URBAN MOBILITY AND LOCATION-BASED SOCIAL NETWORKS: SOCIAL, ECONOMIC AND ENVIRONMENTAL INCENTIVES

by

Ke Zhang

B.S., Huazhong University of Science and Technology, China, 2009

M.S., Huazhong University of Science and Technology, China, 2012

Submitted to the Graduate Faculty of

the School of Information Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

SCHOOL OF INFORMATION SCIENCES

This dissertation was presented

by

Ke Zhang

It was defended on

October 21st 2016

and approved by

Prof. Konstantinos Pelechrinis, School of Information Sciences, University of Pittsburgh

Prof. Prashant Krishnamurthy, School of Information Sciences, University of Pittsburgh

Prof. Yu-ru Lin, School of Information Sciences, University of Pittsburgh

Prof. Christos Faloutsos, Department of Computer Science, Carnegie Mellon University

Dissertation Director: Prof. Konstantinos Pelechrinis, School of Information Sciences, University of Pittsburgh

URBAN MOBILITY AND LOCATION-BASED SOCIAL NETWORKS: SOCIAL, ECONOMIC AND ENVIRONMENTAL INCENTIVES

Ke Zhang, PhD

University of Pittsburgh, 2016

Location-based social networks (LBSNs) have recently attracted the interest of millions of users who can now not only connect and interact with their friends - as it also happens in traditional online social networks - but can also voluntarily share their whereabouts in real time. A location database is the backbone of a location-based social network and includes fine-grained semantic information for real-world places. The footprints captured in a location database represent the socioeconomic activities of city dwellers and urban mobility at scale. LBSNs bridge the gap between the online and offline physical world, providing an unprecedented opportunity for researchers to access information that will allow them to place and understand human movements in the contexts of urban, social and economic activities.

In this dissertation, I design statistical analysis and modeling frameworks to examine how factors, including social interaction, economic incentives and local events, affect human movement across places in urban space. The dissertation first shows that people's visitation to local places exhibit significant levels of homophily, where peer influence can explain up to 40% of a geographically localized similarity between friends. We also find that the social selection mechanism is triggered by non-trivial similarity which is captured by places with specific network characteristics. Next, our quasi-experimental analysis reveals that online promotions in LBSNs are not as effective as anecdotal stories might suggest in attracting customers, and consequently in affecting the underlying city-dweller mobility. These results can have significant implications on advertisement strategies for local businesses. Finally, our developed framework is applied to assess the impact of local government decisions on

urban mobility and economic activities, which can provide a blueprint for future educated policy making. The outcome of this dissertation is envisioned to help better understand human urban movements motivated by social, economic and external environmental factors and further foster applications in sociology, local economy and urban planning.

Keywords: Urban Mobility; Location-based Social Networks; Statistical Modeling; Local Economy; Peer Influence; Quasi-Experimental Analysis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to express my gratitude to many people who have been playing indispensable roles during the entire journey of my PhD study.

Firstly and foremost, I am deeply grateful to my advisor Prof. Konstantinos Pelechrinis. If it is not for him, I would not be studying this degree, not diving into the fantastic world of data science, and not to speak of writing this thesis. During the past years he has been patiently mentoring, directing and supporting my research work. He has always been there to guide me with bright directions and influence me to be an active thinker, whenever I face challenges and obstacles. He has set an example of excellence as a researcher, mentor and instructor.

My sincere gratitude also extends to my committee members, Prof. Prashant Krishnamurthy, Prof. Yu-ru Lin and Prof. Christos Faloutsos, for all their critical and insightful suggestions as well as valuable guidance through my thesis study. I also would like to give my thanks to Prof. David Tipper for his kind offering support since I came to Pitt.

Big thanks also goes to my fellow colleagues. It is all of you who made my life at Pitt enjoyable and memorable. I thanks all the interesting discussions and enjoyable collaborations. I am really grateful for your accompanying me through such a long journey. Your friendship is my best fortune.

Finally, I would like to give my endless gratitude to my parents whose love and support have always been the greatest inspiration for me in my pursuit for betterment. My deepest gratitude goes to my sincere wife, Li Geng, for her dedicated support and encouragement. This dissertation could not be completed without her presence beside me.

1.0 INTRODUCTION

Urban and transportation planners, as well as, city officials have been trying to understand the way people act and behave in our cities for many years now. This will allow them to design cities that can deliver a livable, resilient and sustainable urban environment that is relevant to the city dwellers needs. These efforts can be seen from the recent rise of open data plan (e.g., NYC OpenData [117]) launched by many local government that encourage researchers and data scientists to access and analyze public digital data in order to provide data-driven solutions and guidance for a better urban planning. However, footprints about human movements are still far away from being available at scale in terms of number of participants and geographical reach.

During the last few years a number of location-based services and online social media has emerged mainly due to the technological advancements in mobile computing and networking that have led to the rapid proliferation of powerful mobile devices with location sensors. People can use these devices to obtain a wide range of information related to the geographic area they are currently in. Location-based social networks (LBSNs for short) form a prominent representative mobile service, which allows mobile users to connect and interact with their friends. More importantly, LBSN users can also explore and connect with local establishments through *check-ins*. The latter essentially bridge the gap between offline and online physical worlds.

A typical LBSN has two distinct components; a social network and a location log for each user. The social part of the system resembles any other existing online social networks, where friendships are declared and people can interact with their friends. What differentiates LBSNs from other digital social networks is the type of interaction that are feasible among the users. The main feature of this interaction is location sharing, i.e., users share their

locations with their friends.

Real-world places are at the core of location-based social networks, of which each contains fine-grain location information (e.g., latitude, longitude, street-level address, etc.), rich semantic information (e.g., venue category) and user-generated contents (e.g., reviews, photos, etc.). These places serve as the bridge between mobile users, local economy and urban environment. As users move across places, we can learn about their geographic location, the types of activities they engage in as well as the temporal and social dynamics of these activities. The check-ins of people into real places can potentially indicate their interest and preference in exploring and navigating urban areas. From a user perspective this information can serve as a signal for the aggregate opinion of users with respect to a specific establishment, as well as how their opinions with regard to real-world places are shared and propagated through social connections. From the perspective of a venue LBSNs offer a set of useful mechanisms that can potentially affect the decision of a user to visit the establishment. Of particular interest for our study is the ability of business owners to offer special, Groupon-like, deals through the platform. This is essentially an economic incentive that can affect the mobility and choices of the users.

As millions of user exploit location-based social networks, the user-generated contents associated with mobility traces enjoy an unprecedented scale in terms of number of users engaged, geographical reach and spatio-temporal granularity. Access to such large-scale datasets of mobile activities liberates us from traditional methods used to collect datasets that describe human movement and interactions in real world. For example, the survey-based methods that are typically employed by urban planners incur a high cost both in terms of time and effort. Consequently the data collected are often of small scale. Furthermore, data describing human movement such as Call Detailed Records (CDR) are owned by large telecommunication providers and are usually unavailable to the public. Also, the architecture of the cellular network technology does not allow for fine-grained spatial granularity in CDRs, while spatial semantic information is absent. Finally, the data of origin-destination transitions and trajectory coming from various urban transportation modes, such as taxis with GPS sensors installed, city bicycle system and subway system, can provide an accurate spatial representation of human movement in urban space, but still can not capture the

semantic contexts where human movement emerge.

Given the aforementioned reasons, this dissertation focuses on publicly available data from LBSNs. However, what opportunities exactly does this information really bring? After all there have been a number of studies that attempt to describe human movement and activities in a variety of fields ranging from social to computer and physical sciences. For start, empirical data on human mobility and socioeconomic activities from LBSNs can be helpful for validating models and theories that have been developed by scientists to explain the regular motives behind human mobility (e.g., the gravity model [157], the intervening opportunity model [149] etc.). More importantly though, previous work on human movement has focused on capturing the statistical properties of urban mobility. For instance, a statistical power-law distribution of human displacement has been identified [22, 69], which empirically reveals that human movements are often deterred by geographical distance. However, what are the other factors that can potentially affect human displacement?

Contrary to existing work on modeling of the statistical properties of the human urban mobility patterns, our main contribution in this dissertation is to tie the latter with the context they emerge in as captured through LBSNs. These patterns are affected by social (e.g., social connections), economic (e.g., local business advertisement) and environmental (e.g., local government decisions and urban events such as street festivals, sport events, road constructions, etc.) factors. In this dissertation, I design various statistical analysis frameworks using randomization and quasi-experimental techniques to identify and quantify the force of such factors that potentially motivate human movement across places in urban space. This dissertation is envisioned to fill the gap between statistical mobility models and urban activities, by building modeling frameworks that will enable their joint analysis.

Chapter Outline: The rest of this chapter is organized as follows. I begin in Section 1.1 with a brief overview on how traditional survey-based methods and census dataset initiate the empirical study of human migration patterns, which is followed by more recent methods using mobile communication data from cellular networks as well as the mobility data generated from urban transportation modes. Motivated by the limitations of previously applied methods, in Section 1.2 I introduce the unique characteristics of data coming from location-based social networks and then discuss the opportunities LBSNs bring for human movement

studies in urban contexts. Then in 1.3, I elaborate the research hypothesis examined in this dissertation on human urban mobility, in the contexts of social interaction, local business advertising and urban policy making, which can be captured by data from LBSNs. Finally in Section 1.4, I summarize the contributions of this dissertation by highlighting our findings in each chapter.

1.1 TRADITIONAL METHODS TO STUDY HUMAN MOVEMENT

In this section, I introduce a historical view of human movement studies using data from survey-based method and census as well as phone call records in mobile cellular networks. Then I discuss a more recent emerged type of urban mobility data captured by various transportation modes that have been favored by researchers in the fields of urban computing and planning.

1.1.1 Survey-based and Census Data

The first empirical study on human movement would be traced back to the seminal work published by Ravenstein [128] in 1885, namely "The Laws of Migration". In particular, Ravenstein analyzed the census data in United Kingdom which includes migration movements of million of individuals, where some important patterns of human movement were highlighted: (i) movements are often over only a short distance; (ii) distant migration often go to urban areas with commerce and industry; and (iii) migrations are stimulated by economic factors. These statements indicate that human movement are deterred by geographical distance but economic incentives can stimulate migrations. The work was one of the attempts to frame the understanding of human movement. Since then, a large amount of work has appeared aiming to analyze and model human migration [142, 97, 71]. However, the census data covering a large corpus of population are usually updated every several years, thus only provide a very static viewpoint of human movement. Also the locations reported from participants are typically at country and city levels. There are little knowledge about

what the exact places and contexts people transit between. Therefore, the census data suffers from limited spatial and temporal granularity to understand human movement.

While the research using census data still provide an unprecedented insight into human migration patterns at a large scale, it is not able to capture the large daily snapshots of human movement within urban space. With the rapid urbanization, understanding the pulse of a city through the mobility of its dwellers and visitors has become central to geographical and social sciences as well as to urban and transportation planning. Previously survey-based methods were often applied to acquire knowledge about how people commute from home to work [18] and how urban space are used [139, 34]. These surveys have also made it possible to obtain contextual information about the origins and destinations of trips in a city [80] as well as the transport means employed by commuters. However, survey-based method often conduct over a limited size of population representatives (often due to the high cost) and it is not applicable to capture the temporal dynamics of human movement and activities in urban space.

1.1.2 Mobile Phone Call Records

With the rapid development of telecommunication and mobile devices since the last decade in the 20th century, mobile communications have been brought into the daily life of millions of people. It is the first chance that human movement can be tracked at seconds with a relative high location accuracy, large geographical scale and population size. When people make a call or send a SMS message, their locations associated with the nearest Base Transceiver Station (BTS) are recorded.

However, privacy concerns have been the major barrier for such data to be easily accessed by researchers and scientists. It is only in recent years, the Phone Detailed Records (CDR) are sporadically becoming available to a limited number of research groups. One of the first studies using CDR to model human movement was published in Nature [69]. Together with an earlier novel study [22] by tracking the dollar notes as proxies of human movement, both work verify that human displacement at the country level follow a statistical power-law distribution. This finding aligns well with the statements on human migration patterns in

the early stage that human movements often cover short distances. After that, there are quite a few work followed utilizing CDRs to understand the properties of human movement in various spatial granularities and contexts [145, 155, 26].

Although the CDRs have provided a breakthrough opportunity for human movement study, the data is still only available under some specific agreement with telecommunication providers but not accessible to the public. Also, the spatial granularity of human movement represented by CDRs is not accurate enough, with only up to a few hundred meters depending on the coverage and distribution of mobile cellular towers. Also, the location information represented by the cellular tower only capture the coarse whereabouts of human movement, but no contextual and semantic knowledge are attached to the places where people go. The latter is critical to capture and help researchers understand the social and economic motivations of human movement. To fill these gaps, the recent emergence of location-based services and online geo-social media enable recoding human movement and their associated socio-economic activities at much finer-grain spatial and temporal granularities.

1.1.3 Urban Mobility Data Captured by Transportation Modes

The transportation infrastructures in urban cities often shape the way that people commute and travel. When people move by taking and interacting with a specific transportation mode (e.g., subway, bus, taxi or bicycle), their mobility can be recorded and represented in some formats. For example, when one picks up (drops off) a bike in a city bicycle-sharing system or enters (exits) the subway stations using the metro card, the origin and destination of a trip can be obtained by the station locations. With the GPS sensor deployed in taxis, the transitions or even a full trajectory of the passengers can be tracked. The data sensed from transportation infrastructures can represent urban mobility at a much more accurate granularity (e.g., the typical accuracy of GPS is 10-20m) compared to that of CDRs. As transportation infrastructures become indispensable parts of urban dwellers' daily life and are dedicated to connect every corner across urban space, we can learn intra-urban human movement and dynamics and at large scale in terms of population size and spatio-temporal granularity.

In recent years, these types of data gradually become available to urban researchers and scientists, especially when there are increasing number of local governments launching their "Open Data Plan", e.g., Captical Bikeshare [27] in Washington D.C., Pittsburgh HealthyRide [122], NYC OpenData [117]. It is reported that 119 cities in United Stated have released their open data platforms [1]. The goal of open data plan is to to foster city data transparency, while without losing the privacy, to a broad range of urban researchers and data scientists that they can help with developing a more healthy, more efficient and smarter city life. Since then, a large number of work have begun to leverage such data to understand human movement and activities in urban cities [168, 58, 88, 161, 100] and further facilitate various applications in urban computing [167]. The urban mobility data generated from transportation infrastructures form a great data source to estimate and predict human and traffic flow in urban space cells, which is of particular importance for traffic engineering and resource allocation. However, similar to CDR data, the semantic activities associated with the transitions or trajectories are often hard to be recorded. This lack of information can be compensated by the semantic location information from location-based service and geo-soica media, such as point-of-interests [161] and real-world venues in location-based social networks [111] as discussed later in Section 1.2.

1.2 OPPORTUNITIES FROM LOCATION-BASED SOCIAL NETWOKRS

In this section, I first highlight the unique characteristics of location-based social networks and further elaborate the advantage of utilizing the data from LBSNs to understand human movement and their associated activities in social, economic and urban contexts. After that, I point out some potential limitations and biases of the mobility data generated in LBSNs.

1.2.1 Unique Features of Location-based Social Networks

With the advancements in mobile computing and the rapid proliferation of powerful mobile devices with location sensors, a large number of location-based services and online social

[1] http://us-city.census.okfn.org/

medias have emerged during the last few years. Location-based social networks form a prominent representative mobile service, which allows mobile users to connect and interact with their friends. What differentiate LBSNs from traditional social networks, e.g., Facebook and Twitter, is that the content sharing in LBSNs focuses on real-world places. The unique attributes of a location-based social network form a promising platform for human urban movement study in various contexts. Without loss of generality, I mainly take Foursquare [53], one of the most popular LBSN in the world, as a representative to discuss the specific characteristics of a location-based social network. Some other platforms, like Yelp [160] and Facebook Places [51] resemble Foursquare with similar features.

Venue database: A venue database is at the core of a location-based social network, where each venue corresponds a real-world place. Venues are either initiated by the service provider as seeds at the initial stage, or continuously created and updated by the public crowd. The nature of crowd-sourcing mechanism fosters the explosion of the venue database while its accuracy can be guaranteed by aggregating editing suggestions from the public. Recently, Foursquare has claimed to have more than 65 million venues [2] which span in many countries and cities in the world. To maintain the accuracy of the venue database, Foursquare use a honeypot-based strategy to select a pool of loyal Superusers who can vigilantly maintain a watchful eye over the data for venues in their city or neighborhood [153]. The venue database serves as the location layer for the Internet, helping to connect people with places around the world. Venues do not only have accurate location information (e.g., latitude, longitude point, a full street-level address, etc.), but contain semantic information about their types (e.g., coffee shops, American Restaurant, etc.). When users go and share their presence at places, we can learn their mobility at a much accurate spatial granularity as well as, especially the most novel part compared to CDR and data from transportation modes, what type of activities associated with their movements.

Check-in: Users in LBSNs can voluntarily share their presence at places via *check-ins* in real time whenever Internet connection is available. In order to keep users engaged and willing to share their check-ins, some intriguing "gamification" features are introduced in LBSNs. For example, users in Foursquare are awarded "virtual goods", such as badges,

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

points and mayorship, when they contribute to the community in the network by voluntarily sharing information or accomplishing desired tasks. A pair of friends can participate in a score competition via a "leaderboard" which ranks their point scores in a descending order [85]. A previous study with an analysis of quantitative interviews indicated that Foursquare's gaming elements can impact human mobility decisions [57]. The user can either keep a check-in private (e.g., for self-record purpose) or make it visible directly to their friends via social connections. The latter provides a great opportunity to examine the interplay of human mobility and social interaction at scale.

