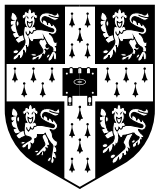
Human Urban Mobility in Location-based Social Networks: Analysis, Models and Applications

Anastasios Noulas



University of Cambridge
Computer Laboratory
St. Edmund's College

2013

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.

Human Urban Mobility in Location-based Social Networks: Analysis, Models and Applications

Anastasios Noulas

Summary

Location-based social networks have attracted the interest of millions of users who can now not only connect and interact with their friends, as in the case of traditional online social networks, but can share their whereabouts in real time exploiting GPS sensors embedded in smartphones with Internet connectivity. Real world places are a core entity of location-based social networks and as users transit between them, urban mobility is represented with unprecedented richness in terms of geographic scale and spatial granularity. As a consequence, location-based services offer new opportunities in the space of mobile applications, but also the potential to allow large scale empirical validation of theories of human movement. However, this new data paradigm comes with the sparsity that is a direct consequence of the heavy-tailed distributions characterising user activity in online social services.

In this dissertation, we perform an analysis of millions of user movements in 34 large metropolitan areas around the world. Our initial observation is that there is significant heterogeneity across cities when considering the statistical properties presented by the movement of users in the urban setting. We identify the source of this heterogeneity to be variations in the geographic density of places across different urban environments. In particular, we discover that in human urban movement it is the relative density between the origin and the destination place that matters - not their absolute geographic distance.

Next, we address a mobility prediction scenario whose application aim is the recommendation of the next place to be visited by a mobile user in real time. Since the limited availability of historic information for each user impedes the use of prediction frameworks that model specifically the movements of an individual, we propose a novel supervised learning training strategy that relies on information built according to the place preferences of user collectives. Finally, we highlight that almost two out of three places visited by users in location-based social networks are new places, not observed being visited by that user historically. In the light of this observation the problem is set to be the recommendation of new venues for mobile users to visit in future time periods. We show how state of the art online filtering algorithms are outperformed by a random walk with restart method that is able to seamlessly combine multiple data signals and cope with the sparse representations of users in the service.

Acknowledgements

I would first like to thank my advisor Cecilia Mascolo. If it was not for her, I would not be studying this degree, not writing this Thesis either and most likely you would not have known me. During these years she has patiently monitored and directed my work, allowing for the expression and development of the good qualities I may possess, but most importantly, she mentored me to become a researcher and taught me to navigate in the landscape of academia in a manner that I now love and appreciate. Thank you.

Subsequently, I would like to express my gratitude to my colleagues in the mobisys group. Salvo has been of untold support since the start of my PhD and the words location, geography and check-in or check in would not have been that familiar to me without him. Big thanks also goes to my PhD roomate Kiran whose positive energy, possibly derived from thousands of years of sacred knowledge acquisition and powerful meditation of generations of Indian monks, has helped me stay calm and feel cozy in my space in the lab. Thanks also to Neal, whose philosophical discourse on recommender systems has helped putting the last, yet important, bricks of this thesis. Chloë whose Oxfordian background has added a touch of classical academic writing style to my writings, but also for the infinite coffee cups we consumed together when the coffee machine was not broken. I would also like to thank Ilias, Christos, Sarfraz, Kharsim, Daniele, Liam, Enzo, Theus, Bence, John, Dmytro, Hao, Sandra, Deborah, Desi, Petko, Amy and the rest of people at NetOS who have formed a magnificent constellation of stars that made the lab not only a good place to work, but also a fun place to be.

I was also lucky to have received great support from people outside the lab in Cambridge. Mirco Musolesi who has walked with me the first steps of my academic existence and showed me the tricks. Massi Pontil at University College London who introduced me to the hardcore science world of Machine Learning and provided crucial help in the beginning of the thesis. Renaud Lambiotte who helped me escape from the reductionist church by revealing the exciting universe of complex systems. Finally, thanks to Enrique Frias-Martinez for the great times at Telefonica, Madrid and Blake Shaw at Foursquare who opened the doors to one of my favourite highlights in the last 4 year period: check in at Foursquare Headquarters in New York.

Last, but not least, thanks to a beautiful flower I discovered when I was wondering around the urban jungle of London sometime in the middle of my PhD.

This Thesis is dedicated to my parents Ntina and Dimitris, and sister Ioanna, for their infinite support.

Contents

1

Introduction

Location-based services correspond to a relatively recent advancement in the space of the World Wide Web and, more broadly, Computer Science. This is an evolution tightly knit to processes driven by the increasingly popular use of computationally powerful mobile devices, commonly known as *smartphones*, that have provided the web users with the ability to access on-line services as they are on the move and, literally, from any place on the planet where Internet connectivity is present. Prominent representatives of the class of location-based services are *location-based social networks* (LBSNs). These are systems that allow users to connect and interact with their friends on-line, as happens in traditional on-line social networks, but with the additional feature that these interactions are focused on real world places.

As millions of users exploit location-based social networks, they generate sequences of digital mobility traces whose scale in terms of numbers of users involved, geographic reach and spatio-temporal granularity is unprecedented. This comes into direct contrast with previous methods used by scientists to collect datasets that describe human movement; population survey methods employed typically by urbanists have been of high economic costs and rather static in recording the temporal dimension of movement, while sensor based methods instrumented by computer scientists in recent years could only be deployed on a small number of participants and for a finite period of time. Furthermore, datasets which describe movement and are owned by large telecommunication providers have become only sporadically available for privacy and economic reasons.

But what are really the opportunities offered by the exploitation of mobility datasets generated in location-based social networks for science, in general, and computer science

more specifically? There has been extensive literature published that attempts to explain human movement and migration patterns. Scholarship on human mobility has been multi-disciplinary in nature and stretches from fields in the social sciences, such as anthropology and sociology, to areas of natural sciences such as physics. Thus, empirical data on human mobility sourced from the new generation of mobile web services can be helpful to compare and validate models and theories that have been enabled by scientists to explain the motives behind human mobility and even predict the future movements of individuals.

Yet, as data from the mobile web is able to put under scrutiny classical models of human movement, there is a feedback process that computer scientists can exploit towards the design and development of a new generation of mobile application and services, in the context of which, location and human mobility play an important role. Indeed, in order to extend well-established applications of computer science such as search and recommender systems and deploy them geographically, we need to understand how the users of these services decide to move across space. To this end, measuring how factors such as the cost of distance or the attractiveness of popular places, to give an example, affect the mobility of individual users is fundamental. Further, it is questionable whether the performance of algorithms and models that have been successful in the online setting will remain intact upon their migration to mobile platforms and systems.

Besides the direct effect of human movement on user interaction with mobile applications, there are challenges that have to do primarily with the characteristics of these new data and their subsequent integration in computational models. The fact that, for instance, in location-based social networks data is generated by humans already implies the potential presence of bursty activity and highly skewed frequency of use distributions [Bar05, VOD+06]. This in turn can lead to extremely sparse representations of users in these systems that could have profound implications for the algorithms employed to provide a mobile service. As a consequence, in this dissertation, my aim will not only be the empirical validation of past mobility theories in the light of new data and understanding how these can benefit computer science models, but in addition, I will look for ways to *mine* these datasets so as to build appropriate features for machine learning algorithms to function effectively in the context of human mobility applications.

So far and despite the fact that location-based social networks are at an early stage of maturity, considering that they have not yet reached the order of hundreds of millions of users similarly to the big players in the market of online social networks, such as Facebook [1] or Twitter [2], their impact in research is noticeable across many areas of computing. Local search [SSSH13], urban computing and neighbourhood modelling [CSHS12], spatial social network analysis [CTH+10, SNLM11], human mobility [CML11], mobility privacy [PZ10] and natural language processing [BNÓS+12] are a few example cases, amongst an existing plethora, where the multi-dimensional signal captured by data in location-based services

[1] `www.facebook.com`

[2] `www.twitter.com`

has already resulted in important contributions. More relevant to this dissertation will be applications in the area of mobile recommendations and, in particular, the recommendations of places for mobile users to visit as they explore a city. The commercial interest of this application area is promising due to its link with location-based advertising and mobile commerce. In order to realise this interest, however, one has to understand how mobile users choose to move to places and subsequently manage this data in such a manner that mobile applications can benefit their users.

1.1 A historic perspective of human movement studies

In this section I provide a brief, data-centric, view of human mobility studies. I begin with an introduction to how survey based methods and census datasets initiated empirical research describing human migration patterns, in Section 1.1.1. Subsequently, I continue with the description of a more recent case that changed the way scientists would study movement, that is, location data extracted from communication interactions observed in cellular networks. As we shall see in Section 1.1.2, the latter, enabled a paradigm shift to the spatio-temporal scale of acquiring mobility records for large populations and led to important breakthroughs regarding our understanding of human movement. However, there were also limitations that mainly related to the accessibility of these data by the research community. I discuss those topics in more detail next.

1.1.1 Human mobility through traditional survey based methods and census data

The first modern attempt to understand human movement in formal terms took place in 1885 when E.G. Ravenstein published his work, The Laws of Migration [Rav85], in the Journal of the Statistical Society. Being provoked by the remarks of Dr. William Farr who stated the fact that human migration appeared to go on without any definitive law [Tob95], Ravenstein analysed census data in the United Kingdom and highlighted important patterns and regularities of population movements amongst the Irish, British and Scottish Kingdoms. Ravenstein supported his work empirically with census data where migration movements of millions of individuals were recorded and, amongst others, he noted the following:

- Most migration is over a short distance.
- Long range migrants usually move to urban areas.
- Migration increases with economic development.

These statements were one of the first attempts to frame the understanding of human mobility in terms of empirical observation. They suggest that distance has a deterring effect on mobility, that the mass of opportunities represent an attractor for movements and that economic growth accelerates it (perhaps due to favouring the means to overcome distance related costs). Since then, a large volume of work aiming to analyse, model and ultimately understand the process of human migration has appeared [Sja62, Lee66, Gre75, Zel71]. Datasets collected through surveys of human participants, however, have provided only a very static viewpoint of human movement. Even today, they may inform us about the city or country that an individual resides in and, perhaps, the year of this occasion, but little do they say of the exact places people go to and the dynamics of their visiting patterns. Thus, traditional work employing census data suffers from limited spatial and temporal granularity in the description of human movement. While research using census data still provides unprecedented insight into large scale human migratory patterns, it is not able to capture a large fraction of human movement activity occurring daily within and between cities.

But how then can one investigate human mobility in urban environments? First, I note that the necessity to understand how people move in cities has been motivated by the process of intensive urbanisation that took place especially in the second half of the twentieth century. As crowds of agricultural populations rushed into cities resulting in a sudden increase in their size, urbanists and city planners became confronted with big challenges ahead of deploying transport infrastructure and providing administrative services to citizens. Knowledge about how people use urban spaces [Col94, Dun78, SW91], how they commute to work [BG77] and where they live became vital at that point. The principal method to acquire this knowledge has been to conduct population representative surveys [OW01]. These surveys have made it possible to acquire information about the origins and destinations of trips in a city [Bec67, Hym69], but also the transport means employed by commuters. Urban transport modellers in the 1970s exploited origin-destination matrices of city commuters and laid the foundations of novel mobility theories whose impact is still evident in today's research [Haz03, ZRW07].

1.1.2 Mobile Phones as Sensors to study movement

In the early 1990s, the launch of the second generation cellular technology (2G) in Finland signalled a massive change in human communications. Although the mobile phone was an idea that had been around for a while, it was this time that it began to become mainstream. It was the first instance in our history when human movement could potentially be tracked with per second accuracy, relatively high geographic precision and at a massive population scale. When a mobile user initiated a call or sent an SMS, her position could be recorded at the nearest Base Transceiver Station (BTS) that was handling the communication event.

Yet, in parallel to the euphoria echoed by the media and the masses of consumers in the light of the potential to communicate on the go using mobile phones, there was also fear. Privacy concerns were raised about the potential misuse of mobile communication datasets by telecommunication providers. Questions about how long Call Detail Records (CDRs) could be kept in databases and under what conditions governmental agencies could have access to them dominated the discussions between legislators, technologists, corporations and the public. In the meantime, the promising power of human mobility datasets was also noticed by the scientific community, but due to privacy concerns, as mentioned above, and also the fact that the data had significant commercial value that had yet to be realised, telecommunication providers were extremely hesitant to share any information on call detail records with scientists.

The ice broke years later when call detail records became sporadically available to various research groups. In 2009 one of the first large scale analyses of human movement using CDR data was published [GHB08]. A little earlier a work [DBG06], also published in Nature, provided a large scale study of human mobility spatio-temporal patterns, but in that case movements were proxied, in a rather creative way, by tracking the spread of dollar notes in the United States; the assumption was that the movement of money was a convolution of movements by individuals, thus similar statistical laws should govern both. The two studies came to verify empirically a power-law statistical distribution characterising human mobility and measured quantitatively, at the country scale, the effect of distance on movement. In the mobility context, the power-law process described, effectively, the frequent existence of short movements and the rare presence of very long movements. In addition to measuring the impact of distance on movement, in the same period, call detail records were also used to estimate the predictability of human dynamics as those are reflected by the mobility patterns of mobile users. Notably, as suggested by the authors in [SQBB10], the location of a given user could be predicted with an expected accuracy score of 93%.

However, the study of human movement in the urban setting was still lacking quantitative evidence; the spatial granularity of call detail records was not standard and was bound to be accurate only up to a few hundred meters. Despite the fact that the appearance of cellular data constituted a big step towards the understanding of human mobility on a large scale (compared to data received from surveys), it did not provide the opportunity for researchers to shed light on the multitude of movements taking place in cities every moment by millions of people. For that, a technology that would enable the recording of movement with spatial granularity on the order of a few tens of metres was required to emerge.

1.2 The rise of location-based social networks

In this section I present how location-based social networks, a new type of online social network where user location and geography acquire a crucial role, present a novel source of mobility data with several advantageous characteristics when compared to survey or cellular data described in the previous section. Prior to this presentation I provide a brief introduction to the systems that preceded LBSNs, online social networks.

1.2.1 The web and online social networks

With the introduction of the World Wide Web in the early 1990s and the emergence of Internet services, it seemed as if modern societies were transcending the constraints imposed by the physical world and were entering a digital era where information storage and exchange was becoming key to everything. Communications such as email let people interact reliably and conveniently with their peers around the world and commercial activity moved also online and money could now flow from one party to the other almost instantaneously. The improvement of web search technology also boosted things along this direction as it solved the problem of navigating effectively through masses of information.

The perception that people would care more about what was happening in the virtual universe rather than their real lives was favoured further by the introduction of online social networks and, in particular, Facebook. Launched in 2004, Facebook, spread like a forest fire through communities of university students and today counts more than one billion registered users. The service provided (and still does) the opportunity for users to generate profiles describing themselves, allowed them to search and connect virtually to new or existing friends, upload and share multi-media content and, put in brief, it encapsulated the social life of a person in an online platform of information. From a new phenomenon, in 2009, Facebook emerged as the site where users spend the most time [Mas10].

1.2.2 Geography and online social networks

The introduction of smartphones in early 2000 though, signalled a massive transition in the way people were accessing the web. The image of the Internet user who is sitting at a desktop machine to access their email account or other online services began to fade and progressively a new a type of web user emerged. That user was mobile, carrying a computational device capable of accessing the online world from almost anywhere. In parallel, software developers went mobile as well and the new, now dominant, ecosystem of mobile applications came into existence.

The idea then to implement a service that adapts a social networking platform to the mobile space blossomed. The first online social networks that explicitly use location as their primary feature appeared in 2008. Foursquare [foua], Gowalla [gow] and Brightkite [bri] took the lead in this new space and their service was based on a rather simple notion: *share with your friends information about the place where you are.* While the services were greeted by some with skepticism and concerns about the sacrifice of private information for no apparent gain, the gamification features of the three location-based social networks together with the thrill of exploring urban space in a completely new way attracted many early adopters who rapidly formed an enthusiastic user base.

The convergence of the virtual, online, ecosystem with the physical, geographic, space is a phenomenon still in progress. Despite the fact that the first glimpse in this direction was (popularly) given by services such as Flickr and Twitter that allowed for the association of photos and messages (tweets) with geographic information provided by GPS modules embedded in smartphones, it is the activity of users in location-based social networks that has brought into existence a completely new paradigm of crowdsourced, large scale, data with mobile and geographic attributes. These new datasets are expected to have a profound impact for the study of human mobility and behaviour, but also the ways we interact and navigate within the physical environment. Moreover, due to its richness, data generated by these systems not only offers the opportunity to lay new foundations on our understanding of movement, but, in addition, it paves the way for the computer sciences and related applications to manifest in a new era where *location* and *geography* acquire a key role.

1.2.3 The importances of places for the study of human movement

If one were to ascribe the novelty of location-based social networks and services to a single source, then that would be the addition of the notion of a place to these systems. A place, or venue, in these services is a virtual entity associated with a physically existing place in the real world. Foursquare users can add information to the Foursquare database about physically existent places by crowdsourcing information about its geographic coordinates and type. Typical examples of places in location-based services are train stations, libraries, coffee shops, restaurants and bars in a city, although the notion has also been associated with cases that are less expected intuitively; people have been checking in to ferries sailing in the sea, moving trains or even in outer space. Popular cases have been that of an astronaut performing a check-in from an international space station [Blo10] or, more recently, the check-ins sourced from NASA's Curiosity robot on Mars that maintains an official Foursquare profile [Blo12].

As users interact with places we learn about their geographic position, the types of activities, the times that they engage with them and, from a social network perspective, with

whom. Hence, the set of places or, put in a more technical jargon, the *venue database*, constitutes the backbone of a location-based social network. It feeds internal application services with data, but also external applications through an API. This database is being continuously updated with new data points in real time and location-based services such as Yelp! and Foursquare have already started to capitalise on this process. Notably, Foursquare currently aspires to become the *location layer* of the Internet [3] exploiting more than 3.5 billion check-in records in order to provide geo-aware intelligence to thousands of developers who are interested in registering activities of their mobile apps with real world venues. Some of these applications, such as Instagram [4], have evolved to become extremely popular on their own and, as a result, the million of users they have acquired produce new data records feeding back to the original location-based service. This data driven feedback loop constitutes the steam engine of a new ecosystem of smartphone applications that currently spreads on the mobile web providing useful services that could range from better map-based urban exploration systems [Dev13b] to food discovery in a city [Dev13a].

1.3 Thesis and its substantiation

As have discussed in the previous paragraphs, location-based social networks constitute a novel, online, platform that generate human mobility data qualitatively different from past datasets in terms of geographic scale, precision and information about the places users visit. They provide the ability to study, for the first time, the movement of individuals in urban environments and compare observations in them, in a process that could be seen as a global mobility experiment where no definitive end or number of participants is provided. Further, the multiple layers of data that concurrently exist in these systems create a new ecosystem of information with promising implications for mobile web services and applications.

Consequently, the **thesis of this dissertation** is that *the study of location-based social networks can progress our understanding about the factors governing human movement in cities and that appropriate mining of data in these systems can lead to the development of effective applications and services for mobile users who navigate the urban space.*

I substantiate this statement with two closely related threads of research. First, I attempt to describe human movements in cities in ways that reveal the common attributes of user behaviour across them, in order to recover underlying universal patterns in the ways humans choose to visit a location positioned at a certain geographic distance. Along these lines, I also investigate the impact of urban geography in human movement as it is encoded

[3] http://marketingland.com/foursquare-wants-to-be-the-location-layer-for-the-internet-40121/
[4] http://instagram.com/

through the spatial distribution of places. Next, I employ insights and realisations offered by the analysis of human mobility in order to build and evaluate prototype applications that aim to predict user movement and foster the exploration of new places by users in location-based services.

1.4 Chapters and contributions

The contribution of this thesis is threefold. After a discussion on the potential sources of bias with respect to LBSN data in Section 3.1.1, I confirm past empirical observations of human movement with LBSN data. Then, I study the distribution of trip distances of LBSN users within cities. At this level of abstraction, I identify a model that captures human mobility in cities in a universal manner, that is, by applying the model with identical parameters across cities I am able to recover the empirically observed statistical properties of human movement. Finally, I present and evaluate two application scenarios for mobile recommendations in location-based social networks. In the first case, the aim is to predict new, previously unvisited, places for mobile users, and in the second scenario my goal is to predict in real time the next place in the city to be visited by a user.

I shall begin by introducing, in Chapter 2, an overview of historically related work in human mobility. As it has been a subject of study in various scientific disciplines (urban planning, sociology, physics, computer science, etc.) I detail the perspectives and influence of each of them on the study of human movement. In this context, I will highlight the most important research questions investigated by scientists in the study of human mobility, but I will also clarify the novelty of my contributions that will then be described in the chapters to follow. Finally, with respect to the introduction of background work and related concepts, I will list the advantages of the data sourced from location-based social networks and I will provide a broad formalisation of the place recommendation problem that is of particular interest for this dissertation. I will close with a discussion of how the understanding of human mobility can benefit mobile applications, and I will present the difference in the requirements of a place recommendation scenario compared to previous applications in human migration and transport modelling.

The rest of the dissertation is organised in the following way:

- In Chapter 3 I study human movement in the urban setting. By analysing the statistical properties of consecutive user check-ins in 34 metropolitan areas around the globe, I initially note that there are fundamental differences between **intra-city movements** and user movement that takes place generally in space, when the geographic constraints of the urban boundary are not taken into account. In more detail, a **power-law** distribution that was employed to model human mobility in the past cannot explain movements in cities. Furthermore, for the first time

movements are compared across cities using the same data source. The significant **heterogeneity** observed in the mobility patterns of users across different urban environments has led me to the deeper investigation of movement in the city in order to identify their causes (cultural variations, differences in transport infrastructure, etc.). Initially, when I analyse human mobility in cities by looking at the **rank-distance** between places (number of places between an origin and a destination), as opposed to absolute geographic distances, universal patterns across cities emerge. I have then exploited this observation to model urban movements according to the rank-distance variable. The single parameter model we devise is able to accurately capture movements in all cities I examine, and highlights the importance of **geography**, as it is encoded by the spatial distribution of places in cities, in human mobility and its pivotal role as the factor in any observed heterogeneities amongst them.

- In Chapter 4 I study the problem of predicting the next venue a mobile user will visit, by exploring the predictive power offered by different facets of user behaviour.**I am interested in predicting a mixture of historically seen and new places in real time.** I initially propose a set of data mining features that aim to capture the factors that may drive user movements. The features exploit information about transitions between types of places, mobility flows between venues, and spatio-temporal characteristics of user check-in patterns. I further extend the study combining all individual features in supervised learning models. After designing a learning strategy for supervised learning algorithms that effectively deals with sparsity issues in check-in data, I discover that the supervised methodology based on the combination of multiple features offers the highest levels of prediction accuracy: **Continuous learning decision trees are able to rank in the top fifty venues one in two user check-ins, from thousands of candidate items in the prediction list**. Moreover, I observe that the prediction power of the tested algorithms is greatly affected by temporal factors, an insight that has implications for the development of more sophisticated prediction frameworks in the future.

- Chapter 5 introduces an alternative application scenario of the thesis; the goal is to predict **new venues** (historically unobserved) to be visited by mobile users in order to foster the discovery of interesting places in the city that are currently unknown to the user. First, I examine venue discovery behaviour in large check-in datasets from two location-based social services, Foursquare and Gowalla. By using large-scale datasets containing both user check-ins and social ties, the analysis reveals that, across 11 cities, between **60% and 80% of users' visits take place at venues that were not visited in the previous 30 days**. I then show that, by making constraining assumptions about user mobility, state-of-the-art filtering algorithms, including latent space models, do not produce high quality recommendations. Fi-

nally, I propose a new model based on personalised random walks over a user-place graph that, by seamlessly combining social network and venue visit frequency data, obtains between 5 and 18% improvement over other models.

Finally, in Chapter 7 I summarise the findings and identify directions for future work in human mobility research and mobile applications.

1.5 Publication List

During my PhD studies I have been involved in many fruitful collaborations that have yielded 16 published works that span the areas of human mobility modelling, location-based place recommendations, topic modelling, mobile social networks and urban activity and neighbourhood modelling. More related to this dissertation, Chapter 3 is based on the work in [NSL+12]. Salvatore Scellato, Renaud Lambiotte and Cecilia Mascolo provided support on the design of the experiments, pointed to useful related on work on human mobility models and assisted in the writing of the paper. I carried out the implementation of the analysis, modeling and evaluation. Chapter 4 builds on [NSMP11a] and [NSLM12a]. I carried out the design, analysis, implementation and evaluation of these works, whereas the co-authors contributed on the writing of the paper and provided support on refining technical aspects of the methodologies exploited on the predictive models employed. Finally, the contributions of [NSLM12b] are described in Chapter 5. Salvatore Scellato, Neal Lathia and Cecilia Mascolo provided support with the experimental design and writing of the paper, whereas code development was carried out by me and Salvatore Scellato. All the studies are listed below.

Papers related to this dissertation

- [NSMP11a] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. In AAAI International Conference on Weblogs and Social Media, 2011.
- [NSL+12] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: universal patterns in human urban mobility. PloS ONE, 2012.
- [NSLM12a] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining user mobility features for next place prediction in location-based services. In IEEE Internationcal Conference on Data Mining, 2012.
- [NSLM12b] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. A random walk around the city: New venue recommendation in location-based social networks. In IEEE International Conference on Social Computing, 2012.

Other works during PhD study

- [NSMP11b] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In 3rd Workshop Social Mobile Web, Colocated with Fifth International AAAI Conference on Weblogs and Social Media, 2011.

- [BNÓS+12] S. Bauer, A. Noulas, D. O Seaghdha, S. Clark, and C. Mascolo. Talking places: Modelling and analysing linguistic content in foursquare. In IEEE International Conference on Social Computing, 2012.

- [BNS+12b] C. Brown, V. Nicosia, S. Scellato, A. Noulas, and C. Mascolo. Where online friends meet: Social communities in location-based networks. In AAAI International Conference on Weblogs and Social Media, 2012.

- [BNS+12a] C. Brown, V. Nicosia, S. Scellato, A. Noulas, and C. Mascolo. The importance of being placefriends: discovering location-focused online communities. In ACM Workshop on Online Social Networks, 2012.

- [SNM11] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011.

- [SNLM11] S. Scellato, A Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. In AAAI Intenational Confernece on Weblogs and Social Media, 2011.

- [KNS+13] D. Karamshuk, A. Noulas, S. Scellato, V Nicosia, and C. Mascolo. Geo-spotting: Mining online location-based services for optimal retail store placement. In ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2013.

- [ZNSM13] A. Zhang, A. Noulas, S Scellato, and C. Mascolo. Hoodsquare: Exploiting location-based services to detect activity hotspots and neighborhoods in cities. In IEEE International Conference on Social Computing, 2013.

- [BNS+13] C. Brown, V Nicosia, S. Scellato, A. Noulas, and C. Mascolo. Social and place-focused communities in location-based online social networks. In European Physics Journal B., 2013.

- [NFMM13] A. Noulas, E Frias-Martinez, and C. Mascolo. Exploiting foursquare and cellular data to infer user activity in urban environments. In IEEE International Conference on Mobile Data Management, 2013.

- [BNMB13] C. Brown, A. Noulas, C. Mascolo, and V. Blondel. A place-focused model for social networks in cities. In IEEE International Conference on Social Computing, 2013.

2

Human mobility and its importance for mobile applications

As we have seen in the previous chapter, the analysis and modelling of human mobility has a history of almost two centuries. The journey began with the seminal work of Ravenstein [Rav85] and concerned the declaration of a set of principal laws that govern human migration, which was then recorded through the first census survey reports. Since then, surveys constituted the primary source for the study of human movement, until the adoption of mobile phones by the public, towards the end of the twentieth century. As has been described in Chapter 1, the latter advancement allowed for the recording of mobile user whereabouts in real time, through call detail records. However, due to their private status CDRs were not exploited to a great extent by the research community.

Chapter Outline In this chapter, I identify and present in detail the main schools of thought that have emerged in human mobility modelling during the 20th century and have been the main drivers of research in the area. In particular, in Section 2.1, I will begin with the presentation of the so-called *gravity models* that were inspired by Newtonian physics. I will then continue with the introduction of the *intervening opportunities models* which have been initially theorised by Samuel Stouffer, an American sociologist who was active in the 1940s. Interestingly, despite their apparent differences, the two models have been shown to be statistically equivalent under certain assumptions (Section 2.1.3). In light of the above theories, I present the most recent evolutions of large scale movement analysis. In Section 2.2.2 I will list the advantages offered by data sourced from location-based

social networks to the study of human mobility, but I will also highlight how these can benefit the development of a new generation of mobile applications, and in particular, location-based venue recommendations. Finally, in Section 2.3 I will provide a broad formalisation of the place recommendation problem in location-based services and discuss its relationship to human mobility, migration and transport modelling.

2.1 Dominant trends in human mobility modelling

Before proceeding with detailed explanations of models about human movement, it would be appropriate to define the problem in a formal manner. In general, mobility modellers have aimed to capture the statistical properties of movement flows given a certain spatial environment where movement can take place. A main goal in this context has been the prediction of the number (or fraction) of movements between an origin and a destination. Formally, given a set of origin points O and a set of destination points D in space, the aim is to provide a model that accurately enumerates the number of movements (equiv. transitions) between a point i in O and and a point j in D.

The specific application scenarios of the theoretical models that aim to achieve the above goal can vary and sometimes are discipline specific. In urban planning, origins and destinations usually correspond to home locations and employment destinations positioned in geographic zones i and j respectively. In migration theory, the goal is to model the number of individuals migrating from one country to another, or equivalently from one city to another depending on the geographic scale at which movement is studied. In this thesis, and in particular in Chapter 3, our aim is to model trips between *any* place in a city to any other and observe the statistical properties in terms of the frequency distribution of distances that emerges from user movements. However, when considering the place recommendation scenarios in Chapters 4 and 5, I will refine the level of abstraction of human mobility, and target prediction of the exact places visited by users.

2.1.1 Gravity models

In analogy to Newton's law of universal gravitation, the objective gravity models is to model the mobility flow between an origin and a destination proportionally to their *masses* and inversely proportionally to their *distance*. Formally, given two objects i and j with corresponding masses m_i and m_j and geographic distance d_{ij}, then the force of attraction between i and j, given by F_{ij} is equal to

$$F_{ij} = \gamma \frac{m_i \, m_j}{d_{ij}^2} \tag{2.1}$$

where γ is a data specific constant. The analogous formulation in the context of transport modelling [Wil67] would be

$$T_{ij} = k\frac{O_i D_j}{d_{ij}^2} \tag{2.2}$$

for a set of origins O and destinations D and where k is again a constant.

2.1.2 Intervening opportunity models

Despite their elegant mathematical form, there is an element of reductionism lying in the theory behind gravity models. Could it be that humans move like small particles whose behaviour is simply governed by the physical laws of gravitual attraction? Sociologists in the 1930s were very concerned about human movement in space [MMR39] and they were looking for a theory that would sufficiently explain migration patterns at various geographic scales (city, state, country, etc.). Not surprisingly, their stand was philosophically different to that of physicists who were seeing human movement from the point of view of particle diffusion. In December 1940, Samuel Stouffer published a work [Sto40] that attempted to explain the relationship between human mobility and distance in terms that placed human decision making, societal factors and cognition at the centre of the movement process. The theory known as *theory of intervening opportunities* states that:

> *The number of persons going a given distance is directly proportional to the number of opportunities at that distance and inversely proportional to the number of intervening opportunities.* Samuel Stouffer, Theory of Intervening Opportunities, 1940.

