WWW 2007) (Banff, Alberta, Canada), 2007, pp. 835–844.

[AJB00] R. Albert, H. Jeong, and A.-L. Barabási, *Error and attack tolerance of complex networks*, Nature **406** (2000), no. 6794, 378–382.

[AMS09] S. Aral, L. Muchnik, and A. Sundararajan, *Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks*, Proceedings of the National Academy of Sciences **106** (2009), no. 51, 21544–21549.

[And06] C. Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More*, Hyperion, 2006.

[BA99] A.-L. Barabási and R. Albert, *Emergence of scaling in random networks*, Science **286** (1999), no. 5439, 509–512.

[Bar03] M. Barthélemy, *Crossover from scale-free to spatial networks*, Euro-

physics Letters **63** (2003), 915.

[Bar11] ———, *Spatial Networks*, Physics Reports **499** (2011), 1–101.

[BBPSV04] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, *The architecture of complex weighted networks*, Proceedings of the National Academy of Sciences **101** (2004), no. 11, 3747–3752.

[BBR+11] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna, *Four degrees of separation*, CoRR **abs/1111.4570** (2011).

[BBV05] A. Barrat, M. Barthélemy, and A. Vespignani, *The effects of spatial constraints on the evolution of weighted complex networks*, Journal of Statistical Mechanics: Theory and Experiment (2005), no. P05003.

[BC96] D. L. Banks and K. M. Carley, *Models for network evolution*, Journal of Mathematical Sociology (1996), 173–196.

[bE07] d. boyd and N. Ellison, *Social network sites: Definition, history, and scholarship*, Journal of Computer Mediated Communication **13** (2007), no. 1, 210–230.

[BGG03] M. Barthélemy, B. Gondran, and E. Guichard, *Spatial structure of the internet traffic*, Physica A **319** (2003), 633–642.

[BGLL08] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *Fast unfolding of communities in large networks*, Journal of Statistical Mechanics: Theory and Experiment (2008), P10008.

[BHW92] S. Bikhchandani, D. Hirshleifer, and I. Welch, *A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades*, Journal of Political Economy **100** (1992), no. 5, 992–1026.

[Bia11] L. Bianchi, *The History of Social Networking*, `http://www.viralblog.com/research-whitepapers/the-history-of-social-networking/`, January 2011, Last accessed in January 2012.

[BIVW10] T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer, *Catching a viral video*, Proceedings of the 10th IEEE International Conference on Data Mining Workshops (ICDM 2010 Workshop) (Sidney, Australia),

IEEE Computer Society, 2010, pp. 296–304.

[BLM+06] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang, *Complex networks: Structure and dynamics*, Physics Reports **424** (2006), no. 4-5, 175–308.

[BMS+08] K. Bhattacharya, G. Mukherjee, J. Saramäki, K. Kaski, and S. S. Manna, *The International Trade Network: weighted network analysis and modelling*, Journal of Statistical Mechanics: Theory and Experiment (2008), P02002.

[Bre01] L. Breiman, *Random Forests*, Machine Learning **45** (2001), no. 1, 5–32.

[BRMA12] E. Bakshy, I. Rosenn, C. A. Marlow, and L. A. Adamic, *The Role of Social Networks in Information Diffusion*, Proceedings of the 21st World Wide Web Conference (WWW 2012) (Lyon, France), April 2012.

[BRST01] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády, *The degree sequence of a scale-free random graph process*, Random Structures & Algorithms **18** (2001), no. 3, 279–290.

[BSH07] P. Bonhard, M. A. Sasse, and C. Harries, *"The devil you know knows best": how online recommendations can benefit from social networking*, Proceedings of the 21st British HCI Group Annual Conference on People and Computers (BCS-HCI 2007) (Swinton, UK), British Computer Society, 2007, pp. 77–86.

[BSM10] L. Backstrom, E. Sun, and C. Marlow, *Find me if you can: improving geographical prediction with social and spatial proximity*, Proceedings of the 19th World Wide Web Conference (WWW 2010) (Raleigh, North Carolina, USA), 2010, pp. 61–70.

[BSW12] A. Brodersen, S. Scellato, and M. Wattenhofer, *YouTube Around the World: Geographic Popularity of Videos*, Proceedings of the 21st World Wide Web Conference (WWW 2012) (Lyon, Paris), 2012.

[But03] C. T. Butts, *Predictability of large-scale spatially embedded networks*, Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers (2003), 313–323.

[Cai01] F. Cairncross, *The Death of Distance: How the Communications Revolution Is Changing our Lives*, Harvard Business School Press, 2001.