Commercial Application: The most important application of a LBSN is to help local businesses attract customers. First of all, users can share their experience associated with their check-ins at venues through writing tips or reviews, rating the visited local business or posting photos. A venue in LBSNs is often regarded as an online "Yellow Page" for a physical place, with both static status information (e.g., hours and menu for a restaurant) and dynamic opinions shared by the crowd. The word-of-mouth [82] effects in LBSNs (e.g., reviews in Yelp) have already been proved to impact the success of local businesses. More directly, LBSNs are offering mechanisms that can serve as an affordable advertisement channel to local businesses for attracting customers. Business owners can claim to manage their venues. In particular, a business joined in LBSNs can promote *special offers* to its customers that connect through the platform. This type of data essentially open the gate for researchers to understand at scale how the economic incentive stimulate individuals' visits to local businesses.

Public Availability: The data in LBSNs can be accessed easily through various methods. For example, with the Foursquare Venue API [3], one can access the detailed venue status, such as number of check-ins, unique users and tips, at the moment of querying. This is also the main methods applied in this dissertation. Meanwhile, the mechanism of cross-platform information sharing enable check-ins in LBSNs be synchronized to the feeds of other third-party online social media platforms (e.g., Twitter and Facebook), by which APIs provided (e.g., Twitter Streaming APIs [4]) can be utilized to access the check-in data

[3]https://developer.foursquare.com/docs/venues/venues
[4]https://dev.twitter.com/streaming/

in real time.

As millions of users adopt LBSNs and share their whereabouts in real-world places, human movements can be learnt at a large scale, in terms of number of participants, geographical reach and spatio-temporal granularity, and, in particular, in a much richer contexts.

1.2.2 Opportunities for Urban Mobility Study

Despite their relatively short life-time - as compared to *traditional* online social media - there has already been an interesting line of work utilizing LBSN datasets. In particular, there are various studies across multiple areas that exploit similar data. Representative examples include human mobility [112, 31, 158], social-spatial network analysis [40, 31], urban computing and neighborhood modeling [116, 39], businesses placement [89, 64], location prediction and recommendation [113, 114] as well as location privacy and security [126, 77, 162, 165] to name just a few. Of particular interest in this dissertation is that the data from LBSNs provide unprecedented opportunities for understanding human movement in social, economic and urban contexts.

Interplay of Human Mobility and Social Interactions: In online social networks, people tend to be involved with others that exhibit similar characteristics and behaviors. This phenomenon is named "homophily" and has been discussed for various types of human behaviors, such as digital product adoption [14], emotion [55], politician opinions [21], etc. However, few work has been done with regard to human mobility. As users' mobile behaviors captured through check-ins are shared with their friends, it becomes feasible to examine whether there is significant correlation between check-ins behaviors and social ties and what are the underlying reasons. On one hand, friends may influence the decision of individuals' movement. For example, previous work [31] indicated that long-distance travel is more likely to be influenced by social ties. On the other hand, when people often go and meet at common places, they are more likely to make a social connection. Mining features from human movement and activities in the physical world would help with social link prediction [135, 155]. As present in Chapter 3, with a longitudinal data from LBSNs available we are able to preform a microscopic study on the interplay of human urban movement and social

connections.

Incentive of Local Business Advertisement: Furthermore, one of the most important *commercial* applications of LBSNs is helping the local businesses attract more customers. User-generated content and aggregate check-in counts serve as signals for the quality of the establishment and can have significant effects on their revenue [102]. In addition, LBSNs provide an immediate way for venues to advertise to potential customers. One of the advertisement mechanisms allows local businesses to provide economic incentives through "special offers" to customers that connect with them through these services. For instance, a venue on Foursquare or Yelp[5] can offer special deals (i.e., discounts) to people that check-in to the locale through the application. There have already been success stories featured in the media for businesses that have benefited from this mechanism [54]. This can potentially be an inexpensive way of advertisement for local businesses to people that are nearby and actually have the potential to visit them. As present in Chapter 4, the time-series of venue check-ins collected from LBSNs allow us at the first time to examine the effectiveness of local business advertisement to attract customers, which can provide valuable implications on its ability to *influence* the mobility of LBSN users through economic incentives.

Urban Events and Local Government Decisions: Finally, check-in information can serve as a proxy to the economic activity of a venue or a neighborhood in general. Hence, it can be used to assess the impact of external events in urban environment and local government decisions (e.g., road closures, street fairs, transportation facility update, etc.) on the local economy, thus the underlying human movement. Models that capture the effects of an "urban intervention" are crucial and not yet studied [167]. For example, understanding how a construction project that requires road closures affects the local economy is crucial for calculating liquidated damages. However, since to date there is no methodology to incorporate these effects in the calculations, they are ignored altogether. By analyzing the mobility data from LBSNs, the local government can quickly obtain an educated guidance on how urban environment and policy change impact local economy and the underlying urban mobility.

[5] www.yelp.com

1.2.3 Limitations and Biases

The datasets from LBSNs I utilize in this dissertation might suffer from a variety of limitations and biases that lead to various challenges. First, there can be demographic bias since these datasets capture the behaviors of a specific part of the population that uses digital social networks (e.g., "tech savvy" people, who are usually the young people). Furthermore, the voluntary nature of location sharing can provide us with an "undersampled" dataset of human urban activities. Given that it is hard to know how people choose to share their locations (i.e., it is not necessarily uniformly at random) it will be extremely hard to account for this bias. Moreover, virtual and real-world rewards can lead to people generate fake check-ins [77] [162], while the creation time of a friendship in the online social network might be different than that in the real world. The latter distorts the study of interplay of human mobility and social interactions. Finally, the number of check-ins or unique customers might not a good proxy for the actual visitation or revenue of local businesses.

Despite the above limitations, I believe our studies using the LBSNs datasets can greatly help understand human movement in a much richer urban contexts and further facilitate social, economic and urban planning applications. Note here that, some specific research questions are not expected to suffer from these biases. For example, in Chapter 4 I am really interested in the users of location-based social networks, since the promotions are offered through these systems and only these people can benefit from them. Hence, even though the number of check-ins might not be representative for the actual revenue, it is a good signal for the visibility of the business to the LBSN ecosystem.

1.3 RESEARCH HYPOTHESIS

Human movements usually follow general properties from a statistical point of view. Inspired by Newtons law of gravity, Gravity Model [28, 49, 93] describes that the flow of people from the origin to the destination is proportional to their population size and inversely proportional to the distance between them with certain magnitude. As another line of work,

Stouffers law of *intervening opportunities* [149] points out 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". These two types of models have been empirically verified and extended to fit human mobility data from different sources. More recently, a variation to gravity model is to model the statistical distribution of displacement distance, where there is no assumption for the population size in the origin and destination zones. These statistical models are often aimed to describe the human movements at an aggregate level in terms of geographical and temporal granularities, thus the contexts where movements take place are often ignored.

Human mobility and activities in urban space often exhibit periodic patterns both temporally and geographically [115, 31]. Although human movement turns out to highly predictable [145], often captured by people's daily commuting [18, 94] or regular travelling between habitats [12], human movement across urban space can be evolved with many other dynamic factors instead of just geographical constrains and regular travel needs. Even from the statistical viewpoint, one can still observe long-distance movements. Such movements may be stimulated by other external factors, e.g., economic incentives as stated in "Laws of Migration" by Ravenstein [128]. As present in Section 1.2, the data coming from location-based social networks provide unprecedented opportunities to capture various urban contexts where a much richer knowledge of human movement can be learnt.

Instead of modeling of the statistical properties of the urban human mobility patterns, this dissertation is to tie the latter with the social, economic and urban contexts they emerge in as captured through LBSNs. The core research hypothesis explored in this dissertation is **"human urban mobility can be affected by social interactions, economic incentives as well as urban events and local government policy making"**. In particular, three types of research hypotheses are studied in Chapter 3, 4 and 5, separately.

In Chapter 3, I examine the interplay between human urban mobility (captured by their check-ins in LBSNs) and their social connections. The research hypotheses include:

- Hypothesis 3.1: A significant correlation exists between human movement across urban space and their social connections.
- Hypothesis 3.2: Users' visitations to real world places are influenced by their friends at

specific geographical scales and the level depends on the type of places.

- Hypothesis 3.3: Non-trivial similarity of users' mobility is likely to trigger formations of social ties.

Inspired by the commercial application of LBSNs, in Chapter 4 I further investigate the effectiveness of a local business advertising mechanism in LBSNs to attract customers' variations to local places, in both long term and short term. In particular, the two hypotheses are:

- Hypothesis 4.1: The presence of a promotion through location-based social media leads to an increase in the visitation of a local business during the duration of the campaign.
- Hypothesis 4.2: The presence of a promotion through location-based social media leads to an increase in the visitation of a local business after the campaign has been completed.

In Chapter 5, I finally study the impact of urban events (initiated by local government and community) on human movement to nearby local places, where I take the street fair events as the study case. The two research assumptions are listed as follows:

- Hypothesis 5.1: Street fair events lead to an increase in customer visitations for nearby business venues.
- Hypothesis 5.2: The impact of street fairs on the customer visitations is geographically contained in a very small area.

1.4 CONTRIBUTIONS AND CHAPTERS

The major contribution in this dissertation is to tie human movement with the social, economic and urban contexts they emerge in as captured through location-based social networks. In particular, it is threefold: (i) I confirm with our findings that homophily, a common phenomenon in social networks, also significantly exist with regard to individuals' movement in LBSNS. Our designed statistical randomization models further quantify to what extent peer influence can explain the geographically local similarity between friends; (ii) inspired by the commercial application of LBSNs, I analyze data of 14 million venues that we collected

from Foursquare to examine the effectiveness of local businesses advertisement to attract the visitations of customers, which essentially represent how the economic incentive affect underlying individuals' mobility. I further design and implement a supervised learning model by extracting three type of features, aiming to provide strategies for improving campaign effectiveness in local marketing; This study had been featured in a number of media press such as Pittsburgh Post-Gazette [1]. (iii) Finally, I apply quasi-experimental techniques to quantify the impact of local government decisions on local economy, by taking LBSN users' check-ins to businesses venues as a proxy.

The rest of the dissertation is organized as follows.

In Chapter 2, I elaborate the background and related work to this dissertation. In Section 2.1, I first summarize two major classes of work on statistical laws and modeling of human movement, which lay the foundations for modern understanding of human movement. Then in Section 2.2, I discuss more recent work using data from location-based social network and media to understand and model human movement and urban activities.

In Chapter 3, I present our study on the interplay of human movement and social interactions. In particular, we use a longitudinal dataset obtained from Gowalla [6] (described in Section 3.1, a location-based social network, to examine the reasons behind the homophilous patterns observed with regards to the actual spots visited by people. After an brief introduction of two of the fundamental mechanisms, peer influence and social selection, that lead to the phenomenon of spatial homophily, I study its significance in Section 3.2. Then in Section 3.3 various randomization models are designed to quantify the levels of peer influence with regard to different geographical scales and location contexts. Finally, I examine the social selection mechanism in the network in Section 3.4. The main findings in Chapter 3 can be summarized in the following:

- There is a significant correlation between social ties and users' movement represented by their check-ins.
- While the similarity of users' geo-trails at a global scale cannot be attributed to peer influence, the latter can explain on average up to 40% of the geographically local similarity between friends.

[6] https://en.wikipedia.org/wiki/Gowalla

- The levels of local peer influence differ depending on the type of the location we consider.
- The social selection mechanism works upon non-trivial similarity which is captured by specific types of venues.

In Chapter 4, I further discuss our study on how economic incentives captured by a groupon-like promotion in LBSNs, affect users' visitations to promoted local businesses. In particular, we conduct a systematic study of the effectiveness of the LBSN advertising paradigm for local businesses to attract customers. This work is the first to address this problem by collecting and analyzing a large longitudinal dataset (as described in Section 4.1) of more than 14 million businesses on Foursquare. In Section 4.2, I first design statistical hypothesis testing experiments to evaluate both the long-term and short-term effects of LBSN campaigns for participating businesses, while taking into consideration the influence of possible confounding factors. In particular, our findings are validated by adopting two alternative methods for statistical testing, which lead to the same conclusions. In addition in Section 4.3, in order to gain a deeper understanding of our results and increase the practical value of our methodology, we design and implement a supervised learning model for predicting the popularity of a venue during and after a campaign by extracting three types of features, that are venue-related, promotion-related, and geographical features. Finally in Section 4.4, I discuss the implications of our work for businesses but also for the LBSN platforms as well. In particular, I describe how our findings can be used to inform strategies for improving campaign effectiveness. The major findings in this chapter are highlighted as follows:

- The effects of special offers through the LBSN platform examined are significantly more limited than what anecdotal success stories seem to suggest.
- Our experiments provide encouraging evidence on the feasibility of this prediction task, which can serve as a practical tool for supporting the design and cost-benefit analysis of LBSN campaigns. Specifically, a simple logistic regression model is sufficient to achieve an 83% accuracy with an 88% AUC.
- The influence of the considered features are fully aligned with our main results, as we find that promotion-related features have only a marginal contribution to the estimation

of popularity.

In Chapter 5, I present our study on examining the impact of urban events and local government decisions on local economy. Given the absence of actual revenue data for the local businesses, I take the human movement to nearby business venues as a proxy. In particular, I collect Foursquare *check-ins* (in Section 5.2) from the city of Pittsburgh over a three-month period (June-August 2015) and evaluate the effect of summer street fairs on customers' check-ins to nearby businesses. Given only the observational data, then in Section 5.1 I design a framework using two quasi-experimental techniques, that are Propensity Score Matching and Difference-in-Differences, to quantify the impact of street fairs. The main findings in this chapter can be summarized as follows:

- We provide quantifiable evidence that support the positive impact of street fairs on human check-ins at nearby local businesses. In particular, the impact decays fast with the spatial distance to events and the level of impact varies depending on different types of locations.
- We show how social media data - despite their potential biases - can be useful to public policy makers and local governments since they are transparent, accessible and are able to provide good evidence when analyzed properly.

Finally, in Chapter 6, I conclude this dissertation with future directions.

1.5 PUBLICATIOIN LIST

Papers related to this disseration

- [Chapter 3] K. Zhang and K. Pelechrinis, "Understanding Spatial Homophily: The Case of Peer Influence and Social Selection", in *ACM WWW*, 2014
- [Chapter 4] K. Zhang, K. Pelechrinis, T. Lappas, "Electronic Promotions via Location-based Social Media: Evidence from Foursquare", to appear in *IJEC*, 2017.
- [Chapter 4] K. Zhang, K. Pelechrinis, T. Lappas, "Analyzing and Modeling Special Offer Campaigns in Location-based Social Networks", in *AAAI ICWSM*, 2015

- [Chapter 5] K. Zhang, K. Pelechrinis, "Do Street Fairs Boost Local Businesses? A Quasi-Experimental Analysis Using Social Network Data", in *ECML-PKDD*, 2016

Other work during my PhD study

- K. Zhang, J. Xu, M.R. Min, G. Jiang, K. Pelechrinis, H. Zhang, "Automated IT System Failure Prediction: A Deep Learning Approach", in *IEEE BigData*, 2016
- K. Zhang, Y.R. Lin, K. Pelechrinis, "EigenTransitions with Hypothesis Testing: The Anatomy of Urban Mobility", in *AAAI ICWSM*, 2016
- K. Zhang, Q. Jin, K. Pelechrinis and T. Lappas, "On the Importance of Temporal Dynamics in Modeling Urban Activity", in *ACM SIGKDD UrbComp*, 2013
- K. Zhang, K. Pelechrinis and P. Krishnamurthy, "ACM HotMobile 2013 poster: detecting fake check-ins in location-based social networks through honeypot venues", in *ACM SIGMOBILE Mobile Computing and Communications Review*, Volume 17, Issue 3, 2013
- K. Zhang, W. Jeng, F. Fofie, K. Pelechrinis and P. Krishnamurthy, "Towards Reliable Spatial Information in LBSNs", in *ACM Ubicomp LBSN*, 2012
- L. Jin, K. Zhang, J. Lu, Y.R. Lin, "Towards Understanding the Gamification upon Users Scores in a Location-based Social Network", in *Multimedia Tools and Applications*, Springer, 2014
- L. Jin, X. Long, K. Zhang, Y.R. Lin and J. Joshi, "Characterizing Users Check-in Activities Using Their Scores in a Location-based Social Network", in *Multimedia Systems*, Springer, 2016

2.0 BACKGROUND AND RELATED STUDIES

Identifying the pulse of a city through the mobility of its dwellers and visitors has been central to geographical and social sciences as well as to urban and transportation planning. Human mobility has been studied for more than a century since the seminal work by Ravenstein [128]. Traditional mobility modellers aimed to capture the statistical properties of movement flows between the origin and destination given a certain spatial environment where movement can take place, such as gravity model [157] and intervening opportunity model [149]. Recent studies suggest a universal power-law distribution of human displacement, at the country level, by tracking the spread of dollar notes [22] and using CDRs [69]. The availability of new mobility data such as Phone Call Records and mobile social media check-ins further facilitate the study of human movement in a finer spatial granularity. The work on modeling human movement in urban cities [112] [158] and human movement prediction [145] [32] [133] [31] are just quite a few examples. In this chapter, I elaborate the related work on human movement studies to the focus of this dissertation. On one hand, I summarize the two major classes of work on statistical modeling of human mobility and the corresponding variations. On the other hand, I explore more related studies on understanding human movement in various urban contexts.