Stouffer suggested that the notion of *opportunities* described in the theory could be adopted appropriately, depending on the specific domain that the theory is applied to.

In his paper [Sto40], the theory of intervening opportunities was empirically tested on a census dataset that described the migration movement of families the city of Cleveland, Ohio, United States. In the study, the spatial distribution of employment opportunities was known and Stouffer demonstrated how family mobility was driven by this distribution. In the sketch depicted in Figure 2.1 I provide a toy example describing the theory's rationale in a similar empirical context to that of Stouffer; given the *origin* geographic position where a family in Cleveland is situated, a set of concentric zones is formed around it and the location of job opportunities in these zones is recorded. Then the probability of migrating to a zone of a target job is inversely proportional to the sum of jobs available in the zones between the origin and the destination zone.

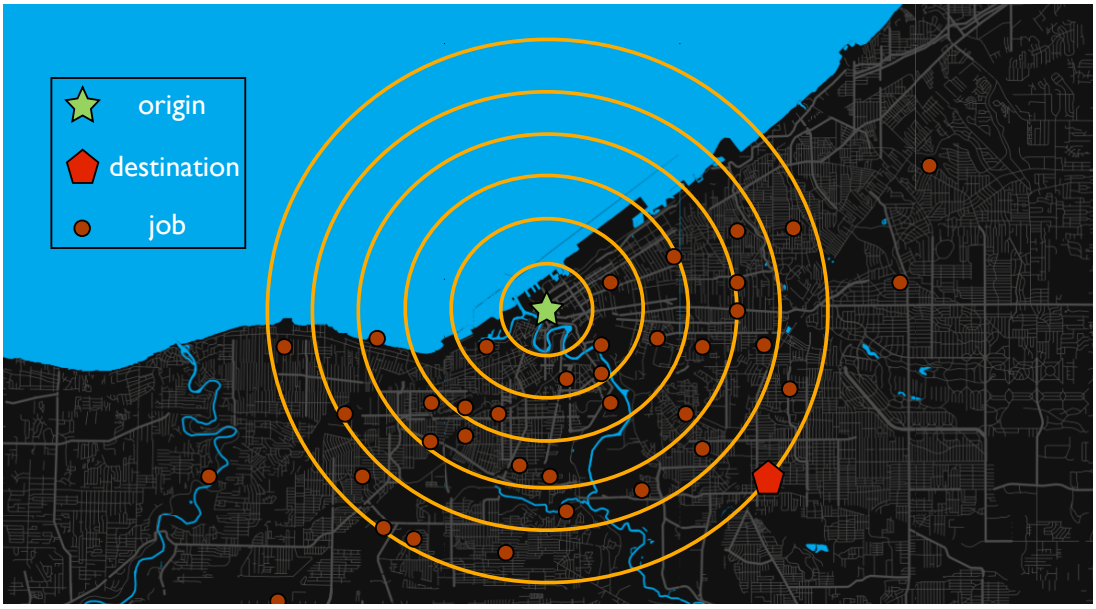


Figure 2.1: Sketch reproducing Stouffer's original approach to empirically verify his theory of intervening opportunities. On a map of Cleveland, Stouffer plotted the positions of jobs that were available during the period of investigation. After splitting the territory of the city into zones, he suggested that the number of families migrating from one zone to another was inversely proportional to the number of jobs between the origin zone (family) and the target zone (job).

2.1.3 Gravity and intervening opportunity models: a proof of convergence

According to the the descriptions of the Gravity and Intervening Opportunities Theories, their fundamental difference is that while the intervening opportunities model explicitly takes into account the existence of *opportunities* between the origin and destination in a potential journey, Equation 2.2 of the Gravity Theory assumes that only the interaction between the two points in question matters. Broadly speaking, the existence of *third parties* cannot influence the flux (volume of trips) T_{ij} between an origin O and a destination D.

Alan G. Wilson, in his 1967 paper [Wil67], identified the deficiencies of a simplistic version of a gravity model and reframed it in a more generalisable form. First, he mentioned that if we double the volume of work trips with origin O_i and also double the number of work trip destinations D_i, the number of trips, T_{ij}, according to Equation 2.2 should quadruple, whereas the expectation would be for T_{ij} to also double. To express this argument formally, Wilson added the following constraints:

$$\sum_j T_{ij} = O_i \tag{2.3}$$

$$\sum_i T_{ij} = D_j \tag{2.4}$$

which effectively guarantee that the number of trips generated at the origin zone O_j and the number of trips attracted towards the destination zone D_j sum up correctly, when two constants A_i and B_j are associated with trip production and attraction zones, respectively. In addition, Wilson provided a more generic representation of the effect of distance in transport modelling by considering a function $f(d_{ij})$ which could be domain specific and relaxes the assumption that T_{ij} is inversely proportional to d_{ij} that is in turn raised to a power of 2. The newly introduced mathematical formulation of the gravity model was then

$$T_{ij} = A_i B_j O_i D_j f(d_{ij}) \tag{2.5}$$

where

$$A_i = [\sum_j B_j D_j f(d_{ij})]^{-1} \tag{2.6}$$

and

$$B_j = [\sum_i A_i O_i f(d_{ij})]^{-1} \tag{2.7}$$

This not only provided a more general integration of gravitational theory to transport models, but it also paved the way for the provision of a new at the time statistical framework, the main contribution of which was that the two models of gravity and intervening opportunities, initially thought as different, were statistically equivalent [Wil67].

2.1.4 Modern approaches to the modelling of human movement

Up until recently the empirical validation of the gravity or intervening opportunities theories was poor in terms of geographic and temporal scale. The lack of mobile devices that could help in tracking human movement at large scale had prevented progress in this direction. As mentioned first in the previous chapter, even when mobile phones became a mainstream communication tool the data was not openly available to research scientists, and thus, the problem persisted. Nevertheless, Brockmann et. al in [DBG06] proposed an alternative way to trace human mobility with high spatio-temporal accuracy. The rather creative idea employed by the authors involved the tracing of marked bank notes carried by humans. In particular, the trajectories of 464 thousand dollar bills were followed after approximately 1 million reports of their sightings were submitted on a purpose built website[1]. The principal assumption made by the authors was that dollar bill movement represents a convolution of human movements (people carry the bills across space) and, therefore, is expected to be governed by similar statistical laws.

For each pair of successive reportings at locations x_i and x_j for a dollar note, the authors measured the corresponding geographic distance $\Delta r = |x_j - x_i|$. Subsequently, the authors

[1] http://www.wheresgeorge.com/

measured the probability density of Δr, $P(\Delta r)$, and found that the following relationship presents a good approximation for the distribution of location datapoints:

$$P(\Delta r) \propto \Delta r^{-\beta} \tag{2.8}$$

with $\beta = 1.59 \pm 0.02$. It was thus hypothesised that the distribution of human displacements (movements) also followed a power-law distribution. A careful inspection of Equation 2.8 will help the reader understand that a possible interpretation is that $P(\Delta r)$ follows a gravity like law, where distance has a decaying effect, and the *masses*, or put differently, the trip volumes between origin and destination are assumed to be unitary. In the scope of the work of Brockmann *et al.* [DBG06], the latter presents a reasonable assumption given that there are no fixed sets of origins O and destinations D which correspond to different geographic zones as stated in the original theory of gravity in transportation literature. Instead, due to the spatial accuracy of dollar bill sightings which were recorded in terms of specific latitude and longitude coordinates, then the mass of a potential origin (resp. destination) can be implied to be equal to 1.0 and then it is assumed that only distance matters in movement.

Soon after the publication of [DBG06], the first large scale study to model the movement of humans carrying mobile phones was published [GHB08]. The statistical relationship described in Equation 2.8 for modelling human mobility was confirmed with a similar exponent measured now at $\beta = 1.75 \pm 0.15$. In the study, which involved a sample of approximately 16 million movements, an exponential cut-off was also proposed and Equation 2.8 was slightly modified in a new formulation:

$$P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} \exp(-\frac{\Delta r}{\kappa}) \tag{2.9}$$

where Δr_0 and κ where dataset specific parameters, in that case $\Delta r_0 = 1.5$ km and $\kappa = 400$ km.

Perhaps due to the fact that these works originated from the physics' community (that has traditionally favoured Newtonian formulations on movement), these findings have not been tested explicitly for an empirical proof of the theory of intervening opportunities. Most importantly, however, they have only modelled movements at long distances. We will see how the new generation of datasets from location-based social networks (to be presented immediately next) will help us not only to look at human movement in its most common habitat, the city, but also, how they may form the source of data for the development of novel mobile applications.

2.2 The importance of location-based social network data for human mobility

I now explain in more detail the principal attributes of location-based social networks, the novel characteristics of the datasets generated by their users and argue why they constitute a qualitatively different paradigm for research on human movement.

2.2.1 Social networks emerging through places

As mentioned earlier in Chapter 1, location-based social networks are, as a concept, a direct descendant of online social networks. While places are their focus point, an important element in these system is the user social network. Foursquare, to give an example, allows users to befriend each other through a number of channels. Besides those that are typically available in *traditional* social network platforms and involve the reception of email recommendations or search mechanisms over an online directory of users, Foursquare allows for users to meet and connect through places. More specifically, one can search for people who have checked in at a venue and subsequently make a request for an online connection with them.

The ability for Foursquare users to meet and connect through places presents an attractive case study for researchers who are interested in the interplay of human mobility and social interaction. Numerous models have recently appeared in the literature that aim to predict human movement through the exploitation of the social connectivity graph [SKB12], or inversely, models that take advantage of movement preferences to propose generative mechanisms for the formation of the social network. Some of these works [SNM11] have specifically targeted applications for location-based services, by exploiting, for instance, check-in patterns at places in order to recommend friends to LBSN users. While the Foursquare social network is not a core subject of study in the present dissertation, I will exploit it to recommend venues to Foursquare users in the two application scenarios that I consider in Chapter 4 and 5.

2.2.2 Why data from location-based social networks matters

The layers of information produced by these services form a core element of the present thesis and next I highlight, firstly, the properties that differentiate them from past datasets of human movement, and secondly, how these properties can shed light on previously inaccessible knowledge about human mobility. Whenever it will be convenient to refer to a specific location-based social network using a concrete example, the choice will be Foursquare. It is currently the most popular location-based service counting more than 35

million users [GNI13] and most relevant to this thesis as it constitutes its primary source of data.

- **The Check-in.** This is the single most central notion in location-based social networks. When a user is at a place (e.g. Coffee Shop) she can communicate her presence there by exploiting the *check-in* feature of the mobile social networking application. Typically the GPS sensor of the mobile phone is exploited to provide the service with information about her geographic position that is encoded via latitude and longitude coordinates. The application then automatically queries a database with millions of recorded venues from around the world and returns to the user a list of nearby places. Subsequently the user *checks in* at the venue where she currently is and can choose to share this information with her social network, push it on Twitter or Facebook, or privately store it in her Foursquare profile.

- **Venue Database.** What makes Foursquare places *special* are the multiple layers of information that are associated with them. As also discussed in Chapter 1 the venue database constitutes the *core wheel* of these services. Not only has knowledge become synchronously available about the *exact* places users go to, but also semantic information about their types (eg, Greek Restaurant or University Lab), user generated linguistic content such as tips, comments or tags and the list continues with multi-media content that includes photos and videos, together with geographic, temporal and social information about user check-ins and meetings. Today there are numerous mobile applications that exploit parts of this information through Foursquare's Venue API[2]. In terms of scientific research, the output that has been yielded is also of unprecedented volume [CRH11, SNLM11, NSMP11a, VRA+12, JTC12]. Besides static content, real time information is also available, such as the number of people who are checked in to the venue over the past hour (*hereNow*) or even special offers available for customers (*specials*). The latter also highlights how Foursquare's strategy of promotion also involves financial incentives to attract the participation of local businesses and their customers to the service. Overall, the service enumerates more than 50 million venues globally, which span geographically the majority of countries and continents around the world. This set is a result of a seed database that included hundreds of thousands of points of interest (POIs) when Foursquare was launched and, since then, it has been augmented through crowdsourcing as users add new venues every day.

- **Geographic Accuracy.** When a user checks in she declares her presence at a place. Location-based services provide for the first time the opportunity to record the geographic position of an individual not only with GPS accuracy (10-20 metres approximately), but also associate her with a specific, real world, venue. This

[2]https://developer.foursquare.com/

provides a clear advantage over cellular datasets where information about the accuracy of the position of the user is given with respect to the nearest BTS tower and there is no knowledge about the userséxact location. As noted previously, the latter is significant to the study of movement within cities, or even neighbourhoods of those, a perspective that could be previously provided only at a small scale through surveys [OW01]. As I elaborate in later chapters, this information about places, besides offering a granular geographic representation of human movement, it may be used as a proxy to infer the type of activity performed by the user (for instance having lunch at a restaurant) and this can be utilised in many related applications including mobile recommendations.

- **Global Accessibility.** The deployment of a mobile social networking application takes place on the web, thus *anyone* at *any place* in the world can access it given the presence of Internet connectivity. This is a remarkable paradigm shift in the way human mobility information can be acquired. With respect to surveys, which could theoretically provide a researcher with fine grained knowledge about the venue a user is visiting, location-based social networks allow for the collection of data that goes beyond the limitations posed by an experimental setting, both in terms of user participation and duration of the experiment. Moreover, the fact that Foursquare, and similar services, are globally accessible offers the opportunity to observe movement at long distances (for instance across countries). While this could also be possible with cellular data, I highlight that in principle mobile providers operate at a country level and if one would like to see movements beyond national borders an aggregation process should take place. Thus, location-based services offer an unprecedented global viewpoint for movement under a single service umbrella. This matter is especially relevant to the present dissertation as it will allow us to compare the movements of users across different cities in the world with datasets that originate from the same system. Further, the geographic scope of Foursquare's accessibility is such that it only allows one to observe movement over very long distances, but it also offers the opportunity to test the validity of proposed mobility models empirically at different geographic scales and areas. Thus, the universality of different theories or models, such as those presented in previous paragraphs, could be tested in various urban settings, countries or continents. In Figure 2.2, I depict the global spread of millions of Foursquare venues. Despite the bias of check-in activity towards certain geographic areas, which could be attributed to privacy and technological adoption variations, the high-resolution recovery of the Earth's land schematic pattern, through crowdsourced check-ins, is staggering.

- **Publicly Available.** Finally, a number of routes through which data sourced from location-based social networks is publicly available can be identified. Our source

in the present thesis is Twitter's Streaming API[3] since it represents the most effective way to access large amounts of data. An alternative route to acquire data is Foursquare's own API, yet the corresponding query limits yield much smaller datasets than Twitter. It is important to note that, in either case, check-ins that users have set as private are not available. The merits of public access to mobility data are two-fold; first, the data becomes available to academic researchers and therefore new analyses, techniques and models are likely to emerge (research innovation), and secondly, research output can be scrutinised by other researchers upon publication (reproducibility).

The novel characteristics for research in human mobility that have been brought by LBSN data cannot limited to the above, however. Besides user movements, expressed through check-ins, other types of data co-exist in these systems. To name a few important ones, the social network of users is also described in parallel to their movements - the opportunity to connect with already known friends or even meet new ones at the places they go to is provided by location-based services. Besides, as users check in to Foursquare venues they can express their views on this experience with comments pushed on Twitter. Further, an alternative source of linguistic content, popular in Foursquare, is *tips* as users can offer advice to future visitors about interesting things to do (or avoid) at a place. Finally, users express themselves using a host of features such as *likes*, photo sharing and videos that have been popularised, in recent years, through online social networking platforms. Therefore, location-based social networks are not only promising to help us understand human movement on its own, but they have also paved the way for the observation of mobility in conjunction with other important layers of data such as place semantics, natural language and the social network.

2.3 Place recommendations

The idea of predicting the whereabouts of mobile users has been around since the first datasets that describe human movement emerged. Numerous prediction frameworks [CSTC12, SMM+11, SK12] have appeared, the specifics of which depend on the types of mobility datasets that have been available. Up until recently the dominant datasets to study movement in computer sciences were either Bluetooth contacts, RFID or WiFi access point sightings characterised by limitations both in geographic scale and user participants. Indeed, the rise of ubiquitous computing in the past decade has given rise to many interesting quantitative studies based on human movement and interactions [EP06, CHC+07, HCY10], but the characteristics of the datasets generated in this context never escaped from the scaling constraints of the experimental/lab setting. Cel-

[3]https://dev.twitter.com/docs/streaming-apis

Figure 2.2: Geographic Representation of Foursquare Venues Around the Globe.

lular data has also been employed [SQBB10] to assess the predictability of human movements. The main finding of these works is that human movements present strong *periodicity* over time, since humans are very likely to return to previously visited locations and moreover, they are very likely to be at one of their *significant* spots (either *Home* or *Work*). The observation that human movement is indeed, to some extent, predictable, together with the mainstream appearance of GPS sensors on mobile phones led to the first works that specifically address the problem of *place recommendations* for mobile users.

This problem can be framed in the classical user-to-item recommender system terms; given a set of users and some information about their past location preferences the goal is to identify the set of places (items) they are more likely to enjoy and rank them accordingly. Ranking should take into account information about the mobility profiles of individual users. A toy example of the place recommendation concept is shown in Figure 2.3, where our imaginative Foursquare user has checked in to a flower shop and a set of places (right) need to be ranked and offered to this user as potential future destinations. Formally, in the place recommendation task, given a set of users U and a set of places L, the goal is to measure the relevance of a location $l \in L$ so that the places a user u_i is more likely to go are put at the top of the *recommendation list* $\hat{L}$. In the urban context, where thousands of potential candidate venues for recommendations may exist, this may prove to be a challenging task. I will analyze and discuss these challenges further when formulating different variations of the place recommendation problem in Chapters 4 and 5.

Figure 2.3: A Foursquare user has checked in at a Flower Shop (and potentially other places in the past). The goal then of a place recommendation algorithm is to recommend the most relevant venues in the system for the user in question. This is usually done by exploiting mobility, social and other information signals.

2.3.1 Human mobility and place recommendations

So far, from an abstract point of view, recommending places to mobile users appears to have similar aims to those of the two models, *gravity* and *intervening opportunities*, that were presented in Section 2.1. After all, in both human migration or urban transport and in the place recommendation task the main process corresponds to the prediction of human movement. As we shall see in the following chapters of this dissertation, analytical and modelling insights will be drawn from human mobility models to build effective place recommendation models. Yet, at this point, I list some important differences between the two tasks:

- **User Versus Area Centric View.** Recall that the principal aim of the gravity and intervening opportunity models, in the context of urban transport and human migration, was the prediction of volume of transitions between sets of origins O and destinations D. Origins and destinations usually correspond to geographic areas in a city or a country depending on the geographic scale being examined. On the other hand, in place recommendations the goal is to predict the whereabouts of a *specific* user. Two users can be in the same area, but they may still want to visit different places depending on their likes. *Personalisation* is a key concept in recommender systems and I will study it in detail in Chapters 4 and 5.

- **Serendipity.** The idea of offering novel items or content to users has been central to the recommender systems literature [HKTR04]. While the regularity in human movement allows for the development of prediction frameworks, user preferences or social connections could also be exploited to recommend new places for mobile users. Thus, the trade-off between *exploration* and *recurrence* is core in mobile recommendations. In contrast, the migration and transport literature features strictly

prediction oriented approaches, since their goal is to model flows of movement for the design of better urban services, infrastructure and policies.

- **Context and Dynamics.** Finally, the temporal scale of events in mobile recommendations is different to that of migration modelling. In the former, the dynamics of time matter a lot and the preferences of users about where they would like to go change every hour of a day or during the week. In the latter, most studies concern predictions with a temporal horizon on the order of months or years. The importance of temporal dynamics for recommendations have also brought the notion of *context* into the spotlight of recommender systems research [ASST05]. Context is a very broad term and is discipline specific; in mobile recommendations refers usually to information about the current geographic position of the user, the time recommendations are performed or even the places that are popular nearby.

These differences are highlighted as they have a direct impact on the choice of information and methodologies (statistical frameworks, prediction models, evaluation metrics, etc.) that handle it when solving either problem. This will be reflected by the three contribution chapters of the thesis as both the general problem of human mobility and mobile venue recommendations will be addressed. Next, I present the composition of this dissertation and its future outlook.

2.4 Present dissertation and future outlook

In this chapter I have reviewed two classical modelling paradigms from the human migration and transportation literature. The *gravity* and *intervening opportunities* models. As indicated in Paragraph 2.1.4, recent studies suggest the existence of a universal power-law distribution $P(\Delta r) \sim \Delta r^{-\beta}$, observed, for instance, in cell tower data concerning humans carrying mobile phones $\beta = 1.75$ [GHB08] or in the movements of "Where is George" dollar bills $\beta = 1.59$ [DBG06]. This universality is, however, in contradiction with observations that displacements strongly depend on where they take place. For instance, a study of hundreds of thousands of cell phones in Los Angeles and New York demonstrate different characteristic trip lengths in the two cities [SI10]. This observation suggests either the absence of universal patterns in human mobility or the fact that physical distance is not a proper variable to express it. These issues will be revisited again in Chapter 3, where we compare again the efficacy of these classical models of human movement for datasets sourced from location-based social networks and assess their relevance with regards to mobility in urban environments.

As pointed in Section 2.2.2, these data allow for the study of human mobility at an unprecedented scale and with fine geographic representations. This is achieved through GPS sensors offering accuracy of a few tens of metres in movement records and thanks to

the Internet reach of location-based services. These advancements also effectively provide the opportunity to study mobility in cities through a single service (e.g., Foursquare) and overcome measurement biases that could be induced by variations in the experimental context (for instance if data from different sources were to be aggregated).

Further, an important application scenario in the context of location-based services that has been identified in this chapter is that of mobile place recommendations. As noted in Section 2.3.1, human movement and mobile recommendation services are strongly influenced by the spatio-temporal dynamics of user behaviour. In Chapter 4 I will build a supervised learning framework whose aim will be the prediction of user whereabouts in real time by combining various sources of contextual information available in location-based social networks. In fact, we will see how the information signals in these systems can be exploited not only to balance out the disadvantages of sparse user representations, but also to allow for the development of models that are able to both predict *historically* visited venues and to recommend *new* venues to mobile users.

Finally, in Chapter 5, I will tackle an alternative place recommendation scenario where our goal is the prediction of the *new venues* visited by mobile users. With an eye towards applications such as digital urban exploration and local search, I will consider a family of recommender system algorithms that has been applied mainly in the online web domain and assess their performance when deployed geographically. From a computer science perspective it is important to understand whether filtering algorithms are resilient upon their migration to location-based services and systems, where factors such as the effect geographic distance come into play.

3

Modelling human mobility in urban spaces

As demonstrated in Chapter 2, location-based social networks allow for the observation of mobile user movement with granular geographic representations and with global scope. The former property will enable us to inspect mobility within a given urban environment, while the latter will allow us to perform a comparison across different cities.

The initial focus in the present chapter will be on the properties of collective user movements, such as for instance, the geographic distances observed in the journey trips of users. Analysing and modelling movement at this level of abstraction, can foster a deeper understanding of the fundamental processes that drive the mobility of users and, as a consequence, it can help the design of appropriate application frameworks such as mobile place recommendations that we investigate in Chapters 4 and 5 of this dissertation.

Movement of people in space has been an active subject of research in the social and geographical sciences. It has been shown in almost every quantitative study and described in a broad range of models that a close relationship exists between mobility and distance. People do not move randomly in space, but instead human movements exhibit high levels of regularity and tend to be hindered by geographical distance [SQBB10]. The origin of this dependence of mobility on distance, and the formulation of quantitative laws explaining human mobility remains, however, an open question especially in the context of urban mobility in cities, where the availability of large scale data has been scarce.

The answer to this could lead to many applications beyond the scope of location-based recommendations [ZZXY10, DQC10, SS11] that we examine here, and improve engineered systems such as cloud computing, enhance research in social networks [JPO06, DC10,

SNLM11, ECL11] and yield insight into a variety of important societal issues, such as urban planning and epidemiology [NN08, LHG04, CBB+07].

Chapter Outline. In this chapter, we focus on the analysis and modelling of human mobility patterns in a large number of cities across the world. More precisely, we make the following contributions:

- **A premier empirical study of human mobility across multiple urban centres around the globe.** We aim to answer the following question: "Do people move in a substantially varied way in different cities or, rather, do movements exhibit universal traits across different urban centres?". To do so, we take advantage of the advent of mobile location-based social services accessed via GPS-enabled smartphones, for which fine granularity data about human movements is becoming available. Moreover, the worldwide adoption of these tools implies that the scale of the datasets is planetary. Exploiting data collected from public *check-ins* made by users of the most popular location-based social network, Foursquare [foua], we study the movements of 925,030 users around the globe over a period of about six months, and study the movements across 5 million places in 34 metropolitan cities that span four continents and eleven countries.

- **An identification of the key role of density in human movement as dictated by the law of intervening opportunities.** In Section 3.1, we discuss how at larger distances we are able to reproduce previous results of [GHB08] and [DBG06] with the aim being the empirical validation of the relevance of data sourced from location-based services to study human movement. Subsequently, in Section 3.2 we focus on urban mobility. We first confirm that mobility, when measured as a function of distance, does not exhibit universal patterns. The striking element of our analysis arises in Section 3.3, where we observe a universal behaviour in all cities when mobility is measured with a different variable to that of geographic distance. We discover that the probability of transiting from one place to another is inversely proportional to a power of their *rank*, that is, the number of intervening opportunities between them. This universality is remarkable as it is observed despite cultural, organisational and national differences. This finding is in agreement with the social networking parallel that suggests that the probability of a friendship between two individuals is inversely proportional to the number of friends between them [DLN05], and depends only indirectly on physical distance.

- **A modelling driven empirical comparison of the intervening opportunities and gravity theories in human urban movement.** Driven by our analytical findings, in Section 3.4, we propose the use of *rank-distance* as a core element for modelling urban movement. We then provide an empirical comparison between the

rank-distance model and a gravity variant. By using only information about the distribution of places in a city as input and by coupling this with a rank-based mobility preference we are able to reproduce the actual distribution of movements observed in real data. Overall, our analysis is in favour of the concept of intervening opportunities rather than gravity models, thus suggesting that trip making is not explicitly dependent on physical distance but on the accessibility of resources satisfying the objective of the trip. Individuals thus differ from random walkers in exploring physical space because of the motives driving their mobility. Further, while our results do not exclude the possibility for the development of gravity models that achieve good fits of urban mobility, we argue that the configuration of parameters in their context is more complex. Finally, in Section 3.5 we quantify the role of geography in human mobility in the light of our observations that at the level of abstraction of aggregate trip distance distribution, the spatial distribution of places it is the primary source of mobility variations observed across cities.

We close this chapter by presenting in Section 3.4 a detailed mathematical explanation of the rank-distance model, and in Sections 3.6 and 3.7 the implications of our findings and related work, respectively, is discussed. In Section 3.8 we summarise our findings.

3.1 Urban movements analysis

In this section we introduce the collection methodology and the properties of the mobility dataset that we employ in the present and the following chapters of the thesis (unless specified). Subsequently, in Section 3.1.2, we perform a statistical validation of the collected data with respect to well established previous work on large scale movement analysis. While our findings are in line with past observations of human mobility over large distances, this confirmation does not hold for shorter movements within cities. Motivated by the latter, we present an in-depth analysis of movements in urban environments in the sections to follow.

3.1.1 Dataset Description.

We draw our analysis upon a dataset collected from the largest location-based Social Network, Foursquare [foua]. The dataset features $35,289,629$ movements of $925,030$ users across $4,960,496$ places collected during 5 months (May 27th to November 2nd 2010). We estimate that this sample contains approximately 20% to 25% of the entire Foursquare user base at the time of collection [1]. Foursquare places or venues are Web 2.0 geo-tagged entities that match a real venue observed in the physical world and here are represented

[1] `http://mashable.com/2010/08/29/foursquare-3-million-users/`

through GPS indicated longitude and latitude geographic coordinates. In this context a movement is the indication of presence at a place that a user gives through the Foursquare system. In the present work we focus on the 34 most active cities in terms of check-in numbers in the dataset. We have matched a Foursquare venue to a city by utilising *locality information* available for all Foursquare venues. The reader can view summary statistics for all cities we have experimented with in Table 3.1. The mobility dataset analysed in the following paragraphs is comprised of *check-ins* made by Foursquare users and became publicly available through Twitter's Streaming API. The collection process lasted from the 27th of May 2010 until the 3rd of November of the same year. By considering only consecutive *check-ins* that take place within the same city we have extracted almost 10 million *intra-city* movements analysed in Figure 3.2. Detailed statistics including the number of check-ins and venues in each city can be found in Table 3.1.

Place Semantics, Time and the Social Network. In addition to the information about user check-in movements and the GPS positioning of Foursquare places, in later chapters, as we tune the level of abstraction of our analysis to fit application scenarios, we will make use of metadata about places in location-based social networks. Specifically, we will exploit semantic information about the categories of places in Foursquare. In [foub] one can inspect the more than 300 different types of places available in the system to characterise a venue. This data will be used, for instance, as a proxy to model the preferences of mobile users in terms of activities they like to perform in the city. Further, in Chapter 4, as we examine real time aspects of user mobility we regard also the check-in times of user movements that are available with per second granularity.

As with user check-ins, the collection of social network data was via Twitter. For every user we requested her list of followers. For every reciprocal follower relationship we then generated an edge between two Foursquare users in the dataset. This process yielded a graph whose giant connected component is comprised $591,146$ nodes and almost $10,573,803$ million edges between those resulting in a rather dense network. The social network formed by Foursquare users by observing their relationship on Twitter does not correspond to the original social network in the actual Foursquare platform. At the time of data collection, Foursquare did not offer the option to crawl the social network formed by its user base via its API. Thus, we had to resort to a publicly available version of a social network for the group of users whose behaviour we explore here. In previous work it has been shown how the statistical properties of this social network match those of other location-based social networks [SNLM11].

Considerations on Data Generation and Collection Biases. Before we proceed with presenting the findings of this chapter, and the thesis, in general, we provide a discussion on potential sources of bias that may associate with datasets collected from location-based social networks. First, the demographics of the users engaging with location-based

social networks are not expected to be representative of the world's human population. This is an aspect that influences much of the data collected in the context of scientific experiments as discussed in detail in [HHN10], where the case of WEIRD datasets (participants are of Western, Educated, Industrialized, Rich and Democratic origins) is brought forward. Such an effect may be more intense in the present datasets given that the check-in data is crowdsourced and location-based social networks are at an early adoption stage. Second, the way in which users check-in can be influenced by numerous factors such as social, privacy and cognitives ones. In plain words, a person may go somewhere and not check in. Despite the fact that there is lack of established scientific conclusions with respect to the factors that motivate (resp. demotivate) users from checking in, there is a line of work that attempts to address these issues specifically [LCW+11, CRH11]. It should be noted that these sources of bias, or similar ones, influence many datasets that are being used to explore human movement. While the analysis in the sections to follow reveals interesting aspects that are related to human movement, the cautious reader should always recall that any dataset used as a proxy to human movement is likely to deviate from the description of the actual human mobility. Last but not least, the fact the Foursquare data was acquired via Twitter's streaming API effectively means that we have only a limited window onto Foursquare user activity; that is, we have access to public user check-ins pushed on Twitter. Despite that this may constitute an additional source of bias, the fact that in Chapter 5 we make similar observations between the Foursquare check-in dataset and the full snapshot of the Gowalla service is encouraging with respect to the conclusions derived by the experimental analysis to follow.