[Car56] V. Carrothers, *A Historical Review of the Gravity and Potential Concepts of Human Interaction*, Journal of the American Institute of Planners **22** (1956), 94–102.

[Car89] J. W. Carey, *Communication as Culture: Essays on Media and Society*, Routledge, January 1989.

[CBC+10] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, *Inferring social ties from geographic coincidences*, Proceedings of the National Academy of Sciences **107** (2010), no. 52, 22436–22441.

[CCH+08] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri, *Feedback effects between similarity and social influence in online communities*, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD 2008) (Las Vegas, Nevada, USA), ACM, 2008, pp. 160–168.

[CFL09] C. Castellano, S. Fortunato, and V. Loreto, *Statistical physics of social dynamics*, Reviews of Modern Physics **81** (2009), no. 2, 591–646.

[CHBG10] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, *Measuring User Influence in Twitter: The Million Follower Fallacy*, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM 2010) (Washington, DC, USA), 2010.

[Che98] L. Cherkasova, *Improving WWW Proxies Performance with Greedy-Dual-Size-Frequency Caching Policy*, Tech. report, HP, 1998.

[CKR+09] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, *Analyzing the video popularity characteristics of large-scale user generated content systems*, ACM Transactions on Networking **17** (2009), 1357–1370.

[CLP06] P. Crucitti, V. Latora, and S. Porta, *Centrality measures in spatial networks of urban streets*, Physical Review E **73** (2006), 036125.

[CMG09] M. Cha, A. Mislove, and K. Gummadi, *A measurement-driven analy-*

sis of information propagation in the Flickr social network, Proceedings of the 18th World Wide Web Conference (WWW 2009) (Madrid, Spain), ACM, 2009, pp. 721–730.

[Col64] J. S. Coleman, *An introduction to mathematical sociology*, Collier-Macmillan, London, UK, 1964.

[Coo94] C. H. Cooley, *The Theory of Transportation*, Publications of the American Economic Association **9** (1894), no. 3, 71–73.

[CS08] R. Crane and D. Sornette, *Robust dynamic classes revealed by measuring the response function of a social system*, Proceedings of the National Academy of Sciences **105** (2008), no. 41, 15649–15653.

[CSBR11] F. Calabrese, Z. Smoreda, V. D. Blondel, and C. Ratti, *Interplay between telecommunications and face-to-face interactions: a study using mobile phone data*, PLoS ONE **6** (2011), no. 7, e20814.

[CSLP06] A. Cardillo, S. Scellato, V. Latora, and S. Porta, *Structural properties of planar graphs of urban street patterns*, Physical Review E **73** (2006), 066107.

[CTH+10] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh, *Bridging the gap between physical location and online social networks*, Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp 2011) (Copenhagen, Denmark), 2010, pp. 119–128.

[CW08] J. Caverlee and S. Webb, *A Large-Scale study of MySpace: Observations and implications for online social networks*, Proceedings from the second AAAI International Conference on Weblogs and Social Media (ICWSM 2008) (Seattle, Washington, USA), 2008.

[Db04] J. Donath and d. boyd, *Public Displays of Connection*, BT Technology Journal **22** (2004), no. 4, 71–82.

[DDADG05] L. Danon, J. Duch, A. Arenas, and A. Dìaz-Guilera, *Comparing community structure identification*, Journal of Statistical Mechanics: Theory and Experiment (2005), P9008.

[DGP07] Y. Dourisboure, F. Geraci, and M. Pellegrini, *Extraction and classification of dense communities in the web*, Proceedings of the 16th

World Wide Web conference (WWW 2007) (Banff, Alberta, Canada), ACM, 2007, pp. 461–470.

[DMW03] P. S. Dodds, R. Muhamad, and D. J. Watts, *An experimental study of search in global social networks*, Science **301** (2003), no. 5634, 827–829.

[EC08] B. M. Evans and E. H. Chi, *Towards a model of understanding social search*, Proceedings of the 11th ACM Conference on Computer Supported Cooperative Work (CSCW 2008) (San Diego, California, USA), ACM, 2008, pp. 485–494.

[EEBL11] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte, *Uncovering space-independent communities in spatial networks*, Proceedings of the National Academy of Sciences **108** (2011), 7663–7668.

[EMB02] H. Ebel, L.-I. Mielsch, and S. Bornholdt, *Scale-free topology of e-mail networks*, Physical Review E **66** (2002), 035103.