2.1 STATISTICAL MODELING OF HUMAN MOBILITY

In general, the goal of mobility modeling is to capture the statistical property of movement flows. At an earlier stage, the mobility modellers aim to predict the movement flow between the origin and the destination. There are two major classes of models following this line.

The first one is inspired by Newtons law of gravity and supports that mobility is impeded by distance. Movements over long distances cost more than moves over short distances. In particular, the flow of people from the origin to the destination is proportional to their population size and inversely proportional to the distance between them with certain magnitude [28, 49, 93]. Formally, given the origin i with population mass O_i and the destination j with population mass D_j as well as their geographical distance d_{ij}, the *Gravity Model* is defined as

$$T_{ij} = k\frac{O_i D_j}{f(d_{ij})} \tag{2.1}$$

where the scaling factor k and the form of function $f(\cdot)$ are often fitted to specific data.

The second class of models is based on Stouffers law of *intervening opportunities* [149]. As Stouffer posits it "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". Simply put, displacements are driven by the spatial distribution of places of interest. While existing literature seems to favor Stouffers theory [107, 75], both models are extensively used. Actually, these two classes of models are eventually proved to be statistically equivalent [157].

With the human movement data becoming available at scale in terms of population size, geographical reach and spatial granularity, both models can be verified empirically and some variations have been proposed. For example, the work by Simini et al. [140] addressed the limitation of Gravity Model and they put forward a parameter-free Radiation model. In this model, the expected flux $\langle T_{ij} \rangle$ from origin i to destination j is defined as

$$\langle T_{ij} \rangle = T_i \frac{o_i d_j}{(o_i + s_{ij})\,(o_i + d_j + s_{ij})} \tag{2.2}$$

where o_i and d_j are the total population of location i and j. T_i is the total number of transits starting from i and s_{ij} the total population in the circle of radius r_{ij} centred at i, but excluding the source and destination population. However, this model is proved

to predict the population movements between countries and cities successfully [140], but does not perform well when applied to intra-city movements [99]. Inspired by the law of intervening opportunities, the work [112] proposed a Rank model to capture the probability of a user in LBSNs transits from a starting place to the destination place given the distribution of places in between, which is provably aligned well with real data in many cities.

Another line of variation to gravity model is to model the statistical distribution of displacement distance, where there is no assumption for the population size in the origin and destination zones. In the work [22], the authors proposed a rather novel way to track human movement with a high spatio-temporal granularity, by tracing 464 thousand marked back notes carried by human. For each pair of successive locations reported for a dollar bill, the displacement distance Δr was calculated. The authors then measured and modeled the probability density of Δr and found that the distribution of human displacements follows a *power-law* distribution, that is

$$P(\Delta r) \propto \Delta^{-\beta} \tag{2.3}$$

where $\beta = 1.59 \pm 0.02$.

Soon after this study, the first large scale study to model the human displacement using phone Call Detailed Record was published [69]. The power-law distribution of human displacement are confirmed with similar exponent at $\beta = 1.75 \pm 0.15$. The authors further proposed an update model with an exponential cut-off, formally,

$$P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} exp(-\frac{\Delta r}{k}) \tag{2.4}$$

where Δr_0 and k are parameters depending on the dataset, in that case, $\Delta r_0 = 1.5$ km and $k = 400$ km.

These models are even empirically studied today using human displacements in geo-social media, e.g., Twitter [87]. The statistical properties discussed above provide a very static viewpoint of human movement that the displacements are often deterred by the geographical distance or intervening opportunities between the origin and destination. To describe the

dynamics of human flow between the origin and destination, some work in the domain of transportation and traffic engineering have been done on O-D flow matrix estimation, modeling and prediction [8, 9, 76, 164]. However, the contexts where movements take place are often ignored. Although human movement turns out to be highly predictable [145], often captured by people's daily commuting [18, 94] or travelling between regular habitats [12], human movement across urban space can be evolved with many other factors instead of just geographical constrains and regular travel needs. Even from the statistical viewpoint, we still can observe long-distance movements. Such movements may be stimulated by other external factors, e.g., economic incentives as stated in "Laws of Migration" by Ravenstein [128]. The unique data (as discussed in Section 1.2) coming from location-based social networks offer an unprecedented opportunity to learn the knowledge of human movement in a much richer context.

2.2 HUMAN MOBILITY IN URBAN CONTEXTS

In recent years, data from a variety of sources (e.g., location-based social networks, CDRs from cellular networks, GPS traces, etc.) have been used to quantify and model the dynamic activities that people engage in the urban space [161, 129, 16, 39, 163]. The common motivation behind these studies lays on the fact that understanding the spatial and temporal properties of urban activities can facilitate data-driven urban planning operations such as urban redevelopment and resource allocation. For instances, the work [161] mined the latent semantic information from taxi transitions to identify the functionality of urban regions. The urban land use can be inferred by analyzing the human movement using phone call records from cellular networks [16]. As another example, the authors in work [39] redefine the urban neighborhood using check-ins behaviors of users in Foursquare. Human movement and activities in urban space essentially capture their responds to the dynamic socioeconomic environment they live, thus understanding the regular patterns and how the patterns change dynamically with the environment is critical and attract tremendous amount of research work.

Human activities in urban cities usually have a temporally periodic pattern [115] and human movements are periodic both spatially and temporally most of the time [31]. However, external factors and incentives can distort such regular patterns. For example, the work [72] found that the periodic rhythm of human behaviors captured by check-ins were disrupted during the hurricane. The authors in [146] analyzed GPS records of 1.6 million users over one year and found that human behavior and their mobility following a large-scale disaster sometimes correlate with their mobility patterns during normal times, but are also highly impacted by their social relationship, intensity of disaster, government appointed shelters, news reporting, large population flow and etc. In the same direction, the work [156] took geo-tagged tweets as a proxy of human displacement and found that the climate change (e.g., Typhoons) had significant influence on human movement patterns. With the data available from location-based social networks, our main contribution in this dissertation is to understand human movement in various urban contexts. In particular, we study how factors including social connections, economic incentive of location business advertisement and external urban events interplay with human movement across urban places. In the following, I summarize the previous work related to each topic and highlight the difference from the work in this dissertation.

Interplay of Social Interaction and Human Movement: The availability of electronic traces of human activities has enabled the study of human behaviors in online social networks. Of particular interest in this dissertation is the phenomenon of Homophily, that is people tend to have similar attributes/behaviors with their friends. When it comes to human mobility, previous work using either phone call records [31, 12] and check-ins in LBSNs [31, 134] find a significant correlation between human movement and social ties. For example, the authors of the work [31] find that individuals' long-ranged travel is more influenced by social network ties while short-distance travel is periodic both spatially and temporally and not effected. They further show that social relationships can explain about 10% to 30% of all human movement. In summary, the work on interplay of social networks and human movement can be divided into two classes. One class of work focus on how the social network structure can help mobility prediction [132, 31] and location recommendation [159, 114]. Another part of work examine the co-location features to help link prediction

[155, 38, 40, 135].

In this dissertation, a microscopic study is provided on two fundamental mechanisms, peer influence and social selection, that can explain mobility similarity between friends. Peer influence and social selection are two main mechanisms that lead to homophilous patterns (e.g., friends are more similar in their mobility than random pairs) in online social networks. Isolating the corresponding effects of these forces is important for several reasons. First, the two processes produce homophily in different ways with regarding to network structure [78]: peer influence facilitates spreading of behaviors through links and produce network-wide uniformity, while social selection drives the network toward smaller clusters of like-minded individuals. Furthermore, the two mechanisms function based on different types of forces: interaction and similarity. In particular, some applications such as virus marketing [44, 105] leverage users social interactions to predict future behaviors (influence), which recommender systems [130] build predictions based on the similarity of peoples behaviors (selection). Quite a few general models, such as Holme-Newman Model [78, 37] and a quantitative model [136], have been proposed to separate and quantify the significance of the two mechanisms.

In this dissertation, we analyze a longitudinal dataset, with users' geo-trails in a LBSN and dynamic social interactions, to examine the two mechanisms that lead to mobility similarity. In particular, we design different randomization tests to examine the effect of one mechanism while eliminating the effect from the other one.. In particular, we delve into the details of peer influence, and we examine both its geographic scope as well as its contextual properties captured by the types of places. For social selection, we investigate the non-trivial mobility similarity that tends to be captured by places with specific network characteristics.

Economic Incentive: Online promotions have gained a lot of attention in recent literature. Such promotions have been a popular strategy for local merchants to increase revenues and/or raise the awareness of potential customers. A detailed business model analysis on Groupon was first presented by [4], while in [43] the authors surveyed businesses that provide Groupon deals to determine their satisfaction. Edelman *et al.* [46] considered the benefits and drawbacks from a merchant's point of view on using Groupon and provided a model that captures the interplay between advertising and price discrimination effects and the potential benefits to merchants. Finally, Byers *et al.* [24] designed a predictive model for the Groupon

deal size by combining features of the offer with information drawn from social media. They further examined the effect of Groupon deals on Yelp rating scores [25] [23]. However, there is no work done yet to examine the effectiveness of the advertisement mechanisms in LBSNs to attract customers, thus affect the underlying human mobility captured by their check-ins at local businesses. Our work in Chapter 4 present the first empirical study on evaluating and modeling the effect of the promotion mechanisms in a location-based social network.

Urban Environment and Local Government Decisions: Human movement in urban space usually exhibits a strong temporal periodicity [145], the dynamics of urban environment with urban events, such as urban road constructions, festival activities, sport events or urban planning, may change the original pattern of human movement to local places, introducing potential economic impacts on local businesses and economy. Previous study focus on the aggregated economic impacts of mega-events [104]. Getz *et al.* [65] investigate the effects of festival events on attracting tourists to attractions and destination areas, in order to facilitate the planning, development and marketing of festivals and special events. Previous work have also indicated a promising economic benefits generated from large sport events in cities, such as FIFA World Cup [96], Commonwealth Games [70] and Olympic Games [19]. Our work presented in Chapter 5 instead investigate the economic impacts of micro-events on users' movement to small-scale local businesses, and how the impact vary depending on spatial distance and type of business venues.

2.3 SUMMARY

In this chapter, I provide a systematic reviews on traditional statistical modeling of human movement, and then summarize recent work on studying human movement and activities in urban contexts. In summary, with movement data becoming available at a boarder population size, larger geographical reach, finer spatio-temproal granularity and richer urban contexts, human movement study has been transiting in both geographical scale, e.g., from country and inter-city levels to intra-city level, and in methodologies, e.g., from traditional statistical modeling of regular laws to more detailed studies in a semantic urban context.

In the following chapter, I present the first part of the dissertation on a microscopic study of the interplay between human movement in LBSNs and their social interactions, that is the phenomenon of spatial homophily.

3.0 UNDERSTANDING SPATIAL HOMOPHILY IN LBSNS

Homophily - also referred to as assortative mixing - is a phenomenon that appears very often in (social) networks. A (positively) mixed network, is one where the number of ties/edges between vertices that exhibit the same characteristics is significantly higher compared to the number that would have been expected if connections were made at random. McPherson *et al.* [106] refer to this phenomenon as "the birds of a feather flock together", and present many instances of homophily in social networks with regards to a large spectrum of people's attributes (e.g., age, religion, education, occupation, behavior etc.).

While mixing patterns in a network with respect to a specific characteristic can be formally and precisely quantified (e.g., assortativity coefficient [110]), the reasons behind their existence are not clearly understood and might differ for different scenarios. Nevertheless, there are three sources of mechanisms that are usually cited as the roots of homophily: (i) peer influence; (ii) social selection; and (iii) confounding variables.

Peer Influence: Peer influence appears specifically when we examine mutable characteristics, such as behavior, political views etc. When this mechanism is in play, people first become friends for reasons that are possibly not related to the characteristic X under examination, and then one *influences* the other on decisions related to X. In this chapter, we are interested in studying mixing patterns with regards to locations visited by people (as we will see in detail in Section 3.2 these patterns are homophilous). Given that this is a mutable characteristic, peer influence can be a possible cause of the observed assortativity. Figure 3.1a illustrates the peer influence mechanism in the context examined in our work. In this figure, we depict a socio-affiliation network, where affiliations (shown as the rectangular nodes) are the actual venues that people visit. We have further timestamped representative edges, with the time of their creation. In particular, Joe and Jack became friends at time

t_k, while Joe has visited "Li's Restaurant" at time t_{k-n} (that is, prior to becoming friends with Jack). On the other hand, Jack has not visited "Li's Restaurant" prior to time t_k. Assuming peer influence between Joe and Jack, an affiliation edge between Jack and "Li's Restaurant" will appear some time after they become friends (e.g., t_{k+m}) as presented in the figure. Simply put, when peer influence operates, people tend to first form a social tie and then become (more) similar.

Social Selection: Social selection can cause assortative mixing in networks, either with regards to mutable or immutable (e.g., age, race, sex etc.) characteristics. When social selection acts, people tend to associate with others that are already similar to them with regards to the characteristic under examination. In other words, people are already similar and this is essentially the cause of the friendship creation. Figure 3.1b illustrates the above concept. As we see, Joe and Alice, became friends at time t_l. Prior to that, they exhibit a large similarity with regards to the places visited, since they both visited "Li's Restaurant" and "Mike's Coffee Shop". Simply put, when social selection operates people first become (or are by nature) similar and then they create a social tie.

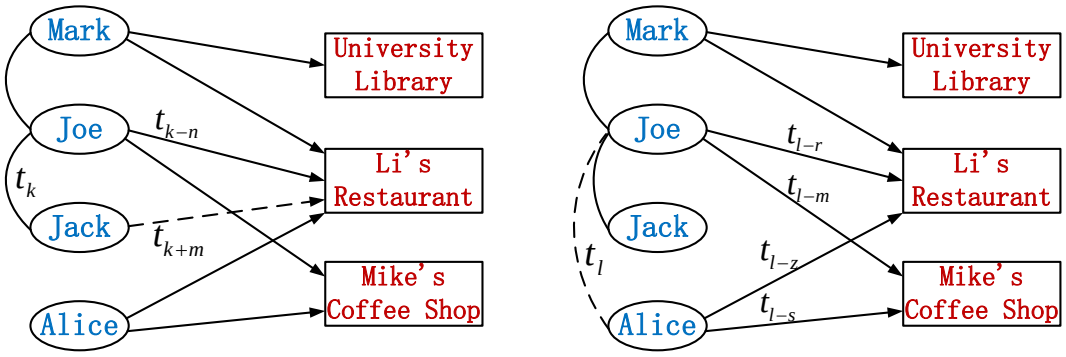
Confounding Variables: It is also called environmental or external influence. There are some unknown factors, such as geographical constrains, that may cause pairs of friends to behave similarly with each other, no matter if social influence or selection plays in a role. For example, people live in the same urban city are more likely to both build social ties and exhibit similar mobility behaviors. Therefore the urban mobility (e.g., check-ins) between pairs of friends tend to exhibit a higher similarity than what is expected if the social tie is created between two people who are randomly selected from the whole world. There is little work [109] examining external influence since it is often unobservable.

In this dissertation, we focus on the first two mechanisms, peer influence and social selection, that lead to the same observable phenomenon, that is, homophily with regards to the locations visited by friends in our setting. However, it might be hard to trace back to its actual roots. As it should be obvious from the above, one needs longitudinal data in order to decompose the reasons behind assortative mixing. This has traditionally been a burden for large scale studies on this topic. However, during the last years there is a rapid penetration of online social media in people's daily activities. This, in turn, has enabled the collection of

massive datasets that can foster social studies on human interactions. For instance, Bakshy *et al.* [13] using data collected from Twitter, examine the way that people adopt and share content, while Goel *et al.* [66] study how content diffuses through the underlying social network.

In this chapter, we use a longitudinal dataset obtained from Gowalla, a location-based social network, in order to examine the reasons behind the homophilous patterns observed with regards to the actual spots visited by people. In particular, we study the existence as well as the levels of peer influence and social selection in the network. Our approach is microscopic, in the sense that we consider the above mechanisms in a variety of granularities. In particular, we consider global versus local influence, the relation between the actual type of location and the underlying peer influence, as well as the impact of the type of location on the *effectiveness* of the social selection process. Our main findings can be summarized in the following:

- While the similarity of users' geo-trails at a global scale cannot be attributed to peer influence, the latter can explain on average up to 40% of the geographically local similarity between friends.
- The levels of local peer influence differ depending on the type of the location we consider.
- The social selection mechanism works upon non-trivial similarity and can be stimulated by specific types of venues.



(a) Peer influence mechanism

(b) Social selection mechanism

Figure 3.1: Two mechanisms as the roots of homophily.

Chapter Outline: The rest of the chapter is organized as follows. Section 3.1 describes the longitudinal dataset we use for our analysis. In Section 3.2 we examine the existence of homophily with regards to the locations visited by people. Section 3.3 provides a detailed, microscopic analysis of peer influence, while Section 3.4 examines the social selection mechanism. Section 3.6 discusses the related work to this work. Finally, Section 3.5 briefly discusses the scope and limitations of our study and concludes our work.

3.1 DATASET AND ANALYSIS SETUP

In this section I first briefly describe the Gowalla data utilized and present the initial setup and definitions before the following statistical analysis. Then I formally introduce the research hypothesis examined in this chapter.

3.1.1 Dateset and Definitions

The longitudinal dataset that we used for our study was provided to us by the authors of [135]. It was crawled from Gowalla, a commercial LBSN[1], between May 05, 2010 and August 18, 2010. The dataset consists of 10,097,713 public check-ins performed by 183,709 users in 1,470,727 distinct places. Every venue is associated with a category, that essentially describes the type of location the user checked-in. There are 283 distinct categories in Gowalla. Every check-in log is a tuple of the form `<User ID, Venue ID, Latitude, Longitude, Time, Category ID>`. 27,895 venues are unclassified (i.e., `Category ID = NULL`) and hence, we discard them. This results in a dataset with 10,062,916 check-ins, in 1,442,832 distinct places by 183,500 users (209 users had check-ins only in unclassified spots).