Table 3.1: **Summary of city statistics**

City Name	**Movements**	**Places**	**Places/km²**	**Area (km²)**	$< \Delta r >$ **(km)**
Amsterdam	32934	8847	275.61	21.63	2.29
Atlanta	63220	10090	214.72	19.94	5.37
Austin	60296	9492	199.32	14.06	5.82
Bangkok	45860	7574	248.32	10.81	3.97
Boston	42196	6795	366.94	13.25	1.57
Chicago	185496	23050	315.16	41.94	4.02
Columbus	32388	7463	181.18	8.88	5.42
Dallas	39380	8177	200.8	13.06	5.21
Denver	30695	6123	215.26	12.81	4.67
Houston	47996	11808	168.68	14.63	7.57
Indianapolis	30382	6417	213.02	5.38	6.99
Kuala Lumpur	62595	14223	268.44	30.88	3.18
Las Vegas	82437	11910	260.39	16.63	4.76
London	62837	15760	290.92	30.5	3.32
Los Angeles	86092	18508	220.92	31.5	4.86
Milwaukee	38697	5318	218.77	9.56	3.15
Minneapolis	29572	5482	228.04	11.13	3.1
New York	371502	43681	715.02	58.0	2.24
Orlando	37783	8060	224.56	8.88	5.44
Paris	38392	12648	261.98	35.94	2.77
Philadelphia	54545	10270	293.2	17.31	2.86
Phoenix	34436	8689	183.1	9.44	6.27
Portland	38409	8413	238.34	15.63	3.08
Rio de Janeiro	25808	6788	248.2	12.31	5.99
San Antonio	33516	8237	144.17	6.0	8.35
San Diego	69152	13365	227.26	22.38	5.7
San Francisco	112168	15970	377.64	32.25	2.36
Santiago	56743	10636	235.17	20.69	4.94
Seattle	66423	10410	294.6	20.75	3.61
Seoul	44303	9271	250.76	18.31	4.8
Singapore	79624	15617	316.67	21.31	5.26
São Paulo	52855	14291	224.68	32.56	4.31
Toronto	77548	13870	322.26	24.81	3.59
Washington	71557	10279	325.11	21.31	1.92

3.1.2 Urban Movements and Power-laws.

In order to confirm the large scale results reported in [GHB08, DBG06], we have computed the distribution of human displacements in our dataset (Figure 3.1) by measuring the geographic distance between the consecutive check-ins of a Foursquare users in the dataset: we observe that the distribution is well approximated by a power law with exponent $\beta = 1.50$ and a threshold $\Delta r_0 = 2.87$ ($p - value = 0.494$). This is almost identical to the value of the exponent calculated for the dollar bills movement ($\beta = 1.59$) [DBG06] and very close to the 1.75 estimated from cellphones calls analysis of human mobility [GHB08]. With respect to these datasets, we note that the Foursquare dataset is planetary, as it contains movements at distances up to 20,000 kilometres (we measure all distances using the great-circle distance between points on the planet [2]).

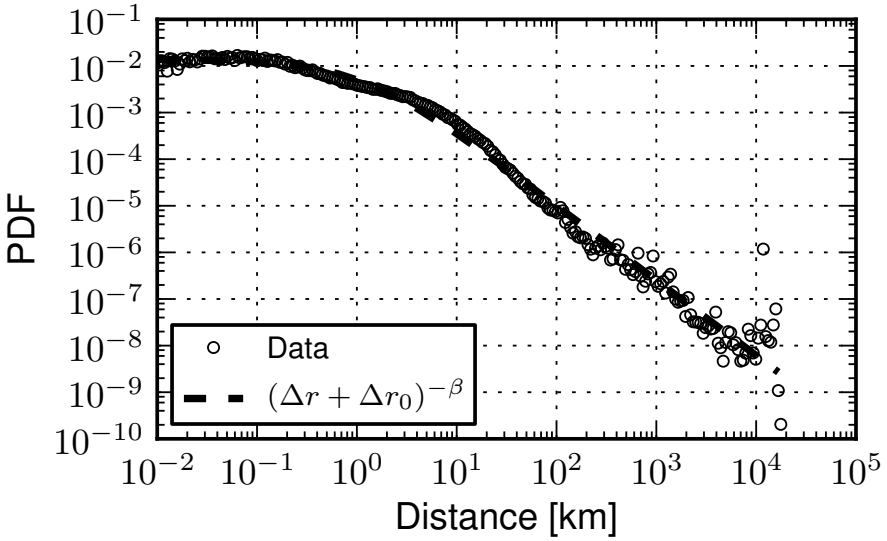


Figure 3.1: **Global movements.** The probability density function (PDF) of human displacements as seen through 35 million location broadcasts (check-ins) across the planet. The power-law fit features an exponent $\beta = 1.50$ and a threshold $\Delta r_0 = 2.87$ confirming previous works on human mobility data. The spatial granularity offered by GPS data allows for the inspection of human movements at very small distances, while the global reach of Foursquare reveals the full tail of the planetary distribution of human movements.

At the other extreme, small distances of the order of tens of metres can also be tracked in the dataset thanks to the fine granularity of GPS technology employed by mobile phones running these geographic social network applications. Indeed, we find that the probability of moving up to 100 metres is uniform, a trend that has also been shown in [DBG06] for a distance threshold Δr_{min}. Each transition in the dataset happens between two well

[2] http://en.wikipedia.org/wiki/Great-circle_distance

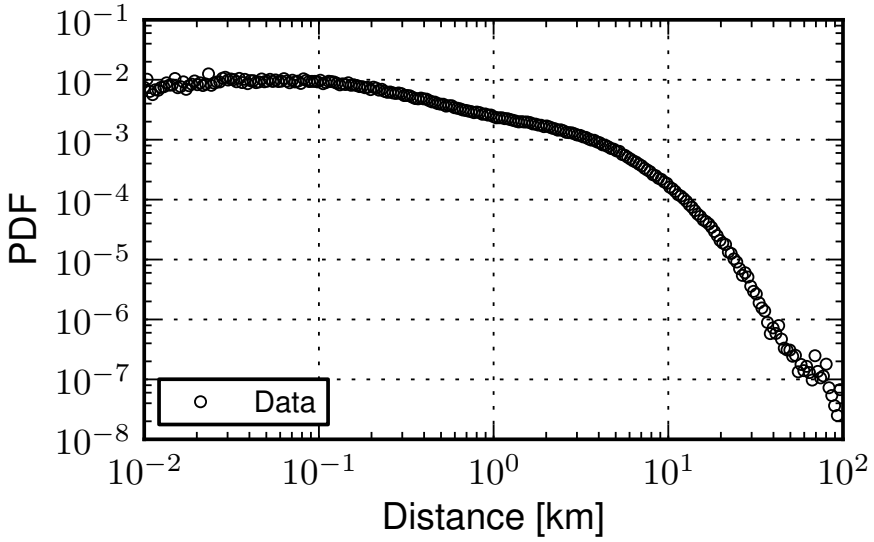


Figure 3.2: **Urban movements.** The probability density function (PDF) of human displacements in cities (intra-city). For two successive location broadcasts (check-ins) a sample is included if the locations involved in the transition belong to the same city. Approximately 10 million of those transitions have been measured. The poor power-law fit of the data ($\beta = 4.67$, $\Delta r_0 = 18.42$) suggests that the distribution of intra-city displacements can not be fully described by a power law. Short transitions, which account for a large portion of the distribution, are not captured by such process.

defined venues, with data specifying the city they belong to. We exploit this information to define when a transition is urban, that is, when both start and end points are located within the same city. Figure 3.2 depicts the probability density function of the about 10 million displacements within cities across the globe. We note that a power-law fit does not accurately capture the distribution. First of all, a large fraction of the distribution exhibits an initial flat trend; then, only for values larger than 10 km the tail of distribution decays, albeit with a very large exponent which does not suggest a power-law tail. Overall, power-laws tend to be captured across many orders of magnitude, and this is not true in the case of urban movements. The estimated parameter values via maximum-likelihood [CSN09] are $\Delta r_0 = 18.42$ and exponent $\beta = 4.67$ ($p - value = 1.0$).

3.2 Comparing human movements across cities

Since the distribution of urban human movements cannot be approximated by a power law distribution nor by a physically relevant functional relation, how can we represent displacements of people in a city more appropriately? We start by comparing human movements across different cities. In Figure 3.3, we plot the distribution of human displacements for Houston, San Francisco and Singapore, noting that similar patterns have been observed across all cases we have considered in the experiments. The shapes of the

distributions, while different, exhibit similarities suggesting the existence of a common underlying process that seems to characterise human movements in urban environments. There is an almost uniform probability of travelling in the first 100 metres, that is followed by a decreasing trend between 100 metres and a distance threshold $\delta_m \in [5, 30]$ km, where we detect an abrupt cutoff in the probability of observing a human transition. The threshold δ_m could be due to the reach of the *borders* of a city, where maximum distances emerge. While the distributions exhibit similar trends in different cities, scales and functional relation may differ, thus suggesting that human mobility varies from city to city. For example, while comparing Houston and San Francisco (see Figure 3.3), different thresholds δ_m are observed. Moreover, the probability densities can vary across distance ranges. For instance, it is more probable to have a transition in the range 300 metres and 5 kilometres in San Francisco than in Singapore, but the opposite is true beyond 5 kilometres. This difference could be attributed to many potential factors, ranging from geographic ones such as area size, or density of a city, to differences in infrastructures such as transportation and services or even socio-cultural variations across cities. In the following paragraphs we present a formal analysis that allows to dissect these heterogeneities.

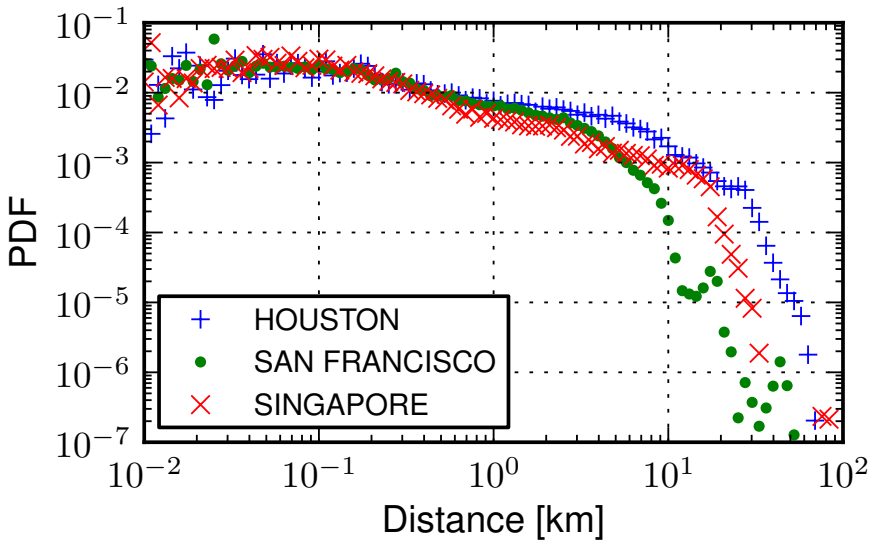


Figure 3.3: **Urban movement heterogeneities.** The probability density function (PDF) of human displacements in three cities: Houston, San Francisco and Singapore (for 47, 112 and 79 thousand transitions, respectively). Common trends are observed, e.g., the probability of a jump steadily decreases after the distance threshold of 100 metres, but the shapes of the distributions vary from city to city, suggesting either that human movements do not exhibit universal patterns across cities or that distance is not the appropriate variable to model them.

3.3 The importance of place density

Inspired by Stouffer's theory of intervening opportunities [Sto40] that as we recall, from its introduction in Section 2.1.2, suggests that *the number of persons travelling a given distance is directly proportional to the number of opportunities at that distance and inversely proportional to the number of intervening opportunities*, we explore to what extent the density of places in a city is related to the human displacements within it.

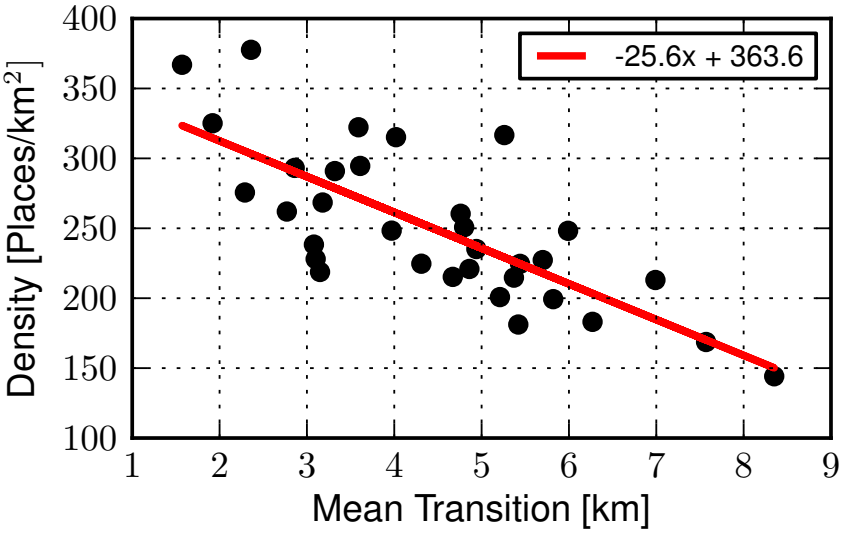


Figure 3.4: **City place densities and mean movement lengths.** Scatter plot of the density of a city, defined as the number of places per square kilometer, versus its mean human transition in kilometres. Each datapoint corresponds to a city, while the red line is a fit that highlights the relationship of the two variables ($R^2 = 0.59$). A longer mean transition corresponds to the expectation of a sparser urban environment, indicating that the number of available places per area unit could have an impact on human urban travel.

First, we define the density of a city in the Foursquare dataset by applying a grid onto each city using squares of area size equal to 0.25 km^2 and filtering out those grid areas that feature less than five Foursquare venues. Then the density is equal to the number of places per square km^2 averaged across the grid. As a next step, we plot the place density of a city, as computed with our check-in data, against the average distance of displacements observed in a number of cities. In Figure 3.4 one observes that the average distance of human movements is *inversely proportional* to the city's density. Hence, in a very dense metropolis, like New York, there is a higher expectation of shorter movements. We have measured a coefficient of determination $R^2 = 0.59$ [Nag91]. Intuitively, this correlation suggests that while distance is a cost factor taken into account by humans, the range of available places at a given distance is also important. In this context, the geographic deployment of places could be thought of as a proxy for the distribution of resources geographically. In a scarce resource environment one may need to travel longer

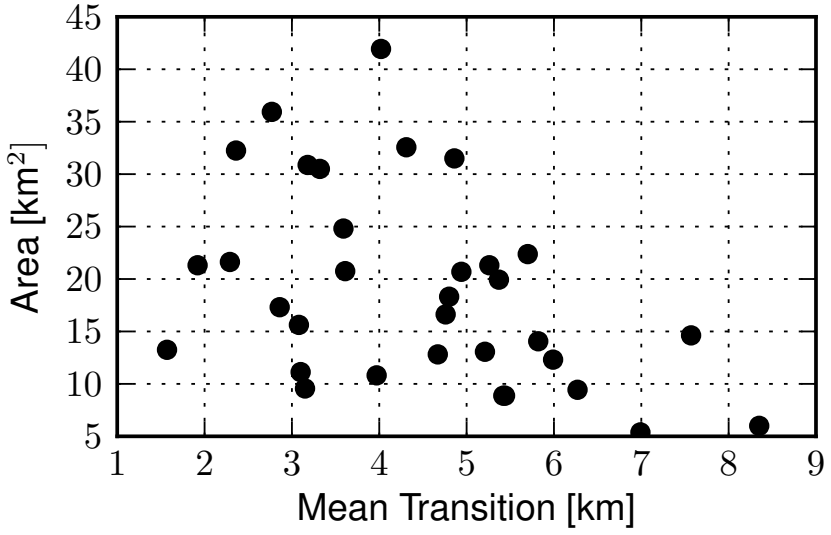


Figure 3.5: **City area sizes and mean movement lengths.** Scatter plot of the area of a city, measured in square kilometres, versus its mean human transition in kilometres. Unlike place density, the area of a city does not seem strongly related to the mean length of its transitions ($R^2 = 0.19$). To measure the area of a city we have segmented the spatial plane around its geographic midpoint in squares of size 250×250 m^2. The area of a city has been defined as the sum area of all squares that feature at least five places.

distances to satisfy a need; it may take a few kilometres to get a sandwich in the desert, yet in the example of the dense metropolis mentioned above hundreds of restaurants may be packed in a small piece of land and, as a result, one would need to travel a few tens of metres to get food. Nevertheless, one could doubt these arguments and suggest that instead it is the area size of a city that is important in movement. After all, a larger city would allow, by definition, for the observation of longer trips. Hence, as a next step, we explore whether the geographic area size covered by a city affects human mobility by plotting the average transition in a city versus its area size (see Figure 3.5). Our data indicates no apparent linear relationship, with a low correlation $R^2 = 0.19$, thus indicating that density is a more informative measure. To shed further light on the hypothesis that density is a decisive factor in human mobility, for every movement between a pair of places in a city we sample its the rank value. The rank for each transition between two places u and v is the number of places that are closer in terms of distance to u than v is. We account for every place w that satisfies that condition and formally we have

$$rank_u(v) = |\{w : d(u, w) < d(u, v)\}|.$$

The rank between two places has the important property of being invariant in scaled versions of a city, where the relative positions of the places are preserved but the absolute distances dilated. In Figure 3.6 we plot, for the three cities, the rank values observed for each displacement. The fit of the rank densities onto a log-log plot shows that the rank

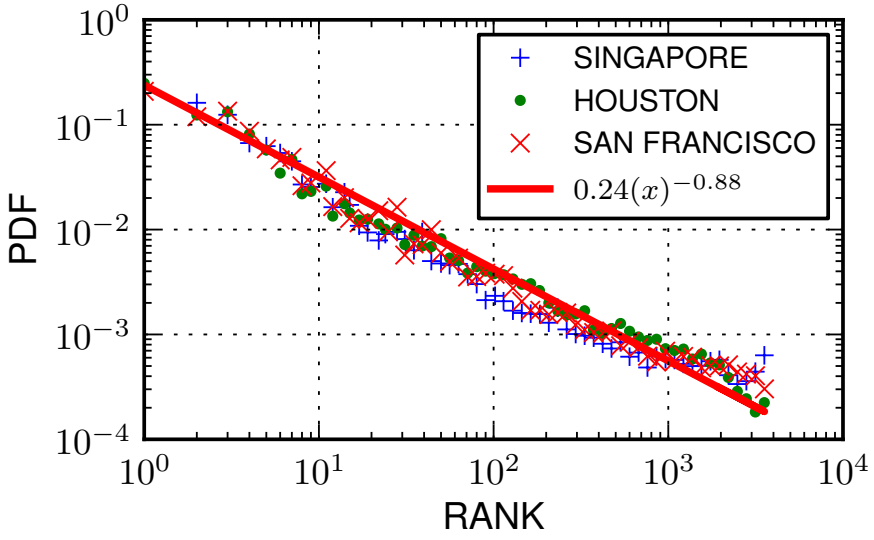


Figure 3.6: **Rank distributions in three cities.** We observe that the distributions of the three cities collapse to a single line, which suggests that universal laws can be formulated in terms of the rank variable. The observation confirms the hypothesis that human movements are driven by the density of the geographic environment rather than the exact distance cost of our travels. A least squares fit (red line) underlines the decreasing trend of the probability of a transition as the rank of a places increases.

distribution follows a linear trend similar to that of a power-law distribution[3].

This observation suggests that the probability of moving to a place decreases when the number of places nearer than a potential destination increases. Moreover, the ranks of all cities collapse on the same line despite the variations in the probability densities of human displacements. Using a *least squares* error optimisation method [LH74], we have fitted the rank distribution for the thirty-four cities under investigation and have measured an exponent $\alpha = 0.84 \pm 0.07$. This is indicative of a universal pattern across cities where density of settlements is the driving factor of human mobility. We superimpose the distribution of ranks for all cities in Figure 3.7.

Interestingly enough, a parallel of this finding can be drawn with the results in [DLN05], where it is found that the probability of observing a user's friend at a certain distance in a geographic social network is inversely proportional to the number of people geographically closer to the user.

3.4 Modelling urban mobility

The universal mobility behaviour emerging across cities, shown in Figures 3.6 and 3.7 where we have plotted the probability distributions of the rank value for each transition,

[3]Strictly speaking a power-law is not well defined for exponent regimes α smaller than 1.

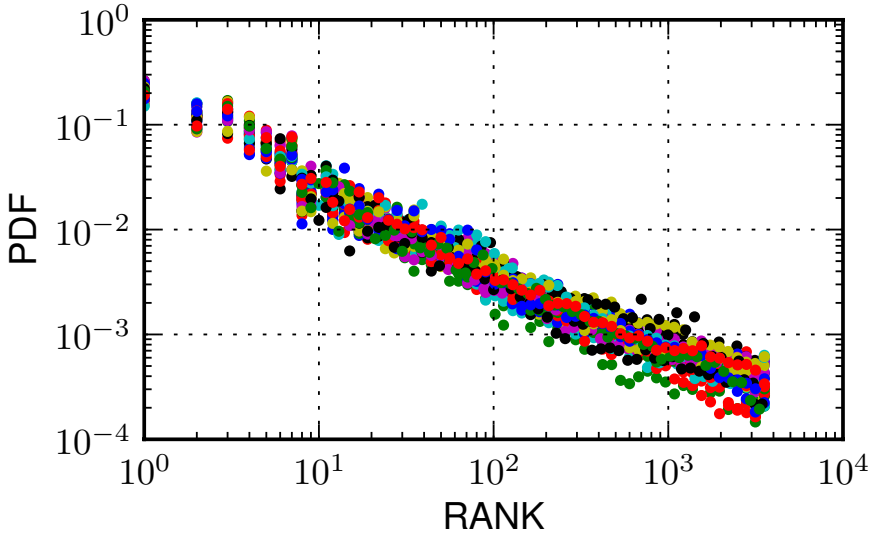


Figure 3.7: **Rank distributions in urban environments.** Superimposition of the probability density functions (PDF) of rank values the thirty-four cities analysed in the Foursquare dataset. A decreasing trend for the probability of a jump at a place as its rank value increases is common. The trend remains stable despite the large number of plotted cities and their potential differences with respect to a number of variables such us number of places, number of displacements, area size, density or other cultural, national or organisational ones.

paves the way for a new model of movement in urban environments. If the *rank-distance* is a better variable than pure geographic distance in terms of expressing movement in cities in a manner that effectively reduces variations in measurements across them, then we may as well integrate it into a human mobility model and empirically verify its efficacy using Foursquare data. This follows closely the spirit of Samuel Stouffer who validated statistically the theory of intervening opportunities in the city of Cleveland as we discussed in Section 2.1.2.

Formalising urban mobility models. We formalise this next, where given a set of places $\mathcal{U}$ in a city, the probability of moving from place $u \in \mathcal{U}$ to a place $v \in \mathcal{U}$ is defined as

$$P_r[u \to v] \propto \frac{1}{rank_u(v)^a}$$

where

$$rank_u(v) = |\{w : d(u,w) < d(u,v)\}|.$$

In addition to the *rank-distance* model presented above, we have adopted a gravity-based model of human urban movement in order to perform a direct comparison with the alternative popular theory where movement depends on the absolute *geographic distance*

between places in cities introduced previously in Section 2.1.1. In this context such model should incorporate two factors. On one hand, the deterring effect of distance on movement, and on the other hand, the attractiveness of places due to a gravitational force. The former factor is captured by measuring the geographic distance, $d(u, v)$, between two places u and v. To quantify the *gravitational mass* of a place u, we measure the number of nearby venues assuming that the denser the area that surrounds a place, the higher its attractiveness. This has required the use of an additional parameter r_u, which corresponds to the radius of the disc centred on the geographic position of place u. We can now define the mass m_u of u, simply by enumerating the number of places that fall within the disc's surface. The probability of a transition between two places u and v in the gravity-based model is set to be proportional to the product of the places' masses and inversely proportional to their geographic distance. Formally:

$$P_g[u \to v] \propto \frac{m_u.m_v}{d(u,v)^b}$$

Agent-based simulations. We run agent based simulation experiments where agents transit from one place to another according to the probabilities defined by the two models, respectively. Averaging the output of the probability of movements by considering all possible places of a city as potential starting points for our agents, we present the human displacements resulting from the models in Figure 3.8: as shown, despite the simplicity of the rank model, this is able to capture with very high accuracy the real human displacements in a city. It is interesting to note that the model is able to reproduce even minor anomalies, such as the case of San Francisco where we have 'jumps' in the probability of a movement at 20 and 40 kilometres. In contrast, the gravity model does not offer a precise fit, since small distances are overestimated. A potential explanation for this behaviour could be given by the fact that in urban environments most places are positioned in a central, highly dense, core of a city. In this case, not rare in an urban context, the probability of a transition to a nearby place may rise dramatically when considering a gravity model, as density reaches a maximum and geographic distances are minimised.

Model parameterisation. Besides comparing the performance of the two models in the task of fitting the empirical distributions of human movement, it is worth discussing their parameterisation too. In the case of the *rank-distance* model, a common parameter $\alpha = 0.84$ has been set for the simulations of all cities. That is the empirical average observed by fitting the distributions of the rank values observed in cities as depicted in Figure 3.7. Given the small standard deviations observed across cities, it would be sufficient to observe movements in one city to fit accurately the transitions of others, provided we have knowledge of the geographic position of their venues. On the other hand, the identification of the parameters for the gravity model is a more complex task. Initially, we had to choose a radius r_u to define the mass m_u of a place u. While this

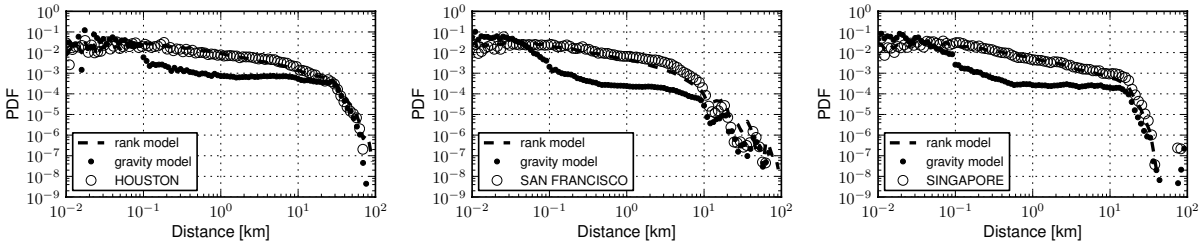


Figure 3.8: **Fitting urban movements.** Probability Density Functions (PDF) of human movements and corresponding fits with the rank-distance and gravity models in three cities (Houston, San Francisco and Singapore). In the rank-distance model the probability of transiting from a place u to a place v in a city, only depends on the rank value of v with respect to u. In the case of the gravity model, the deterrence effect of distance is co-integrated with a mass based attractiveness of a place u. The associated mass, m_u, has been defined according to the number of neighbouring places.

would have been easier to perform if we were considering movement across countries, or across cities, by considering for instance the size of their populations, it is much harder to define a similar geographic or organisational scope within a city. In our experiments we tested exhaustively r_u values ranging from 0.1 to 1 kilometres (the parameters for the depicted fit of the gravity model are $b = 1.0$ and $r_u = 100$ metres). Equally, selecting an exponent b to control the effect of distance in movements required again a brute-force exploration of values (we have experimented for values within the range 0.5 to 2.5). We note that our aim is not to exclude the possibility that more complex gravity models could be devised achieving potentially better fits of urban movement. In the light of the evidence that our experiments have provided, the use of a rank-distance variable qualifies better for divising a universal urban mobility model. Moreover, it is worth noting that the rank model does not take into account other parameters such as individual heterogeneity patterns [GHB08] or temporal ones [DBG06] that have been studied in the past in the context of human mobility and yet it offers very accurate matching of the human traces of our dataset. Plots with the performance of the models for all thirty-four cities that we have evaluated can be found in Figure 3.9.

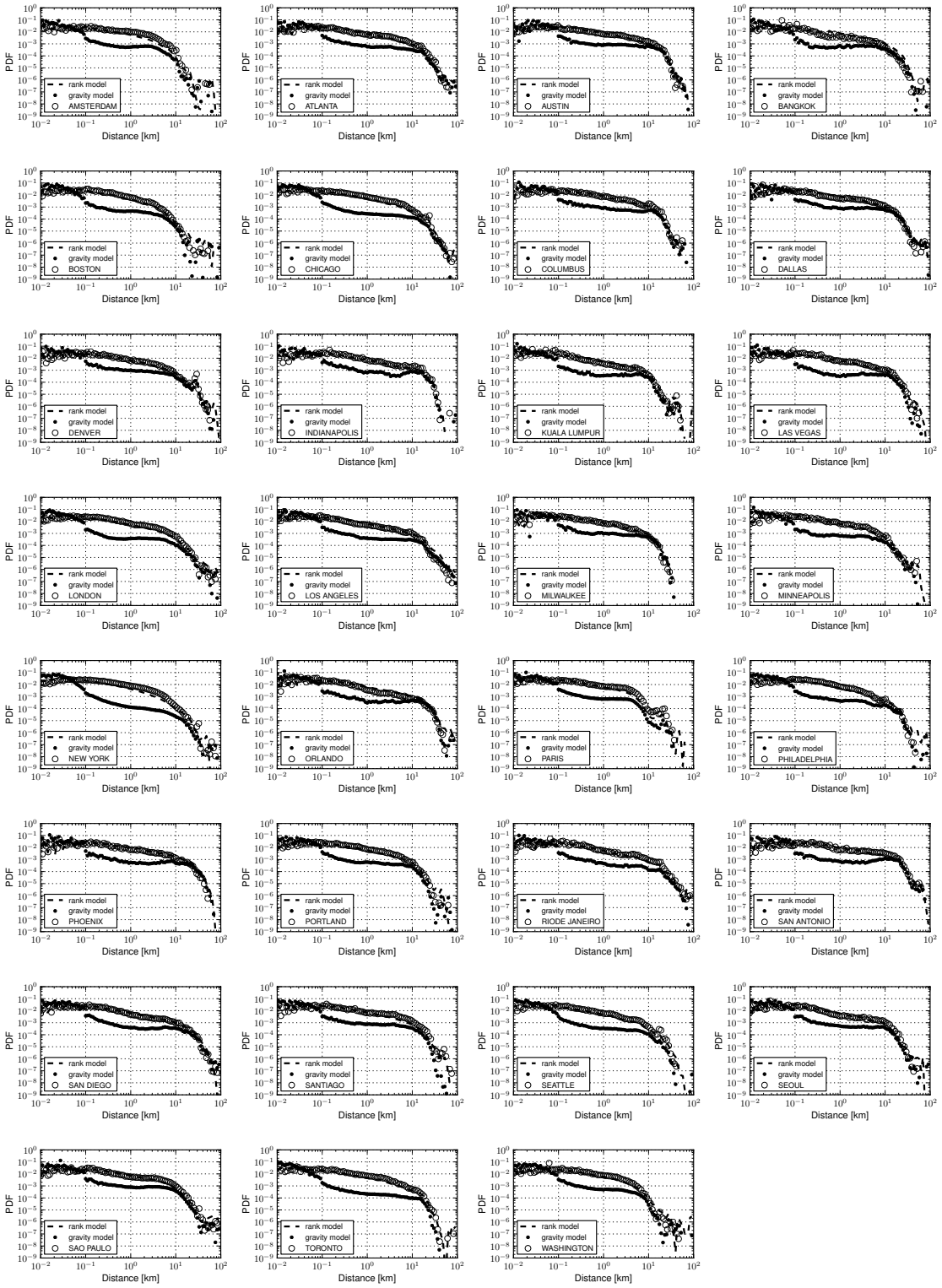


Figure 3.9: **Fitting urban movements for all cities in the Foursquare dataset.** The dominance of the rank-distance model over the gravity case extends to the rest of the cities (34 in total) we have experimented with in the Foursquare dataset. The results depicted here correspond to the gravity model with parameters $b = 1.0$ and $r_u = 100$ metres, and in the case of the rank-distance model an exponent $a = 0.84$ has been used to simulate movement in all cities and corresponds to the empirical average of the exponents resulting from the fit of the rank value distributions.