[EP05] N. Eagle and A. Pentland, *Social serendipity: Mobilizing social software*, IEEE Pervasive Computing **4** (2005), no. 2, 28–34.

[EPL09] N. Eagle, A. S. Pentland, and D. Lazer, *Inferring friendship network structure by using mobile phone data*, Proceedings of the National Academy of Sciences **106** (2009), no. 36, 15274–15278.

[ES90] S. Erlander and F. S. Stewart, *The Gravity Model in Transportation Analysis: Theory and Extensions*, Brill Academic Publishers, Utrecht, Netherlands, 1990.

[Fac12] *Facebook statistics*, `https://www.facebook.com/press/info.php?statistics`, 2012, Last accessed in January 2012.

[FBA11] F. Figueiredo, F. Benevenuto, and J. M. Almeida, *The tube over time: characterizing popularity growth of YouTube videos*, Proceedings of the fourth ACM International Conference on Web Search and Data Mining (WSDM 2011) (Hong Kong, PRC), ACM, 2011, pp. 745–754.

[Fel81] S. L. Feld, *The Focused Organization of Social Ties*, American Journal of Sociology **86** (1981), no. 5, 1015–1035.

[FFF99] M. Faloutsos, P. Faloutsos, and C. Faloutsos, *On power-law relationships of the Internet topology*, Proceedings of the fifth International Conference on Computer and Data Communication Networks (SIGCOMM 1999) (Cambridge, Massachusetts, United States), ACM, 1999, pp. 251–262.

[FLGC02] G.W. Flake, S. Lawrence, C.L. Giles, and F.M. Coetzee, *Self-organization and identification of Web communities*, IEEE Computer **35** (2002), no. 3, 66 –70.

[FLM06] M. J. Freedman, K. Lakshminarayanan, and D. Mazières, *OASIS: Anycast for any Service*, Proceedings of the third Symposium on Networked Systems Design and Implementation (NSDI '06) (San Jose, CA), USENIX, 2006.

[FMNW03] D. Fetterly, M. Manasse, M. Najork, and J. Wiener, *A large-scale study of the evolution of Web pages*, Proceedings of the 14th World Wide Web Conference (WWW 2003) (Budapest, Hungary), ACM, 2003, pp. 669–678.

[For10] S. Fortunato, *Community detection in graphs*, Physics Reports **486** (2010), 75–174.

[FSB63] L. Festinger, S. Schachter, and K. Back, *Social pressures in informal groups: A study of human factors in housing*, Stanford University Press, 1963.

[FWI+98] E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. H. Witten, *Using model trees for classification*, Machine Learning **32** (1998), 63–76.

[Gal10] G. Gale, *Location vs. place vs. poi*, `http://www.vicchi.org/2010/11/16/location-vs-place-vs-poi/`, November 2010, Last accessed in February 2012.

[GGCM08] M. Gashler, C. Giraud-Carrier, and T. Martinez, *Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous*, Proceedings of the seventh International Conference on Machine Learning and Applications (ICMLA2008) (San Diego, CA), IEEE, December 2008, pp. 900–905.

[GGLNT04] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, *Information*

diffusion through blogspace, Proceedings of the 13th World Wide Web conference (WWW 2004) (New York, New York, USA), ACM, 2004, pp. 491–501.

[GGM+09] T. Gupta, S. Garg, A. Mahanti, N. Carlsson, and M. Arlitt, *Characterization of FriendFeed - A Web-based Social Aggregation Service*.

[GKF+06] S. Garriss, M. Kaminsky, M. J. Freedman, B. Karp, D. Mazières, and H. Yu, *RE: Reliable Email*, Proceedings of the third Symposium on Networked Systems Design and Implementation (NSDI '06), May 2006, pp. 297–310.

[GL04] D. Garlaschelli and I. Loffredo, *Patterns of link reciprocity in directed networks*, Physical Review Letters **93** (2004), no. 26.

[GN02] M. Girvan and M. E. J. Newman, *Community structure in social and biological networks*, Proceedings of the National Academy of Sciences **99** (2002), no. 12, 7821–7826.

[GN06] M. T. Gastner and M. E. J. Newman, *The spatial structure of networks*, European Physical Journal B **49** (2006), no. 2, 247–252.

[Gol08] J. Golbeck, *Weaving a web of trust*, Science **321** (2008), no. 5896, 1640–1641.

[Gra73] M. S. Granovetter, *The strength of weak ties*, American Journal of Sociology **78** (1973), no. 6, 1360–1380.