Gowalla users also participate in a friendship network with reciprocal relations, which consists of 765,871 links. Gowalla was crawled every day for the aforementioned period, and hence the formation time of every friendship that was created after May 05, 2010 was able to be obtained. For the purposes of our work, we will use only the pairs of friends for which

[1]Gowalla has been acquired from Facebook and ceased its operations in March 2012.

we have the actual friendship creation time. There are 289,888 such links in total. Some of the edges may also have been deleted (e.g., Jack "de-friends" Alice). For these links we also have a deletion time. Hence, the friendship edges have the following 4-tuple form `<User ID, Friend ID, Formation Time, Deletion Time>`. From the 289,888 links above, only $\approx 2\%$ of them were deleted afterwards, and thus, we can safely discard them. If we further keep only pairs of friends for which both users have at least one check-in we have a final number of 202,424 links that we use for our study.

Home Location of a User: Our dataset does not include home location information for the users. However, we are interested in examining the mechanisms of peer influence and social selection in relation to the distance between the home locations of the users. In order to infer the home locations of the users, we apply a density clustering algorithm (DBSCAN [50]) on the check-in history of each user. The check-in points are then grouped into clusters each of which is in general of different size. We select the dominant cluster (say C_1), i.e., the one with the maximum number of the data points (i.e., check-ins), and we re-apply DBSCAN on C_1 to improve the estimation accuracy. Finally, we pick again the dominant cluster (say $C_{1,1}$) and we estimate the home location of the user as the centroid of the data points (lat/lon) in $C_{1,1}$.

Definitions: Before moving on to our analysis, we wish to introduce some terminology that we will be used throughout the rest of the paper. Two users u and v are said to have been ***check-in co-located***, if they have both checked-in to at least one common venue, regardless of the actual check-in time. Furthermore, u and v are said to have been ***area co-located at*** v***'s home location***, if u has checked-in to at least one venue, within 25kms from v's home location.

In the above definitions, we do not impose a constraint of co-location both in time and space[2]. When Alice and Bob influence each other with respect to some behavior, e.g., adopting a specific product, it does not necessarily mean that they will buy it at the same time. This is exactly what Figure 3.1a depicts. Furthermore, similarity is related with the actual actions and not necessarily when these actions take place (e.g., two people that write on a specific Web blog can still be considered similar, regardless of whether they both write

[2]Note that this would require knowledge of a "check-out" time as well.

on Thursday nights or not.).

3.1.2 Hypothesis Development

The existence of homophily has been confirmed in large online social networks with regards to people's interests, opinions and many other behaviors [141]. In this dissertation, we study the homophilous patterns with regards to human check-ins in physical places, and investigate the two underlying roots, peer influence and social selection. On one hand, previous study [31] has indicated that peer influence plays in a role only for distant movement but not for short displacements. Meanwhile, the level of influence can vary depending on type of activities associated with their movements. For example, pairs of friends are more likely to hand out to nightlife places but less frequently visit a subway station together. One the other hand, similarity of mobility behaviors can lead a pair of users to form a social connections. However, common check-ins (i.e., co-locations) do not always lead them to be similar in an important way, and further trigger them to be connected.

Therefore, for the rest of this chapter we provide a microscopic study on the two mechanisms behind spatial homophily. In particular, we examine peer influence by considering different spatial scales and types of venues, and we further investigate specific network characteristics of co-locations that can capture the non-trivial similarity between friends. In particular, we examine the following three hypotheses:

Hypothesis 3.1 (Spatial Homophily Existence)**.** *A significant correlation exists between human movement across urban space and their social connections.*

Hypothesis 3.2 (Spatial Peer Influence)**.** *Users' visitations to real world places are influenced by their friends at specific geographical scales and the level depends on the type of places.*

Hypothesis 3.3 (Spatial Social Selection)**.** *Non-trivial similarity of users' mobility is likely to trigger formations of social ties.*

3.2 SIGNIFICANCE OF SPATIAL HOMOPHILY

Traditionally, vertices in a network are annotated with scalar or enumerative characteristics and metrics for quantifying the level of homophily in these scenarios are very well defined [110]. Nevertheless, in our case, we want to evaluate the mixing patterns in the network with regards to the spatial behavior of users, that is, the places they visit, which cannot be described by a single number or label.

A user u of our LBSN is associated with a vector $\mathbf{c}_u$ capturing the places he has visited. In particular the i^{th} element of vector $\mathbf{c}_u$, is equal to the number of check-ins that u has in venue i. Since we cannot directly compare vectors and directly apply the assortativity coefficient [110], we rely on a different methodology. For our purposes, we will need to define a similarity measure between vectors. In this work, we will utilize the cosine similarity. In particular, the similarity between two users u and v is defined as:

$$sim_{u,v} = \frac{\mathbf{c}_u \cdot \mathbf{c}_v}{\|\mathbf{c}_u\|_2 \|\mathbf{c}_v\|_2} \tag{3.1}$$

In order to identify the existence of assortative mixing - or not - in the network we will follow the same line of thought as in the definition of the assortativity coefficient, tailored though in our context. The assortativity coefficient essentially estimates the difference between the actual number of edges in the network that fall between vertices of the same type (enumerative characteristic) or of similar attribute value (scalar characteristic) and those that would have been expected if connections were made at random. Adopting this idea in our context, we ought to calculate the average spatial similarity between friends in the real network, sim_{real}, and compare it with the expected average similarity if connections were made at random. In order to calculate the latter we will rely on Monte Carlo simulations. In particular, we will sample the ensemble of $G(n, m)$ Erdős-Rènyi random graphs[3] and calculate the average spatial similarity between friends in the sampled networks, sim_{rnd}. If $sim_{rnd} \ll sim_{real}$, the network essentially exhibits homophily.

Nevertheless, sampling the pure $G(n, m)$ model might lead to under-estimation of the

[3] A brief background on Erdős-Rènyi random graphs is provided at Appendix A.

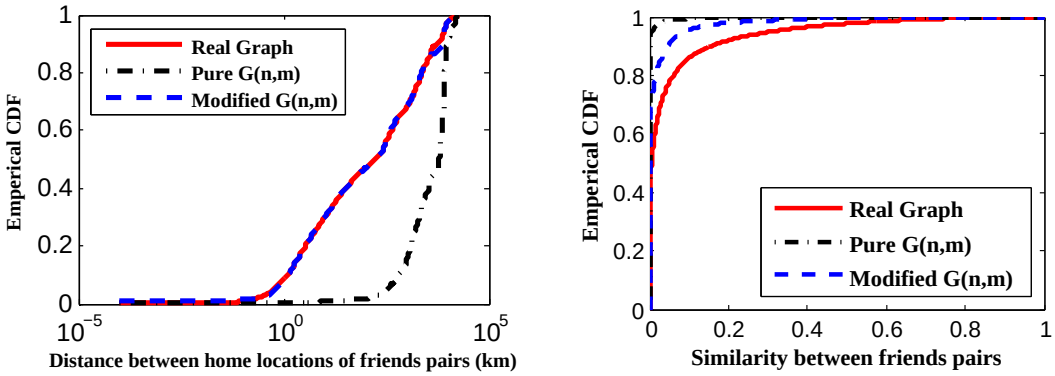
average similarity value. In particular, it has been found that the majority of one's friends live in nearby locations [134] [31]. In other words, the probability distribution of the home location distance between two friends d_f has the majority of its mass concentrated into small distances. We have also verified this is true in the dataset we are using (see Figure 3.2a). This can have implications on the sim_{rnd} value as computed above. In particular, since the majority of the user pairs live far from each other (the number of pairs of users living in the same city is much less compared to all possible pairs of users), $G(n, m)$ sampling will lead to edges between users that live far away. Such pairs though are also expected to have much lower similarity, since they simply do not have many chances to visit the same places. Hence, we also perform a second series of Monte Carlo simulations, where we sample from a modified, location-aware, $G(n, m)$ ensemble. In particular, we pick the first end of an edge uniformly at random, and we use the distribution of d_f to randomly select the other end of the edge. In other words, while we randomly sample the edges, we make sure to preserve the distribution of the friends' home distance. Using these randomized networks we can calculate the average similarity between friends, $sim_{l,rnd}$.

We sample the two randomized networks 100 times and then calculate the 95% confidence interval (CI) for the average similarity between friends. Our results are presented in Table 3.1. As we can see the value of the friends' average similarity in the real network lays outside the 95% confidence intervals for both random network models and is significantly higher as compared even to the upper bound of these CIs. This leads us to the conclusion that the network under consideration exhibits strong assortative patterns with respect to the spatial trails of the users. This strengthens the results reported by Wang *et al.* [155] where a correlation between spatial trajectory similarity and network closeness is reported using call detail records as well as those in [120], where a different, less rigorous, method was used along with a different similarity metric between users. Finally, Figure 3.2a, presents the cumulative distribution function of the home distance between two friends for the real network, a pure $G(n, m)$ representative sample and a modified $G(n, m)$ representative sample. As we can see the spatially modified $G(n, m)$ model exhibits a similar home distance distribution with the one of the real network. On the contrary the pure Erdős-Rènyi random graph exhibits longer home distances overall as expected. Figure 3.2b further depicts the cumulative distribution

function of the similarity values for the connected vertices in the real network, and the representative random graph samples used in Figure 3.2a. Results verify the ones we obtained in Table 3.1.

Table 3.1: There is a clear homophily with regards to the spatial trails of Gowalla users.

sim_{real}	$sim_{l,rnd}$	sim_{rnd}
0.05425	[0.01836, 0.01837]	[0.00236, 0.00237]



(a) (b)

Figure 3.2: (a) Our modified random graph ensemble retains the distribution of home location distances observed in the real network. (b) Similarity between friends in a real network is much higher compared to that in the randomized networks.

3.3 PEER INFLUENCE

Having established the existence of spatial homophily in the network we turn our attention to decomposing the reasons behind this phenomenon. In this section, we examine the peer influence mechanism with regards to spots visited by people. Our analysis considers both (i) the geographical scope of peer influence (i.e., whether people are influenced at a global/local

scale) and (ii) the context of peer influence (i.e., whether people are influenced - or not - at the same degree with regards to different types of places).

3.3.1 Global Influence

We begin by examining the **global influence** between people. By the term global, we essentially refer to possible effects people can have on their friends' decisions related with their check-ins at *any* part of the world. In other words, if Bob, who is from New York City, and his friend Alice, who is from Boston, visited "Restaurant X" in San Francisco, was it a result of peer influence between each other? We would like to emphasize here that, while from a sociological perspective the question of whether peer influence affects the check-ins of a pair of friends anywhere in the world might seem absurd, we begin with this question in order to smoothly introduce the various tests we will use in our analysis[4].

Knowing the time of friendship formation between Bob and Alice enables us to calculate a similarity value for Bob and Alice before and after becoming friends. A similarity increase *might* be a signal for peer influence. Using the cosine similarity metric, the global similarity between users u and v prior to becoming friends is:

$$gsim^b_{u,v} = \frac{\mathbf{c}^b_u \cdot \mathbf{c}^b_v}{\|\mathbf{c}^b_u\|_2 \|\mathbf{c}^b_v\|_2} \tag{3.2}$$

where, $\mathbf{c}^b_u$ and $\mathbf{c}^b_v$ are defined analogously to Section 3.2, as vectors describing the venues that u and v checked-in to before they became friends. In particular, the i^{th} element of $\mathbf{c}^b_u$, is equal to the number of check-ins that u had in venue i, prior to becoming friends with v. Since $gsim^b_{u,v}$ is computed over the check-ins that took place before u and v got connected, it can be thought as an *inherent* global similarity between these two users.

Once u and v got associated, we can compute a new similarity value as follows:

$$gsim^a_{u,v} = \frac{\mathbf{c}^a_u \cdot \mathbf{c}^a_v}{\|\mathbf{c}^a_u\|_2 \|\mathbf{c}^a_v\|_2} \tag{3.3}$$

[4]Furthermore, the methodology presented in this section itself could be applicable in different settings and hence, beneficial to other researchers.

where now vectors $\mathbf{c}_u^a$ and $\mathbf{c}_v^a$ are created as above but using the whole check-in histories of u and v respectively (i.e., both before and after they became friends).

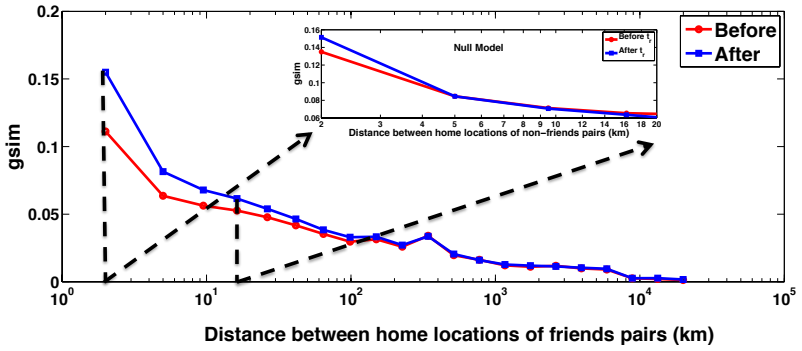


Figure 3.3: Global influence can possibly explain only up to 2.32% of the global similarity between friends.

Using our data, we can compare the global similarity between a pair of users before and after becoming friends. Figure 3.3 presents our results. The value on the x-axis is the distance between the home locations of friends pairs, and the y-axis is the corresponding average global similarity of the pairs. We use logarithmic binning for the home location distance to reduce - to the extent possible - the noise due to fewer samples at the right end of the x-axis. As we can observe, the global similarity does not change significantly after the friendship creation when the home locations of the friends are more than 10km apart. However, for smaller home distances, there is a non-negligible increase in the *gsim* between the friendship pairs formed. Furthermore, it is worth noting that global similarity values reduce with an increase in home location distance, as one might have expected (the more distant the homes of two friends the less possible is for them to check-in to common venues).

If we further consider the area under the "blue" line to represent the overall global similarity between friends, we can see that on average 88.68% of it can be explained from the *inherent* similarity between pairs of users (as captured by the area under the "red" curve), while the rest 11.32% can be attributed to peer influence. However, we would like to emphasize here that this number serves only as an upper bound for the amount of global peer influence. With the above statistical test we can only quantify the contribution of the

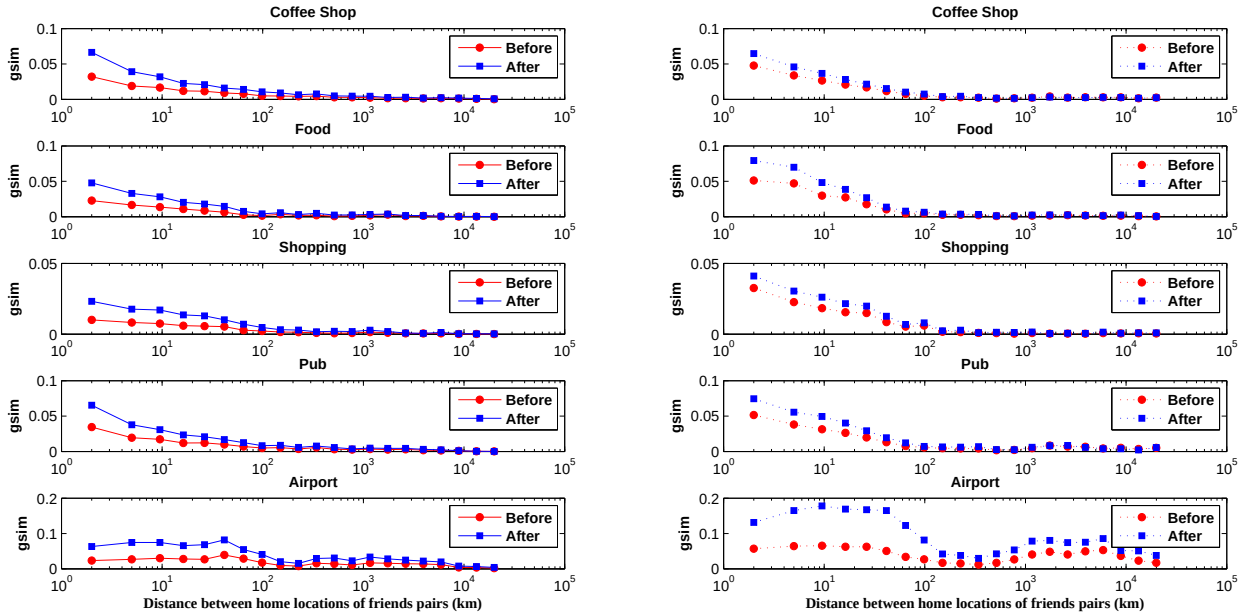


Figure 3.4: Levels of global peer influence are very small regardless of the venue context.

inherent similarity between pairs of users on the overall global similarity. Nevertheless, there can be other reasons that can explain the additional 11.32% in the global similarity.