Detailed Description of Rank Model's Computation. We now describe the rank-based model we have devised with the aim to fit human movements. Our aim is to calculate the displacement probability distribution over a given city, which is described by a set of M places $\mathcal{U} = \{u_1, u_2....u_M\}$. We measure the pairwise transition probability from a starting place $u \in \mathcal{U}$ to a destination place $v \in \mathcal{U}$ as

$$P_{uv} = \frac{\text{rank}_u(v)^{-\alpha}}{\sum_{u \in \mathcal{U}} \text{rank}_u(v)^{-\alpha}}$$

where, recall that $\text{rank}_u(v) = |\{w \in \mathcal{U} : d(u,w) < d(u,v)\}|$ and we use the convention that $\text{rank}_u(u) = 0$ for every $u \in \mathcal{U}$. The above configuration takes into account all places in the city away from u and suggests a probabilistic setting that the sum of the probabilities of transition to any destination place is equal to 1.

Elaborating further, we define the probability of observing a movement of length Δr away from an initial place u as

$$P_u(\Delta r) := \sum_{v: d(u,v) \in [\Delta r, \Delta r + \epsilon]} P_{uv}$$

where ϵ is some prescribed "resolution" parameter. We can now measure the probability of observing a transition of length within $[r, r+\epsilon]$ considering an arbitrary starting place $u \in \mathcal{U}$ as

$$P(\Delta r) = \frac{1}{M} \sum_{u \in \mathcal{U}} P_u(\Delta r).$$

We note that the parameter α of the model has been set equal to 0.84 in all cases. This is the empirically calculated average of the rank value distributions, observed across the cities of the Foursquare dataset. The parameter ϵ has been set by binning the x-axis logarithmically using 100 bins in the range $[10^{-2}, 10^2]$. To obtain the Probability Density Functions (PDF) shown in the figures, we have divided $P(\Delta r)$ with the size of each bin, that is $[\Delta r, \Delta r + \epsilon]$.

3.5 Controlling urban geography

This analysis provides empirical evidence that while human displacements across cities may differ, these variations are mainly due to the spatial distribution of places in a city instead of other potential factors such as social-cultural, organisational, transportation or related to human cognition. Indeed, the agent based simulations are run with the same rules and parameters in each city, except for the set of places $\mathcal{U}$ that belong to a city, which

is taken directly from the empirical dataset. The variation in the spatial organisation of cities is illustrated in Figure 3.10, where we plot thermal maps of the density of places within cities and in Figure 3.11, where we plot the probability density function that two random places are at a distance Δr.

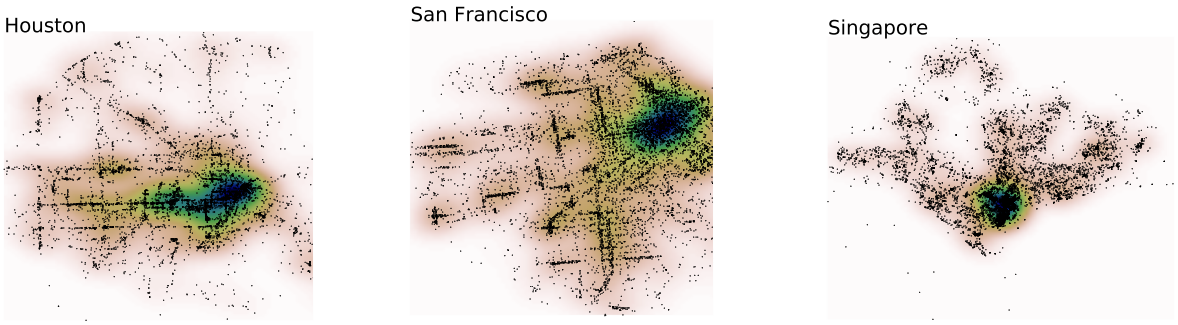


Figure 3.10: **Geographic distribution of places in cities.** Gaussian kernel density estimation (KDE) applied to the spatial distribution of places in three cities (Houston, San Francisco and Singapore). Each dark point corresponds to a venue observed in the Foursquare dataset encoded in terms of longitude and latitude values. The output of the KDE is visualised with a thermal map. A principal core of high density is observed in the three cities, but point-wise density and spatial distribution patterns may differ.

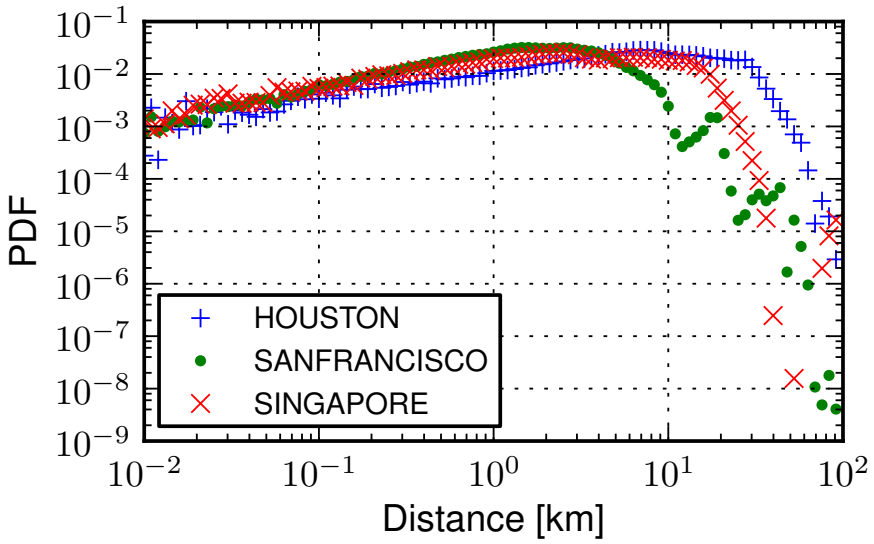


Figure 3.11: **Probability density function (PDF) of observing two randomly selected places at a distance Δr in a city**. We have enumerated 11808, 15970 and 15617 unique venues for Houston, San Francisco and Singapore respectively. The probability is increasing with Δr, as expected in two dimensions, before falling due to finite size effect.

Both figures highlight large heterogeneities in the distribution of places across cities. The rank-based model can cope with these heterogeneities as it accounts for the relative density for a given pair of places u and v. These differences in the geographic distribution of Foursquare venues across cities have enabled us to examine further how the geography of a city, encoded by the longitudinal positions of its settlements, impacts human mobility. Could we then *alter* the spatial distribution of settlements in a city and quantify the effect of this process on human movement?

Randomising the geographic coordinates of places. The methodology we have put forward to demonstrate this is based on the spatial randomisation of places, $\mathcal{U}$, of a city. We do so by iterating through all places in $\mathcal{U}$ and randomizing the coordinates, $lat_u, long_u$, of a place u with probability P_{rand}. A new pair of latitude and longitude coordinates is chosen, $(lat_{u'}, long_{u'})$, by considering a uniform sample in a predefined range, where $lat_{u'} \in [lat_u \pm 0.1]$ and $long_{u'} \in [long_u \pm 0.1]$. In Figure 3.12, we present the Kullback-Leibler Divergence (KL-Divergence), $D_{KL}(H||R)$, between the empirically observed distribution of human displacements, H, and the distribution R obtained by the *rank-distance* model for different values P_{rand}. The KL-Divergence [SK51] is a non-symmetric measure of the difference between two probability distributions and is formally defined as

$$D_{KL}(H||R) = \sum_i H(i) \ln \frac{H(i)}{R(i)}$$

The reader may observe that as the probability of randomizing the position of a place increases, the quality of the fit attained by the *rank-distance* model, on average, drops. This observation becomes statistically significant only for $P_{rand} \geq 0.7$. We note that any alternative randomisation process which, instead, preserves the relative density between all pairs of places (or $P_{rand} = 0.0$ equivalently) would not have an impact on the performance of the model on the original set of places $\mathcal{U}$. That is expected as the probability of a transition in the *rank-distance* model is dependent exclusively on this factor (provided that the exponent *alpha* remains unchanged and equal to 0.84 as it is the case in all simulations). Overall, this analysis highlights the impact of geography, as expressed by the spatial distribution of places, on human movements, and confirms at a large-scale the seminal analysis of Stouffer [Sto40] who studied how the spatial distribution of employment opportunities in the city of Cleveland affected the migration movements of families.

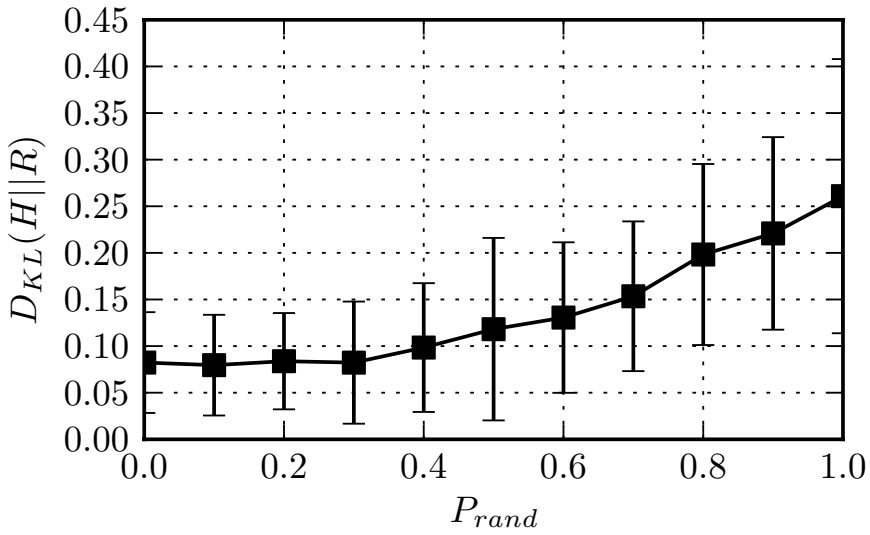


Figure 3.12: **Effect of place coordinate randomisation on the performance of the rank-distance model**. On the y-axis we present the KL-divergence, $D_{KL}(H||R)$, between the empirically observed distribution of displacements in a city H and R which is the one obtained by the *rank-distance* model. On the x-axis the probability of randomisation, P_{rand}, is depicted. $P_{rand} = 0$ corresponds to the case that the original distribution of displacements within a city is maintained, while the opposite extreme where P_{rand} equals 1.0 means that all places have been randomized. The error bars correspond to standard deviations across cities.

3.6 Discussion and implications

The empirical data on human movements provided by Foursquare and other location-based services allows for unprecedented analysis both in terms of scale and the information we have about the details of human movements. The former means that mobility patterns in different parts of the world can be analysed and compared across cultural, national or other organisational borders. The latter is achieved through better location specification technologies such as GPS-enabled smartphones, but also with novel online services that allow users to lay out content on the geographical plane, such as the existence of places and semantic information about those. As those technologies advance our understanding on human behaviour can only become deeper.

In this chapter, we have focused on human mobility in a large number of metropolitan cities around the world to perform an empirical validation of past theories of the driving factors of human movements. As we have shown, Stouffer's [Sto40] theory of intervening opportunities appears to be a plausible explanation for the observed mobility patterns of users in location-based social networks. The theory suggests that the distance covered by humans is determined by the number of opportunities (i.e., places) within that distance,

and not by the distance itself. This behaviour is confirmed in our data where we observed that physical distance does not allow for the formulation of universal rules for human mobility, but a universal pattern emerges across all cities when movements are analysed through their respective rank values: the probability of a transition to a destination place is inversely proportional to the relative rank of it, raised to a power α, with respect to a starting geographical point. In addition, α presents minor variations from city to city.

Our approach opens avenues of quantitative exploration of human mobility, with several applications in urban planning and computer science. The identification of rank as an appropriate variable for the deterrence of human mobility is in itself an important observation, as it is expected to lead to more reliable measurements in systems where the density of opportunities is not uniform, e.g. in a majority of real-world systems. The realisation of universal properties in cities around the globe also goes along the line of recent research [LBW07, BW10] on urban dynamics and organisation, where cities have been shown to be scaled versions of each other, despite their cultural and historical differences. Contrary to previous observations where size is the major determinant of many socio-economical characteristics, however, density and spatial distribution are the important factors for mobility. Moreover, the richness of the dataset naturally opens up new research directions, such as the identification of the needs and motives driving human movements, and the calibration of the contact rate, e.g. density- vs. frequency-dependent, in epidemiological models [MJSB09]. Finally, we note that there may be a strong demographic bias in the community of Foursquare users. While this is an inherent characteristic of many telecommunications services and corresponding datasets, it is encouraging that the analysis and models developed in the context of the present work demonstrate strong similarities across multiple urban centres and different countries.

3.7 Related work

As we have discussed in Chapter 2, in classical studies, two related but diverging viewpoints have emerged. The first camp argues that mobility is directly deterred by the costs (in time and energy) associated wit physical distance. Inspired by Newton's law of gravity, the flow of individuals is predicted to decrease with the physical distance between two locations, typically as a power-law of distance [Wil67, ES]. Besides distance, more complex versions of gravity models may also consider a parameter that captures the "volume" between the starting point and the destination of a trip. These so-called "gravity models" have a long tradition in quantitative geography and urban planning and have been used to model a wide variety of social systems, e.g. human migration [Lev10], inter-city communication [GMKB09] and traffic flows [JWS08]. The second camp argues instead that there is no direct relation between mobility and distance, and that distance is a surrogate for the effect of *intervening opportunities* [Sto40]. The migration from origin to destination is assumed to depend on the number of opportunities closer than this

destination. A person thus tends to search for destinations where to satisfy the needs giving rise to their journey, and the absolute value of their distance is irrelevant. Only their ranking matters. Displacements are thus driven by the spatial distribution of places of interest, and thus by the response to opportunities rather than by transport impedance as in gravity models. The first camp appears to have been favoured by practitioners on the grounds of computational ease [Eas93], despite the fact that several statistical studies have shown that the concept of intervening opportunities is better at explaining a broad range of mobility data [Mil72, KEHS73, Wad75, FR85, CB05]. This long-standing debate is of particular interest in view of the recent revival of empirical research into human mobility. Our work has studied two versions of these two diverging schools of thought on human movement. While our present findings support the theory of intervening opportunities due to its elegance in modelling movements in a simple way (as opposed to a less straight forward parameterisation of the corresponding gravity model), we most certainly cannot derive definitive conclusions as there is still room for more empirical studies that will consider both alternative modelling formulations and new datasets. Besides, in the light of this discourse one should recall the proof about the models' statistical convergence presented by Alan Wilson in [Wil67] and discussed in detail in Section 2.1.3.

The current study also shares the interests in determining the universal laws governing human mobility and migration patterns of [SGMB12]. We concentrate on modelling movement at the city scale, using the distribution of places in cities while the radiation model presented in [SGMB12] exploits population densities to model larger scale mobility patterns across states or municipalities. In fact, its applicability in the urban setting has been doubted empirically recently in [YG13].

3.8 Summary

In Section 2 we presented the fundamental mechanics of the most popular theoretical models in the literature of human mobility; the *gravity* and *intervening opportunities* models. As discussed, their main difference lies in the former's suggestion that movement between places depends only on the absolute distance between them, whereas in the latter case, the relative density of opportunities between a trip's origin and destination matter.

In this chapter, we have empirically reviewed the two models by conducting a large scale study in 34 cities around the world. Using a check-in dataset from location-based social networks, we initially revealed the key role of density in human movement. The denser an urban environment is, the shorter is the expected trip length in that environment. Base on this insight, we have subsequently demonstrated how the *rank-distance* variable allows for an *elegant* representation of movement in cities that effectively dissects the heterogeneity observed when *geographic distance* is employed to measure trip length frequency distributions. Equipped with these analytical observations, we employed the

rank-distance as the core element of an agent based human mobility model. The fact that the model was able to reproduce accurately the real mobility patterns of Foursquare users in cities has confirmed the plausibility of the theory of intervening opportunities in urban mobility. Without excluding the case that a gravity based model could present an equivalent alternative, we have argued that its parameterisation is not straightforward.

In the next chapter we will refine the level of abstraction at which we observe human mobility. Instead of aiming to model emergent properties in human mobility from a complex system perspective, we will attempt to predict the exact places visited by mobile users through the development of an application oriented framework. In addition, we will assess how the *rank-distance* compares to absolute *geographic distance* when they are being integrated as features into machine learning algorithms.

4

Next place prediction in location-based services

In Chapter 3 we studied the mobility patterns of users in location-based social networks by examining the statistical properties of their movements in urban environments. We have seen how the density of places between an origin and a destination, the *rank-distance*, matters greatly and is, in fact, a more informative variable than absolute geographic distance when modelling human mobility in the city. In this chapter, we will fine tune the level of abstraction at which we treat human movement and instead of targeting the modelling of the general statistical properties of user trips in the city, we will focus on the prediction of the *exact* place a user visits *next* given their current geographic position and time. Put more generally, we are aiming to predict the place where the *next check-in* of the user will occur.

Insights about the type and time of users' visits can greatly improve the development of recommendation systems. For instance, advertisers who want to push offers to users would greatly benefit from *knowing the next location a user is going to visit*, so to offer the right coupon or the right recommendation in a timely manner. Location-based services that target on venue recommendations can also benefit from effective predictions of where mobile users will go next. However there are many challenges involved in the prediction of the next visited location, relating to user preferences and place properties as well as the spatio-temporal context in which movement occurs. More technically, a major challenge posed in this context is to rank all the potential target places in the prediction scenario, which could easily contain thousands of candidates, so that the actual place visited *next*

by the user is ranked as highly as possible. This represents a highly imbalanced prediction scenario, where a single correct instance has to be found (the place a user is going to) amongst thousands of candidate instances (all places in the city).

Chapter Outline In the work to be presented in the following sections, we follow a methodology that focuses on interpreting *why* users choose to visit a place. Without making causal claims, our goal is to see movement through the lens of the different signals of information that are available in location-based services. In addition to focusing on the *interpretability* of user mobility prediction, our goal is to achieve high prediction performance. We will see that Foursquare users are likely to visit new venues with a probability above 60% (roughly two out of three times). Adding to this the observation that the user movement information is extremely sparse with only a few check-ins per user, then exploiting probabilistic frameworks which simply exploit historical information to make predictions is not a plausible direction to consider. We will thus see that quantifying how the different layers of data available in location-based services play a role in user movement can not only help us in understanding it better as a process, but also in building powerful mobility prediction models.

Specifically, in this chapter, we make the following contributions:

- **A formulation of the factors that drive user movement in location-based services.** First, we extend the analysis conducted in Chapter 3 on the check-in dataset of Foursquare users by inspecting the spatio-temporal dynamics of their movements. Driven also by the conclusions drawn by previous literature in human mobility, we define a set of prediction features that exploit different information dimensions of user mobility: these include information tailored specifically to an *individual user*, such as historical visits or social ties, and features extracted by mining *global knowledge* about the system such as the popularity of places, their geographic distance and user transitions between them. We employ a set of features that leverage explicitly *temporal information* about users' movements. We assess the predictability of individual features and we discover that the most effective features are those which leverage the popularity of target venues and user preferences. Moreover, by assessing the mobility prediction features across different evaluation metrics an interesting *duality* pattern in their performance emerges; features built on personalised user information are superior in predicting historically visited venues, whereas, movements towards new venues are better predicted when employing features that mine global information about mobility in the system.

- **An effective training strategy for applying supervised learning methods in the next check-in prediction problem.** We combine the predictive power of individual features in a supervised learning framework. By training two supervised regressors, a ridge linear regression [LCVH92] model and a continuous learning

decision tree [Qui92], on past user movements, we demonstrate how a supervised approach can significantly outperform single features in the prediction of future user movements, indicating that user behaviour in location-based services is driven by multiple factors that may act simultaneously. Notably, the decision tree model ranks consistently one in two user *check-ins* in the top 50 predicted venues. Furthermore, our supervised training strategy is able to offer predictions even for users with little check-in data (the majority of users); the proposed training methodology is built on information collected by groups of users moving in a city and is based on a technique of implicit label extraction that *teaches* supervised learning models what are the *good* (preferred) and the *bad* (non preferred) places for users to go in the city.

- **Offering insights on the temporal variations in movement prediction.** We study the performance of features and classifiers over time, finding that prediction performance is higher over lunchtime and weekdays. In all cases, a strong temporal periodicity is apparent in the prediction task, but features based on the geographic distance amongst places achieve higher scores at nighttime, unlike other features. This shows how the factors influencing human mobility can vary over time and highlights the importance of adding spatio-temporal context to the prediction task.

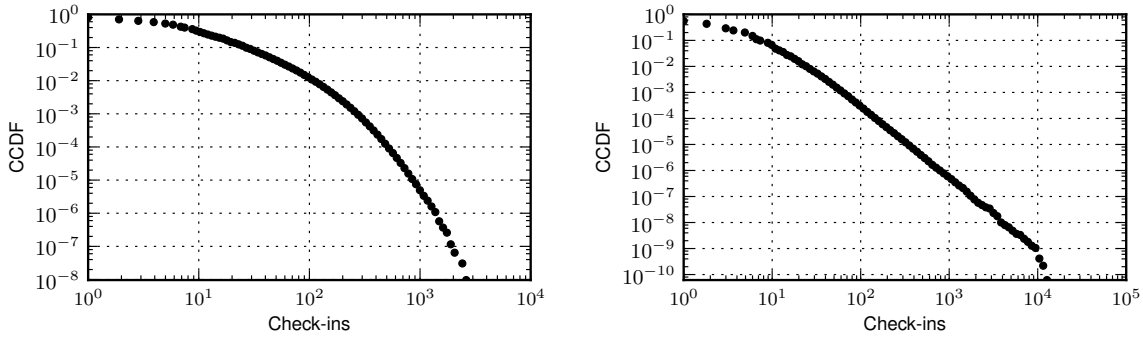
In the following sections, we begin by formulating the next check-in prediction problem (Section 4.2.1). Subsequently, we define twelve mobility prediction features (Section 4.2.2) and evaluate their performance individually (Section 4.3) and in a supervised learning framework (Section 4.4). We close with a discussion of the implications of our findings followed by related work and a summary.

4.1 Data and preliminary analysis

We employ the same dataset used in Chapter 3 comprised of publicly-available check-ins crawled via Twitter's streaming API[1]. Table 3.1 shows the 34 cities with the highest number of check-ins observed in the dataset: New York, Chicago and San Francisco top the list with 371, 185 and 112 thousand check-ins respectively. The table also includes the numbers of places that have been checked into in the cities as well as the average distance between consecutive check-ins, which ranges from 1.9 to approximately 6 kilometres. These values reflect the geographic distribution of check-ins in each city, as well as the urban sprawl of different areas. The analysis we present next concerns the full dataset (i.e., all cities and users in Foursquare), as we measure aggregate statistics, but the mobility prediction algorithms are evaluated for the set of the 34 *active* cities, that is,

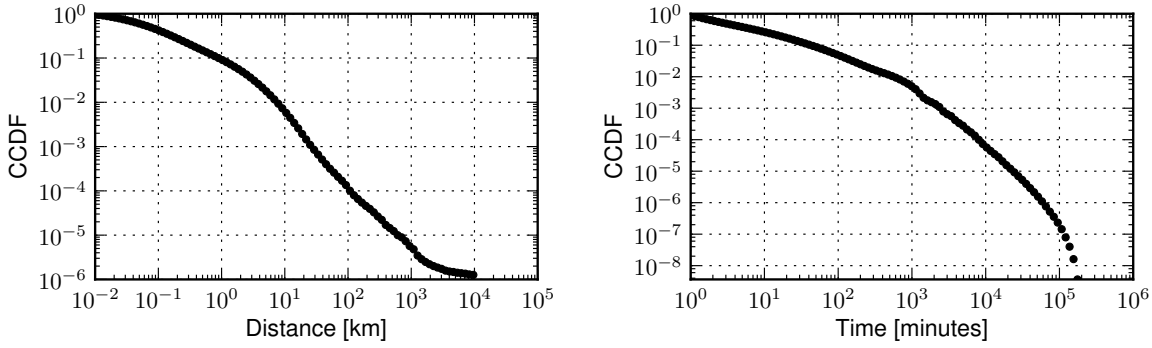
[1] `https://dev.twitter.com/docs/streaming-api`

cities with at least 25,000 check-ins. The choice to focus on the top cities has been motivated by the fact that the supervised learning algorithms shown in Section 4.4.2 require the provision of sufficient data during the learning phase and our models are trained on a per city basis for reasons whose details we clarify in Section 4.4.1.



(a) Users (b) Places

Figure 4.1: Complementary Cumulative Distribution Function (CCDF) of (a) number of check-ins per user and (b) number of check-ins per place in the Foursquare dataset.



(a) Inter Check-in Distances (b) Inter Check-In Times

Figure 4.2: Complementary Cumulative Distribution Function (CCDF) of (a) spatial distance and (b) time elapsed between consecutive user check-ins.

With respect to user engagement in the Foursquare service as reflected by the collected dataset, the number of check-ins made by users is highly heterogeneous: the probability distribution exhibits a heavy tail, with about 50% of users having fewer than 10 check-ins. In particular, about 15% of users have only a single check-in contained in the entire dataset. A similar pattern arises when considering the number of check-ins made in each place: only 10% of places have more than 10 check-ins. The Complementary Cumulative Distribution Function (CCDF) of the number of check-ins per user and per place are shown in Figures 4.1(a) and 4.1(b). The distribution of check-ins per place, which is indicative of the popularity of Foursquare venues, is highly skewed and provides, already, a strong signal that exploiting this information in order to rank venues in a city may be a promising direction to predict the next place a user will move to. However, we will

show that popularity alone, as with distance, does not guarantee maximum prediction performance.

In order to analyse user movements between venues, we have to consider only users who have more than one check-in (85% of the dataset users). For those remaining, we focus on the sequence of check-ins they make over time and study two aspects: the spatial distance between two consecutive check-ins and the amount of time elapsed between them. The data reveals the importance of both space and time in determining where the user will check in to next.

The distribution of time intervals between consecutive check-ins is shown in Figure 4.2(b). Longer intervals are less likely than shorter ones, meaning that sequences of more frequent check-ins might arise, together with long periods of inactivity. This reveals that users exhibit bursts of check-ins that can be mined to understand how they choose where to go next. There are two different trends that become prominent: the first is formed by consecutive check-ins within 1440 minutes (a day) and a second, steeper trend when consecutive check-ins happen across different days. As consecutive check-ins become separated by longer time intervals they might become also less related; hence, we will focus our prediction efforts on check-ins that happen within 24 hours of the previous one.

Finally, Figure 3.3 highlights the heterogeneity observed in distance between check-ins across different cities. While smaller distances appear to be more probable in all cities, the effect of distance at longer ranges can greatly differ between cities with different urban and spatial properties. As a consequence, we will frame the next venue prediction problem as a separate problem in each city for reasons we detail in Section 4.4.1.

4.2 Next check-in venue prediction in Foursquare

In this section, we formalise the *Next Check-in Problem*. Given the current check-in of a user, we aim to predict the next place in the city that the user will visit, considering thousands of candidate venues in the prediction list (a sample of venues in the centre of London is depicted in Figure 4.3). Subsequently, we will propose a set of features that leverage upon a wide spectrum of user mobility patterns in Foursquare, in order to effectively predict the whereabouts of mobile users.

4.2.1 The next check-in problem

Notation and problem formulation

We define a set of users U and a set of locations L. Each check-in c by $u \in U$ is defined as a tuple $\{l, t\}$, where $l \in L$ represents a venue and t is the check-in's timestamp. The total set of check-ins is denoted as C and the set of check-ins for a specific user u as C_u.

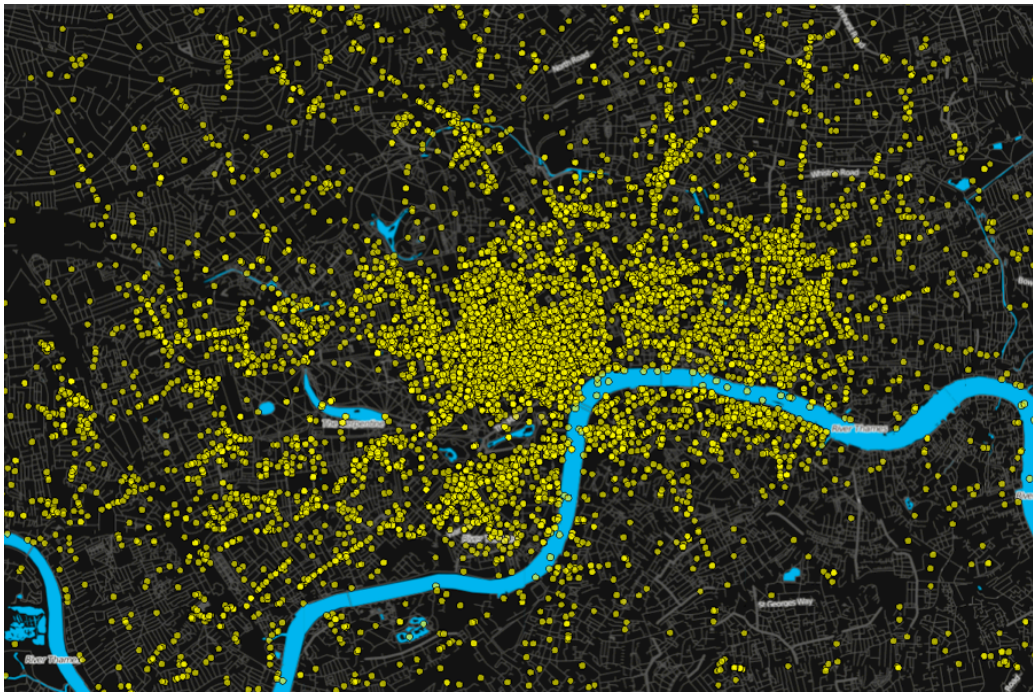


Figure 4.3: Spatial distribution of Foursquare venues in the area of central London. In the next check-in prediction task, given the appearance of a mobile user at one of the places in the city, thousands of places need to be ranked appropriately so as to predict the next location a user will check in to.

We then formalise the next check-in prediction problem as follows. Given a user u whose current check-in is c (to venue l' at time t'), our aim is to rank the set of venues L so that the next venue to be visited by the user will be ranked at the highest possible position in the list. According to the setting described above, the *next check-in* problem is essentially a *ranking task*, where we compute a ranking score $\hat{r}$ for all venues in L.