[Gra87] Mark. S. Granovetter, *Threshold Models of Collective Behavior*, American Journal of Sociology **83** (1987), no. 6, 1420–1443.

[Gra98] S. Graham, *The end of geography or the explosion of place? Conceptualizing space, place and information technology*, Progress in human geography **22** (1998), no. 2, 165–185.

[Gua90] J. Guare, *Six degrees of separation: a play*, Random House, 1990.

[GWH07] S. A. Golder, D. M. Wilkinson, and B. A. Huberman, *Rhythms of Social Interaction: Messaging Within a Massive Online Network*, Proceedings of the third International Conference on Communities and Technologies, Springer London, 2007, pp. 41–66.

[HA99] B. A. Huberman and L. A. Adamic, *Growth dynamics of the World-Wide Web*, Nature **401** (1999), no. 6749, 131–131.

[Hay08] B. Hayes, *Cloud computing*, Communications of the ACM **51** (2008), 9–11.

[Hei46] F. Heider, *Attitudes and cognitive organization*, Journal of Psychology **21** (1946), 107–112.

[HHSC11] B. Hecht, L. Hong, B. Suh, and E. H Chi, *Tweets from Justin Bieber's Heart: The Dynamics of the "Location" Field in User Profiles*, Proceedings of the 30th Conference on Human Factors in Computing Systems (CHI 2011) (Vancouver, British Columbia, Canada), May 2011.

[HK02] P. Holme and B. J. Kim, *Growing scale-free networks with tunable clustering.*, Physical Review E **65** (2002), no. 2 Pt 2, 026107+.

[HK10] D. Horowitz and S. D. Kamvar, *The Anatomy of a Large-Scale Social Search Engine*, Proceedings of the 19th World Wide Web Conference (WWW 2010) (Raleigh, North Carolina (USA)), ACM, 2010.

[Hum07] L. Humphreys, *Mobile Social Networks and Social Practice: A Case Study of Dodgeball*, Journal of Computer-Mediated Communication **13** (2007), no. 1.

[HW01] K. Hampton and B. Wellman, *Long Distance Community in the Network Society*, American Behavioral Scientist **45** (2001), no. 3, 476–495.

[HWL+11] Y. Hu, Y. Wang, D. Li, S. Havlin, and Z. Di, *Possible Origin of Efficient Navigation in Small Worlds*, Physical Review Letters **106** (2011), no. 10, 108701.

[HWLR08] C. Huang, A. Wang, J. Li, and K. W. Ross, *Measuring and Evaluating Large-scale CDNs* , Proceedings of the eighth Internet Measurement Conference (IMC 2008) (Vouliagmeni, Greece), ACM, 2008.

[HZ09] Y. Hu and D. Zhu, *Empirical analysis of the worldwide maritime transportation network*, Physica A **388** (2009), no. 10, 2061–2071.

[IKK+00] C. L. Isbell, M. J. Kearns, D. Kormann, S. P. Singh, and P. Stone, *Cobot in LambdaMOO: A Social Statistics Agent*, Proceedings of the National Conference on Artificial Intelligence, 2000, pp. 36–41.

[JIB07] A. Jøsang, R. Ismail, and C. Boyd, *A survey of trust and reputation systems for online service provision*, Decision Support Systems **43** (2007), 618–644.

[JSFT07] A. Java, X. Song, T. Finin, and B. Tseng, *Why we twitter: understanding microblogging usage and communities*, Proceedings of the ninth WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (WebKDD/SNA-KDD 2007) (San Jose, California), ACM, 2007, pp. 56–65.

[JWS08] W.-S. Jung, F. Wang, and H. E. Stanley, *Gravity model in the Korean highway*, Europhysics Letters **81** (2008), no. 4, 48005+.

[KB78] P. Killworth and H. Bernard, *Reverse small world experiment*, Social Networks **1** (1978), 159–192.

[KCRB09] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel, *Urban Gravity: a Model for Intercity Telecommunication Flows*, Journal of Statistical Mechanics: Theory and Experiment (2009), no. L07003.

[KGA08] B. Krishnamurthy, P. Gill, and M. Arlitt, *A few chirps about Twitter*, Proceedings of the first Workshop on Online Social Networks (WOSN 2008) (Seattle, Washington, USA), ACM, 2008, pp. 19–24.

[KGNR10] T. Karagiannis, C. Gkantsidis, D. Narayanan, and A. Rowstron, *Hermes: clustering users in large-scale e-mail services*, Proceedings of the first ACM Symposium on Cloud computing (SoCC 2010) (Indianapolis, Indiana, USA), ACM, 2010, pp. 89–100.