One possible reason for the increase in the global similarity, especially for friends whose home locations are nearby, is the fact that as people accumulate more activity and visit more places, their similarity to other users (friends or not) can increase just by the chance of visiting the same locations. In other words, (the inherent similarity) $gsim_{u,v}$ might be an increasing function with time, regardless of whether u and v form a tie or not. To examine such a possibility, we consider pairs of users (w, z) that have not formed a social tie during the period that our dataset spans. We will pick a reference time t_r at random, and compute their global similarity prior and after t_r. We are especially interested in pairs whose home location distance is less than 20km. Our results from this *null* model are overlayed and zoomed in, within Figure 3.3. As one can observe, even for pairs of non-friends their global similarity increases with time. If we further calculate the areas under the curves, we can see that approximately 9% of the change in the global similarity of a pair of users after becoming friends can be explained by its *natural* increase with time[5].

[5]We would like to emphasize on the fact that this result is only approximate, since an exact result (i) depends on the accurate choice of t_r and (ii) would require the estimation of a function $gsim_{u,v}(t)$. Both

Factoring in the above percentile temporal increase of global similarity, we re-calculate the upper bound on global peer influence, and eventually only at most 2.32% of the global similarity can be attributed to global peer influence. Hence, it appears that ***there is no global influence between friends***.

Contextual dependencies: In the above we have considered the check-ins of users at all possible types of venues. However, influence can clearly be context dependent; while in aggregate there is no (or very small) global influence among friends, it is possible that global peer influence exists for certain types of places. For instance, while our friends' visits at restaurants might not affect us because we have our own taste in food, the same might not be true for nightlife spots. More general, people might have an impact on friends' decisions about specific types of venues.

Table 3.2: Even after considering specific context (i.e., type of places), there appears to be no global peer influence.

Venue type	Upper bound on global influence
Coffee Shop	2.08%
Food	1.05%
Shopping	-4.60%
Pub	-3.13%
Airport	0.04%

To quantify any context dependencies on global influence we perform the same statistical test as above, but instead of considering check-ins to all the venues in vectors $\mathbf{c}_u$, we only consider check-ins to venues of the specific category under examination. Figure 3.4 presents our results for five representative, distinct categories of spots in Gowalla. In particular, we consider "Coffee Shops", "Food" joints, "Pubs", "Shopping" venues, "Airports"(results for the rest 278 categories are omitted but they do not differ significantly). The left column of figures are the results for the pair of friends, while in the right column are the corresponding

are beyond the scope of our work.

results for the null model as above. Table 3.2 further presents the upper bound on the percentage of global similarity between friends that can be assigned to the global peer influence for the different types of venues. Note that in some cases we obtain negative values for the upper bound of global peer influence. This is essentially an artifact of the time t_r picked for the null model. Nevertheless, even with the most accurate choice of t_r the upper bound on the global peer influence is not expected to be a much larger positive number[6]. Hence, even if we consider types of places in isolation, there appears to be no global peer influence on average.

3.3.2 Local Influence

Our previous results support - as one might have expected - the absence of global influence between pairs of friends. However, peer influence mechanism might operate in smaller spatial scales, and hence we seek to examine in this section the existence of a localized version of peer influence. In other words, while Jack is not influenced by his friend Jill (who possibly lives more than 20kms away) at a global scale, he might be influenced when he is around Jill's home location. In order to examine whether there exists **local influence** or not using the above procedure, we would need pairs of friends who had been area co-located (see Section 3.1) in each others home location both prior and after they become friends. This would allow us to estimate both an inherent local similarity as well as a local similarity after becoming friends. However, there are not many such pairs to yield statistically significant results. Hence, we devise a different test.

In particular, we consider pairs of friends, (u, v), who have been area co-located at u's and/or v's home location after becoming friends. In addition, we filter out pairs (u, v) that have been check-in co-located before becoming friends (however, they can have been area co-located). The reason for the latter is to remove from our test-set users that have non-zero inherent similarity and essentially to rule out one possible reason for the observed (if any) local similarity. Note here that, the test we will use in what follows cannot account for that. The above filters finally give us 43,618 pairs of friends.

[6]We have examined a variety of other strategies for choosing t_r, and they all give values close to zero.

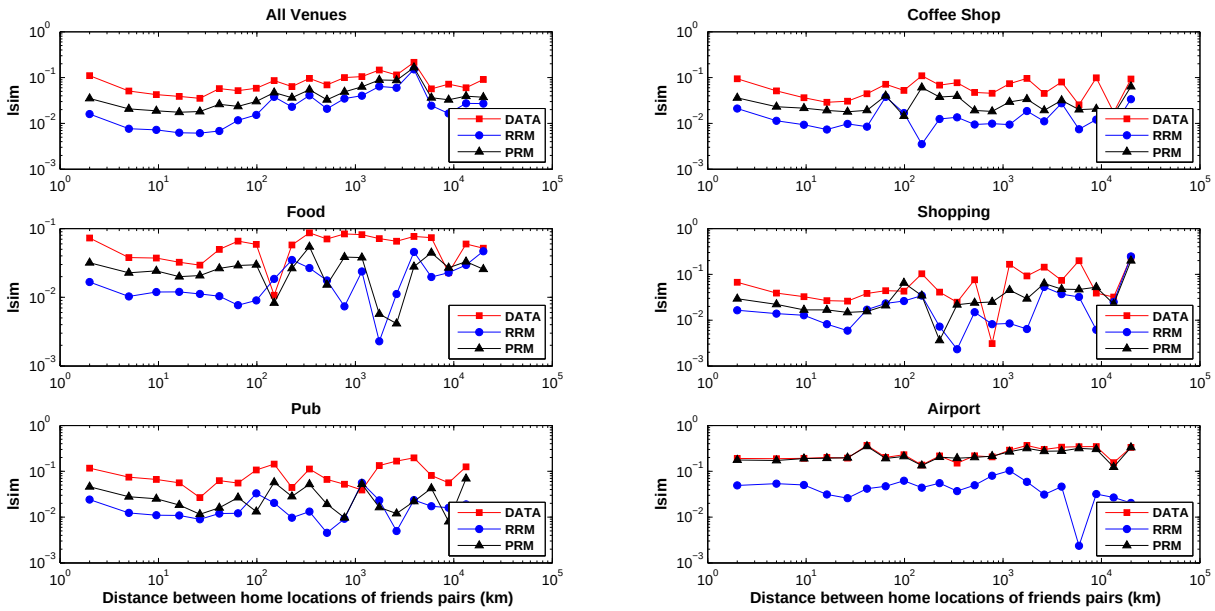


Figure 3.5: Local similarity as obtained through data and two randomized reference models.

Using these pairs we calculate the local similarity between u and v as follows:

$$lsim_{u,v}^{data} = \frac{\mathbf{c}_u^v \cdot \mathbf{c}_v^v}{\|\mathbf{c}_u^v\|_2 \|\mathbf{c}_v^v\|_2} \tag{3.4}$$

where, $\mathbf{c}_u^v$ and $\mathbf{c}_v^v$ are as in Equations (3.2) or (3.3) but now we are considering only the check-ins of u and v respectively, in venues *near* v's home location (i.e., within a radius of 25km from v's home location).

To reiterate, given our setup, this value of local similarity (large or small) cannot be attributed at any part to inherent similarity between the users, since the specific pairs we examine were not check-in co-located prior to becoming friends (i.e., their similarity - global or local - was zero)[7]. However, we need to compare this local similarity with some benchmark values that capture other possible reasons that lead to the observed behavior. In particular, we devise two reference models that aim in capturing (i) the expected local similarity if u was checking-in at random (Random Reference Model - `RRM`), and (ii) the expected local

[7]Of course, as we also further explain in Section 3.5, there can be missing "check-in co-locations" of users, not captured from the dataset, due to the voluntary fashion of location sharing in Gowalla.

similarity if u was choosing his check-ins based on the popularity of the venues (Popularity-based Reference Model - PRM). In both models we retain the structure of the real check-ins, that is, the total number and their categories.

Simply put, let us assume that u has visited v's home location, and he performed z number of check-ins ten of which where in coffee shops and the rest in restaurants. For our RRM, we will uniformly at random sample ten times all the coffee shops in v's home location, and $z - 10$ times the local restaurants. This process will generate a synthetic dataset for the check-ins of u in v's home location, and consequently will give a corresponding vector $\mathbf{c}^v_{u,\texttt{RRM}}$. Similarly, for the PRM, we will follow exactly the same process, but instead of picking venues uniformly at random, we will bias the sampling probabilities based on the popularity of each venue π as captured from the total number of users that have checked-in at π. This will further give us another vector $\mathbf{c}^v_{u,\texttt{PRM}}$.

Using our reference models' vectors for user u we can obtain the reference local similarities between u and v as:

$$lsim^{\texttt{RRM}}_{u,v} = \frac{\mathbf{c}^v_{u,\texttt{RRM}} \cdot \mathbf{c}^v_v}{\|\mathbf{c}^v_{u,\texttt{RRM}}\|_2 \|\mathbf{c}^v_v\|_2} \tag{3.5}$$

$$lsim^{\texttt{PRM}}_{u,v} = \frac{\mathbf{c}^v_{u,\texttt{PRM}} \cdot \mathbf{c}^v_v}{\|\mathbf{c}^v_{u,\texttt{PRM}}\|_2 \|\mathbf{c}^v_v\|_2} \tag{3.6}$$

We want to emphasize on the fact that v's check-in vector in Equations (3.5) and (3.6), is obtained from the real data. Furthermore, we run RRM and PRM 100 times for each pair of users and obtain the average reference local similarity.

Figure 3.5 presents our results. As with global similarity, we present the results obtained both by considering all the check-ins (top left subplot) as well as considering check-ins to specific types of locations (rest of the subplots). As we can see there is some level of local similarity that would have been expected even if people where checking-in completely at random. The percentage of local similarity that can be explained by PRM is even higher, and many times it appears to be the main reason for the levels of local similarity. For instance, for "Airports", the curve obtained from the real data is almost on top of the curve for PRM. Each city typically only has a few airports, among of which, even less support many connections,

and therefore, being popular. People will pick these airports not because they are influenced by their peers, but because they are more convenient.

Table 3.3 summarizes the progressive percentage of local similarity between friends that can be explained by the two reference models (`RRM` and `PRM`), as well as the *maximum* possible effect of local influence. As we can see `RRM` can explain approximately 44% of the observed local similarity when considering all the check-ins, while an additional 16% can be attributed to `PRM`. Consequently, local peer influence mechanisms can explain up to almost 40% of the local similarity between friends. Hence, ***friends appear to be influenced more easily when they are in proximity as compared to a global scale***. Moreover, the levels of local influence are context dependent. For instance, there appears to be no local peer influence in "Airports" (upper bound of local peer influence is only 10%), but significant levels are observed in "Pubs" (approximately 64%).

Table 3.3: Progressive percentage of local similarity that can be attributed to `RRM`, `PRM` and local peer influence.

Venue type	% Explained by `RRM`	Additional % explained by `PRM`	Upper bound on local peer influence
All Venues	43.93%	16.29%	**39.78%**
Coffee Shop	26.32%	21.47%	52.21%
Food	56.02%	8.63%	35.35%
Shopping	47.14%	1.31%	51.55%
Pub	12.55%	23.63%	63.82%
Airport	6.84%	82.94%	**10.22%**

Our previous results support the existence of *local* influence between friends. Nevertheless, we have explicitly focused on the activities of pairs of friends around each other's home location. Now, we seek to examine the existence of peer influence in a *third location*. In particular, Jack and Jill can have been area co-located in a third location, not specifically

at one of their home locations. We consider a third location, L_{3rd}, as a distant area (25kms far away) from both of the pairs' home locations. For example, Jack's home location is New York City and his friend's Jill's is San Francisco. If they have been area co-located in a third city (e.g., Boston), one might still influence the other. Note here that these situations are included in the global influence study. Nevertheless, they might have been *lost* in the aggregation of all different locations around the globe that users have visited. Hence, we examine them separately in what follows.

For every pair of friends (u, v) we use their combined check-ins to locations different than their home locations. We then exert the DBSCAN clustering method on these check-ins (features are the latitude/longitude pairs). If a cluster identified by the algorithm includes check-ins from both u and v, then we say this pair has been area co-located at L_{3rd}. Furthermore, L_{3rd} is considered to be the centroid of all the check-ins in the specific cluster. Similar to the above local similarity analysis, we use these area co-located pairs in a third location and we calculate a similarity score (which we refer to as remote similarity - $rsim$) as follows:

$$rsim_{u,v}^{data} = \frac{\mathbf{c}_u^{L_{3rd}} \cdot \mathbf{c}_v^{L_{3rd}}}{\|\mathbf{c}_u^{L_{3rd}}\|_2 \|\mathbf{c}_v^{L_{3rd}}\|_2} \tag{3.7}$$

where, $\mathbf{c}_u^{L_{3rd}}$ and $\mathbf{c}_v^{L_{3rd}}$ are as in Equation (3.4) but now we are considering only the check-ins of u and v respectively, in venues *near* L_{3rd} (i.e., within a radius of 25km from L_{3rd}). Similarly, we consider two reference models as in Equations (3.5) and (3.6):

$$rsim_{u,v}^{\texttt{RRM}} = \frac{\mathbf{c}_{u,\texttt{RRM}}^{L_{3rd}} \cdot \mathbf{c}_v^{L_{3rd}}}{\|\mathbf{c}_{u,\texttt{RRM}}^{L_{3rd}}\|_2 \|\mathbf{c}_v^{L_{3rd}}\|_2} \tag{3.8}$$

$$rsim_{u,v}^{\texttt{PRM}} = \frac{\mathbf{c}_{u,\texttt{PRM}}^{L_{3rd}} \cdot \mathbf{c}_v^{L_{3rd}}}{\|\mathbf{c}_{u,\texttt{PRM}}^{L_{3rd}}\|_2 \|\mathbf{c}_v^{L_{3rd}}\|_2} \tag{3.9}$$

where, $\mathbf{c}_{u,\texttt{RRM}}^{L_{3rd}}$ and $\mathbf{c}_{u,\texttt{PRM}}^{L_{3rd}}$ are randomly generated by considering venues around the center of the L_{3rd}. In the case of remote similarity, user v is the user that checked-in first in a venue around location L_{3rd}.

Figure 3.6 and Table 3.4 present our results, where we can see peer influence also exists in these so-called third locations. Compared to the local influence, the upper bound of similarity explained by peer influence is smaller. Nevertheless, this can be flipped when considering specific context (i.e., categories of venues). To sum up, even though global peer influence does not appear to be significant, if we focus our attention to remote geographic areas that both friends have visited - not necessarily their home locations - peer influence can possibly explain a large part of their similarity.

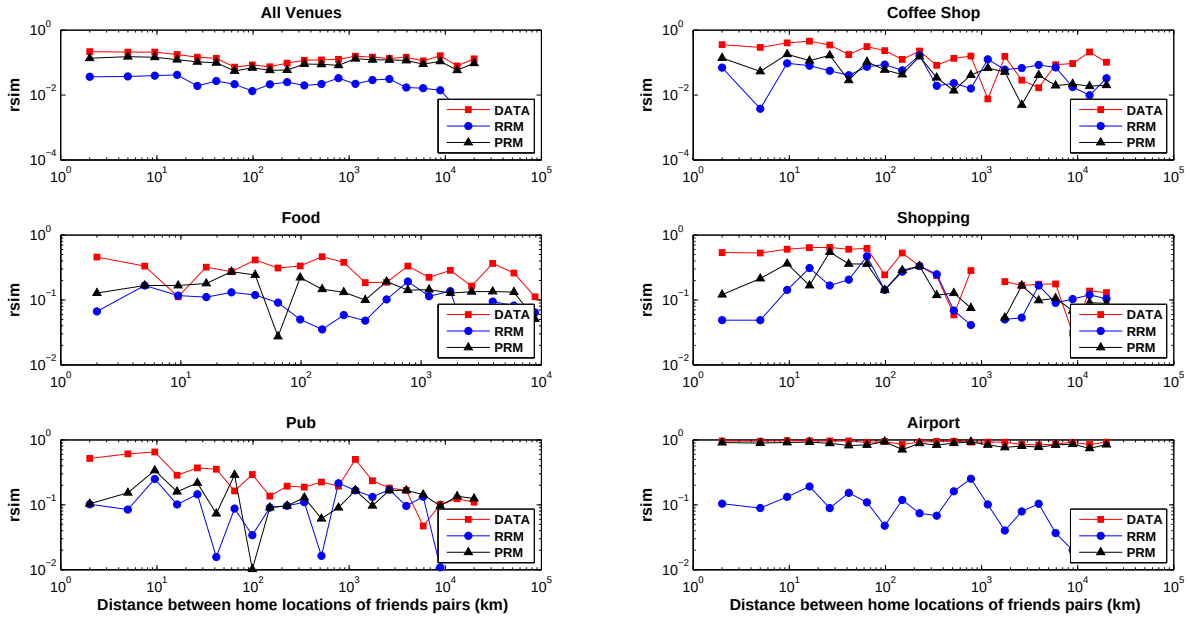


Figure 3.6: Similarity in a third location as obtained through data and two randomized reference models.

3.4 SOCIAL SELECTION

Social selection works between people that have high levels of similarity and can cause the creation of friendships. We further saw in the previous section that users exhibit some

Table 3.4: Progressive percentage of similarity in a third location that can be attributed to RRM, PRM and local peer influence.

Venue type	% Explained by RRM	Additional % explained by PRM	Upper bound on local peer influence
All Venues	9.73%	64.9%	25.37%
Coffee Shop	12.94%	0.51%	86.55%
Food	17.62%	9.04%	73.33%
Shopping	38.78%	19.63%	41.58%
Pub	46.49%	20.72%	32.8%
Airport	5.76%	85.77%	**8.47%**

inherent similarity, which also increases with time. Also by observing the absolute global similarity values of the null model in Section 3.3.1, we find that pairs of non-friends exhibit significant levels of global similarity as well. Why then social selection works with specific pairs and not with others?