Filtering venues by city

We constrain the selection of candidate venues to the set of places L within a given a city, decreasing dramatically the cardinality of the prediction set with respect to the entire set of Foursquare places. This is computationally desirable, since we decrease the number of potential venue targets from the order of millions, observed globally, to the order of thousand venues usually included in a city (see Table 3.1). This approach is justified, if one bears in mind that almost 99% of consecutive *check-ins* feature a distance smaller than 10 kilometres, as shown in Figure 4.2(a), suggesting that the vast majority of user activity in Foursquare occurs within the urban boundary. Further, this choice allows us to avoid requiring the introduction of distance as an explicit parameter (for instance if we were to filter the prediction list by applying a bounding box around the user's current position) and we can examine its effect as a prediction feature in an unbiased way.

4.2.2 Mobility prediction features

We now describe in detail the set of prediction features employed to tackle the next check-in problem. The twelve features we mine here can be loosely grouped into different categories depending on the source of mobility information exploited. Formally the three categories are:

- **User Mobility Features** that mine data straight from the target user.
- **Global Mobility Features** that are formulated based on the behaviour of groups of users in the city.
- **Temporal Features** that seek to exploit temporal information from check-ins at venues.

For all cases, we note as t' and l' the time and location of the current check-in respectively. We set t' as the current prediction time and we compute the ranking scores of all features assuming knowledge up to that time.

User Mobility Features

This class refers to features tailored to the check-ins generated by the user under prediction or by her social network. We aim to capture the likelihood that a user will return to a place visited in the past, but also the likes of the user in terms of types of places she likes to hang out.

Historical Visits. By measuring the number of past visits of user u at a target venue k, we are aiming to assess to what extent the next check-in of a user is likely to emerge at a place that has been visited by the user in the past. Formally we have

$$\hat{r_k}(u) = |\{(l,t) \in C_u : t < t' \wedge l = k\}| \tag{4.1}$$

We will see, in Section 4.3 when we evaluate the features, that there is a high chance of users visiting new places, however, the probability of returning to previously seen venues remains high (around 30%) and constitutes a strong signal of a user's probable whereabouts, especially in the case of active users.

Categorical Preferences. Another source of information based on historical behaviour is the number of check-ins user u has performed at a place that belongs to category z. In this way, we identify the importance of different categories of places (cinema, coffee shop, football stadium, etc.) for a given user and rank them accordingly:

$$\hat{r_k}(u) = |\{(l,t) \in C_u : t < t' \wedge z_l = z_k\}| \tag{4.2}$$

We note that we subsequently rank venues that belong to the same category by their popularity in terms of number of check-ins. Thus between coffee shops for instance, those with most check-ins are ranked more highly. This feature corresponds to a content filtering based approach, a popular strategy in the literature of recommender systems.

Social Filtering. Previous works [CML11, SKB12] have suggested a strong relationship between the places users visit and those visited by their friends. Thus we build the corresponding feature, considering a user u and his set of friends Γ_u. We rank a target venue k by summing the total number of check-ins that any friend v of the user has performed at place k:

$$\hat{r_k}(u) = \sum_{v \in \Gamma_u} |\{(l,t) \in C_v : t < t' \wedge l = k\}| \tag{4.3}$$

As expected for users who have not got any friends in the system there is no utility in this case. However, knowledge of the whereabouts of friends can be very useful in the cold start prediction scenario, where a user joins the system, declares social connections but has yet to check in to any venue.

Global Mobility Features

Now, we demonstrate how we can exploit global information about the check-in patterns of Foursquare users going beyond an individual user and her social network. In this category we will include popularity and geographic features together with features that exploit transitions between venues. Our choice to seek information sources that do not immediately relate to the patterns observed about the movements of the user under prediction stems from our willingness to predict venues that the user has not visited before. Past approaches in mobility prediction have been mainly focused on the development of frameworks that depend on historical data and do predictions only on previously visited places for a given user. Yet, in location-based social networks there are thousands of real world places a user may visit. This corresponds to a novel class of predictions and an important characteristic of location-based social networks that we will address specifically in Chapter 5.

Popularity. The distribution of check-ins per place (see Figure 4.1(b)), which is indicative of the popularity of Foursquare venues, is highly skewed and provides, already, a strong signal that exploiting this information in order to rank venues in a city may be a promising direction to predict the next place a user will move to. In response to this observation we define the corresponding feature by counting the total number of check-ins performed by the total set of users U in the dataset to a venue k:

$$\hat{r_k}(U) = \sum_{u \in U} |\{(l,t) \in C_u : t < t' \wedge l = k\}| \tag{4.4}$$

In a lot of recommender systems item popularity is considered to be a very strong baseline predictor and in the evaluation Section 4.3 we will see that this is also the case here. Nonetheless, we will also see how this baseline is clearly outperformed by more sophisticated prediction methods.

Geographic Distance. The role of geographic distance in human movements has been investigated in various works using mobile phone based datasets [DBG06, GHB08] and, as also has been studied in this dissertation (Chapter 3 and Figure 4.2(a)) it is also important in location-based social services. Considering the current location l' of user u we measure the distance $dist(l', k)$ to all other places based on their geographic coordinates. Venues are subsequently ranked in ascending order so the nearest place will be at the top of the prediction list.

$$\hat{r_k}(l') = dist(l', k) \tag{4.5}$$

Rank Distance. Similarly to geographic distance, we define for the next place prediction problem, *rank distance* that we initially investigated in Chapter 3. We may recall that the *rank distance* measures the relative density between the current place of the user, l', and all other places. Formally, considering all places $l \in L$ we define

$$\hat{r_k}(l') = |\{l \in L : dist(l', w) < dist(l', k)\}| \tag{4.6}$$

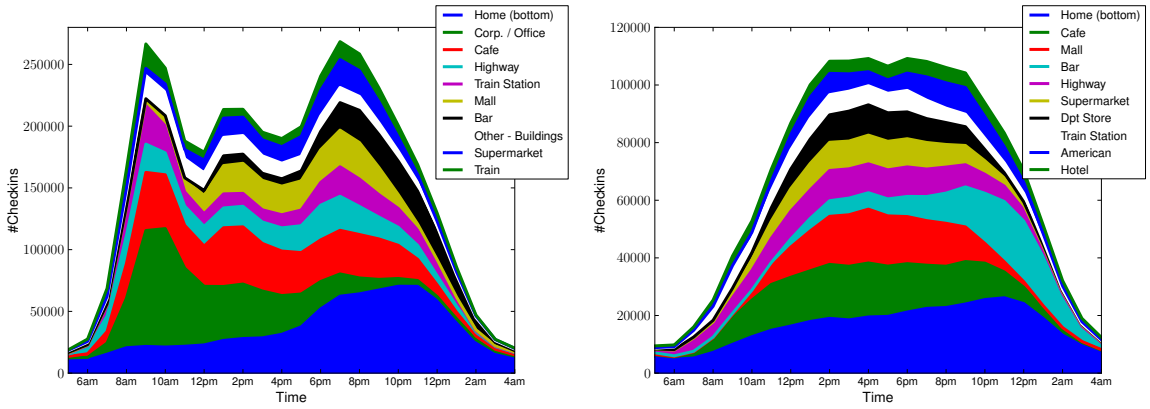
which in plain words translates to the enumeration of venues that are geographically closer to l' than the destination k. Our assumption here is that the movement of people is not based on absolute distance values, but rather by the density of opportunities or resources nearby. This approach has been motivated by Stouffer's *theory of intervening opportunities* presented in the literature on human migration [Sto40] and described in detail in Chapters 2 and 3. It is worth mentioning that the ranking scores output by the *geographic distance* and *rank distance* features are equal (since geographic distance reduces it to a ranking task such as the one dealt with here). Yet, as we will see in Section 4.4, when

Origin Venue Category	Destination Venue Category	$\mathbf{P_t^{10}}$	$\mathbf{P_t^*}$
Train	Train Station	0.48	0.30
Terminal	Airport	0.46	0.17
Gate	Airport	0.45	0.22
Moroccan	Theme Park	0.39	0.06
Train Station	Train	0.38	0.22
Rental Car	Airport	0.36	0.18
Plane / In-flight	Airport	0.33	0.19
Tram	Airport	0.33	0.19
Cineplex	Mall	0.30	0.08
Plane	Airport	0.28	0.15
Bridge	Highway / Traffic	0.28	0.10
Lab	University	0.26	0.09
Surf Spot	Beach	0.25	0.06
Trade/Tech School	Other - Buildings	0.25	0.07
Emergency Room	Hospital	0.25	0.08
Hotel Bar	Hotel	0.24	0.07
Engineering	University	0.24	0.07
Movie Theater	Mall	0.24	0.06
Other - Travel	Highway / Traffic	0.23	0.11
Taxi	Highway / Traffic	0.23	0.09

Table 4.1: Top-20 Activity Transition Probabilities for consecutive check-ins by a user. P_t^{10} refers to check-ins that took place within 10 minutes of each other and P_t^* for the transition probability without considering a temporal threshold.

the two features are incorporated in machine learning models in the light of optimisation criteria they are not equivalent. This has important implication for the integration of distance measurements in location-based social networks and we shall discuss them in detail in Section 4.4.

Activity Transitions. Sequences of human activities are not random, as for instance we may visit the supermarket after work or go to a hotel after landing at an airport. The non-uniformity in the probability of transiting from one Foursquare venue type to another can also been seen in Table 4.2.2 where we plot the highest transition probabilities from one type of place to another in the system. As a consequence, we are defining the corresponding feature which enables us to capture this signal in Foursquare check-in data. Formally, by writing as a tuple, (m, n), the places $m \in L$ and $n \in L$ involved in two



(a) Weekdays

(b) Weekend

Figure 4.4: Stacked plot of the 10 most popular categories over weekdays and weekends. Popularity decreases bottom-up.

consecutive check-ins, with z_m and z_n being their corresponding categories, we have

$$\hat{r_k}(l') = |\{(m,n) \in L_c : z_m = z_{l'} \wedge z_n = z_k\}| \tag{4.7}$$

where L_c denotes the set of tuples for places involved in consecutive transitions before current prediction time t'.

Place Transitions. By definition of the *next check-in problem* we seek to predict consecutive transitions of users between venues. Thus, we build a feature that directly exploits this information, by measuring the direct transitions between all pairs of venues in the city. Accordingly, the rank score of a target venue k is obtained by enumerating the *past* transitions observed by *any* user from the current location l' to location k, which we formally define as

$$\hat{r_k}(l') = |\{(m,n) \in L_c : m = l' \wedge n = k\}| \tag{4.8}$$

This is one of the currently deployed recommendation techniques of the Foursquare application [Eng12](*people who went to the place where you are, they also went to place X*). Recall that during data collection there were no recommendations featured in Foursquare and therefore the results to be seen in the following paragraphs are not biased from this point of view.

Temporal Features

Time has been an important dimension in systems where human behaviour is central. Here, by exploiting the fact that every Foursquare check-in is timestamped with per second accuracy, we define time aware features that capture information both about user

activity in terms of visiting categories of places, but also temporal patterns of visits to specific places.

Category Hour. In Figures 4.4(a) and 4.4(b) we plot the number of check-ins across the ten most popular categories of places observed during weekdays and weekends respectively. The two curves present considerably different patterns. On weekdays activity presents three peaks: in the morning when people go to work, at lunchtime, and between 6pm and 8pm when they commute, return home or go to malls and bars. On the other hand, during weekends user activity presents a smoother evolution course, reaching a long lasting plateau between 12pm and 10pm. Another difference to note between the two is that category Corporate/Office disappears from the top set of user activities and is substituted by leisure related activities such as *American* (Food) and *Hotel*, while categories such as *Bar* and *Mall* also show increased preference rates among users. In both cases, however, checkins at the *Home* category show a continuous rise throughout the day, with a steeper increase at 6pm during weekdays.

This shows that Foursquare user activity is driven by temporal rhythms both on a *daily* and on a *weekly* basis (Figure 4.4). In the light of these observations we formulate features that enable the realisation of these patterns. More specifically, given that z_k denotes the type of the target place k, we define the *Category Hour* popularity as the sum of past check-ins at a place of type z_k in a given hour h of the day.

$$\hat{r_k}(t') = |\{(l,t) \in C : z_l = z_k \wedge tod(t) = tod(t')\}| \tag{4.9}$$

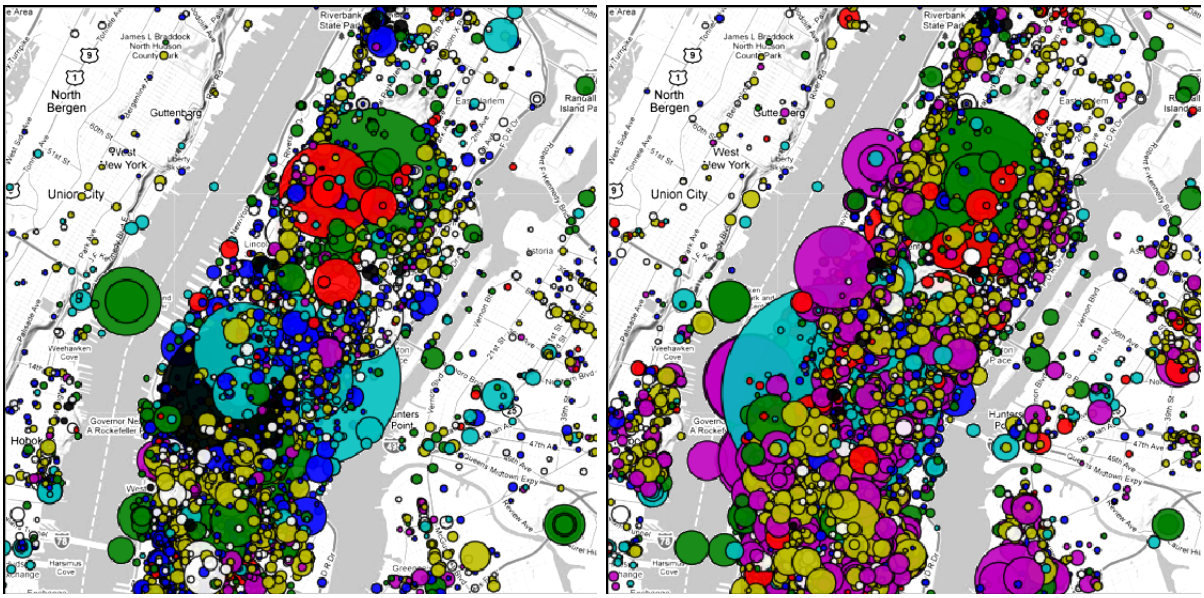
where $tod(t) \in [0, 1 \ldots 24]$ returns a value corresponding to the hour of the day of time t.

Category Day. Similarly, we set *Category Day* popularity as the sum of check-ins at a place of type z at a given hour of a week:

$$\hat{r_k}(t') = |\{(l,t) \in C : z_l = z_k \wedge tow(t) = tod(t')\}| \tag{4.10}$$

where $tow(t) \in [0, 1 \ldots 167]$ returns a value corresponding to the *hour* of the week of time t. While the only difference between *Category Day* and *Category Hour* is their temporal granularity (24 hours of the day versus 168 hours of the week respectively), there is a trade-off between specificity and ability to generalise in a prediction setting that we are willing to explore in the context of Foursquare check-in data. In general, higher dimensional formulations in machine learning imply a more sophisticated and informative model. Nonetheless this advantage is traded off with the requirement for more training data, an issue that is known as *the curse of dimensionality* [HTF03].

Place Day. A spatio-temporal snapshot of the collected corpus is depicted Figures 4.5(a) and 4.5(b), where user activity in New York is shown for morning and night respectively.



(a) Morning (b) Night

Figure 4.5: New York in the morning and at night. We show the 8 eight categories of the top hierarchical level: Arts & Entertainment (red), College & Education (black), Shops (white), Food (Yellow), Parks & Outdoors (green), Travel (cyan), Nightlife (magenta), Home/ Work/ Other (blue).

A circle represents a venue and its radius its poularity in terms of number of checkins. Each color corresponds to one of the 8 general categories introduced by Foursquare (described in caption). The *mosaic* created by user check-in data highlights the diversity of human activity across the spatial plane.

In general, user check-ins at places are known to have characteristic temporal signatures as discussed in [YSL+11], where information about temporal visiting patterns are exploited to infer the semantic tags of places. Thus, as with venue categories, we also define the temporal check-in activity at specific venues. We measure the number of check-ins place k has during a day of the week (*Place Day*) defined as:

$$\hat{r_k}(t') = |\{(l,t) \in C : l = k \wedge dow(t) = dow(t')\}| \tag{4.11}$$

where $dow(t)$ returns the day of the week of time t.

Place Hour. A similar definition follows for the number of check-ins that place k has at a given hour of a day (*Place Hour*), aiming to capture weekly and daily patterns, respectively:

$$\hat{r_k}(t') = |\{(l,t) \in C : l = k \wedge tod(t) = tod(t')\}| \tag{4.12}$$

Foursquare uses *trending places* information [2] in its local search engine, in order to promote in real time venues where a lot of users have recently (past hour) checked in. Effec-

[2] https://developer.foursquare.com/docs/venues/trending

tively this corresponds to venue popularity features that are time aware. We note that due to the sparsity (approximately 300 categories compared to millions of venues) of check-in data on a per place basis in Foursquare, here we do not use the 168 hour granularity formulation as we have done for venue categories.

4.3 Evaluating mobility features

We now evaluate the performance of each individual prediction feature. We first describe the evaluation metrics we adopt; then, we compare the performance of each feature across these metrics. Finally, we assess how prediction performance changes over time.

4.3.1 Methodology and metrics

Given each user check-in eligible for prediction, we have a set L of candidate places to rank. The features compute a numeric value $\hat{r}_k$ for each candidate venue k, which are subsequently used to produce a personalised ranking of the venues. We then write as $rank(k)$[3] the rank of venue k, obtained after sorting in decreasing order all venues in L according to $\hat{r}_k$. We aim to measure the extent to which the future venue that will be visited is highly ranked by the prediction algorithms. We use two metrics to measure the performance of the features and algorithms employed for the prediction of the next place.

First, the *Percentile Rank* [HKV08] (PR) of the visited place k: $PR = \frac{|L|-rank(k)+1}{|L|}$. The PR score is equal to 1 when the place that will be visited next is ranked first and it linearly decreases to 0 as the correct place is demoted down the list. The *Average Percentile Rank* (APR) is obtained by averaging across all user check-in predictions: this measure captures the average normalised position of the correct instance in the ranked list of instances. We also use *prediction accuracy* to assess the performance when using different prediction list sizes N. In this case, we successfully predict the next check-in venue if we rank a venue in the top-N places. Average accuracy is the fraction of successful instances over the total number of prediction tasks, which we note as Accuracy@N. Features that achieve high APR scores do not necessarily also excel in terms of accuracy. The implications regarding the *duality* in the predictive performance of machine learning features across different evaluation metrics will be discussed in the paragraphs to follow.

[3]For a venue k, $rank(k)$ is the position of the venue in the prediction list and should not be confused with the notion of *rank-distance* introduced previously.

4.3.2 Feature based venue prediction

APR results

The APR results for all features are presented in Table 4.2. From the class of **User Mobility** features, we can distinguish the *Categorical Preference* feature which achieves a score of 0.84, which is considerably higher than the *Historical Visits* ($APR = 0.68$) and *Social Filtering* ($APR = 0.61$). This provides an indication that the types of places users tend to visit (cinema, nightclub, coffee shops etc.) can be highly informative about user mobility preferences and could be employed in mobile applications such as place recommendation systems.

With respect to features mined exploiting **Global Mobility** patterns of Foursquare users, *Place Popularity* which ranks venues according to the number of past check-ins is the most promising predictor with an APR score that averages 0.86. This is the highest APR score across all features and confirms our observations in Section 4.1 (Figure 4.1(b)) about the highly skewed distribution of visit frequencies at Foursquare venues where a few hub venues absorb a large fraction of user movements. The *Geographic Distance* and *Rank Distance* attain an average score 0.78, highlighting that spatial distance is an important factor in the way users decide which venue to visit next. Continuing in the same class of features, the *Activity Transition* and *Place Transition* features achieve lower scores with $APR = 0.60$, though remaining higher than the *Random Baseline* which would achieve 0.50.

We close the APR score analysis by looking at the performance of features that exploit **Temporal Information** about the check-in patterns of Foursquare users. The *Place Hour* feature, which ranks target venues according to the frequency of visits by any user observed in the past at the current check-in *hour*, achieves the highest score, 0.79. The *Place Day* ranking, which instead ranks venues by the past number of visits at the *day* of the current user check-in, follows closely with an $APR = 0.76$, perhaps due to its lower temporal specificity (day of week instead of hour of day). However, both features signify that temporal activity around venues constitutes a source of high quality signal in the venue prediction task. Finally the *Category Hour* and *Category Day* features trail in performance with scores 0.56 and 0.57 respectively and offers only a marginal improvement over the random baseline.

The effect of prediction list size

The APR scores denote how well, in general, a prediction feature ranks the next visited venue amongst all candidate venues L. However, in the context of a real mobile application where a finite set of places may be recommended to a user, due to interface or other

Feature	APR	ACC@10	ACC@50
Random Baseline	0.5	0.0001	0.0005
User Mobility			
Historical Visits	0.68	0.30	0.36
Categorical Preference	0.84	0.006	0.05
Social Filtering	0.61	0.17	0.24
Global Mobility			
Place Popularity	0.86	0.07	0.16
Geographic Distance	0.78	0.08	0.19
Rank Distance	0.78	0.08	0.19
Activity Transition	0.60	0.03	0.06
Place Transition	0.60	0.17	0.20
Temporal			
Category Hour	0.56	0.01	0.02
Category Day	0.57	0.01	0.03
Place Day	0.76	0.07	0.16
Place Hour	0.79	0.09	0.20

Table 4.2: **Average APR, Accuracy@10 and Accuracy@50 results for all mobility features.**

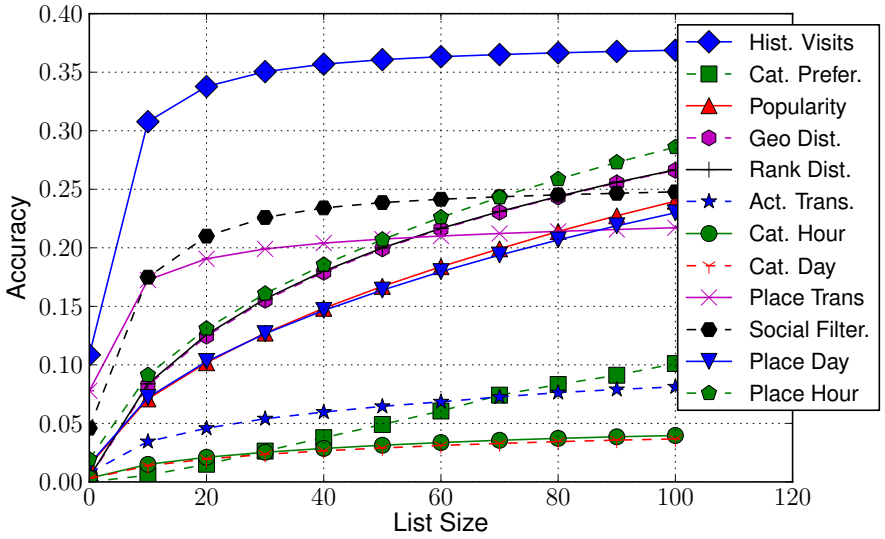


Figure 4.6: **Feature Predictability.** Mean Accuracy for all features when they are being tested on an individual basis for different prediction list sizes N.

constraints, one would be interested to examine how prediction approaches perform when the size of the prediction list N is limited.

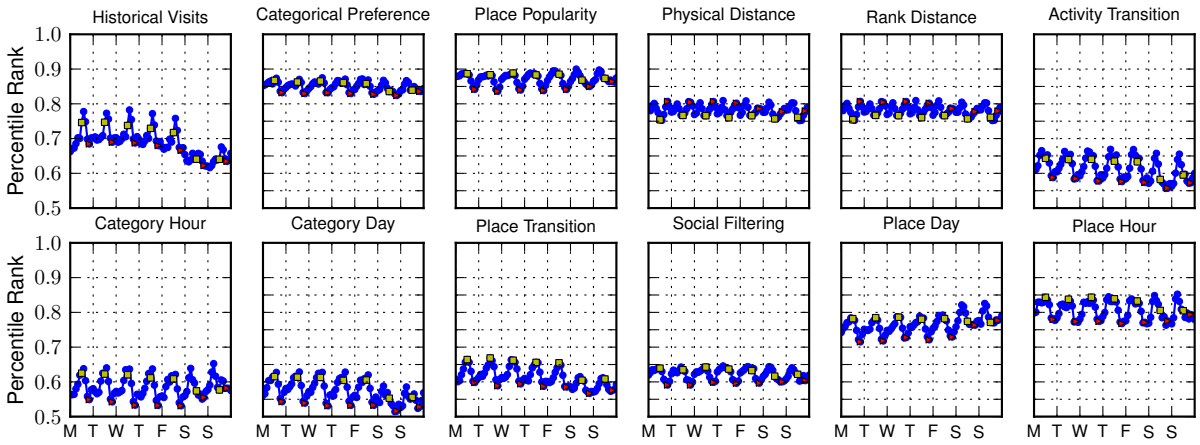


Figure 4.7: **Feature Weekly Predictability.** Average Percentile Rank for all features for different hours of a week. On average, distance- and popularity-based features outperform the rest. Strong daily periodicities are also observed: notice the yellow circles and red squares which correspond to noon and dinner times respectively.

We have evaluated all algorithms across various top-N list lengths using the Accuracy@N metric. We show the full set of results in Figure 4.6 and we report the results of Accuracy@10 and Accuracy@50 in Table 4.2. The principal observation is that features that rank low in APR can potentially demonstrate good performance in accuracy terms, in contrast to the results presented in the previous paragraph.

Overall, the results in Figure 4.6 suggest features tailored specifically to **User Mobility** patterns, such as *Historical Visits* and *Social Filtering,* dominate in accuracy for list sizes smaller than $N = 60$. In particular, *Historical Visits* continues to perform well over larger list sizes, up to $N = 100$. We note that both features had relatively low APR scores. On the other hand, features that harvest **Global Mobility** information, such as *Place Popularity* or *Geographic Distance,* fail to achieve high accuracy scores for small N values. This *duality* in the performance of the various predictors can be explained by the fact that some features can predict *exactly* the next place a user is going to when, for instance, the user returns to a previously visited place or visits places that their friends go to. However, the same features fail to rank appropriately the thousands of previously unseen by the user. This explains the low APR scores achieved by features personalised to the user (**User Mobility** class of features); they are completely agnostic in ranking new venues and, thus, the mean scores observed for these data points are biased to very low values for a large fraction of user check-ins. We shall see though that the heterogeneities observed in the performance of features will be dissected when we will combine them in a unifying supervised learning framework in Section 4.4.

An exceptional case with respect to performance across metrics is the case of the *Place Transition* feature. Despite being mined from **Global Mobility** information it achieves

a relatively high accuracy score, unlike most other features that belong to that class. An explanation for this performance behaviour is that *Place Transition* exploits the relative popularity of venues with respect to the current location of the user. Then the fact that each venue is, in principle, connected to only to a small subset of venues in the city allows for the emergence of a very accurate prediction mechanism. This also confirms that there is a strong signal generated by the sequences of visits made by mobile users to Foursquare venues.

Finally, with regard to the class of features that exploit **Temporal** information about places, we note that the performance of the *Place Day* and *Place Hour* features are in line with features such as *Geography Distance*, *Activity Transitions* and *Place Popularity,* which begin with small accuracy values but constantly improve for larger values of N and do thousands of times better than the random baseline as shown in Table 4.2 for Accuracy@10 and Accuracy@50.

Predictability over time

We have demonstrated the overall performance of various features in the light of two different metrics, APR and Accuracy@N. Another interesting aspect to consider is how well the different prediction strategies perform at different temporal instants throughout the day or the course of a week. Figure 4.7 compares the performance of the various features by showing the temporal evolution of the APR score on a weekly basis. To retrieve those scores we have measured performance at different hours of the week. Given that the check-ins of each user are timestamped with per second granularity, we simply retrieve the hour of the corresponding prediction task and we assess how predictability evolves over time. Overall, we note that the effectiveness of each feature over time is not constant: predictions are more accurate at noon and less accurate in the evening. This suggests that people might be more habitual during the day and more likely to alter their regular patterns and try something new in the evenings.

Interestingly, in the cases of *Geographic Distance* and *Rank Distance* performance is inverted with a clear implication: users are more likely to cover shorter distances at night between consecutive check-ins. Further, the variation between the minima and maxima in the temporal results is more prominent for some features. More specifically, features such as *Historical Visits* and *Place Transition* score significantly lower over the weekend, whereas *Categorical Preference*, *Place Popularity* and the distance based features have a more stable behaviour. Finally, an interesting case in that respect is also the performance of *Place Day*, whose APR score rises over the weekend highlighting how knowledge of the temporal patterns of venue visits becomes prominent during this period. We will discuss the implications of the temporal variability of machine learning features in the next place prediction task in Section 4.5.

4.4 A Supervised learning approach to venue prediction

In this section, we combine the individual prediction features presented previously into a supervised learning framework. Our aim is to exploit the union of individual features in order to improve predictions, assuming that user mobility in Foursquare is driven by multiple factors acting synchronously. After discussing a novel training strategy for machine learning prediction models in the next place prediction setting, we will see how the performance of individual features can be outperformed by unifying prediction frameworks. These are able to attain good performance in light of both evaluation metrics employed here, APR and Accuracy@N.

4.4.1 Training Strategy: Learning to rank from populations of mobile users.

To predict the next check-in venue of a user we train supervised models assuming knowledge up to prediction time t'. For every check-in that took place before t', we build a training example $\mathbf{x}$ which encodes the values of the features of the visited venue (e.g., popularity, distance from previous venue, temporal activity scores) and whose label y is positive. Then, we retrieve a negative labelled input by sampling at random across all other places in the city. Essentially, we are aiming to teach the model what the crucial characteristics are that would allow the differentiation of places that attract user check-ins from those that do not. This method of training a model by providing feedback in the form of user preference judgements has been established in the past [CSS99] and corresponds to an effective reduction of the *ranking problem* to a *binary classification task*. We choose to train supervised regression models that compute a real valued numeric score for each instance; this will subsequently allow us to rank target venues. A potential alternative would be the use of a probabilistic classifier that would return for each input a probability score and then places could be ranked accordingly.