[KH04] M. Kaiser and C. C. Hilgetag, *Spatial growth of real-world networks*, Physical Review E **69** (2004), 036103.

[KH06] D. Krackhardt and M. S. Handcock, *Heider vs Simmel: emergent features in dynamic structures*, Proceedings of the 23rd International Conference on Machine Learning (ICML 2006) (Pittsburgh, Pennsylvania, USA), Springer-Verlag, 2006, pp. 14–27.

[Kle00] J. M. Kleinberg, *Navigation in a small world*, Nature **406** (2000), 845.

[KLPM10] H. Kwak, C. Lee, H. Park, and S. Moon, *What is Twitter, a social network or a news media?*, Proceedings of the 19th World Wide Web Conference (WWW 2010) (Raleigh, North Carolina (USA)), ACM, 2010.

[KMS+09] R. Krishnan, H. V. Madhyastha, S. Srinivasan, S. Jain, A. Krishnamurthy, T. Anderson, and J. Gao, *Moving beyond End-to-end Path Information to Optimize CDN Performance*, Proceedings of the seventh ACM/USENIX Internet Measurement Conference (IMC 2009) (Chicago, Illinois, USA), ACM, 2009.

[KNT06] R. Kumar, J. Novak, and A. Tomkins, *Structure and evolution of online social networks*, Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD 2006) (Philadelphia, Pennsylvania, USA), 2006, pp. 611–617.

[KSS97] H. Kautz, B. Selman, and M. Shah, *Referral Web: combining social networks and collaborative filtering*, Communications of the ACM **40** (1997), 63–65.

[KT06] M. Kurant and P. Thiran, *Extraction and analysis of traffic and topologies of transportation networks*, Physical Review E **74** (2006), 036114.

[KW06] G. Kossinets and D. J. Watts, *Empirical analysis of an evolving social network*, Science **311** (2006), no. 5757, 88–90.

[LAH07] J. Leskovec, L. A. Adamic, and B. A. Huberman, *The dynamics of viral marketing*, ACM Transactions on the Web **1** (2007).

[LBD+08] R. Lambiotte, V. D. Blondel, C. Dekerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Vandooren, *Geographical dispersal of mobile communication networks*, Physica A **387** (2008), no. 21, 5317–5325.

[LBKT08] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, *Microscopic evolution of social networks*, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD 2008) (Las Vegas, Nevada, USA), 2008, pp. 462–470.

[LC09a] N. Li and G. Chen, *Analysis of a location-based social network*, Proceedings of the 12th International Conference on Computational Science and Engineering (CSE 2009) (Vancouver, British Columbia, Canada), vol. 4, IEEE Computer Society, 2009, pp. 263–270.

[LC09b] ———, *Multi-Layered Friendship Modeling for Location-Based Mobile Social Networks*, Proceedings of the sixth International Conference on Mobile and Ubiquitous Systems (Mobiquitous 2009) (Toronto, Ontario, Canada), 2009.

[LCB10] V. Leroy, B. B. Cambazoglu, and F. Bonchi, *Cold start link prediction*, Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD 2010) (Washington, DC, USA), ACM, 2010, pp. 393–402.

[Lei09] T. Leighton, *Improving Performance on the Internet*, Communications of the ACM **52** (2009), 44–51.

[Lev10] M. Levy, *Scale-free human migration and the geography of social networks*, Physica A **389** (2010), no. 21, 4913–4917.

[LH08] J. Leskovec and E. Horvitz, *Planetary-scale views on a large instant-messaging network*, Proceedings of the 17th World Wide Web Conference (WWW 2008) (Beijing, PRC), 2008, pp. 915–924.

[Lin12] *LinkedIn: About Us*, `http://press.linkedin.com/about`, 2012, Last accessed in January 2012.

[LKF05] J. Leskovec, J. Kleinberg, and C. Faloutsos, *Graphs over time: densification laws, shrinking diameters and possible explanations*, Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD 2005) (Chicago, Illinois, USA), ACM, 2005, pp. 177–187.

[LLC10] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, *New perspectives and methods in link prediction*, Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD 2010) (Washington, DC, USA), ACM, 2010, pp. 243–252.

[LM02] V. Latora and M. Marchiori, *Is the Boston subway a small-world*

network?, Physica A **314** (2002), no. 1-4, 109–113.

[LNK07] D. Liben-Nowell and J. Kleinberg, *The link prediction problem for social networks*, Journal of the American Society for Information Science and Technology **58** (2007), 1019–1031.