When examining similar questions, we need to be cautious and in particular to avoid confusing *actual* similarity with what we refer to as "trivial" - or expected - similarity in this study. For instance, there are places that most of the people living in a city will visit, e.g., subway station(s), city hall etc. Such places introduce *trivial* similarity and it does not necessarily mean that the social selection mechanism will be triggered and these people will form social ties. In order to avoid confusions, we would like to emphasize here that trivial similarity is also part of the inherent similarity between two people[8]. However, it does not add valuable information.

In this section we seek to answer the question posed above and investigate the dynamics of the social selection mechanism. In particular, we examine the (network) characteristics of the venues - see Figure 3.1b - that appear to trigger the selection mechanism. More

[8]Actually, it might be the case that the "trivial" part of the inherent similarity is the major component that varies with time. Nevertheless, further examining this is beyond the scope of our work.

specifically, we consider pairs of friends that were check-in co-located prior to becoming friends, and hence they exhibited a non-zero level of similarity. In total, we have 84,460 such pairs. For these pairs of friends, we analyze the categories of the places that they were co-located prior to becoming friends, the *degree* - i.e., number of check-ins - of these places, their clustering coefficient as well as their entropy (to be defined later). As we will see in what follows, all of the above metrics exhibit the same properties for the friends pairs considered.

However, the above results alone are not conclusive. In particular, the common locations of other pairs of users that have been check-in co-located but never formed social ties, can also exhibit the same features. Hence, in order to avoid this sampling bias and to be able to draw safe conclusions, we need a *reference* group for comparison. We randomly pick 84,460 pairs of users that have been check-in co-located (hence, having some non-zero levels of similarity), but never became friends, and we calculate the same statistics for their common locations. Note here that, our random sampling retains in the reference group the same distribution of the home location distances as that for the check-in co-located pairs that eventually became friends. In particular, if there are μ pairs of friends with home distances in the range $[\chi_1, \chi_2]$, we sample uniformly at random μ reference pairs with home distances in the same range.

To preview our results, the features of the common venues of the reference group exhibit significant differences as compared with those of the friends pairs. Furthermore, the common venues of the reference group pairs manifest characteristics of locations that introduce trivial - or expected - similarity (e.g., high degree, low clustering coefficient, large entropy)! These results clearly indicate that **the (online) social selection mechanism works upon non-trivial similarity and can be stimulated by venues with specific features.** In other words, people tend to generate social ties with their peers with whom they exhibit non-trivial/unexpected similarity.

In what follows we introduce the venue metrics we examine and present the details of our results (Appendix B includes statistical significance results for the conclusions).

Venue Category: As mentioned in Section 3.1 every spot in Gowalla is labeled with a category depending on the type of place, and there are 283 possible categories. Hence, we

compute the category probability mass function of the common locations for the pairs in the two groups (friends and reference). Figure 3.7 depicts our results. As we can see, for pairs of users in the reference group, the mass function exhibits a clear *4-modal* distribution. On the contrary, the corresponding mass function for the pairs of friends is closer to a uniform distribution. The four categories-modes of the reference mass function are: "Convention center", "Interactive", "Airport" and "Travel/Lodging". As we can see these are types of places that people can co-locate at, not necessarily because of their similarity or common interests. For instance, there are many reasons that people go to a convention center. Airports are also potentially visited by all people (at the minimum all people that travel). On the contrary, friends tend to co-locate to a variety of places with fairly equal probabilities. Nevertheless, the top-4 places for friends are: "Corporate office", "Pub", "Food" and "Coffee shop". Corporate office is mainly visited by people that work there every day and hence they create tight bonds. Of course "Pub", "Food" and "Coffee shop" locations can also attract a diverse crowd, especially if these places are popular. Hence, while there is a clear difference at the category distributions between friends and reference pairs, we cannot claim that these results are absolutely conclusive. We will now focus on network characteristics of the venues, which are not tied to the category of the place. Such metrics can possibly make further distinctions even between places of the same category (e.g., two restaurants) and therefore, lead to stronger conclusions.

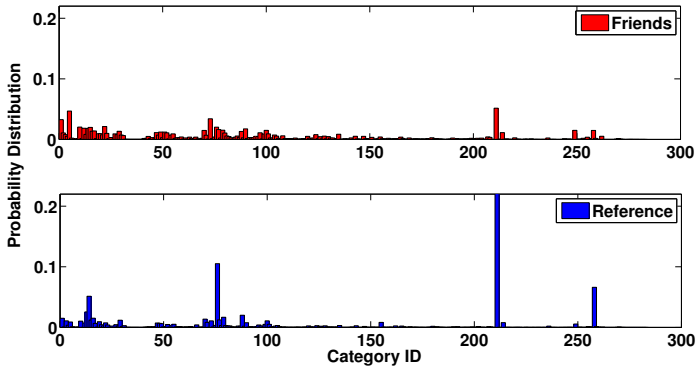


Figure 3.7: Friend and non-friend pairs have only 4 categories in common in their top 10 categories.

Venue Degree: Next, we examine the degree of the common venues. In particular, we define as the degree $deg(\pi)$ of a place/venue π, the number of total check-ins in this place. Figure 3.8 presents our results. As we can see the pair of users that eventually become friends have co-located prior to that to venues with lower average degree as compared to that of the common venues for the pairs in the reference group.

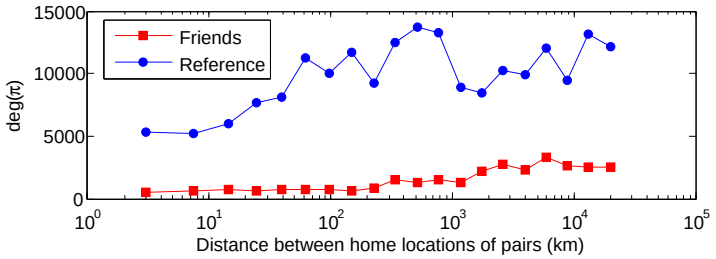


Figure 3.8: Users that form social ties co-locate to venues with low degree.

Venue Clustering Coefficient: If there are n_π unique people that have checked-in at place π, we define the clustering coefficient of π as follows:

$$CC_\pi = \frac{k}{n_\pi(n_\pi - 1)/2} \tag{3.10}$$

where k is the number of friendship pairs between the n_π users that have checked-in at π. Equation 3.10 is essentially the direct extension of the definition of the local clustering coefficient of a graph in our context. A high clustering coefficient for a venue translates to a location where people who visit it form a tightly connected social group. Figure 3.9 depicts our results, and as we can see pairs who become friends tend to co-locate to venues with higher CC as compared to the common places of the non-friends pairs of our reference group.

Entropy: Cranshaw *et al.* [40] use the notion of entropy of a location as a measure of its diversity. In particular, if $P_\pi(u)$ is the fraction of check-ins at place π contributed by user u, then the entropy of π is given by:

$$e(\pi) = -\sum_{u:u\in\mathcal{S}} P_\pi(u) log(P_\pi(u)) \tag{3.11}$$

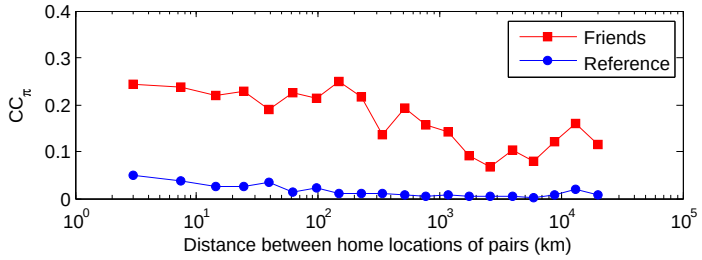


Figure 3.9: Venues with higher CC are more likely to form friendships.

where $\mathcal{S}$ is the set of users that have checked-in venue π.

From the definition of $e(\pi)$ we can see that when a place is visited by many people in fairly equal (and hence, small) proportions, its entropy will be high. Simply put, high entropy corresponds to places such as transportation hubs and malls that exhibit large diversity with regards to people they "attract". On the other hand, when the mass of $P_\pi(u)$ is concentrated only to a few people, the diversity in this location is small and so is the entropy.

In Figure 3.10 we present the average entropy of the common locations of the pairs in our reference, non-friends group, and that of the friends pairs prior to forming their tie. As we can see pairs who become friends after being co-located, have co-located to venues with lower average entropy compared to the average (common venues) entropy of our control group pairs.

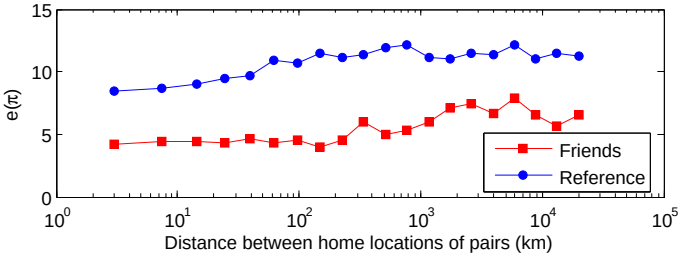


Figure 3.10: Users that form social ties co-locate to venues with lower average entropy compared to the reference pairs.

Note here that places with high degree, low clustering coefficient and high entropy are

essentially places that are responsible for trivial similarity. These are locations with large (high $deg(\pi)$), diverse (high $e(\pi)$) crowds that are disconnected in the social plane (low CC_π). These are places that people visit not essentially because they are in their preference but because they have to (e.g., airports, train stations, big university campus, medical centers etc.). On the other hand places with low degree, high clustering coefficient and low entropy, profile venues where people that visit them know each other and are actually similar, since they do not attract a large diverse public. Hence, based on our results we can clearly see that in order for the social selection mechanism to be triggered, actual similarity between users is required. Trivial similarity caused by co-locations in places with high degree, low clustering coefficient and high entropy, is not enough.

3.5 DISCUSSION AND IMPLICATIONS

We acknowledge that our study is limited by the information available in the dataset used. For instance, while we know the friendship creation time on the system between a pair of friends, we use the implicit assumption that this is also the actual time of the real-world friendship formation. Of course, this might not be always true. Furthermore, our dataset can exhibit biases with regards to the demographics of people that are using systems like Gowalla. Our results inevitably do not extensively take into consideration the behavior of parts of the population that are possibly underrepresented in the dataset (e.g., older people that might not be as technology savvy). Finally, the voluntary nature of location sharing can possibly introduce another type of bias for our analysis. More specifically, the activities shared in Gowalla (or any other similar social media platform - e.g., Foursquare etc.) are only partially reflecting people's trajectory. Nevertheless, these are essentially limitations shared - partially or entirely - by any study that is based on digital trails of human activities.

Despite the above limitations, we believe that our findings can stimulate further research on the topic and will contribute to eventually obtaining a more clear understanding on how people create social ties and move in real space. This understanding can facilitate a variety of applications. For example, it can drive enhancements in socially-aware recommender systems

or assist venue managers identify potential users for targeted advertisement. In the future, we seek to identify methods for controlling to the extent possible for the above biases and exactly quantify the time varying global similarity between people and decompose it to its various parts (e.g., trivial inherent, actual inherent etc.). We also opt to examine groups of friends (rather than only pairs as in our current work) and the ways peer influence operates in such settings.

3.6 RELATED WORK

The availability of electronic traces of human activities has enabled the study of topics related to homophily, information diffusion, social selection and peer influence. For instance, Kossinets and Watts [92] utilize a dataset comprising of e-mail communications between members of a large U.S. university to study the origins of the homophily in the underlying communication network. The same authors [91] study the effects of social selection on friendship formation using again an e-mail communication network of a U.S. university. They show that the friendship probability between two students increases up to a certain number of common interests - as captured from the number of common classes between students - and then remains constant. In the same direction, Lewis *et al.* [98] examine the mechanisms of social selection and peer influence in a group of students of a U.S. institution by studying the co-evolution of the friendships and tastes in music, movies and books, over a period of four years.

Other studies provide some basic intuition on how innovations propagate through the network and they further show that social ties can facilitate information contagion and consequently influence users' actions. For instance, Bakshy *et al.* [14] use data from Second Life to examine the social influence on the adoption of content by the users. Among other findings, they show that adoption rate increases with an increased number of friends that have already adopted a content. However, as one might expect not all adoptions can be attributed to peer influence and in-network effects. Myers *et al.* [109] studied the effect of external influences on information diffusion using data collected from Twitter. They further

developed a model through which they can quantify external influence over time. Very recently, Tang *et al.* [152] studied conformity, which can be thought of as a special type of social influence. In particular, conformity is the action of matching one's actions to the norms of the groups he belongs to. The authors' results indicate that conformity exists in all four digital social network datasets they examined.

A different line of work, studies statistical methodologies for unveiling the roots of homophily. For instance, Anagnostopoulos *et al.* [3] design a statistical test, the shuffle test, for deciding whether peer influence is a likely source of the observed homophily. In brief, the key idea behind the shuffle test, is that if influence is not a possible source of the assortative mixing, timing of actions should not matter. Hence, *reshuffling* of the timestamps of the events, should not significantly change the assortativity level in the network. Another statistical framework for distinguishing between influence and selection effects in dynamic networks was presented in [5]. In particular, the authors develop a dynamic matched sample estimation framework and apply it on a large-scale network dataset that captures the adoption of a specific product. In parts of our work, we use an approach similar to that followed by Crandall *et al.* [37]. The authors study the social selection and influence in online communities such as WikiPedia and LiveJournal. In order to distinguish between social selection and influence, they examine the temporal evolution of the friends similarity prior and after they became associated.

In this chapter, we focus on a novel type of online social media, namely LBSNs, that only recently has attracted attention from the research community. The key difference between traditional online social media and LBSNs, is that the latter directly relates online interactions with physical space, and hence, studies of LBSNs can have stronger implications of actual real-world behaviors of people. The work from Cho *et al.* [31] is the closest one to our study. In particular, the authors examine the relationship between friendship and users' mobility. They show that the actual geographic location (latitude/longitude) that users travel to is influenced by the presence of a friend or not. They further show that this influence increases with an increase in the distance between the home locations of friends. In our work we further delve into the details of peer influence, and in particular we examine both its geographic scope, as well as its contextual properties (existing literature seems to support

the general connection between influence and *topics* [151]). In particular, we do not simply consider geographic locations, but actual venues, and examine the strength of influence for different types of places. Furthermore, we examine the social selection mechanism and how this is realized through the different types of venues (e.g., are there specific characteristics of the venues that *promote* friendship creation by stimulating social selection?).

3.7 SUMMARY

In this chapter, we examine the peer influence and social selection mechanisms in the context of locations visited by friends using a longitudinal dataset from a location-based social network. We find that strong evidence for peer influence existence with regards to human movement in urban space as long as friends are in proximity, and it is context dependent. In particular, for specific types of places (e.g., nightlife spots) users can influence their peers more as compared to other types of locations (e.g., airports). We also reveal, that social selection works upon non-trivial similarity and there are particular venues - with specific network characteristics - that can trigger the social selection mechanism.

In the following chapter, I further present our study on evaluating and modeling the impact of economic incentives, captured by the local business advertising strategy in LBSNs, on people's movement to business places.

4.0 EFFECTIVENESS OF LOCAL BUSINESS ADVERTISEMENT

Location has been identified as a critical factor that can decisively affect the success of a business [83, 84, 124]. Specifically, previous relevant work has verified the intuitive causal connection between the location of a business and the volume of potential customers that it has access to. For instance, a business in a crowded urban neighborhood is exposed to more potential customers than a business in a sparsely populated location [86, 125]. Similarly, the reach of a business can benefit from its proximity to a popular landmark or busy hub [123]. The potential of such benefits makes some locations more desirable than others. Predictably, the increased demand raises the setup cost (e.g rent, taxes) and makes prime locations unattainable for most businesses [144, 56]. Further, even if a businesses manages to secure a favorable location, it is likely to face fierce competition by businesses that also had the means to pay the necessary costs. Previous work has repeatedly verified that, in such competitive settings, it is highly likely to observe "rich-get-richer" phenomena, in which a small subset of the competing businesses claim the lion's share of the customer base [148]. These could be older businesses that had more time to build their reputation and connect with customers, wealthy businesses with superior marketing capabilities, or simply elite competitors that deservedly attract customers with their service quality. On the other hand, businesses outside this "winner's circle" face an uphill battle in their effort to expand their reach and increase their market share, even if they are in a privileged location with access to a large customer base.

Motivated by such challenges, a new location-based advertising outlet has emerged, supported by the advancement and establishment of mobile technology. This outlet is implemented via Location-Based Social Networks (LBSNs) with millions of users, such as Foursquare, Yelp, and UrbanSpoon. After partnering with an LBSN, a business gains access

to a vast user base. This channel is then utilized in two ways. First, after users reveal their location to the LBSN, they can use the platform to locate nearby businesses of different types. Thus, a business with a profile on the LBSN can be discovered by potential customers who might otherwise be unaware of its presence. For instance, in the example shown in Figure 4.1, Alice finds herself in a central business district (CBD, marked with purple), which includes a mall, a park, and other amenities. During her visit, Alice is likely to walk by the restaurants within CBD, given their privileged, central location. However, by using an LBSN, Alice can specify an acceptable range (marked as a blue circle) and discover less obvious options, such as restaurants r_1, r_2 and r_3. This is a mutually beneficial discovery, which increases both the reach of these restaurants and the number of Alice's options. However, Alice could arguably be less willing to seriously consider these less prominent options, as she may be uncertain of their quality. Thankfully, a business can decrease this uncertainly by maintaining an attractive professional profile on the LBSN, including up-to-date information, pictures, and informative reviews [30].