Justification and implications

The proposed training strategy to learn across movements of user collectives has been motivated by a number of factors. Building a model on a user by user basis suffers from the extreme sparsity of the check-in data. The of **User Mobility** features evaluated above, which directly exploit information about the user, offer a relatively good performance, but constrain the prediction task to only a handful of venues amongst the thousands of potential targets in the city. Further, the binary classification labelling of the training instances is a result of the need to resort to *implicit* user feedback since in the Foursquare dataset

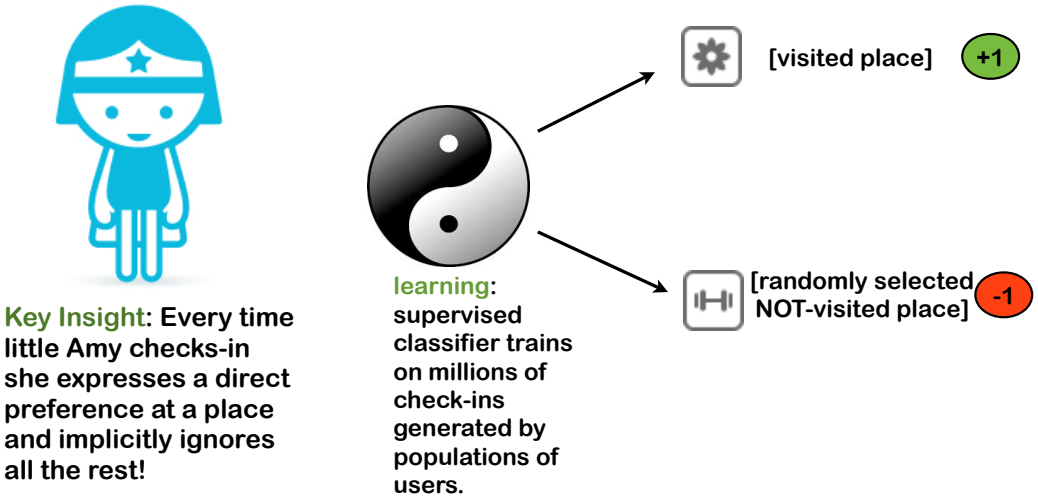


Figure 4.8: Toy example of the learning to rank venues process in a binary classification setting. Exploiting thousands of user movements in a city, machine learning models are able to learn the feature based representations of places preferred (respectively ignored) by Foursquare users.

there were not explicit user to venue ratings as it happens typically in recommendation settings. The absence of ratings provided by mobile users makes it difficult to exploit learning to rank approaches [Liu09] that are explicitly designed to address ranking tasks in recommender systems. Finally, we note that we have trained the supervised models on a per city basis as this has experimentally yielded better results than training across the general Foursquare user population. This could be viewed as a very broad form of personalisation as models are trained on groups of users who check in in the same urban environment as opposed to training on the global set of Foursquare users.

4.4.2 Supervised learning algorithms.

We consider two different supervised models to learn how feature vectors $\mathbf{x}$ correspond to positive and negative labels: linear ridge regression and M5 decision trees [Qui92]. The choice of a linear and a non-linear model respectively, has been motivated by the need to investigate what relationship holds amongst the features studied in the previous paragraphs in the light of the next place prediction problem; linear models are by definition less complex, yet many real world problems and systems are better solved exploiting non-linear representations of input features.

We have used the corresponding implementations that are publicly available through the WEKA machine learning framework [WF05]. As hinted already, the linear model assumes that the relationship between the vector $\mathbf{x}$ of input features and the output label y is linear. The goal is to estimate a vector $\mathbf{w}$ that minimises the error between actual and predicted

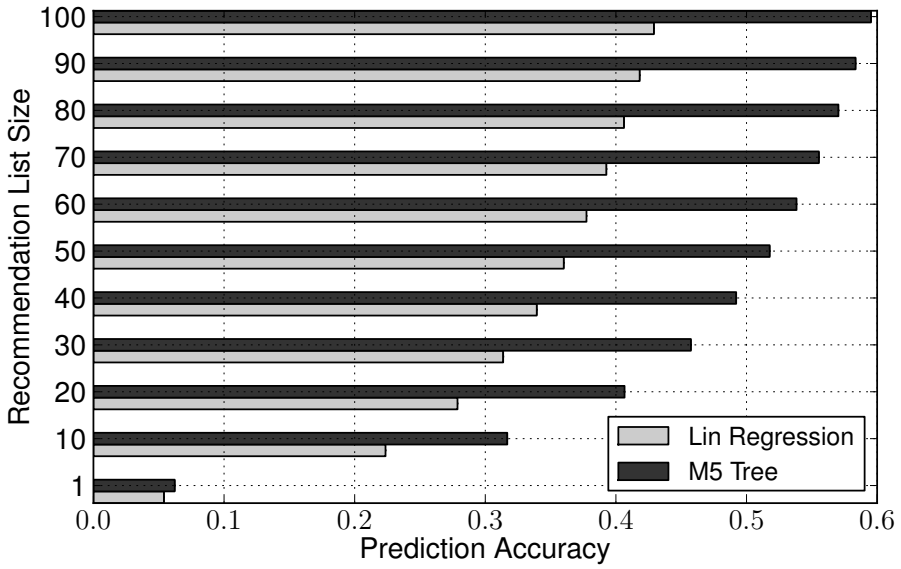


Figure 4.9: Average accuracy obtained by the supervised learning algorithms using linear regression and M5 model trees for different recommendation list sizes.

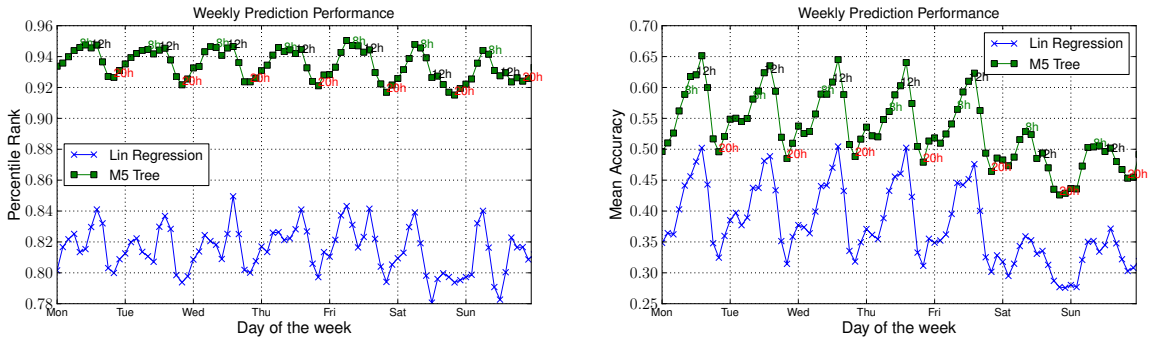
outputs, formally

$$\min_{\mathbf{w}} \|\mathbf{x}^{\mathsf{T}}\mathbf{w} - y\|^2 + \gamma\|\mathbf{w}\|^2$$

with γ being the regularisation parameter set here equal to 10^{-8}, its default value in WEKA and close to optimal in cross-validation testing. The M5 model tree is an approach based on continuous decision-tree learning [Qui92]. The principal advantage offered by M5 trees is their ability to produce continuous numerical outputs, rather than binary categories as usually inferred by decision trees. This is desirable in the present context as we want to rank venues according to real valued scores. That is achieved by creating a decision tree which splits learning instances according to their features: on each leaf, a subset of the features contained in $\mathbf{x}$ is used in a linear regression model to output a numeric score. Typically, nodes in decision trees employ a threshold on a certain feature to split the training set T. Unlike other decision trees where the information gain criterion is used to choose the attribute on which to split, M5 model trees split on the attribute which maximises the expected error reduction (i.e., the attribute that yields the most *homogeneous* branches upon splitting). Formally, the standard deviation reduction (SDR) is defined as

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} \times sd(T_i)$$

where T is the set of training examples that reach the node and $T_1, T_2...$ are the sets that result from splitting the node to the chosen attribute and $sd(T)$ returns the standard deviation for the set of instances belonging to training set T.



(a) Average Percentile Rank (APR)

(b) Average accuracy

Figure 4.10: Temporal evolution of APR scores (a) and average accuracy (b) obtained by the supervised learning algorithms using linear regression and M5 model trees.

4.4.3 Results

We are now presenting the prediction results obtained when we train and test the two supervised learning models. The M5 trees have the best performance across all models, with an APR of 0.94 and a clear performance margin compared to all single feature prediction strategies that achieve, at best, 0.86 when venues are ranked according to *Place Popularity*. On the other hand, the linear regression model achieves an APR score equal to 0.81 which ranks it lower than the popularity and categorical preference features. If we consider the performance of the models in terms of prediction accuracy (see Figure 4.9), we notice that M5 model trees dominate with Accuracy@10 equal to 0.31 and Accuracy@50 equal to 0.51. In the latter case, the next place visited by the user is ranked in the top 50 positions of the prediction list one out of two times on average, which is remarkable performance if one considers the sheer number of places being ranked in a city. Compared to the *Historical Visits* feature that does best in terms of accuracy, M5 model trees present consistently better performance: *Historical Visits* offer good accuracy scores which, however, reach an upper bound when prediction list size $N = 10$, whereas for larger N values no improvement is observed. As the reader may notice by inspecting Figure 4.9, M5 model trees accuracy performance ceases to increase rapidly only when $N = 100$. That means that their predictive power is not biased by a small set of candidate venues as in the cases of *Historical Visits* and *Social Filtering*. The linear model presents similar trends in terms of how its accuracy scores improve relative to list size N, but it fails to achieve high absolute scores, although it still does better than *Historical Visits* for N bigger than 50. Overall, M5 model trees attain peak performance both in APR and Accuracy terms, showing not only that a supervised approach which combines multiple features is more effective, but also that this combination is more effective in a non-linear embedding such as that of decision trees.

Finally, Figure 4.10 plots the prediction performance of the combined approaches over the

week, both using APR (Figure 4.10(a)) and prediction accuracy (Figure 4.10(b)). Model trees excel in terms of prediction accuracy (shown here for $N = 50$, with all N shown in Figure 4.9), scoring above 0.5 in general, denoting that one in two user check-ins are successfully predicted. The evolution of temporal predictability presents similar patterns to those observed for individual algorithms. In the morning and noon prediction accuracy nears 0.65, whereas at night the performance drops almost by 25%. Notably, accuracy also drops during weekends, as the *Historical Preferences* and *Place Transitions* algorithms, which also score highly in this metric also did (see Figure 4.7). This signifies that the predictability of user movements may decrease at given times, perhaps when they are more likely to deviate from their regular mobility patterns by making more randomised choices of places, as we have also indicated in Section 4.3.2 in the evaluation of single features. Interestingly, the APR scores of the supervised learning algorithms, as shown in Figure 4.10(b), present a more robust performance over time, with smaller variations on a daily basis and equally good performance during the weekend. This observation supports the idea that, although, supervised models drop in accuracy performance (as also do the corresponding features being integrated in them), they continue ranking the venues visited by mobile users relatively high in the prediction list during temporal periods such as weekends and night when users are more likely to deviate from their standard behaviour.

Distance versus rank-distance

In Chapter 3 we have shown how the *rank-distance* presents a better alternative to *Geographic distance* for modelling movement in cities. Our goal in that chapter was to model cities through a common variable, that would dissect any heteregoneity observed in the aggregate frequency distribution of movements in urban environments. Further, as discussed in Section 4.2.2, the two features yield the same scores since *Geographic Distance* becomes effectively equal to *Rank Distance* in the venue ranking task. Given these observations, we delved deeper into the comparison of the two features (variables) and we measured their relative importance when integrated in machine learning models.

Learning algorithms typically make use of an optimisation criterion in order to assess the relevance of a given feature. A very common metric established in information theory and is employed very often in decision tree models is the Information Gain (also known as Mutual Information) [Qui86]. The variable X with the highest information gain $IG(X, \alpha)$ for a value α is elected to generate a new branch in the tree. In Table 4.3, we present the average information gain values measures for the training sets available by the 34 cities we evaluate. As we observe the two distance features are correlated, however when they are put together *Rank Distance* appears stronger suggesting that the relative density between two places l_i and l_k may be a more informative measure compared to their absolute distance. While difference in the absolute values are apparent across cities we

note that *Rank Distance* was ranked higher than *Geographic Distance* in all cases.

Feature	Average Information Gain	Standard Deviation
Popularity	0.310	0.030
Place Hour	0.225	0.032
Rank Distance	0.223	0.043
Historical Visits	0.204	0.031
Geographic Distance	0.204	0.045
Place Day	0.175	0.031
Social Filtering	0.121	0.043
Place Transition	0.100	0.023
Categorical Preference	0.090	0.015
Category Day	0.054	0.021
Category Hour	0.027	0.013
Activity Transition	0.026	0.022

Table 4.3: Average Information Gain scores for each feature and Standard Deviations measured across the 34 cities in the dataset.

4.5 Discussion and implications

Mobile location-based services present new challenges, as they reveal not only where users are but also an additional layer of information about the physical places they visit. Thus, service providers can now access data about the multitude of factors that may influence users when deciding which place to visit, ranging from personal interests, social influence, spatial proximity, and temporal context. This makes it possible to extend existing techniques beyond the prediction of spatial trajectories, computing instead the exact place a user will visit. However, together with the new opportunities offered by the additional information layers included in these data, also come challenges.

First, predictability can be more difficult, as prediction algorithms need to be more precise to compute the exact venue a user will be at - amongst thousands - instead of generic geographic positions. Our approach has explored this trade-off: thanks to the extensive data available about users' movements, many different predictive features can be defined and mined to compute how likely a user is to visit any given place. Yet, each single feature offers only a limited window on user behaviour and, thus, is not able to provide a single good answer to the prediction task. We have shown that an effective way to address this problem is to train supervised models that can exploit the combined power of multiple features. Furthermore, due to the extreme sparsity in user check-in data, the training set built for the supervised learning task requires the combination of data mined from multiple users in the city.

Second, as our analysis has demonstrated (Section 4.1), most users in these systems have few check-ins. Being able to keep those users engaged is a key issue in location-based services. Many of the features we have mined in the present work that do not employ user specific information could be exploited to improve recommendations or content delivered to this particular class of users. This strategy effectively corresponds to a solution of the cold start problem [SPUP02] in the context of mobile place recommendations.

Third, our observations highlight that not only does user predictability change over time, but also that the way different factors drive user mobility may have temporal variations. For instance, users tend to move towards more geographically proximate venues over night and they are less likely to visit their historically observed venues during the weekend. This has two important consequences. First, new models that capture and reproduce mobile user behaviour need explicitly to include and exploit these temporal variations. Second, service providers and application developers who aim to offer place recommendation, or any other system that benefits from foreseeing future places visited by users, have to take into account that different facets of user behaviour dynamically and heterogeneously influence users' movements. The supervised learning models presented in this chapter combine features in a static temporal representation (all features are incorporated with the same *weighting* over time). If time were to be incorporated in this context, for example by building different models for different hours of the week, then larger quantities of training data would be required. While this was not feasible during informal experiments conducted using the present Foursquare dataset, future efforts for the creation of more dynamic models in light of better data are not to be ruled out.

Finally, it is important to note our strategy has been to train machine learning models in an off-line manner and subsequently test them on-line on a per user check-in basis. This represents a scheme that guarantees quick computational responses by processing only input of a handful of features encoding information about past user check-ins and the current place of the user. This is because the next check-in prediction scenario requires a quick and dynamic computational response in terms of computing venue rankings in real time. Approaches such as matrix factorisation and random walk models are not specifically tailored for on-line scenarios and in this context could hurt the quality of service provided to the user. We will see however how these families of algorithms perform in a more static prediction setting during Chapter 5 when we attempt to predict the new venues visited by mobile users in future temporal periods.

4.6 Related work

Even though location-based services have only recently enjoyed mainstream popularity, they have already attracted research efforts thanks to the new wealth of social and spatial data they offer [CCLS11]. In particular, the additional information coming from the

places visited by users has been successfully used to improve social link prediction systems [SNM11]. A recent work presents a mobility model that combines social and spatial factors to reproduce user movements [CML11]. The main difference between this work and our approach is that we focus on the places visited by users to extract features, while the statistical model in [CML11] ignores places and does not offer insights into the importance of different factors as space and time vary. A different approach that exploits social networking information to infer the current location of a user has been proposed in [SKB12]. In this case the authors propose a supervised learning model based on the places visited by a user's friends and they test via cross validation. Instead, our supervised models consider a much larger set of candidate places and, thanks to our longitudinal data, we train them in a more realistic prediction setting on past check-ins, testing on future movements (as opposed to a standard statistical cross-validation performed by the authors in [SKB12]). This also makes our approach suitable for prediction on new users with few or zero check-ins or friends.

Several predictive frameworks for mobile users have been designed and tested, often with the aim of forecasting future mobile traffic load. A large category of prediction frameworks are based on Markov models [NN08, LD04], while other methodologies include sequence pattern matching [MPTG09] and time series analysis [SMM+11]. An interesting related work that exploits multivariate nonlinear timeseries has recently been presented in [DDLM12]. In that work, information about the movement of friends, or generic users whose mobility is correlated to that of the user under prediction, is exploited to predict movement. The exploitation of information beyond the target user makes it possible to predict movements towards new locations similar to the present work. Also using cellular data, the authors in [IBC+12] model human mobility at a metropolitan scale. It should be highlighted that in both [DDLM12] and [IBC+12] mobility prediction takes place at the level of geographic coordinates (ie, latitude and longitude values) and not in terms of exact venues as the present work.

While the approaches mentioned above only focus on location prediction, more recent work takes advantage of the user-generated knowledge about places to build location and activity recommender systems [ZZXY10]. The problem of recommending places and events to mobile phone users has also been investigated, adopting predictive features such as place popularity and geographic distance [FCSP06, QLC+10]. Our problem is different to spatial mobility prediction, as we focus on places and not on spatial areas. Furthermore, we adopt a supervised learning framework rather than probabilistic models, leveraging the large amount of data available on location-based services to learn complex patterns. This allows us to focus our attention on the distinctive predictive power of social, local, individual and temporal data.

4.7 Summary

In Section 2 we discussed how urban planners and transportation modellers have traditionally posed the problem of predicting mobility flows between areas in a city or between cities in a country depending on the geographic scale being investigated. As noted in Section 2.3.1, while these traditionally posed problems bear similarities to the problem of mobile venue recommendations, the latter requires the development of user centric, personalised approaches.

The challenge in the context of mobile venue recommendation however, is the extremely sparse data available on a per user basis in these systems. This difficulty becomes more apparent in the next place recommendation problem since a single venue has to be predicted considering thousands of places in a city that a mobile user could choose to check in to. Driven by our findings in the analysis we have performed on location-based social network data in this dissertation, but also on conclusions from the literature on human movement, we have mined numerous *signals* about user venue preference in Foursquare. Subsequently we have presented a learning strategy that exploits venue preferences by user collectives in order to integrate mobility features in supervised learning frameworks. Despite the inherent challenges, the unified approach has yielded prediction scores that are relevant for mobile applications that require knowledge of visit patterns of individual users. The temporal perspective put on the prediction task and the subsequent results suggest how our modelling efforts in this chapter present only the first step towards more sophisticated dynamic models that are able to balance the factors driving user movements in a temporally aware manner.

5

New venue discovery in the city

In the previous chapter we developed a host of machine learning features and a supervised learning framework for the prediction of the next place to be visited by a user in location-based social networks. The prediction space in that case involved a mixture of historically visited and unvisited venues by a target user. The accuracy scores of the algorithms have, however, hinted that a large fraction of visits occurs at new places; specifically, in Section 4.3 we observed that historical knowledge about a user's profile can, on average, allow the correct prediction of only a third of check-ins. Further, in typical recommender systems settings the goal has been the prediction and subsequent recommendation of *new items* to users which could be, to name a couple of well known examples, new movies or new products, in the cases of Netflix [BL07] or Amazon respectively. The aim there has been to foster user exploration into a space of new items, offering this way new experiences to the individual, but also potentially large return of investment to the merchant. It would therefore be of interest for researchers and practitioners, in the space of recommender systems and mobile applications, to see whether prediction algorithms that have worked successfully in the online setting can be *migrated* to the geographic domain gracefully and, moreover, what would be the technical prerequisites and performance implications upon this transition.

The most prominent problem that challenges building a recommender system in this setting is that the relationship between check-in, social, and spatial data—in terms of understanding how these properties relate to people discovering new places to visit—remains unclear. This has two implications: first, while recommender systems have been proven to excel in web settings [AT05], they have historically operated with ordinal rating

data where spatial properties tend not to matter and users have the ability to provide negative feedback. Instead, check-in data only counts users' visits to venues, which are also inherently spread over geographic space. Second, recommender systems have traditionally operated under the sole assumption of like-mindedness (i.e., historically similar users will continue to have shared preferences). Instead, there are a wide range of reasons why mobile users may want to visit a place (e.g., visiting friends, attending an event, touring culturally significant locations); applying the state-of-the-art in web recommendation to this new context will inevitably exclude a host of features that this data contains.

Chapter Outline In this chapter, we tackle the problem of building a recommender system for previously unvisited venues from behavioural, social, and spatial data. To do so, we seek to answer the following questions:

- **How often do people tend to visit new places?** In Section 5.1, we analyse two datasets from check-in services. We discover that between 60-80% of users' check-ins are to venues that have *not* been visited in the previous month; these datasets contain granular representations of irregular behaviour beyond daily routines.
- **What assumptions do web recommender system algorithms make about human mobility?** After formalising the *new venue recommendation* problem in Section 5.2, in Section 5.2.2 we describe a host of algorithms—ranging from content-based, social, and collaborative filtering (with neighbourhood and latent space models)—that have been used to build web recommender systems. We demonstrate that each method has a unique underlying assumption about how people move, which necessarily excludes alternative information signals when computing recommendations. Furthermore, we show that none of these methods outperform a simple popularity baseline.
- **How can recommendation quality be improved by combining the different sources of data?** In Section 5.3 we propose a generalisable method based on a personalised random walk with restart on a user-place network. It seamlessly and simultaneously combines all the available signals into a high-dimensional graph: such structure takes into account the variety of means through which users are exposed to new venues.

Finally, through an extensive evaluation, we discuss how our approach based on random walks obtains between 5 and 18% improvement over those machine learning algorithms used in web contexts (Section 5.4).

In the sections to follow, we first describe the publicly available check-in data that we collected for 11 cities across the world (Section 5.1.1) in two location-based social networks. The nature of the problem we are about to tackle has allowed for the employment of an

additional dataset sourced from the erstwhile rival of Foursquare in the location-based service arena, Gowalla [1]. We analyse the properties of our data sets and we investigate to what extent users visit new places when they use these location-based services. Our main finding is that a large fraction of the visited places are *new* places, which highlights the importance of offering high quality recommendations of new venues in those systems. (Section 5.1.3).

5.1 New venue mobility analysis

Next, we begin with the description, in Section 5.1.1 of the additional dataset that we employ specifically for the purposes of the present chapter, Gowalla[2]. Then in Sections 5.1.2 we formalise necessary notation for the next paragraphs and, finally, in Section 5.1.3 we motivate our approach by performing a thorough analysis on the ways users in location-based social networks visit new places over monthly temporal periods.

5.1.1 Dataset Description

The check-in data we employ in this chapter spans two different popular location-based services: Foursquare and Gowalla. We restrict our analysis to the 11 most popular cities across both services: this allows us to (a) focus on where these services are most used and (b) restrict our prediction space to areas with the highest venue availability, which maximises the number of candidate venues that can be recommended. Since, we have already introduced the Foursquare service in previous chapters (see Section 3.1.1) we focus on the description of Gowalla.

Gowalla is a location-based social service created in 2009, which has been discontinued since its acquisition by Facebook at the end of 2011. The Gowalla dataset is a complete snapshot of the service obtained in August 2010, collected via the public API. The entire dataset contains 12,846,151 check-ins made by 216,734 active users, that is, users with at least one check-in made since they joined the service; these check-ins took place across 1,421,262 million venues over about 18 months. It also contains all social links between users, which amounts to 736,778 friendships.

Each check-in contains the following fields: the unique user id, the date with accuracy limited to the day of the check-in, its geographic location encoded as latitude and longitude coordinates and the venue's category. In order to assign places from the two venue databases to a specific city we have followed the following methodology. Foursquare venues specify locality information (city, province, street), available through the service's

[1] en.wikipedia.org/wiki/Gowalla

[2] The Gowalla dataset does not contain information about time or sequence of check-ins and hence was not suitable for the analysis conducted in the previous chapters.

City	N	M	C	$\langle c_u \rangle$	$\langle c_l \rangle$
Austin	2144	3758	15665	7.3	4.2
Boston	3830	2763	14730	3.8	5.3
Dallas	2418	3338	13779	5.7	4.1
Denver	2097	2342	10402	5.0	4.4
London	7242	6609	24778	3.4	3.7
Los Angeles	8178	6918	32025	3.9	4.6
New York	16131	16554	93309	5.8	5.6
Paris	3091	4345	13086	4.2	3.0
San Francisco	6493	6478	31070	4.8	4.8
Seattle	3493	4398	20128	5.8	4.6
Seoul	9491	4284	35540	3.7	8.3

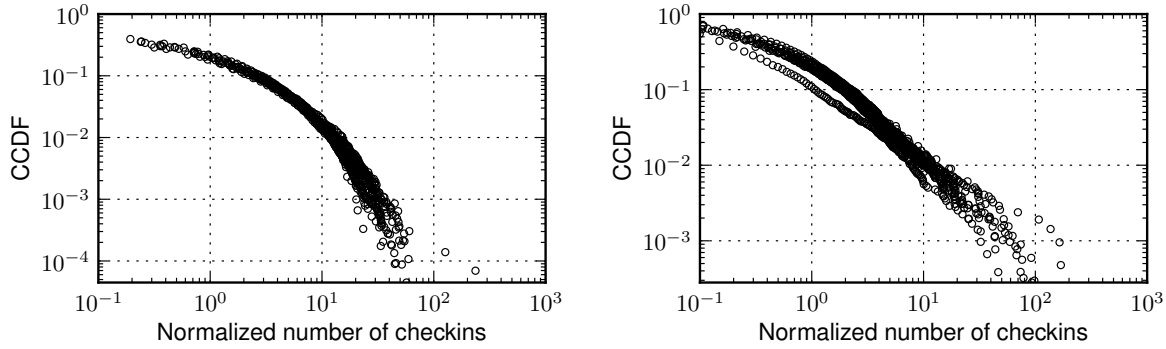
Table 5.1: Average properties observed in Foursquare over a period of one month: total number of users (N), places (M) and check-ins C, average number of check-ins per user ($\langle c_u \rangle$) and per place ($\langle c_l \rangle$).

City	N	M	C	$\langle c_u \rangle$	$\langle c_l \rangle$
Austin	4008	13110	76151	19.0	5.8
Boston	1050	5214	15026	14.3	2.9
Dallas	3494	14586	52403	15.0	3.6
Denver	1139	3188	10219	9.0	3.2
London	2139	13510	39696	18.6	2.9
Los Angeles	2868	12172	34659	12.1	2.8
New York	2659	10903	32467	12.2	3.0
Paris	480	1875	4165	8.7	2.2
San Francisco	3199	13389	47128	14.7	3.5
Seattle	1595	8090	26538	16.6	3.3
Seoul	336	2222	3973	11.8	1.8

Table 5.2: Average properties observed in Gowalla over a period of one month: total number of users (N), places (M) and check-ins C, average number of check-ins per user ($\langle c_u \rangle$) and per place ($\langle c_l \rangle$).

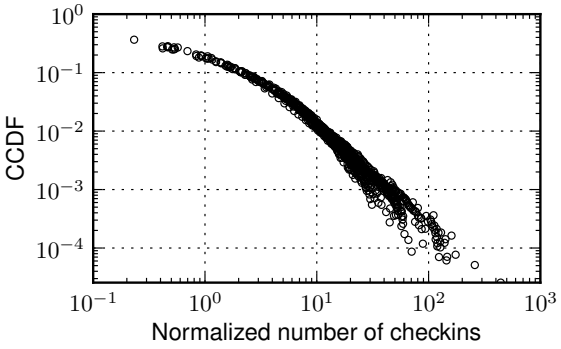
API, thus the assignment was straightforward. In the case of Gowalla, we have assigned a place to a city if it lies within 30 km of its geographic centre.[3] This procedure allows us to compare spatially similar sets of places across the two services.

[3]The geographic centre has been set according to the median latitude and longitude values across the city's places in the Foursquare dataset.

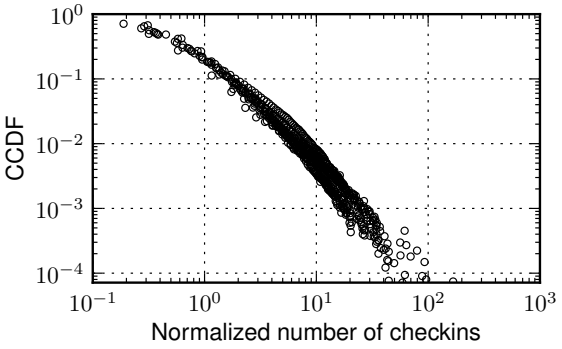


(a) Foursquare Users

(b) Gowalla Users



(c) Foursquare Places



(d) Gowalla Places

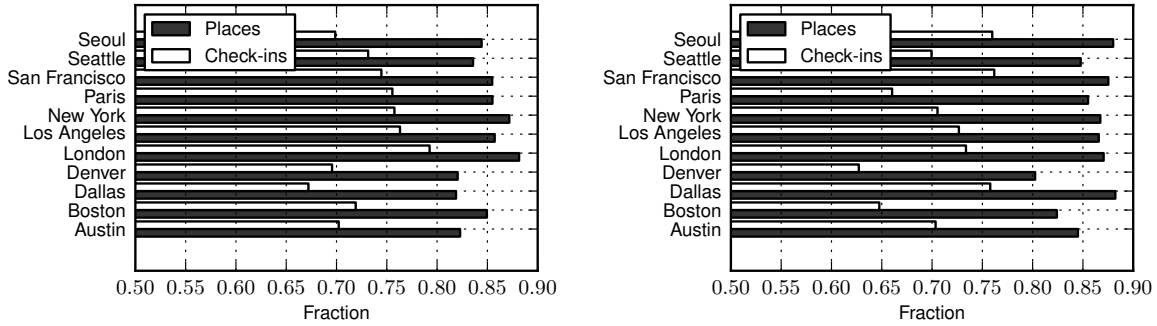
Figure 5.1: Complementary Cumulative Distribution Function of the number of check-ins per user and per-place 1-month of Foursquare (a, c) and Gowalla (b, d) data. Each distribution is normalised with respect to the average value. We consider 1 month of data, but the probability distributions do not change significantly across different snapshots.

5.1.2 Notation

We now introduce the notation that we will use in the next paragraphs. We consider a sample of check-in data over a pre-determined temporal period. Each such temporal *snapshot* t contains a set U of N of users ($N = |U|$) and a set L of M places ($M = |L|$), with each place belonging to a category crowdsourced by the services' users. We represent by c_{ij} the number of check-ins that user i has made at place j. The entirety of a user's check-ins in the sample are represented by the vector $\vec{c_i} = (c_{i1}, c_{i2}, \ldots, c_{iM})$. We use Φ_j to indicate the set of users who have checked in at place j and Θ_i for the set of all places where user i has checked in.

Social ties between users are represented as an undirected graph $G = (V, E)$, with the set of nodes $V = U$ and the set of edges E composed of pairs of users who are present in each other's friend lists in snapshot t. We denote with Γ_i the set of users connected to user i in graph G, with $|\Gamma_i|$ being the number of friends of i in the snapshot.

5.1.3 The Importance of New Places



(a) Foursquare

(b) Gowalla

Figure 5.2: Fraction of visited places that were not visited in the last 30 days and fraction of check-ins to such new places, for cities in Foursquare (a) and Gowalla (b).

We now examine a number of properties that the datasets share and study how check-ins to places are distributed. We discover that users tend to visit places they have not visited in the past: between 60% and 80% of check-ins occur at places which were not visited before by an individual user.