[LNNK+05] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins, *Geographic routing in social networks*, Proceedings of the National Academy of Sciences **102** (2005), no. 33, 11623–11628.

[Lon11] A. Long, *A brief history of social networking 1930-2011*, `http://www.slideshare.net/peoplebrowsr/a-brief-cartoon-history-of-social-networking-19302011`, March 2011, Last accessed in January 2012.

[Mas10] C. Mascolo, *The power of mobile computing in a social era*, Internet Computing **14** (2010), no. 6, 76 –79.

[McL62] M. McLuhan, *The Gutenberg Galaxy: the making of typographic man*, University of Toronto Press, 1962.

[ME95] D. Maltz and K. Ehrlich, *Pointing the way: active collaborative filtering*, Proceedings of the 14th Conference on Human Factors in Computing Systems (CHI 1995) (Denver, Colorado, United States), ACM Press/Addison-Wesley Publishing Co., 1995, pp. 202–209.

[ML76] B. Mayhew and R. Levinger, *Size and the density of interaction in human aggregates*, American Journal of Sociology **82** (1976), no. 1, 86–110.

[MMG+07] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, *Measurement and Analysis of Online Social Networks*, Proceedings of the fifth ACM/USENIX Internet Measurement Conference (IMC 2007) (San Diego, CA), October 2007.

[Mor10] J. B. Morris, *The privacy implication of commercial location-based services.*, `https://www.cdt.org/files/pdfs/CDT-MorrisLocationTestimony.pdf`, 2010, Last accessed in May 2012.

[MSLC01] M. McPherson, L. Smith-Lovin, and J. M Cook, *Birds of a Feather:*

Homophily in Social Networks, Annual Review of Sociology **27** (2001), 415–444.

[MW07] D. Mok and B. Wellman, *Did distance matter before the internet? interpersonal contact and support in the 1970s*, Social Networks **29** (2007), no. 3, 430–461.

[MW09] ———, *Does distance still matter in the age of the Internet?'*, Urban Studies **46** (2009).

[MZL+11] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, *Recommender systems with social regularization*, Proceedings of the fourth ACM International Conference on Web Search and Data Mining (WSDM 2011) (Hong Kong, PRC), ACM, 2011, pp. 287–296.

[New02] M. E. J. Newman, *Assortative mixing in networks*, Phys. Rev. Lett. **89** (2002), 208701.

[New04] ———, *Detecting community structure in networks*, European Physical Journal B **38** (2004), no. 2, 321–330.

[NG04] M. E. J. Newman and M. Girvan, *Finding and evaluating community structure in networks*, Physical Review E **69** (2004), no. 2, 026113.

[Nic09] C. Nickson, *The History of Social Networking*, `http://www.digitaltrends.com/features/the-history-of-social-networking/`, January 2009, Last accessed in January 2012.

[NL75] L. Nahemow and M. Lawton, *Similarity and propinquity in friendship formation*, Journal of Personality and Social Psychology **32** (1975), no. 2, 205–213.

[NP03] M. E. J. Newman and Juyong Park, *Why social networks are different from other types of networks*, Physical Review E **68** (2003), no. 3, 036122.

[NWS02] M. E. J. Newman, D. J. Watts, and S. H. Strogatz, *Random graph models of social networks*, Proceedings of the National Academy of Sciences **99** (2002), no. Suppl. 1, 2566–2572.

[OAG+11] J.-P. Onnela, S. Arbesman, M. C. González, A.-L. Barabási, and N. A. Christakis, *Geographic constraints on social network groups*, PLoS ONE **6** (2011), no. 4, e16939.

[PBMW99] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report 1999-66, Stanford InfoLab, November 1999.

[PES+10] J. M. Pujol, V. Erramilli, G. Siganos, X. Yang, N. Laoutaris, P. Chhabra, and P. Rodriguez, *The little engine(s) that could: scaling online social networks*, Proceedings of the 16th International Conference on Computer and Data Communication Networks (SIGCOMM 2010) (New Delhi, India), ACM, 2010.

[PF01] F. Provost and T. Fawcett, *Robust Classification for Imprecise Environments*, Machine Learning **42** (2001), 203–231.

[PG04] E. Paulos and E. Goodman, *The familiar stranger: anxiety, comfort, and play in public places*, Proceedings of the 23th Conference on Human Factors in Computing Systems (CHI 2004) (Vienna, Austria), ACM, 2004, pp. 223–230.