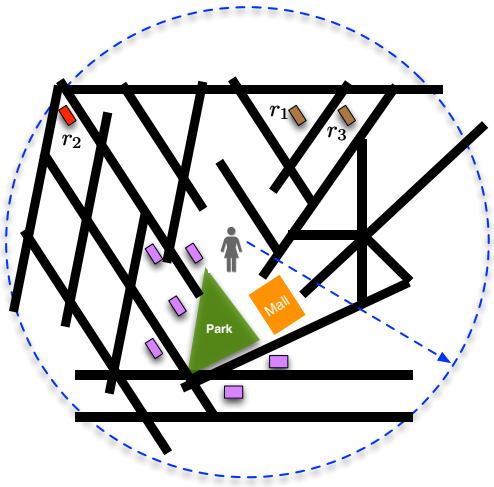


Figure 4.1: Mobile and spatial computing allows customers to discover establishments in non-prime locations (e.g., within the blue range). Moreover, it allows venues (e.g., r_2) to offer monetary incentives through special offers to gravitate customers towards them.

In addition to increasing foot traffic, LBSNs offer business owners additional mecha-

nisms for affordable advertisement. Specifically, a business can use the LBSN's network to promote special offers. Mainstream media is rich with stories on successful LBSN promotions [54, 45, 61]. In 2013, a burger joint in Philadelphia reportedly experienced an increase in customers via a campaign on Foursquare, which offered a free beer to users that used the popular LBSN to publicly state their presence in the restaurant, an action referred to as a "check-in". Earlier, in 2010, a Milwaukee restaurateur used a promotional campaign to attract 161 Foursquare members into his burger restaurant at the same time. Customers were lured by the promise of the coveted "Swarm badge", which Foursquare awards if more than 50 users check-in at a venue at the same time. In the same year, popular fast-food chain McDonalds launched a Foursquare campaign that offered gift certificates to users who checked-in at certain randomly selected McDonalds locations. Given that the selected locations were not released, users were motivated to visit multiple McDonalds restaurants, leading to a 33% increase in the number of check-ins. Despite the plethora of promising anecdotal evidence, a systematic study of the effectiveness of the LBSN advertising paradigm has not been conducted mainly due the lack of appropriate data. Our work is the first to address this challenge by studying a large longitudinal dataset of about 14 million businesses on Foursquare. Our study formally evaluates both the long-term and short-term effects of LBSN campaigns for participating businesses, while taking into consideration the influence of possible confounding factors.

Our main result indicates that the positive effects of special offers through the LBSN platform examined are significantly more limited than what anecdotal success stories seem to suggest. In particular, we find no evidence of a statistically significant advantage, in terms of either the number of daily check-ins or that of new customers, for venues that participate in LBSN campaigns in the *platform examined*. We validate our findings by adopting two alternative methods for statistical testing, which lead to the same conclusions. **In addition, in order to gain a deeper understanding of our results and increase the practical value of our methodology, we design and implement a model for predicting the popularity of a venue during and after a campaign.** Our models consider venue-related, promotion-related, and geographical features. Our experiments provide encouraging evidence on the feasibility of this prediction task, which can

serve as a practical tool for supporting the design and cost-benefit analysis of LBSN campaigns. Specifically, we find that a simple logistic regression model is sufficient to achieve an 83% accuracy with a 0.88 AUC. Further, our findings on the influence of the considered features are fully aligned with our main results, as we find that promotion-related features have only a marginal contribution to the estimation of popularity. In Section 4.4, we discuss the implications of our work for businesses but also for the LBSN platforms as well. In particular, we describe how our findings can be used to inform strategies for improving campaign effectiveness.

Chapter Outline: The rest of the chapter is organized as follows. Section 4.1 describes the time-series dataset we collected, some basic analysis on the data and our hypothesis setting. In Section 4.2 we present the details of our statistical test framework and the results for local promotion effectiveness, which in Section 4.3 we further present a supervised learning model to predict the short and long effect. We then discuss with implications in Section 4.4. We elaborate the related work in Section 4.5 and finally summarize this chapter in Section 4.6.

4.1 DATASET AND ANALYSIS SETUP

In this section I first briefly describe the Foursquare data collected and present a basic analysis of the acquired dataset. Then I formally introduce the research hypothesis we will examine.

4.1.1 Data Collection and Analysis

Using Foursquare's public venue API during the 7-month period between *22th October, 2012 and 22th May, 2013* we queried and obtained information for **14,011,045** venues once every day. This essentially gives us a multi-dimensional time-series for every venue, with daily readings, where each reading has the following tuple format: <**ID, time, # check-ins, # unique users, # specials, # tips, # likes, tip information, special information**>.

During the data collection period, there are 206,163 venues in total that have published at least one special. Approximately 45% of these venues publish only one special. Furthermore, there are in total 735,034 unique special deals, with 88.68% of them being provided by venues in the US.

At the time, Foursquare had 7 types of specials, namely, "Newbie", "Flash", "Frequency", "Friends", "Mayor", "Loyalty" and "Swarm", each requiring different conditions to be earned [60]. Table 4.1 presents the description of the different types and their popularity in our dataset. As we can observe, "Frequency" is the most popular type of special in our dataset, possibly because compared to other types appears to be the easiest one to be *unlocked*, covering approximately 86.5% of all the offers we collected. Compared to other types, "Frequency" appears to be the easiest one to be *unlocked* from many perspectives. For example, a user does not need the *help* of other users as is the case for "Friends" or "Swarm" special deals. Furthermore, the user does not need to compete with other frequent users checking-in to this venue as is the case for the "Mayor" special offers. Similarly, he/she is not constrained by time (as in the "Flash" special).

Another parameter of interest for the special offers is their time duration. Figure 4.2 presents the empirical CDF of the offer duration. As we can see, "Frequency" and "Flash" special offers usually are active for a short duration, while "Friends" and "Swarm" usually last for a longer time possibly due to their stricter requirements. The "Mayor" special often lasts even longer, since a customer needs to become the Foursquare *mayor* of the venue to unlock the deal. The *mayorship* is only awarded to the user who has the most check-ins in the venue during the last two months.

As alluded to above, a venue might offer multiple specials during the 7-month data collection period. These multiple specials can be fully overlapped (i.e., they start and end at the same time), partially overlapped, or sequential. We further define a **promotion period** of a venue to be a continuous time period that the venue provides at least one offer and does not include more than two consecutive days without a special offer. In our dataset, approximately half of the promotions last for more than a week. While a promotion as defined above can include multiple individual offers, for simplicity we will use the terms promotion, offer, campaign and deal interchangeably in the rest of the paper.

Table 4.1: Type of specials in Foursquare. "Frequency" is the most common type provided by Foursquare venues in our 7-month dataset.

Type	Count	Description
Count/Newbie	57,710	This type of special is unlocked on a user's first time ever visiting the venue. The objective is to drive new traffic to the venue.
Flash	5,989	Venue sets the number of specials that can be unlocked per day, in a first-come, first-serve fashion, or defines an active time-window for the special live. When the unlock limit is reached, there are no more specials for the day.
Frequency	636,119	Unlocked after every or several check-ins. The objective is to reward users on their routine check-ins.
Friends	5,469	Venue sets the minimum threshold for a group of Foursquare friends. The objective is to reward friends for visiting the establishment together.
Mayor	22,021	This special is awarded to the Foursquare mayor of the venue.
Regular/Loyalty	6,488	Venue rewards a user every X times they visit, or for coming in X times in total, or for being loyal within a certain period. The objective is to encourage the user to keep coming back to the venue.
Swarm	1,238	A swarm special is aiming at many people checking-in at the same time. The venue can set a minimum number of Foursquare users (not necessarily friends) that need to check-in within a time-window in order to unlock this special.

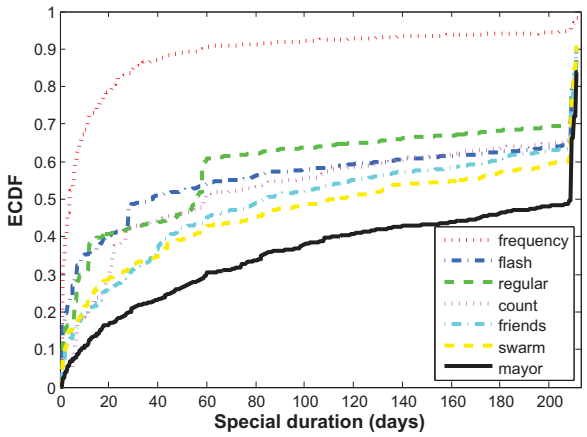


Figure 4.2: "Frequency" and "Flash" specials are usually shorter than other types of specials. The "Mayor" special often lasts for a longer period time.

Finally, Foursquare associates each venue v with a category/type $T(v)$ (e.g., cafe, school etc.). This classification is hierarchical, in the sense that an Italian restaurant belongs to the category "Italian restaurant", which can belong to the higher level category "Restaurants", which can itself belong to the category "Food" and so on. At the top level of the hierarchy there were 9 categories during the time of data collection; *Nightlife Spots*, *Food*, *Shops*

& Services, *Arts & Entertainment*, *College & University*, *Outdoors & Recreation*, *Travel & Transport*, *Residences* and *Professional & Other Places*. From these types, we examine the fraction of venues in each top-level category that offer at least one special deal during the data collection period (Table 4.2). As we can see "Food", "Nightlife Spots" and "Shops & Services" have the highest chances of offering a special deal (0.025, 0.04 and 0.016 respectively). This can be attributed to the fact that the majority of the venues in these categories are commercial and hence, advertisement is most probably among their priorities. While non-commercial venues can also publish specials with the ultimate goal of increasing their visibility, it is certainly less expected and our data verify this.

Table 4.2: *Food*, *Nightlife* and *Shops & Services* venues exhibit the highest probabilities to publish a special offer in our dataset.

Category	# venues	# (%) venues with specials
Nightlife Spots	558,156	6,493 (**1.16**%)
Food	2,604,408	66,136 (**2.54**%)
Shops & Services	2,693,300	107,517 (**3.99**%)
Arts & Entertainment	491,426	5,050 (1.03%)
College & University	493,600	1,923 (0.39%)
Outdoors & Recreation	936,943	1,370 (0.15%)
Travel & Transport	897,404	8,178 (0.91%)
Residences	2,902,492	489 (0.02%)
Professional & Other Places	2,354,975	8,311 (0.35%)

4.1.2 Hypothesis Development

The objective for every business behind offering coupons, discounts and any other type of offers is ultimately to drive revenues up. This could be either through returning customers or through attracting new customers. The same is true for participating businesses in electronic

promotions through a variety of platforms (e.g., Groupon, Living Social, etc.) including location-based social media as well.

One of the benefits of running electronic promotions is that they can be objectively evaluated since - in the majority of the times - this is the only channel through which customers could have learned about the promotion. On the contrary, when offline advertisements and promotions are run, it is not clear what influenced the decision of the customers. For promotions through LBSNs specifically the customers need to check-in through the corresponding platform in order to obtain the discount and hence, the effectiveness of the campaign can be better tracked.

In the rest of this chapter we will therefore examine the following two hypotheses:

Hypothesis 4.1 (Short Term Effectiveness)**.** *The presence of a promotion through location-based social media leads to an increase in the visitation of a local business during the duration of the campaign.*

Hypothesis 4.2 (Long Term Effectiveness)**.** *The presence of a promotion through location-based social media leads to an increase in the visitation of a local business after the campaign has been completed.*

In the next section we will analyze the data we collected in order to verify or reject the above hypothesis. In order to capture the success of a promotion we will rely on the number of check-ins in the corresponding business as well as the number of new customers. We would like to emphasize here that while these hypotheses target a generic LBSN platform, our conclusions are inevitable more relevant to the platform used in the study (i.e., Foursquare). Nevertheless, as we discuss in Section 4.4 our study provides lessons and knowledge that can be applicable to any similar platform.

4.2 STATISTICAL ANALYSIS

Evaluation metric: As alluded to above in order to assess the effectiveness of LBSNs promotions we will rely on the number of check-ins and unique visitors in the venues. Our

data are in a time-series format and we also know the start (t_s) and the end (t_e) times of the promotion period. Using these points we split each time-series to three parts that span the following periods: (i) **before** the special campaign, $[t_0, t_{s-1}]$, (ii) **during** the special campaign, $[t_s, t_e]$, and (iii) **after** the special campaign,$[t_{e+1}, t_n]$. The key idea is to examine and analyze the changes that occur at the daily check-ins and unique visitors across these three time periods.

Data processing: Let us denote the original time-series collected for the check-ins in venue v with $c_{av}[t]$ and that for the unique visitors in v with $p_{av}[t]$. Simply put, $c_{av}[t]$ ($p_{av}[t]$) is the accumulated number of check-ins (unique visitors) in v at time t. As aforementioned we obtain one reading every day for every venue. However, consecutive readings might not be exactly equally-spaced in time due to a variety of reasons (e.g., network delays, API temporal inaccessibility etc.). Hence, we transform each time series to the intended reference time-points using interpolation. For the rest of the paper $c_{av}[\tau]$ ($p_{av}[\tau]$), will represent the interpolated time-series for the total number of check-ins (unique visitors) in v with $\tau_{i+1} - \tau_i = 24$ hours.

In our analysis we focus on campaign periods of venues in the US, since almost 90% of the special deals are offered by US venues, that last for at least 7 days and for which we have enough points in the time-series before the special offer (i.e., at least 4 weeks). This allows us to build a representative baseline for the venue popularity prior to the promotion. The above filters provide us with a final dataset of 40,071 promotion periods offered by 36,567 venues. We refer to this dataset as the *promotion* dataset. Note here, that only a subset of those can be used for studying the long-term effect of the promotion. In particular, for 26,355 of them we have enough points in the time-series after the special offer, and we use them for the long-term effect study.

Since our metric of interest is the daily check-ins, we will utilize the first-order difference of the aggregate time series:

$$c_v[\tau] = c_{av}[\tau] - c_{av}[\tau - 1] \tag{4.1}$$

Similarly, for the daily new customers we have:

$$p_v[\tau] = p_{av}[\tau] - p_{av}[\tau - 1] \tag{4.2}$$

The raw data we have collected might exhibit biases that affect our analysis. For instance, a change in a venue's daily check-ins might simply be a result of a change in the popularity of the social media application. Moreover, seasonality effects can distort the contribution of the campaign on $c_v[\tau]$ and/or $p_v[\tau]$. To factor in our analysis similar potential sources of bias we use a matched reference group of venues that can account for the effects of similar externalities.

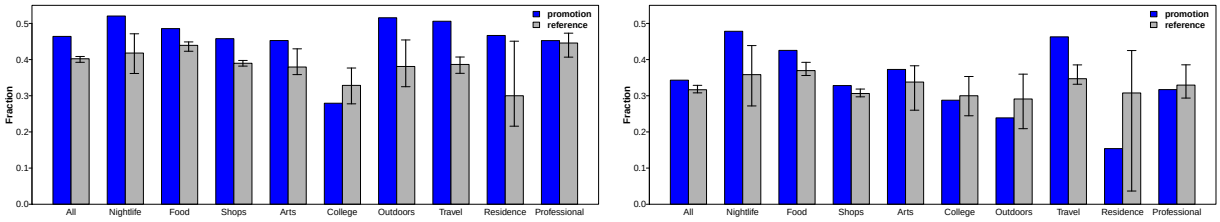
4.2.1 Promotion Dataset Analysis

We begin by examining the fraction of promotions that enjoy an increase in the mean number of check-ins per day. Let us denote the mean check-ins per day[1], in venue v before the promotion (i.e., during the period $[t_{s-k}, t_{s-1}]$) with $m^b_{c_v}$. We similarly define the average check-ins per day in v during (i.e., in the time period $[t_s, t_e]$) and after (i.e., in the time period $[t_{s+1}, t_{s+w}]$) the promotion campaign as $m^d_{c_v}$ and $m^a_{c_v}$ respectively. To reiterate, in order to build a concrete baseline for the period prior to the promotion we set $k = 28$ *days*. In order to study the long term effect of the promotion we would like to have a stabilized time interval after the campaign is over. Hence, we include in our analysis only the venues for which we have data for at least 7 days after the end of the promotion. Consequently, we set $w = k$, if we have 28 days of data after the promotion. Otherwise we set w equal to the number of time-points available (i.e., $7 \leq w \leq 28$).

Given this setting we first compute the difference $m^d_{c_v} - m^b_{c_v}$ $(m^a_{c_v} - m^b_{c_v})$. A positive sign essentially translates to an increase in the average daily check-ins during (after) the promotion period. Figure 4.3 depicts our results. As we can see, the fraction of venues in the promotion group that enjoy an increase in their check-ins during the special offer is approximately 48%, while a smaller fraction (about 35%) exhibits an increase after the promotion is ceased. There is also some variation observed based on the venue type, with some categories exhibiting larger fraction of venues with an increase (e.g., nightlife). However, part of this variability might be attributed to the fact that for some categories we have a very small sample in the promotion set (e.g., we only have 128 promotions in *Outdoors* and

[1]Exactly the same analysis set-up is followed for the mean number of new users per day.

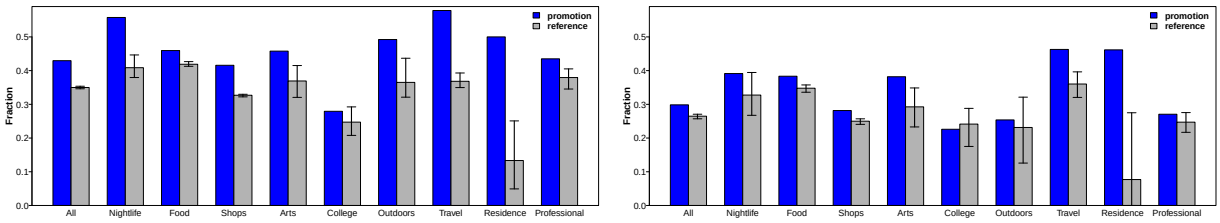
30 in *Residence*).