Tables 5.1 and 5.2 present the basic properties of the Foursquare and Gowalla datasets respectively during the same month (August, 2010). The 11 cities differ widely in terms of monthly users: Foursquare has about $22,000$ active users each month in New York but only $3,200$ in Denver; Gowalla's most popular city, instead, is Austin (where the company was launched) with about $4,000$ active users.

In general, Gowalla has a smaller number of users and places in each city compared to Foursquare, which reflects the latter's overall popularity. However, the average number of check-ins per user is higher in Gowalla: this could be due to the fact that our Foursquare dataset only contains a sample of user check-ins, namely those which were pushed to Twitter. The average number of check-ins per place, though, remains comparable across the two services.

When considering the entire temporal duration of the two datasets, 18 months in Gowalla and 5 months in Foursquare, there are about 10% of users and 25% of venues with only a single check-in in Gowalla; similarly, 20% of users and 35% of venues have a single check-in in Foursquare. This skew in the popularity of places and in user activity is reflected also in single cities. In fact, although each city exhibits different levels of user activity, the normalised distributions of check-ins across users and places are strikingly similar, as reported in Figure 5.1(a) and Figure 5.1(b). When each distribution is normalised by dividing each sample by the average value, all distributions collapse to a similar heavy-tailed pattern. In particular, about 80% of users have fewer check-ins than the average

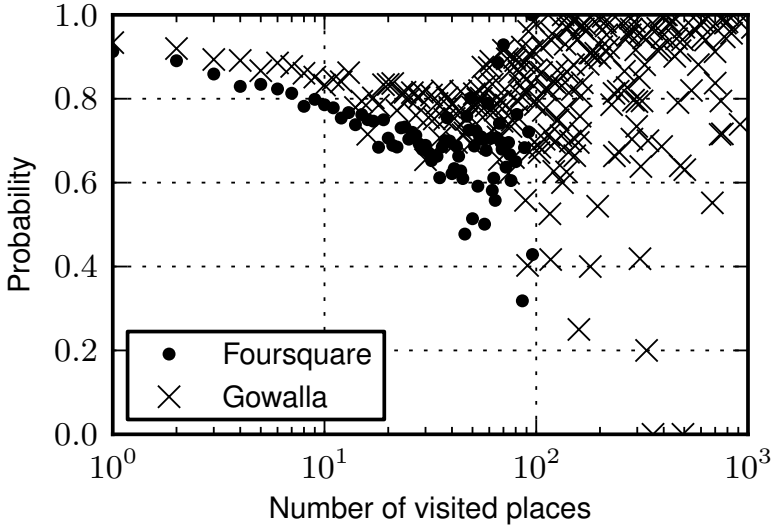


Figure 5.3: Average probability of visiting a new place as a function of the number of places visited by users. A decreasing trend can be observed: this suggests that more active users are less likely to visit a new venue. The noise at the tail appears due to the low number of users with more than 50 check-ins.

across all cities, both in Gowalla and Foursquare. Similar patterns appear when considering the normalised distribution of check-ins at each place, presented in Figure 5.1(c) and Figure 5.1(d). Overall, the tail of each distribution reaches values of up to thousands of check-ins: a bulk of users with low activity coexists with a few extremely active users.

In Figure 5.2 we consider two monthly samples of the data, taken over two consecutive months t and $t+1$; we then define as $\Psi_i^t = \Theta_i^{t+1} \setminus \Theta_i^t$ the set of new places visited by user i in sample $t+1$. Then, we compute for each city across Foursquare and Gowalla two quantities: the probability P_v that a visited venue was not previously visited

$$P_v = \frac{\sum\limits_i \|\Psi_i^t\|}{\sum\limits_i \|\Theta_i^{t+1}\|} \tag{5.1}$$

which is effectively the ratio of the sum of newly visited places by any user over the set of total set of places visited in month $t+1$. We also measure the probability P_c that a check-in takes place in one of these new places defined as:

$$P_c = \frac{\sum\limits_i \sum\limits_{j \in \Psi_i^t} c_{ij}}{\sum\limits_i \sum\limits_{j \in \Theta_i^{t+1}} c_{ij}} \tag{5.2}$$

which in plain words translates to the ratio of the sum of all check-ins in new venues over the sum of check-ins that occurred at any venue during month $t+1$.

Between 80% and 90% of visited places are *new* places, while between 60% and 80% of check-ins happen at these new venues. This demonstrates how *recommending new,*

unvisited places to users has a pivotal value, as they often seek to discover new locations. More in detail, we explore how this probability changes for users with a different number of visited places in Figure 5.3. Users who have a history of 10 visited places or less, have 80% probability of visiting a new place. As we consider more active users, who have checked in to several places over the last 30 days, this fractions drops significantly, yet it remains relatively high even for these active users.

We proceed next by formalising the task of recommending new venues to users as a prediction problem.

5.2 New venue recommendation

We begin by introducing the problem of new venue recommendation (Section 5.2.1). We then describe a number of algorithms that are suitable for the task, with a particular focus on the assumptions that they make about human mobility (Section 5.2.2): popularity baselines capture herding behaviour; content-based filters assume that people will only be interested in a small set of venue categories; nearest-neighbour and matrix factorisation-based collaborative filtering compute recommendations under the like-mindedness assumption; social filters model users exclusively based on their friends; lastly, spatial-filtering, by pruning candidates on physical distance, is tailored towards those who will not venture outside a pre-defined geographic space.

5.2.1 Problem Formulation

We formally define the *new venue recommendation problem* as follows: given a sample of check-in data taken over a time period t, a set of users U and their check-ins across a set of venues L, we aim to predict the values of the set $\Psi_i^t = \Theta_i^{t+1} \setminus \Theta_i^t$, that is, the set of *new* places visited by each user i in next time period $(t+1)$. Thus, we couple a training data set to a test data set which belongs to the following and non-overlapping temporal period. Note that we only predict check-in values for locations and users that have already previously appeared in our data at least once. A toy example for the prediction task we solve can be seen in Figure 5.4. Our imaginative user, little Amy, has check-in in a Bakery, a Hostel and a Casino, amongst others, in her first month since she joined Foursquare (resp. Gowalla). Can we then predict the new places that she would like to visit in the following month?

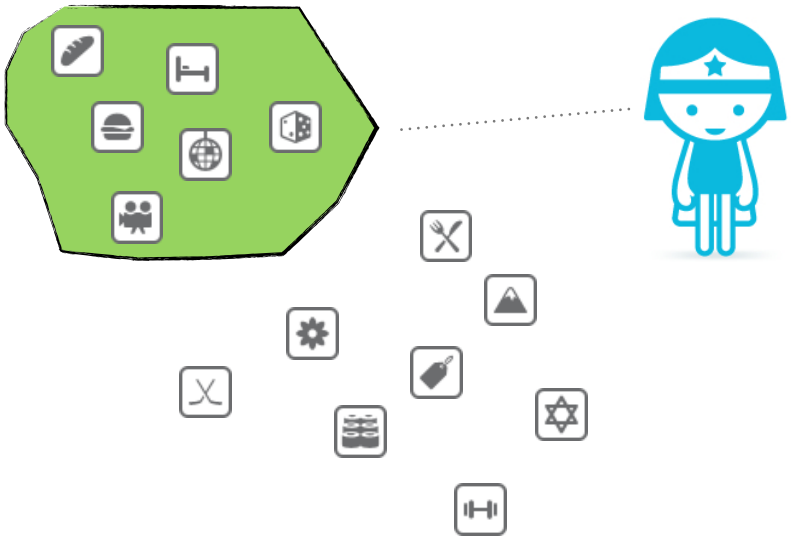


Figure 5.4: In the new venue recommendation problem, given a user and her historically observed venue preferences, we are aiming to predict the new places she would be willing to visit during a future temporal period.

5.2.2 Recommendation Strategies and their assumptions about human movement.

We now describe the set of algorithms that we examined for the new venue prediction problem; the following section will outline our random walk-based method.

Visiting Popular Venues The first (non-personalised) baseline ranks each user's unvisited venues based on their historical popularity: the `popularity` score $\hat{r}_k$ of place k is computed as:

$$\hat{r}_k = \sum_{i \in U} c_{ik} \tag{5.3}$$

In doing so, this method assumes that the likelihood of checking in is proportional to how many people have checked in before; *users will check in at the most popular places.* Recall, in fact, that Figure 5.1 showed the highly skewed distributions of user check-ins at venues; there are a few venues that receive the majority of the check-ins, while many places remain relatively unvisited.

Venue Category preferences drive User Mobility The next method is a content-based filtering approach [PB07]. The Foursquare data contains 313 place categories[4], whereas the Gowalla data contains 293. Given a user, we rank all the categories based on the number of check-ins made by the user in venues of each kind. Then, we populate a

[4]https://developer.foursquare.com/docs/venues/categories

list of recommendations by ranking unvisited venues according to their category; within-category venues are further ordered by their popularity. The underlying assumption is that *user preferences can be captured in a succinct group of categories.* This method further differs from global popularity by taking a first step into learning from user preferences: for example, users who frequently visits coffee shops will be recommended the most popular coffee shop, rather than the most popular venue in the city.

Following Friends The availability of users' social ties allows for the possibility of recommending venues visited by friends. The social filtering approach we consider ranks venues by summing the number of check-ins performed by a user's friends at each place. Formally, the `socialnet` score for a user venue pair is:

$$\hat{r}_{ik} = \sum_{j \in \Gamma_i} c_{jk} \tag{5.4}$$

which operates solely on user i's set of friends Γ_i check-ins to place k. This approach is based on the assumption that *users will exclusively visit the places visited by their friends* and builds on research exploring the interplay between human mobility and social factors[CML11, SNLM11, WPS+11]; the discovery of new events will thus propagate socially.

Staying Close to Home Previous work has suggested that the *home* location of a user may constitute a good predictor of mobility and social event attendance [QLC+10]. Since we do not know the exact location where users live, we set their "home" location to the venue where they check in most frequently; we then rank potential new venues at increasing `distance` from the identified home. Although we may not infer their actual home location, this method assumes that *capturing the locality that users tend to frequently visit will increase the likelihood of finding new venues.* In other words, people will go to places near those that they already visit.

Like-Mindedness and Similarity Collaborative Filtering (CF) has, to date, been the focal point of recommender system algorithm research [AT05]. Recommendations are computed based on the assumption that *historically like-minded users will continue to have shared preferences in the future.* Users are represented as a vector of check-ins that they have historically made to places, and items are viewed as a set of check-ins by users. In other words, these techniques assume that all the important information (both relating to preference as well as spatial dependencies) will be captured in the check-in frequency data. There are three approaches that we consider here: a user-based k-Nearest Neighbour, a item-based approach (which we denote `placenet`), and matrix factorisation based on the Singular Value Decomposition of the user-venue check-in data.

User-based kNN directly compares user profiles to quantify the extent to which pairs of users check in at the same venues. We measure the similarity s_{ij} between a pair of users i and j based on the cosine similarity of each users' check-in vectors. With this similarity matrix at hand, we can compute `kNN` recommendations for each user. Given a user i we compute a prediction score $\hat{r}_{ij}$ for place j as the sum of a baseline estimate and weighted mean of normalised check-ins to that venue by similar users:

$$\hat{r}_{ij} = \frac{\bar{c_j}}{|\Phi_j|} + \frac{\sum_{n \in U} ((c_{nj} - \bar{c_n}) \times s_{in})}{\sum_{n \in U} s_{in}} \tag{5.5}$$

The baseline estimate is the average check ins to venue j ($\bar{c_j}$) divided by the number of unique users to have visited place j ($|\Theta_j|$); i.e., the average number of check-ins per user to that venue. Neighbour check-ins are first normalised by subtracting each user's mean check-ins, $(c_{nj} - \bar{c_n})$, and then weighted by the shared similarity with user i.

An alternative neighbour-based approach is to compute similarity across pairs of venues (rather than users) [SKKR01]: this variation captures the complementary assumption that *places are similar if visited by the same users*. To model this idea in the context of new venue recommendations, our aim is to *connect* two places when they are visited by the same users and assign a weight to this connection by considering the number of distinct users that visit both. We thus form a graph, which we call the `placenet`, whose nodes are places and the edge weight p_{jk} between places j and k is defined as:

$$p_{jk} = |\Phi_j \cap \Phi_k| \tag{5.6}$$

This graph allows us to rank a place j according to the sum of the weights that connect it to the set of places visited by a user i:

$$\hat{r}_{ij} = \sum_{k \in \Theta_i} p_{kj} \tag{5.7}$$

Finally, we also examine the effectiveness of using a CF algorithm based on a latent factor model `MF`. We represent the relationship between users and places as a matrix R, whose dimensionality is $N \times M$: that is, each row represents a user and each column represents a place, with $r_{ij} = c_{ij}$. This method maps both users and places to a joint latent factor space of dimensionality $F \ll N, M$, such that check-ins are modelled as inner products between vectors in that space. User i is associated with a row vector $\mathbf{p}_i \in \mathbb{R}^F$ and place j is associated with a column vector $\mathbf{q}_j \in \mathbb{R}^F$. The estimate for the number of check-ins made by user i at place j is thus $\hat{r}_{ij} = \mathbf{p}_i \mathbf{q}_j$. We learn the mapping from users and places to latent vectors by minimising the regularised squared error E over all the existing check-ins:

$$E = \sum_{i \in U} \sum_{j \in \Theta_i} (c_{ij} - \mathbf{p}_i \mathbf{q}_j)^2 + \lambda(\|\mathbf{p}_i\|^2 + \|\mathbf{q}_j\|^2) \tag{5.8}$$

where the constant λ regularises the learned parameters, whose magnitudes are penalised. We adopt a stochastic gradient descent optimisation algorithm to minimise the error [Fun06]. In our implementation we set $F = 20$, since we have found this value to be a reasonable trade-off between scalability and accuracy: higher values of F provide only diminishing returns.

5.3 A Random walk around the city

Each method that we have presented above leverages one unique aspect of the data: CF approaches capture like-mindedness and venue similarity, social filtering computes on friends' data, and spatial filtering considers solely physical distance. In this section we aim to achieve a better recommendation quality with an approach that can automatically combine each of these features: we define a network which connects places and users and we perform personalised random walks with restart to compute recommendations for individual users.

5.3.1 Random Walk Models

A random walk over a linked structure is based on the idea that the connections between items encode information able to rank them in a useful way; as the random walker jumps across the graph's nodes according to transition probabilities, it will spend a different amount of time on each node: under certain assumptions, the random walk will approach a steady-state, resulting in a vector of steady-state probabilities for each node. These probabilities represent the desired output of a random walk model and are a function of both the structure of the network and of the transition probabilities assigned to links. A notable example in this domain is the use of PageRank [PBMW99] to rank Web pages. Personalised versions of PageRank have been designed in order to alter the ranking according to other factors, such as the topic of a page [Hav02] or users' preferences. More generally, a *random walk with restart* can be adopted to personalise rankings: at any step there is a constant probability of jumping back to a target node, thus nodes that are closer to the target tend to be ranked more highly than distant nodes, providing a personalised view of the network [TFP06].

In a random walk over a network, the transition probabilities can be arranged in a matrix $Q = \alpha W + (1 - \alpha)R$, formed by two factors, a structural one and a random one: W encodes the transition probabilities according to the network structure, while R models a random probability of jumping to any other node. The parameter α is used to tune the behaviour[5]. The steady-state probability of node i is p_i, and the steady-state probability vector $\mathbf{p}$ can be defined as the solution of the matrix equation $\mathbf{p} = \mathbf{p}Q$. A popular

[5]It is usually set to $\alpha = 0.85$ [PBMW99].

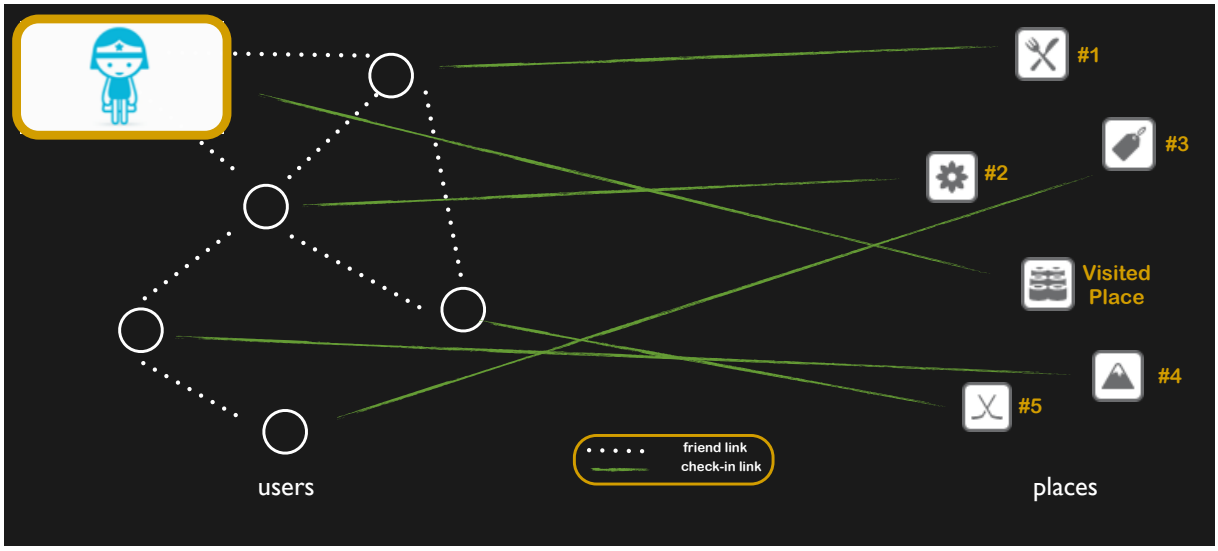


Figure 5.5: Visual representation of the graph formed by users and venues in a city. A random walk with restart algorithm is applied and returns a personalised recommendation list of venues for a given user.

approach to compute $\mathbf{p}$ is to repeatedly iterate this equation until the vector converges, exploiting the sparsity of Q to reduce memory requirements.

5.3.2 Recommending with Random Walks

We represent the data as an undirected graph whose nodes are users and venues. A toy representation of the graph is shown in Figure 5.5. A user i is linked to venue j if c_{ij} is non-zero; furthermore, a user is linked to another user if the pair are friends. This graph is used to define the structural transition matrix W which contains a uniform transition probability for each edge. For every user i we define a *random walk with restart*: at every step there is a constant probability of jumping back to the node of the user. In each case, the matrix R encodes the probability of randomly jumping back from every node to the node of the user. In order to compute predictions for user i we compute the steady-state probabilities of the related random walk: then we rank each place in decreasing order of steady-state probabilities. This favours places that are more connected to the user: through friends, through visited places or through any combination of factors. The restart probability maintains the random walk in the user's neighbourhood, thus biasing recommendation results towards venues that are more connected, in any sense, to the user. This simultaneously promotes places with several connections (i.e., popular) that are also reachable through friends and through already visited places. This feature is referred as `rwr`.

5.3.3 Weighted version

We also introduce a *weighted and directed* version of the random walk approach, denoted as `wrwr`, where each link is annotated with a weight that biases the transition probabilities, rather than having uniform probabilities on all the links going out from a node. Further, weights between two nodes can be different in opposite directions. A link from user i to user j is weighted as $\frac{1}{|\Gamma_i|}$; each friend of user i is given an equal weight, inversely proportional to the total number of friends. A link from user i to place k is weighted as $\frac{c_{ik}}{\|\Theta_i\|}$, or proportionally to the user's check ins to that venue over the total number of check-ins for that user. Finally, a link from place k to user i is weighted as $\frac{c_{ik}}{\|\Phi_k\|}$, or the user's check-in frequency at that place over the total number of check-ins for that place. All weights are normalised so that they represent transition probabilities: this is achieved by computing the sum of all the weights on the links going out from a node and then dividing every weight by this value.

5.4 Evaluation

We now evaluate the recommendation algorithms and compare results across predictors, datasets and cities. In Section 5.4.1, we describe our experimental methodology and the three metrics that we use to evaluate recommendation quality. The results in Section 5.4.2 then show that the sole method to outperform a popularity-based baseline is the Random Walk approach; finally, we discuss the implications of our results in Section 5.5.

5.4.1 Methodology and metrics

We partition the check-in data temporally into multiple training/test splits (each consisting of 30 consecutive days) in order to obtain cross-validated results. We then filter any check-in c_{ij} from each test set if the training set has a non-zero entry for c_{ij}, i.e., if the user has already visited the venue. We note that users with no check-ins in the test set are not included in the performance evaluation. The output of each prediction algorithm is a per-user personalised ranked list of venues.

We use three metrics to quantify the quality of these recommendations. The first two, Precision@N and Recall@N, convert the outcome of each predictor into binary values: either the user will visit the top-N venues and will not visit venues ranked below N. Precision (p) and Recall (r) are measured as proportions of true positives (tp), false positives (fp), and false negatives (fn):

$$p = \frac{tp}{tp + fp}; r = \frac{tp}{tp + fn} \tag{5.9}$$

Method	APR	Precision@10	Recall@10
Random	0.500	0.000	0.003
Popular	**0.772**	**0.026**	**0.089**
Activity	**0.772**	**0.025**	**0.087**
Home	0.617	0.008	0.026
Social	0.607	0.015	0.049
kNN	0.557	0.003	0.011
PlaceNet	0.663	**0.026**	**0.077**
MF	0.719	0.004	0.014
RW	**0.783**	**0.028**	**0.094**
Weighted-RW	**0.771**	**0.025**	**0.088**

Table 5.3: Foursquare Results: Average APR, Precision@10 and Recall@10 results.

Since each algorithm outputs an ordered list, we also verify the extent to which the ranking reflects users' interests; i.e., that venues that are highly ranked are indeed those that will be more frequently visited. To do so, we first define the *interest* i that a user u has in a venue s as the proportion of times that the user checks into that venue during the test period. We then define $rank_{u,s}$ as the percentile ranking of venue s for user u in the ranked list of venues; if $rank_{u,s} = 1$, then the venue appears first in the list, while $rank_{u,s} = 0$ implies that the venue was the last in the list. We combine these with each user's interest in the station $interest_{u,s}$ and average the results to measure the Average Percentile Ranking (APR):

$$\overline{APR} = \frac{\sum_{u \in U} \sum_{s \in L} i_{u,s} \times rank_{u,s}}{\sum_{u \in U} \sum_{s \in L} i_{u,s}} \tag{5.10}$$

In the following sections, we report and discuss the empirical results we obtained following the above methodology and metrics.

5.4.2 Results

In order to put the following results into an appropriate context, we also compare them to a random predictor, which simply shuffles the candidate set of unvisited venues for each user. In this case, APR results are 0.5 and both Precision and Recall are near zero. We further note that better results are obtained with *higher* APR, Precision and Recall values.

Method	APR	Precision@10	Recall@10
Random	0.500	0.002	0.001
Popular	**0.722**	**0.043**	**0.090**
Activity	**0.720**	0.032	0.073
Home	0.660	0.023	0.042
Social	0.582	0.029	0.054
kNN	0.574	0.005	0.012
PlaceNet	0.662	**0.043**	**0.077**
MF	0.657	0.009	0.025
RW	**0.768**	**0.048**	**0.095**
Weighted-RW	**0.756**	**0.045**	**0.095**

Table 5.4: Gowalla Results: Average APR, Precision@10 and Recall@10 results

Performance across methods

Tables 5.3 and 5.4 show the APR, Precision@10 and Recall@10 for the Foursquare and Gowalla datasets, respectively. The most eminent result is that nearly all methods, including social filtering and all (`kNN` and `MF`) versions of collaborative filtering—which were supposed to better model users' preferences— fail to outperform the popularity-based baseline. The `Activity` predictor, which ranks venues based on the categories of the venues visited by each user, also ranks amongst the top performing approaches. The random walk variants are the only approaches that outperform popularity, and are amongst the top performing methods for both datasets: `rwr` achieves an improvement of 5% in Foursquare and of 18% in Gowalla with respect to `popularity`, the best performing among the other methods.

In general, the three metrics agree with each other in terms of algorithms' relative performance, with one exception. When considering Precision and Recall, both `placenet` and, to a lesser extent, `socialnet` achieve results similar to the best four methods. This difference between APR and Precision/Recall is due to the fact that `placenet` and `socialnet` are the only two methods that do not rank all the available places, but only a subset of places specific to the target user. As their recommendation lists may thus contain fewer items, they are penalised in the APR score, which is agnostic to list size, but they benefit in Precision and Recall, where list size is important.

Performance across cities

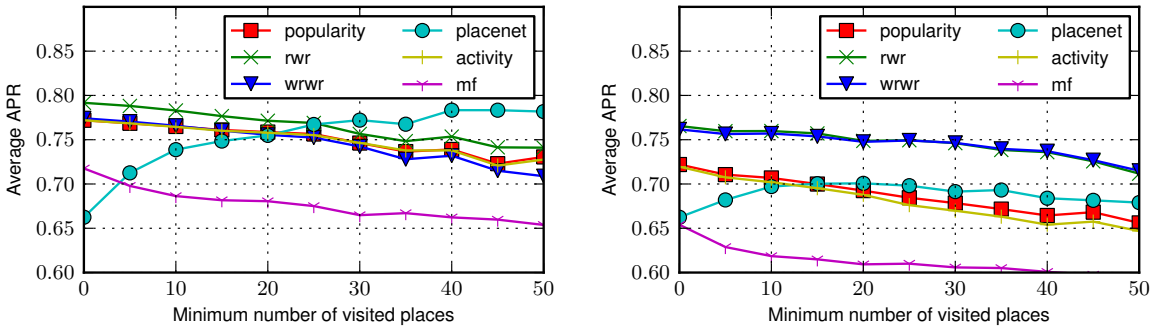
Although there are performance variations across cities, our method based on a random walk consistently outperforms the other approaches. This observation is reflected in Table 5.5, which presents the APR scores for `popularity` and `rwr` achieved across different cities

	Foursquare		Gowalla	
City	`popularity`	`rwr`	`popularity`	`rwr`
Austin	0.765	0.778	**0.825**	**0.856**
Boston	0.796	0.804	0.687	0.747
Dallas	0.753	0.768	0.752	0.802
Denver	0.767	0.800	0.715	0.764
London	0.736	0.738	0.689	0.756
Los Angeles	0.788	0.804	0.719	0.758
New York	**0.808**	**0.815**	0.720	0.758
Paris	0.735	0.744	0.729	0.796
San Francisco	0.792	0.800	0.780	0.817
Seattle	0.762	0.782	0.736	0.774
Seoul	0.790	0.774	0.590	0.619
Average	0.772	0.783	0.722	0.768

Table 5.5: APR achieved by the `popularity` and `rwr` prediction methods across cities in both datasets. For each service we highlight the city with the best APR score.

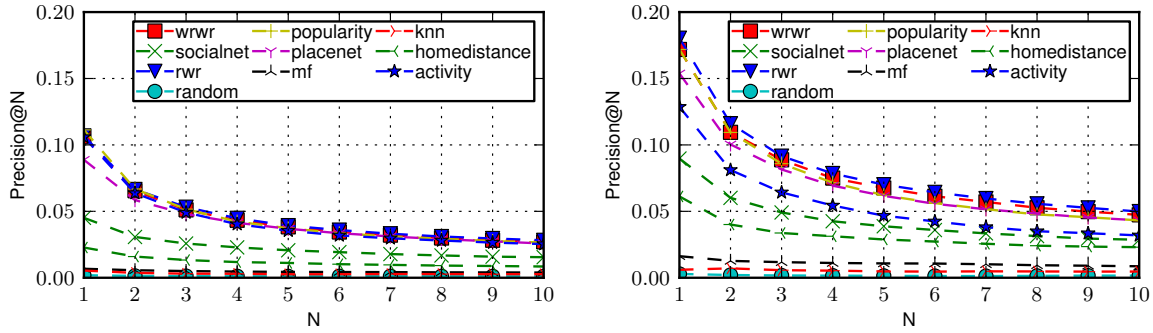
in both services; we observed a similar pattern for the Precision and Recall metrics (not shown). Moreover, this analysis suggests that there is no strong correlation between the individual city statistics presented in Tables 5.1 and 5.2, such as the number of active users and places, and the prediction performance. One outlier seems to be Seoul in Gowalla, with much lower performance: this might be due to the fact that this case has the lowest level of user activity across all the considered cities. On the other hand, New York in Foursquare and Austin in Gowalla, which are the cities with most check-ins in the corresponding datasets, show the best performance.

Furthermore, the results obtained across the two datasets agree with one another; this is a notable result since both systems have different interfaces and incentives for user participation. We must note that no location recommendation engine was put in place by either service during the data collection period. The most prominent difference in performance is obtained with the `home distance` feature, that is consistently better in Gowalla than in Foursquare across all the performance metrics. This could be due to the fact that the average number of user check-ins in Gowalla is higher than that observed in Foursquare, thus allowing the "home" location inference to be more accurate. In addition, in Gowalla, when the entirety of user social links and check-ins are present, the random walk models achieve a larger performance gain, as they are able to exploit higher quality data to build the network structure.



(a) Foursquare (b) Gowalla

Figure 5.6: Average APR of the best approaches for users with a different minimum number of visited places in the training snapshot in Foursquare (a) and Gowalla (b).

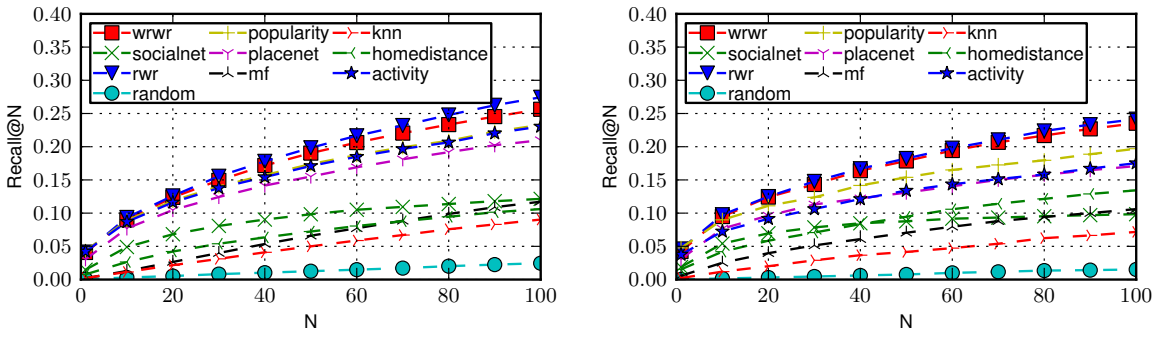


(a) Foursquare (b) Gowalla

Figure 5.7: Average Precision@N obtained by each algorithm on all cities for various values of N in Foursquare (a) and Gowalla (b).

Impact of user activity

We have discussed how users who have visited more places tend to visit fewer new places, as presented in Figure 5.3. Thus, we investigate how prediction performance changes when we consider users with progressively higher amount of visited places. We filter out users who have visited less than a certain number of places in the train snapshot and we compute how the average APR over all the remaining users changes when we progressively increase the minimum number of visited places. As shown in Figure 5.6, where we depict some of the best performing methods, prediction performance decrease as we focus on the most active users. A noticeable difference is the `placenet` feature, which achieves better results when we filter out less active users, while overall its performance hardly competes with the best methods. However, as the vast majority of users have visited only a few places, any improvement for active users is not likely to impact the performance across the entire user base.



(a) Foursquare

(b) Gowalla

Figure 5.8: Average Recall@N obtained by each algorithm on all cities for various values of N in Foursquare (a) and Gowalla (b).