[PSV01] R. Pastor-Satorras and A. Vespignani, *Epidemic spreading in scale-free networks*, Physical Review Letters **86** (2001), 3200–3203.

[QC09] D. Quercia and L. Capra, *Friendsensing: recommending friends using mobile phones*, Proceedings of the third ACM Conference on Recommender Systems (RecSys 2009) (New York, New York, USA), ACM, 2009, pp. 273–276.

[Qui93] J. R. Quinlan, *C4.5: programs for machine learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[QWB+09] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, *Cutting the electric bill for internet-scale systems*, Proceedings of the 15th International Conference on Computer and Data Communication Networks (SIGCOMM 2009) (Barcelona, Spain), ACM, 2009.

[Rap53] A. Rapoport, *Spread of information through a population with socio-structural bias*, Bulletin of Mathematical Biology **15** (1953), 523–533, 10.1007/BF02476440.

[Rat08] F. Ratiu, *Facebook blog: People you may know*, `https://blog.facebook.com/blog.php?post=15610312130`, May 2008, Last accessed in February 2012.

[RB02] A. Reka and A.-L. Barabási, *Statistical mechanics of complex networks*, Reviews of Modern Physics **74** (2002), 47–97.

[RBC+11] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and Virgílio Almeida, *On word-of-mouth based discovery of the Web*, Proceedings of the ninth ACM/USENIX Internet Measurement Conference (IMC 2011) (Berlin, Germany), ACM, 2011, pp. 381–396.

[RCC+04] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, *Defining and identifying communities in networks*, Proceedings of the National Academy of Sciences **101** (2004), no. 9, 2658–2663.

[RD02] M. Richardson and P. Domingos, *Mining knowledge-sharing sites for viral marketing*, Proceedings of the eigth ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD 2002) (Edmonton, Alberta, Canada), ACM, 2002, pp. 61–70.

[Rog95] E. G. Rogers, *Diffusion of innovations*, New York Free Press, 1995.

[Sal83] G. Salton, *Introduction to Modern Information Retrieval*, McGraw-Hill Companies, September 1983.

[SC10] N. Sastry and J. Crowcroft, *SpinThrift: saving energy in viral workloads*, Proceedings of the first ACM SIGCOMM Workshop on Green Networking (New Delhi, India), ACM, 2010, pp. 69–76.

[Sch69] T. Schelling, *Models of segregation*, American Economic Review **59** (1969), no. 2.

[Sie10] MG Siegler, *One millionsquare: Foursquare hits the big number*, `http://techcrunch.com/2010/04/22/foursquare-one-million-users/`, April 2010, Last accessed in July 2012.

[Sim08] G. Simmel, *The sociology of Georg Simmel*, The Free Press, New York, USA, 1908.

[Sim09] M. Simon, *The Complete History of Social Networking – CBBS to Twitter*, `http://www.maclife.com/article/feature/complete_history_social_networking_cbbs_twitter`, December 2009, Last accessed in January 2012.

[SRCCMV08] R. V. Solé, M. Rosas-Casals, B. Corominas-Murtra, and S. Valverde, *Robustness of the European power grids under intentional attack*, Phys. Rev. E **77** (2008), 026102.

[SRML09] E. Sun, I. Rosenn, C. Marlow, and T. M. Lento, *Gesundheit! Modeling Contagion through Facebook News Feed*, Proceedings of the third International AAAI Conference on Weblogs and Social Media (ICWSM 2009) (San Jose, California, USA), 2009.

[SS98] D. Strang and S. A. Soule, *Diffusion in Organizations and Social Movements: From Hybrid Corn to Poison Pills*, Annual Review of Sociology **24** (1998), no. 1, 265–290.

[STCR10] A. Silberstein, J. Terrace, B. F. Cooper, and R. Ramakrishnan, *Feeding frenzy: selectively materializing users' event feeds*, Proceedings of the 36th International Conference on Management of Data (SIGMOD 2010) (Indianapolis, Indiana, USA), ACM, 2010, pp. 831–842.

[Ste41] J. Q. Stewart, *An inverse distance variation for certain social influences*, Science **93** (1941), no. 2404, 89–90.

[Sti02] S. M. Stigler, *Statistics on the Table: The History of Statistical Concepts and Methods*, Harvard University Press, 2002.

[SW93] M. F. Schwartz and D. C. M. Wood, *Discovering shared interests using graph analysis*, Communications of the ACM **36** (1993), 78–89.

[SWB+08] G. Swamynathan, C. Wilson, B. Boe, K. Almeroth, and B. Y. Zhao, *Do social networks improve e-commerce? A study on social marketplaces*, Proceedings of the first Workshop on Online Social Networks (WOSN 2008) (Seattle, Washington, USA), ACM, 2008, pp. 1–6.

[SYC09] N. Sastry, E. Yoneki, and J. Crowcroft, *Buzztraq: predicting geographical access patterns of social cascades using social networks*, Proceedings of the second ACM EuroSys Workshop on Social Network Systems (SNS 2009) (Nuremberg, Germany), ACM, 2009, pp. 39–45.

[THT+12] S. Traverso, K. Huguenin, V. Trestian, I. Erramilli, N. Laoutaris, and K. Papagiannaki, *TailGate: Handling Long-Tail Content with a Little Help from Friends*, Proceedings of the 21st World Wide Web Conference (WWW 2012) (Lyon, France), April 2012.

[TL88] N. Thrift and A. Leyshon, *The gambling propensity: Banks, developing country debt exposures and the new international financial system*, Geoforum **19** (1988), no. 1, 55–69.

[TM69] J. Travers and S. Milgram, *An Experimental Study of the Small World Problem*, Sociometry **32** (1969), no. 4, 425–443.

[Tob70] W. R. Tobler, *A Computer Movie Simulating Urban Growth in the Detroit Region*, Economic Geography **46** (1970), 234–240.

[TRW09] M. Torkjazi, R. Rejaie, and W. Willinger, *Hot today, gone tomorrow: on the migration of MySpace users*, Proceedings of the second Workshop on Online Social Networks (WOSN 2009) (Barcelona, Spain), ACM, 2009, pp. 43–48.

[Tun09] D. Tunkelang, *A Twitter Analog to PageRank*, `http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank/`, January 2009, Last accessed in January 2012.

[VMCG09] B. Viswanath, A. Mislove, M. Cha, and K. Gummadi, *On the evolution of user interaction in Facebook*, Proceedings of the second Workshop on Online Social Networks (WOSN 2009) (Barcelona, Spain), ACM, 2009, pp. 37–42.

[Wax88] B. M. Waxman, *Routing of multipoint connections*, Selected Areas in Communications **6** (1988), no. 9, 1617–1622.

[WBHS06] C. Wiuf, M. Brameier, O. Hagberg, and M. P. H. Stumpf, *A likelihood approach to analysis of network data*, Proceedings of the National Academy of Sciences **103** (2006), no. 20, 7566–7570.

[WBS+09] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao, *User interactions in social networks and their implications*, Proceedings of the 4th ACM European Conference on Computer Systems (EuroSys 2009) (Nuremberg, Germany), ACM, 2009, pp. 205–218.

[WC06] J. Wang and J. Canny, *End-user place annotation on mobile devices: a comparative study*, Proceedings of the 26th Conference on Human Factors in Computing Systems - Extended abstracts (CHI 2006) (Montréal, Quebec, Canada), ACM, 2006, pp. 1493–1498.

[WF94] S. Wasserman and K. Faust, *Social network analysis: methods and applications*, 1 ed., Cambridge University Press, November 1994.

[WF05] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, second ed., Morgan Kaufmann, June 2005.

[WH99] B. Wellman and K. Hampton, *Living Networked On and Offline*, Contemporary Sociology **28** (1999), no. 6, 648–654.

[WPD+10] M. P. Wittie, V. Pejovic, L. Deek, K. C. Almeroth, and B. Y. Zhao, *Exploiting locality of interest in online social networks*, Proceedings of the sixth International Conference on Emerging Networking Experiments and Technologies (CONEXT 2010) (Philadelphia, Pennsylvania), ACM, 2010, pp. 25:1–25:12.

[WS98] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks.*, Nature **393** (1998), no. 6684, 440–442.

[WSS05] B. Wong, A. Slivkins, and E. Sirer, *Meridian: a lightweight network location service without virtual coordinates*, Proceedings of the 11th International Conference on Computer and Data Communication Networks (SIGCOMM 2005) (Philadelphia, Pennsylvania, USA), vol. 35, ACM, August 2005, pp. 85–96.

[YJB02] S.-H. Yook, H. Jeong, and A.-L. Barabási, *Modeling the Internet's large-scale topology*, Proceedings of the National Academy of Sciences **99** (2002), no. 21, 13382–13386.

[Zha04] H. Zhang, *The Optimality of Naive Bayes*, Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004) (Miami Beach, Florida), AAAI Press, 2004.

[Zip49] G. K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, Massachusetts, USA, 1949.