Impact of recommendation list size

Finally, we explore the effect of a varying recommendation list size on the final performance. As the number of venues recommended to a user increases Recall tends to improve, while Precision might suffer. In Figure 5.7 we plot the average Precision@N obtained by each algorithm on all cities: we find that Precision decreases as we increase list size, with the different features maintaining their relative ranking in performance. Similarly but with an opposite trend, Recall@N quickly increases as the recommendation list grows larger, as depicted in Figure 5.8; again, the dominating features outperform the others over the entire range of list size. This analysis highlights the trade-off between Precision and Recall that each feature faces: real systems should tune their results according to what users require.

5.5 Discussion and implications

The data that we obtained has a number of characteristics that differentiate it from the typical recommender system scenario. An important difference is that while in other scenarios users reveal their preferences through ordinal ratings, in our case check-ins only capture numeric frequencies: as a consequence, there is no negative feedback provided by users. Furthermore, the data is highly sparse, with many users and venues having a single check-in. At the same time, across both datasets there are few places with extremely high numbers of check-ins, while the majority of them have only a few user check-ins. Thus, there is a high heterogeneity across how users check in at places, with some venues reaching high levels of popularity. This may go some way towards explaining the high performance of popularity-based recommendations.

The previous section also uncovered that the worst performing algorithms for this task were the ones that are typically associated with recommender systems (i.e., nearest neigh-

bours and matrix factorisation). There are a number of reasons why this may be the case. First, check-in data may not be sufficient to fully capture users' preferences. In fact, unlike web ratings, it leans towards capturing habitual behaviour well and does not allow for negative feedback. Second, like-mindedness may not suffice to model why people visit venues; the random-walk approaches outperform standard collaborative filtering algorithms by simultaneously leveraging several sources of data, encoding them in the network structure. Another interpretation to the fact that that collaborative filtering has under performed in this context is the extreme sparsity observed in the check-in data (very few check-ins per user) as discussed in Section 5.1.3. Finally, it is worth noting that computing mobile recommendations using a random walk is not tied to the datasets employed in the present work: it is generalisable to any situation where signals of user preference can be encoded into a graph.

5.6 Related work

There is a wide range of research on the data mining algorithms that form the basis of recommender systems [AT05]. From a data mining perspective, these systems take as input a set of users' preferences, such as ratings, and aim to predict the preference values for items that have yet to be rated, and fall under the umbrella term *collaborative filtering*. In recent years, work has centred around datasets from the web, namely movies (e.g., Netflix [BL07]) and music (Yahoo Music[6]); these datasets have characteristic features (such as high sparsity) that pose challenges to the design of accurate preference estimators. Matrix factorisation has become a popular approach for collaborative recommender systems [BK07], due to its robustness in the face of sparse data; moreover, hybrid composites of predictors have recently been awarded for their ability to improve rating prediction [Kor09].

While, historically, users' ratings were considered the sole necessary input data for building recommender systems, there is increasing attention to a variety of other signals that may aid learning algorithms' accuracy. These include temporal features of the data [Kor09] and social network links between users [Gol08]. This broad approach, which relies on augmenting recommendation systems with a more granular picture of the (social, temporal, spatial) setting of the users, has been named *context-aware* recommendation [AT05]. In the domain that we investigate here, recent work has used both sensors and user-activity data to improve recommendations. For the former, GPS data has been used for location and travel recommendation [ZZXM09], and mobile phone call records have been used for social event recommendation [QLC+10]. The latter group, instead, includes using geotagged photos for itinerary recommendation, mining interesting locations, and inferring users' trips [PG09]. The data that we examine in this chapter, which is based on explicit

[6] `http://kddcup.yahoo.com`

check-ins to locations, falls into this latter category (although GPS sensor data may be used to validate the check-in).

The recent literature, overall, reflects the success of online recommender systems; it is clear that smartphones, as they gain greater traction and popularity, will be the bridge that enables recommender systems to be used in the wild as much as they are used online. However, whether algorithms that have been successful online can be applied to these new domains remains an open question: here, by investigating one facet of location recommendation, we have shown that state-of-the-art collaborative filtering can be outperformed by a hybrid model that aggregates and learns from a range of data about users.

5.7 Summary

In this chapter we have provided a framework that integrates movement and social information to recommend new venues for mobile users. Effectively, we have solved a *specialised* version of the problem encountered in Chapter 4, where our goal had been the prediction of a mixture of historically visited and not visited venues, by tailoring it appropriately in the classical new item recommendation setting.

We have seen how popular online recommender algorithms such as Matrix Factorisation or k-Nearest Neighbour approaches have failed to model accurately the preferences of mobile users, as they were not able to outperform a popularity based ranking strategy that has been the most powerful of all baselines. We have also presented a graph-based random walk with restart model which has provided superior prediction performance that remained consistent in the vast majority of cities that we tested using two datasets from Foursquare and Gowalla, respectively. These findings suggest that random walk models may be more effective in the context of mobile recommendations due to their resilience in extremely sparse data representations that are common in these systems.

Next, we summarise the findings of this and the previous chapters of this dissertation and project on directions for future work.

6

Reflections and outlook

Human mobility, due to its important role in numerous economical, societal and behavioural processes, has been a core subject of study in multiple disciplines. Social scientists, ecologists, geographers and urbanists have long been long theorising with models that aim to explain migration trends, home to work commuting patterns or resource consumption norms in ecosystems. Despite the large number of manuscripts published, there has been a lack of datasets of appropriate scale and geographic granularity to allow an extensive validation of theoretical models.

The rise of computational social science in the past three decades has brought a revolution to large scale empirical studies about the structure and dynamics of social networks. The recent introduction of geographic social networks, driven by the proliferation of the mobile web, is expected to bring advancements of a similar scale in all academic fields where human movement studies are central. The *digital breadcrumbs* that are laid on the geographic plane by millions of users every day can not only help the validation validation of proposed theories, but also constitute a novel primary resource for the development of new mobile application and services. This represents a foundation for the cities of the future; urban exploration, local search and information discovery are envisioned to be powered by data streams emerging from ubiquitous technologies deployed in the city.

In this dissertation I have taken a step forward in both the abstract modelling of human mobility and the development of frameworks that could support mobile applications. I have empirically tested the plausibility of two theories on human migration with data from 34 cities around the world. I then proposed two different application scenarios for the recommendation of venues to mobile users. In all cases, I have exploited the

rich, granular geographic representations and multiple layers of information available in datasets generated by users in location-based social networks.

6.1 Summary of contributions

A classic question in the studies of human mobility has been the measurement of the deterring role of distance in movement. While empirically-backed insights into migration patterns within or across countries had been offered in the past, there was lack of evidence concerning the role of distance in the context of urban mobility. In Chapter 3 I aimed to shed light on this problem by analysing the distribution of distances in a large set of urban centres around the world. I have seen how when two cities are compared by looking at the absolute geographic distances of trips within them, then strong heterogeneities become apparent. However, when the same data is transformed so that each trip is characterised by the relative density between the respective origin and destination, then a universal pattern emerges. This finding has led towards the devising of a new model for urban movements that reveals the important role of geography. Indeed, the representation of the human cognitive factor in the agent based model presented is common across all cities. The only difference in the model's input between any two cities is the spatial distribution of places in them, and thus, any apparent variation in movements should arise from this element.

The study of the abstract properties of human movement in terms of frequency distribution of distances has treated movement in cities as a complex system where the interaction of individuals with places gives rise to large scale properties that govern the urban system and are common across urban environments. These studies favour our deeper understanding of the process of movement in cities, but in the sphere of real mobile applications and services one needs to refine our abstraction and model mobility in a more detailed manner.

The first step in this respect was taken in Chapter 4, where our goal was the the prediction of the next place where a mobile user will check in in real time. A major challenge in this context has been to incorporate appropriately multiple signals available about movement in these systems in order to rank effectively the thousands of candidate places where a user could go to in the city. The sparse representations of user movements and preferences, in terms of number of data points available per user, has also been a catalyst in the modelling approach I have followed to solve this problem. Since the generation of a model that exploits knowledge only about a single user has been prohibitive in this context, I have resorted to a learning method which trains a prediction model using information on the check-ins of user collectives. A non-linear decision tree model has offered good results, predicting the next place to be visited by a user one in two times, greatly outperforming individual features such as historical venue preferences or venue popularity. Further, the large temporal variations in the performance of various prediction features has suggested

the need for future models that will be able to incorporate these dynamics in order to build more accurate recommender systems, or indeed to build any application that could benefit from knowledge of the exact whereabouts of mobile users over time.

Subsequently, in Chapter 5, I have treated a specialised version of the movement prediction problem where the aim has been the prediction of *new venues* to be visited by mobile users in future time periods. The formulation of this problem has been tailored according to the classical recommender system setting where *new items* are being recommended to a target user based on their past preferences. Our analysis there has shown how in location-based services users check in to new places with higher than expected probability, highlighting this way the importance for the deployment of mobile application that foster urban exploration and discovery of new venues and activities in the city. After reviewing a number of web filtering algorithms that were previously employed in the online domain, I designed new versions of them for their exploitation in the setting of mobile place recommendations. Our results have suggested that the extremely sparse representations of human movements in location-based social networks have inhibited the performance of these recommendation algorithms which have failed to outperform a simple popularity based baseline. To deal with this challenge, I have proposed an alternative random walk with restart model that seamlessly combines social and user to place preference signals in order to provide effective recommendations for all users.

6.2 Future directions

The identification of the *rank-distance* variable as a means to view movement in cities in a universal manner and the isolation of the spatial distribution of places in a city as the source of the variations observed in movements across urban environments have strong implications with respect to future studies in human mobility and behavioural studies in general. These may become more apparent if we imagine movement as a process composed of two factors; *human cognition* being represented by the probability of travelling from an origin to a destination, and *geography* being represented as the set of places and their geographic coordinates.

Primatologists and behavioural scientists have argued in the past that animal, and by extension human, mobility is driven by a spatial cognition system embedded in the hippocampal part of the brain [OB05, JWM+13]. It is thus a possibility that human navigation is governed by mental processes that have been shaped through evolution over thousands of years. In that respect, our findings in Chapter 3 that support a universal cognitive response in city movements do not come as a surprise. From an evolutionary perspective, the way the mammalian brain is wired is likely to be similar for humans who live in different countries or cities, and therefore their behavioural responses when interacting with space are expected to be similar. In terms of future work it would be in-

teresting to investigate whether this hypothesis holds for our navigation in virtual spaces, for instance when we move from one web page to another. In a recent work [WL12] that studied navigation patterns of Wikipedia web visitors it was pointed out that the probability of transiting from one Wikipedia article to another was inversely proportional to their *semantic* rank-distance. In the light of this evidence and our findings in this dissertation it would be interesting to explore the parallel of human navigation in virtual and physical spaces. Such studies could not only reveal important information about the way our brain and spatial cognition system functions, but also could help towards the design of better navigation environments perhaps both for urban and web spaces.

In addition to human cognition, our work has highlighted the important role of geography in movement. While there has been a large volume of research focusing on movement, little attention has been paid to the growth patterns of cities. In [MAB+98] the authors proposed a spatial percolation model to simulate urban growth in the United Kingdom, while in [MNR98] different classes of urban forms are being identified for cities. Data about places and urban activities that becomes available through location-based services could augment our understanding about city size and shape patterns. Figure 2.2 is indicative of the scale at which these data are available. We now have the opportunity to quantify the urban sprawl of the whole planet through the same lens. It would therefore be interesting to answer questions about the possible existence of universal patterns in the way humans organise urban settlements.

While our findings in Chapter 3 have the potential to ignite new pathways of research on human mobility in various disciplines, the implications of the two applications scenarios I have considered in Chapters 4 and 5 are more closely related to advancements in the areas of mobile computing, applied machine learning and the computer sciences in general. First, the next place prediction task described in Chapter 4 has brought forward the importance of the temporal dimension for frameworks that target the modelling of user movement in the city. We have seen that users are more likely to move over short distances during nighttime and that they will deviate from their regular visitating patterns with higher chance during the weekends. Models that explicitly take into account how different factors govern human movement over time are expected to yield more accurate predictions of user whereabouts. Besides time however, it is questionable whether all users are being influenced by the forces of attraction or repulsion to (resp. from) places in the same ways. For instance I have described how the popularity of places is a strong predictor of human movement, but it is unlikely that all users are being influenced by it in the same way. Therefore, personalisation methods that account for differences in the ways that various groups of users attach to places could be another direction to explore so as to achieve better quality recommendations through the delivery of relevant content to mobile users.

Finally, the observation that users in location-based social networks prefer to check in at new venues supports the understanding that urban exploration and the discovery of new activities in the city are an important priority for mobile users. Therefore, algorithmic

models and filtering methods that exploit effectively the multi-dimensional information signal available from user activity in mobile systems and web applications are expected to stay at the forefront of academic research in order to support applications such as mobile activity recommendations and local search. The latter is in fact a new source of competition for tech giants such us Microsoft, Google and Facebook amongst many start-ups, which currently invest a lot in areas related to geo-commerce [Rep13].

6.3 Outlook

The arrival of a new generation of mobile web services and applications has generated large amounts of mobility data of unprecedented geographic scale and spatial granularity. Moreover, mobility data is now becoming available in parallel to other layers of information including social interactions between users, natural language expressions of users or the distribution and consumption of digital content across time and space using smartphone devices. The fact that every piece of online information is now being geo-tagged brings not only new opportunities to answer important research questions or offer better services for users, but comes with challenges that concern the general case deploying computer science services on the geographic plane. It will take some time until the streams of *Big Data* are fully digested by academics, government institutions and industry, and when this happens, as history teaches, more data and more questions shall emerge.

In this dissertation, I have attempted to take a step towards a better understanding of human urban movement and the development of effective frameworks for mobile recommender systems. Besides the strictly quantitative aspects of our findings that may be volatile in light of future experimentations with new types of data or algorithmic methods, I hope that the approach I have taken in modelling movement by positioning the places of a city at the centre of the process, will inspire researchers in various academic disciplines and also practitioners in the area when they design their own models or build new applications.

Bibliography

[ASST05] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin. Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach. *ACM Transactions on Information Systems*, 2005.

[AT05] G. Adomavicius and A. Tuzhilin. Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, 2005.

[Bar05] A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 2005.

[Bec67] Martin J Beckmann. On the theory of traffic flow in networks. *Traffic Quarterly*, 1967.

[BG77] B. J. L. Berry and Q. Gillard. *The changing shape of metropolitan America: commuting patterns, urban fields, and decentralization processes, 1960-1970.* Ballinger Publishing Company Cambridge, MA, 1977.

[BK07] R. Bell and Y. Koren. Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In *IEEE International Conference on Data Mining*, 2007.

[BL07] J. Bennett and S. Lanning. The netflix prize. In *KDD Cup*, 2007.

[Blo10] Foursquare Blog. One small step for man, one giant check-in for mankind. `http://blog.foursquare.com/2010/10/22/foursquare-nasa-check-in/`, October 2010.

[Blo12] Foursquare Blog. Nasa's curiosity rover checks in on mars using foursquare. `http://www.nasa.gov/mission_pages/msl/news/msl20121003.html`, March 2012.

[BNMB13] C. Brown, A. Noulas, C. Mascolo, and V. Blondel. A place-focused model for social networks in cities. In *IEEE International Conference on Social Computing*, 2013.

[BNÓS+12] S. Bauer, A. Noulas, D. Ó Séaghdha, S. Clark, and C. Mascolo. Talking places: Modelling and analysing linguistic content in foursquare. In *IEEE International Conference on Social Computing*, 2012.

[BNS+12a] C. Brown, V. Nicosia, S. Scellato, A. Noulas, and C. Mascolo. The importance of being placefriends: discovering location-focused online communities. In *ACM Workshop on Online Social Networks*, 2012.

[BNS+12b] C. Brown, V. Nicosia, S. Scellato, A. Noulas, and C. Mascolo. Where online friends meet: Social communities in location-based networks. In *AAAI International Conference on Weblogs and Social Media*, 2012.

[BNS+13] C. Brown, V Nicosia, S. Scellato, A. Noulas, and C. Mascolo. Social and place-focused communities in location-based online social networks. In *European Physics Journal B.*, 2013.

[bri] Brightkite. http://en.wikipedia.org/wiki/Brightkite.

[BW10] L. Bettencourt and G. West. A unified theory of urban living. 2010.

[CB05] C. Cheung and J. Black. Residential location-specific travel preferences in an intervening opportunities model: Transport assessment for urban release areas. 2005.

[CBB+07] V. Colizza, A. Barrat, M. Barthelemy, A-J. Valleron, and A. Vespignani. Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *PLoS Med*, 2007.

[CCLS11] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring millions of footprints in location sharing services. In *AAAI Conference on Weblogs and Social Media*, 2011.

[CHC+07] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, 2007.

[CML11] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2011.

[Col94] F. Colombijn. *Patches of Padang: The history of an Indonesian town in the twentieth century and the use of urban space.* Research School CNWS Leiden, 1994.

[CRH11] H. Cramer, M. Rost, and L. E. Holmquist. Performing a check-in: emerging practices, norms and conflicts in location-sharing using foursquare. In *International Conference on Human Computer Interaction with Mobile Devices and Services*, 2011.

[CSHS12] J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *AAAI International Conference on Weblogs and Social Media*, 2012.

[CSN09] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-Law Distributions in Empirical Data. *SIAM Review*, 2009.

[CSS99] W. Cohen, Robert E. Schapire, and Yoram Singer. Learning to order things. *Journal of Artificial Intelligence Research*, 1999.

[CSTC12] Y. Chon, H. Shin, E. Talipov, and H. Cha. Evaluating mobility models for temporal prediction with high-granularity mobility data. In *IEEE International Conference on Pervasive Computing and Communications*, 2012.

[CTH+10] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the Gap Between Physical Location and Online Social Networks. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2010.

[DBG06] L. Hufnagel D. Brockmann and T. Geisel. Understanding individual human mobility patterns. *Nature*, 2006.

[DC10] D. Cosley S. Suri D. Huttenlocher J. Kleinberg D. Crandall, L. Backstrom. Inferring social ties from geographic coincidences. *Proceedings National Academy of Sciences*, 2010.

[DDLM12] M. De Domenico, A. Lima, and M. Musolesi. Interdependence and predictability of human mobility and social interactions. *arXiv preprint arXiv:1210.2376*, 2012.

[Dev13a] Foursquare Developer. Foodspotting food reviews by dish - and powered by foursquare's location database. `https://developer.foursquare.com/appsshowcase/foodspotting`, September 2013.

[Dev13b] Foursquare Developer. Thrillist helping users navigate cities with foursquare's location layer. `https://developer.foursquare.com/appsshowcase/thrillist`, September 2013.

[DLN05] R. Kumar P. Raghavan A. Tomkins D. Liben-Nowell, J. Novak. Geographic routing in social networks. 2005.

[DQC10] F. Calabrese G. Di Lorenzo D. Quercia, N. Lathia and J. Crowcroft. Recommending social events from mobile phone location data. 2010.

[Dun78] James Duncan. Men without property: the tramp's classification and use of urban space. *Antipode*, 1978.

[Eas93] S. M. Easa. Urban trip distribution in practice. *Journal of Transportation Engineering*, 1993.

[ECL11] S. A. Mayers E. Cho and J Leskovec. Friendship and mobility: User movement in location-based social networks. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2011.

[Eng12] Foursquare Engineering. Machine learning with large networks of people and places. `http://engineering.foursquare.com/2012/03/23/machine-learning-with-large-networks-of-people-and-places/`, March 2012.

[EP06] N. Eagle and A. S. Pentland. Reality mining: sensing complex social systems. In *Personal and Ubiquitous Computing*, 2006.

[ES] S. Erlander and N. F. Stewart. The gravity model in transportation analysis: Theory and extensions. In *Brill Academic Publishers*.

[FCSP06] J. Froehlich, M. Y. Chen, I. E. Smith, and F. Potter. Voting with Your Feet: An Investigative Study of the Relationship Between Place Visit Behavior and Preference. In *International Conference on Ubiquitous Computing*, 2006.

[foua] Foursquare. http://www.foursquare.com/.

[foub] Foursquare Venue Categories. `http://aboutfoursquare.com/foursquare-categories/`.

[FR85] R. H. Freymeyer and P. N. Ritchey. Spatial distribution of opportunities and magnitude of migration: An investigation of stouffer's theory. 1985.

[Fun06] S. Funk. Netflix Update: Try This At Home. `http://sifter.org/~simon/journal/20061211.html`, 2006.

[GHB08] M. C. González, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 2008.

[GMKB09] C. Ratti G. M. Krings, F. Calabrese and V. D. Blondel. Urban gravity: A model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009.

[GNI13] GNIP. Full coverage of anonymized foursquare check-in data now available exclusively from gnip. http://blog.gnip.com/gnip-foursquare-partnership/, May 2013.

[Gol08] J. Golbeck. *Computing With Social Trust*. Springer, 2008.

[gow] Gowalla. http://en.wikipedia.org/wiki/Gowalla.

[Gre75] M. J Greenwood. Research on internal migration in the united states: a survey. *Journal of Economic Literature*, 1975.

[Hav02] T. H. Haveliwala. Topic-sensitive pagerank. In *International World Wide Web Conference*, 2002.

[Haz03] M. L. Hazelton. Some comments on origin–destination matrix estimation. *Transportation Research Part A: Policy and Practice*, 2003.

[HCY10] P. Hui, J. Crowcroft, and E. Yoneki. Bubble rap: social-based forwarding in delay tolerant networks. *IEEE Transactions on Mobile Computing*, 2010.

[HHN10] J. Henrich, S. J. Heine, and A. Norenzayan. Most people are not weird. *Nature*, 2010.

[HKTR04] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 2004.

[HKV08] Y. Hu, Y. Koren, and C. Volinsky. Collaborative Filtering for Implicit Feedback Datasets. In *IEEE International Conference on Data Mining*, 2008.

[HTF03] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2003.

[Hym69] G. M. Hyman. The calibration of trip distribution models. *Environment and Planning*, 1969.

[IBC+12] S. Isaacman, R Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, 2012.

[JPO06] J. Hyvnen G. Szab D. Lazer K. Kaski J. Kertsz A.-L. Barabsi J.-P. Onnela, J. Saramki. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 2006.

[JTC12] K. Joseph, C. H. Tan, and K. M. Carley. Beyond local, categories and friends: clustering foursquare users with latent topics. In *ACM Conference on Ubiquitous Computing*, 2012.

[JWM+13] J. Jacobs, C. T. Weidemann, J. F. Miller, A. Solway, J. F. Burke, X.-X. Wei, M. Suthana, M. R. Sperling, A. D. Sharan, and I. Fried. Direct recordings of grid-like neuronal activity in human spatial navigation. *Nature Neuroscience*, 2013.

[JWS08] W.-S. Jung, F. Wang, and H. E. Stanley. Gravity model in the korean highway. *European Physics Letters*, 2008.

[KEHS73] D. Poston K. E. Haynes and P. Sehnirring. Inter-metropolitan migration in high and low opportunity areas: Indirect tests of the distance and intervening opportunities hypotheses. *Economic Geography*, 1973.

[KNS+13] D. Karamshuk, A. Noulas, S. Scellato, V Nicosia, and C. Mascolo. Geo-spotting: Mining online location-based services for optimal retail store placement. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2013.

[Kor09] Y. Koren. Collaborative filtering with temporal dynamics. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2009.

[LBW07] D. Helbing C. Khnert L. Bettencourt, J. Lobo and G. B. West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 2007.

[LCVH92] S. Le Cessie and J. C. Van Houwelingen. Ridge Estimators in Logistic Regression. *Applied Statistics*, 1992.

[LCW+11] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman. I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011.

[LD04] S. Libo and K. David. Evaluating Location Predictors with Extensive Wi-Fi Mobility Data. In *IEEE Intenrational Conference on Computer Communications*, 2004.

[Lee66] E. S. Lee. A theory of migration. *Demography*, 1966.

[Lev10] M. Levy. Scale-free human migration and the geography of social networks. *Physica A*, 2010.

[LH74] C. L. Lawson and R. J. Hanson. *Solving least squares problems*. SIAM, 1974.

[LHG04] D. Brockmann L. Hufnagel and T. Geisel. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences*, 2004.

[Liu09] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 2009.

[MAB+98] H. A. Makse, J. S. Andrade, M. Batty, S. Havlin, and H. E. Stanley. Modeling urban growth patterns with correlated percolation. *Physical Review E*, 1998.

[Mas10] Mashable. Web users now spend more time on facebook than google. `http://mashable.com/2010/09/10/facebook-overtakes-google/`, September 2010.

[Mil72] E. Miller. A note on the role of distance in migration: Costs of mobility versus intervening opportunities. *Regional Sciences*, 1972.

[MJSB09] E. R. Kallio S. Burthe A. R. Cook X. Lambin M. J. Smith, S. Telfer and M. Begon. Host-pathogen time series data in wildlife support a transmission function between density and frequency dependence. *Proceedings National Academy of Sciences*, 2009.

[MMR39] H. Makower, J. Marschak, and H. W. Robinson. Studies in mobility of labour: analysis for great britain, part i. *Oxford Economic Papers*, 1939.

[MNR98] F. Medda, P Nijkamp, and P. Rietveld. Recognition and classification of urban shapes. *Geographical Analysis*, 1998.

[MPTG09] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2009.

[Nag91] N. Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 1991.

[NFMM13] A. Noulas, E Frias-Martinez, and C. Mascolo. Exploiting foursquare and cellular data to infer user activity in urban environments. In *IEEE International Conference on Mobile Data Management*, 2013.

[NN08] A. J. Nicholson and B. D. Noble. BreadCrumbs: Forecasting Mobile Connectivity. In *ACM International Conference on Mobile Computing and Networking*, 2008.

[NSL+12] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: universal patterns in human urban mobility. *PloS ONE*, 2012.

[NSLM12a] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining user mobility features for next place prediction in location-based services. In *IEEE Internationcal Conference on Data Mining*, 2012.

[NSLM12b] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. A random walk around the city: New venue recommendation in location-based social networks. In *IEEE International Conference on Social Computing*, 2012.

[NSMP11a] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. In *AAAI International Conference on Weblogs and Social Media*, 2011.

[NSMP11b] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *3rd Workshop Social Mobile Web, Colocated with Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[OB05] J. O'Keefe and N. Burgess. Dual phase and rate coding in hippocampal place cells: theoretical significance and relationship to entorhinal grid cells. *Hippocampus*, 2005.

[OW01] J. de D. Ortúzar and L.G. Willumsen. *Modelling transport*. Wiley Chichester, 2001.

[PB07] M.J. Pazzani and D. Billsus. Content-Based Recommendation Systems. *The Adaptive Web*, 2007.

[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, 1999.

[PG09] A. Popescu and G. Grefenstette. Deducing Trip Related Information from Flickr. In *World Wide Web*, 2009.

[PZ10] K. Puttaswamy and B. Y. Zhao. Preserving privacy in location-based mobile social applications. In *Eleventh Workshop on Mobile Computing Systems & Applications*, 2010.

[QLC+10] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. Recommending social events from mobile phone location data. In *IEEE International Conference on Data Mining*, 2010.

[Qui86] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1986.

[Qui92] J.R. Quinlan. Learning with continuous classes. In *Australian Joint Conference on Artificial Intelligence*, 1992.

[Rav85] E. G. Ravenstein. The laws of migration. *Journal of the Royal Statistical Society*, 1885.

[Rep13] Web Republic. Tech safari 2013: Geocommerce on the rise. `http://www.webrepublic.ch/blog/2013/tech-safari-2013-geocommerce-1155`, February 2013.

[SGMB12] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási. A universal model for mobility and migration patterns. *Nature*, 2012.

[SI10] R. Caceres S. Kobourov J. Rowland A. Varshavsky S. Isaacman, R. Becker. A tale of two cities. In *In 11th Workshop on Mobile Computing Systems and Applications*, 2010.

[Sja62] L. A. Sjaastad. The costs and returns of human migration. *The Journal of Political Economy*, 1962.

[SK51] R. A. Leibler S. Kullback. On information and sufficiency. *The Annals of Mathematical Statistics*, 1951.

[SK12] A. Sadilek and J. Krumm. Far out: predicting long-term human mobility. In *AAAI Conference on Artificial Intelligence*, 2012.

[SKB12] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *ACM International Conference on Web Search and Data Mining*, 2012.

[SKKR01] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *ACM World Wide Web Conference*, 2001.

[SMM+11] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell. NextPlace: a spatio-temporal prediction framework for pervasive systems. In *IEEE International Conference on Pervasive Computing*, 2011.

[SNLM11] S. Scellato, A Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. In *Proceedings of AAAI Intenational Confernece on Weblogs and Social Media*, 2011.

[SNM11] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011.

[SPUP02] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002.

[SQBB10] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 2010.

[SS11] M. Musolesi J. Crowcroft S. Scellato, C. Mascolo. Track globally, deliver locally: Improving content delivery networks by tracking geographic social cascades. In *Proceedings of ACM World Wide Web Conference*, 2011.

[SSSH13] B. Shaw, J. Shea, S. Sinha, and A. Hogue. Learning to rank for spatiotemporal search. In *Proceedings of ACM international Conference on Web Search and Data Mining*, 2013.

[Sto40] S. Stouffer. Intervening opportunities: A theory relating mobility and distance. *American Sociological Review*, 1940.

[SW91] V. Shukla and P. Waddell. Firm location and land use in discrete urban space: a study of the spatial structure of dallas-fort worth. *Regional Science and Urban Economics*, 1991.

[TFP06] H. Tong, C. Faloutsos, and J.-Y. Pan. Fast random walk with restart and its applications. In *IEEE International Conference on Data Mining*, 2006.

[Tob95] W. Tobler. Migration: Ravenstein, thornthwaite, and beyond. *Urban Geography*, 1995.

[VOD+06] A. Vázquez, J. G. Oliveira, Z. Dezsö, K. Goh, I. Kondor, and A.-L. Barabási. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 2006.

[VRA+12] M. A. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto, and V. Almeida. Tips, dones and todos: uncovering user profiles in foursquare. In *ACM International Conference on Web search and Data Mining*, 2012.

[Wad75] W. J. Wadycki. Stouffer's model of migration: A comparison of interstate and metropolitan flows. 1975.

[WF05] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, second edition, June 2005.

[Wil67] A Wilson. A statistical theory of spatial distribution models. 1967.

[WL12] R. West and J. Leskovec. Human wayfinding in information networks. In *ACM Conference on World Wide Web*, 2012.

[WPS+11] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *ACM SIGKDD International Conference on Data Mining and Knowledge Discovery*, 2011.

[YG13] Y. Yang and M. C. González. A multi-scale multi-cultural study of commuting patterns incorporating digital traces. *NetMob*, 2013.

[YSL+11] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz. On the semantic annotation of places in location-based social networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011.

[Zel71] W. Zelinsky. The hypothesis of the mobility transition. *Geographical Review*, 1971.

[ZNSM13] A. Zhang, A. Noulas, S Scellato, and C. Mascolo. Hoodsquare: Exploiting location-based services to detect activity hotspots and neighborhoods in cities. In *IEEE International Conference on Social Computing*, 2013.

[ZRW07] J Zhao, A. Rahbee, and N. Wilson. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering*, 2007.

[ZZXM09] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *ACM Conference on World Wide Web*, 2009.

[ZZXY10] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with gps history data. In *ACM Conference on World Wide Web*, 2010.