(a) Short-term (b) Long-term

Figure 4.3: Fraction of venues exhibiting an increase in the **mean daily check-ins**.

Similar results are obtained when examining the average daily new users that visit a venue. In particular we examine the difference $(m^d_{p_v} - m^b_{p_v})$ and $(m^a_{p_v} - m^b_{p_v})$. Figure 4.4 depicts our results, where again we observe that there is a large fraction of venues in the promotion group that enjoys an increase in the new users checking-in per day.



(a) Short-term (b) Long-term

Figure 4.4: Fraction of venues exhibiting an increase in the **mean daily unique customers**.

In summary, a large fraction of venues exhibits increase in their check-ins as well as their new customers during and after the special offer. However, an equally large proportion of venues does not enjoy an increase in the average daily check-ins. Next we delve further into the details of the effectiveness of LBSN promotions.

4.2.2 Reference Venues

Our results above clearly cannot establish any causal relationship between the promotion campaign and the observed changes in the daily check-ins and/or new customers. This would

require careful design of field experiments. In such randomized experiments the covariates distribution between the treatment and control groups are matched in expectation. However, this is not possible in our work since we only have access to observational data. The direct comparison between venues that offer promotions and those that do not, can be affected by a confounding bias introduced by the systematic assignment of the treatment. This would lead to comparing two groups with unbalanced covariates distributions.

In order to account for these confounding factors and other externalities, we opt to get a baseline for comparison by utilizing techniques for quasi-experimental studies. A typical approach used in these cases where observational data are only available is to match the covariates distributions in the two groups [73]. In particular, we sample a reference group from the set of venues with no promotion, such that the distribution of specific *observed* features of this sample *matches* that of the promotion group. The matched venues can thus be interpreted as the counterfactuals. We perform the matching on a venue-basis and we use four features/covariates. Specifically we use (i) the location of the venue, (ii) the type of the venue, (iii) the popularity of the venue prior to the promotion (i.e., the number of total check-ins) and (iv) the rate of change in the daily check-ins of the venue the period prior to the promotion. The reference group also ensures that on average the venues at both groups will experience similar externalities (e.g., seasonal effects, effects related to the popularity of Foursquare etc.). Once the reference group is obtained, we sample the empirical promotion period distribution of the promotion venues and assign pseudo-promotion periods to the reference group venues. Consequently we perform the same analysis described in the previous section on the reference group.

Our results from 20 non-overlapping reference groups for the daily check-ins are also depicted in Figure 4.3, where the 95% confidence intervals are also presented. As we can see the fraction of venues enjoying an increase in the promotion group is higher compared to that in the reference group. If we denote with $I_{c,d}$ ($I_{c,a}$) the event of an increase for $m^d_{c_v}$ ($m^a_{c_v}$), with S the event of a venue offering a special deal and with E the various environmental externalities that are present, the reference group opts to obtain an estimate for the probability $P(I_{c,d}|E)$. On the other hand, the promotion group includes an additional externality, the presence of a promotion. Hence, with the promotion group we are able to

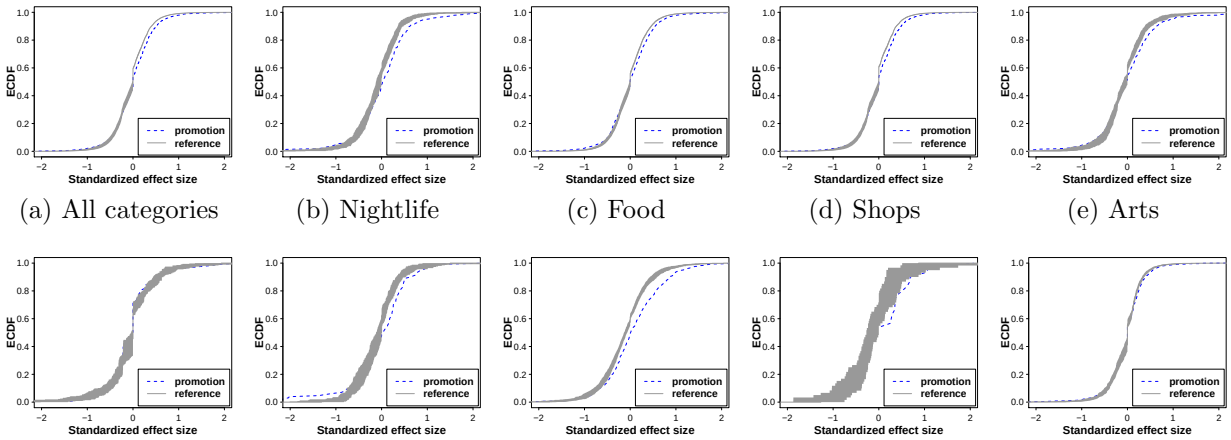
estimate $P(I_{c,d}|S,E)$. Our results indicate that $P(I_{c,d}|S,E) > P(I_{c,d}|E)$ and $P(I_{c,a}|S,E) > P(I_{c,a}|E)$ when considering all types of venues. However, the difference between these two probabilities is less than 0.1 for both the short and long term. Similarly, Figure 4.4 depicts the results from the reference groups with respect to the new customers. If we denote with $I_{p,d}$ ($I_{p,a}$) the event of an increase for $m^d_{p_v}$ ($m^a_{p_v}$), we observe that again $P(I_{p,d}|S,E) > P(I_{p,d}|E)$ and $P(I_{p,a}|S,E) > P(I_{p,a}|E)$ when considering all types of venues. Again the difference is smaller than 0.1 for both the short and long term.

Another aspect related with the potential effectiveness of the promotion campaign is the actual effect size of the observed change. The degree of this change can be captured through the standardized effect size of Cohen's d. For example, when considering the effect during the promotion on the daily check-ins:

$$d_{c,d} = \frac{m^d_{c_v} - m^b_{c_v}}{\sigma_{pooled}} \tag{4.3}$$

where σ_{pooled} is the pooled standard deviation of the two samples (before and during the promotion). Similar definitions of course are used for the daily new customers as well as the effects after the promotion. Figures 4.5 and 4.6 present the empirical CDF for the observed standardized effect sizes on the daily number of check-ins in both the promotion and the reference groups for the short and long term respectively. For the reference groups we also present the 95% confidence intervals of the distributions. As we can observe there is a shift in the distribution for the promotion group, which is different for different categories. However, this shift is very small. Furthermore, an interesting point to observe is the jump at the reference groups' ECDF at $d = 0$. This means that there is a non-negligible fraction of venues in the reference group that have exactly the same mean for the two periods compared. We will come back to this observation in the following section.

Figures 4.7 and 4.8 present the ECDF for the standardized effect sizes $d_{p,d}$ and $d_{p,a}$. As we can see the results are very similar to the ones for the standardized effect size on the mean daily check-ins. Also note here that, the jump at the reference groups' ECDF at $d = 0$ is observed as well.



(a) All categories (b) Nightlife (c) Food (d) Shops (e) Arts

(f) College (g) Outdoors (h) Travel (i) Residence (j) Professional

Figure 4.5: Both the promotion and reference groups enjoy similar effect sizes $d_{c,d}$ on the daily check-ins.

4.2.3 Bootstrap Tests

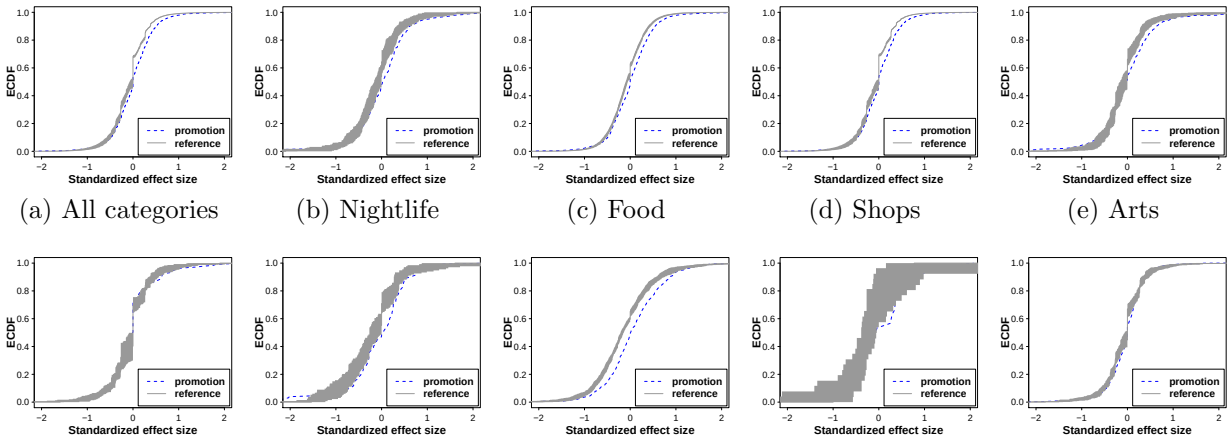
Our results above indicate that a large number of venues exhibit small effect sizes, which might not represent robust observations. Therefore, in this section we opt to identify and analyze the promotions in our dataset that are associated with a statistically significant change in their check-ins and/or their new customers.

Given our setting, the following two-sided hypothesis test examines whether their is a statistically significant change observed in the short-term with respect to the daily check-ins:

$$H_0 : \quad m^b_{c_v} = m^d_{c_v} \tag{4.4}$$

$$H_1 : \quad m^b_{c_v} \neq m^d_{c_v} \tag{4.5}$$

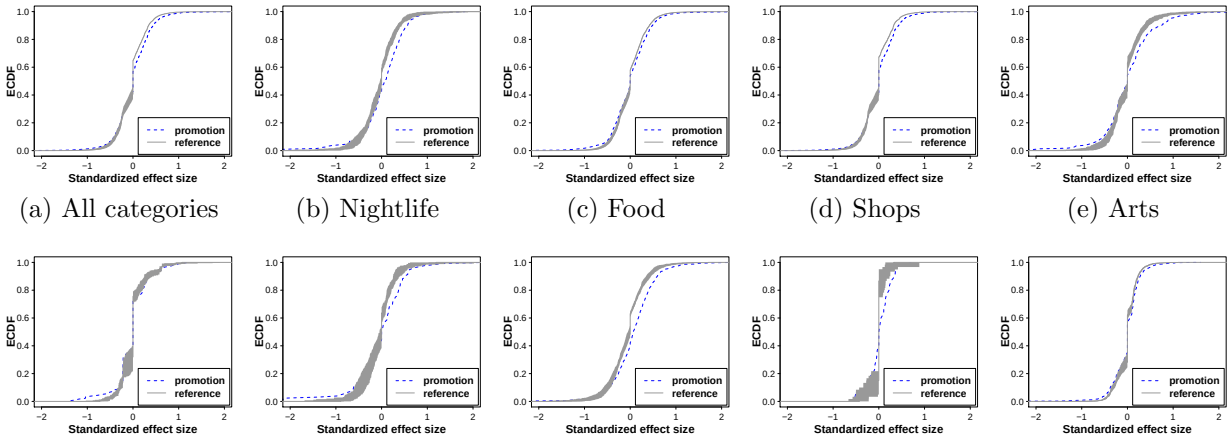
The sign of the observed difference will further inform us if the change is positive. In our analysis we pick the typical value of significance level $\alpha = 0.05$. If we want to examine the long-term effectiveness of special deals on the daily check-ins we devise the same test as in Equations (4.4) and (4.5), where we substitute $m^d_{c_v}$ with $m^a_{c_v}$, while similar tests are performed for the daily new customers. We choose to rely on bootstrap for the hypothesis



(a) All categories (b) Nightlife (c) Food (d) Shops (e) Arts

(f) College (g) Outdoors (h) Travel (i) Residence (j) Professional

Figure 4.6: ECDF of the standardized effect size $d_{c,a}$ on the daily check-ins after the promotion.

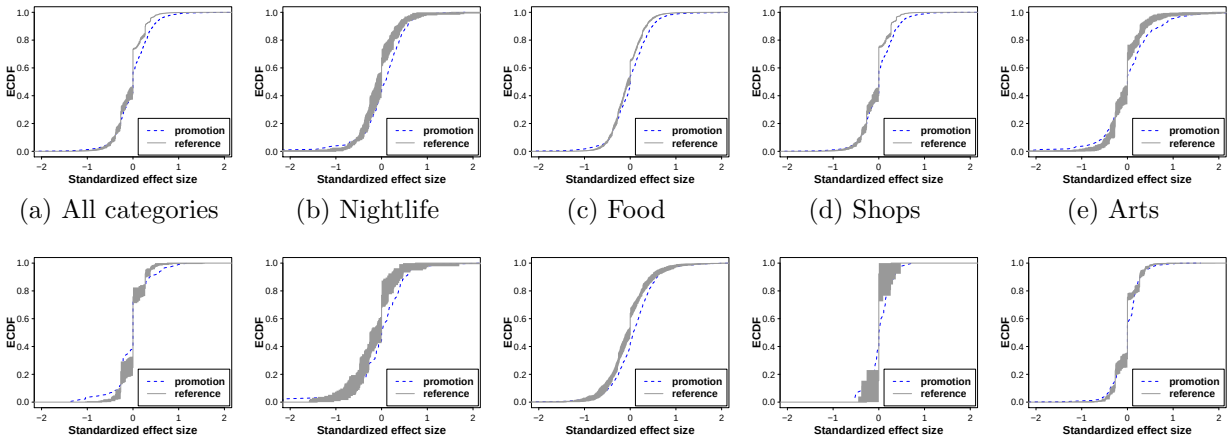


(a) All categories (b) Nightlife (c) Food (d) Shops (e) Arts

(f) College (g) Outdoors (h) Travel (i) Residence (j) Professional

Figure 4.7: Both the promotion and reference groups enjoy similar effect sizes $d_{p,d}$.

testing rather than on the t-test to avoid any assumption for the distribution of the check-ins (or the new users). Bootstrap also allows us to estimate the statistical power π of the performed test. This is important since an underpowered test might be unable to detect statistically significant changes especially if the effect size and/or the sample size are small. Consequently, this can lead to underestimation of the cases where the alternative hypothesis



(a) All categories (b) Nightlife (c) Food (d) Shops (e) Arts

(f) College (g) Outdoors (h) Travel (i) Residence (j) Professional

Figure 4.8: The difference between the effect size $d_{p,a}$ for the promotion and reference groups is the largest observed. Nevertheless, it is still fairly small.

is true.

Statistical bootstrap [47] is a robust method for estimating the unknown distribution of a population's statistic when a sample of the population is known. The basic idea of the bootstrapping method is that in the absence of any other information about the population, the observed sample contains all the available information for the underlying distribution. Thus, resampling with replacement is the best guide to what can be expected from the population distribution had the latter been available. Generating a large number of such resamples allows us to get a very accurate estimate of the required distribution. Furthermore, for time-series data, block resampling retains any dependencies between consecutive data points [95].

In our study we will use block bootstrapping with a block size of 2 to perform the hypothesis tests. When performing a statistical test we are interested in examining whether under the null hypothesis, the observed value for the statistic of interest was highly unlikely to have been observed by chance. In our setting, under H_0 the two populations have the same mean, i.e., $m^d_{c_v} - m^b_{c_v} = 0$. Hence, we first center both samples, before and during the special, to a common mean (e.g., zero by subtracting each mean respectively) in order to force the null hypothesis to be true. Then we bootstrap each of these samples and calculate

the difference between the new bootstrapped samples. By performing $\mathcal{B} = 4999$ bootstraps, we are able to build the distribution of the difference $m^d_{c_v} - m^b_{c_v}$ under H_0. If the $(1-\alpha)$ confidence interval of $m^d_{c_v} - m^b_{c_v}$ under the null hypothesis does not include the observed value from the data, then we can reject H_0. An empirical p-value can also be calculated by computing the fraction of bootstrap samples that led to an absolute difference greater than the one observed in the data.

With statistical bootstrapping we can further estimate the power π of the statistical test performed. π is the conditional probability of rejecting the null hypothesis given that the alternative hypothesis is true. For calculating π we start by following exactly the same process as above, but without centering the samples to a common mean. This will allow us to build the distribution of $m^d_{c_v} - m^b_{c_v}$ under H_1. Then the power of the test is the overlap between the critical region and the area below the distribution curve under H_1.

We have applied the bootstrap hypothesis test on our promotion and reference groups. Figure 4.9 presents our results for all types of venues, while similar behavior is observed for specific venue categories. In particular, we calculate the fraction of promotions associated with a statistically significant increase in the average daily check-ins. Note that we consider only the promotions whose p-value is less than $\alpha = 0.05$ or $\pi \geq 0.8$ (the latter is a typical value used and increases our confidence that failure to reject H_0 was not due to an under-powered test). As we can see, in this case the fraction of venues that exhibit an increase in the average daily check-ins is almost the same for both groups and for both short and long term, i.e., $P(I_{c,*}|S,E) \approx P(I_{c,*}|E)$. This suggests that the presence of a local promotion and the increase in the average check-ins are conditionally independent given the externalities E! To be more precise, during the promotion period the probability of increase in the check-ins for the treated venues appears to be larger than that of the control venues. However as one can see in Figure 4.9, where the confidence intervals for $P(I_{c,d}|E)$ are also presented, this increase is very small from a practical perspective.

More importantly though, in the previous section we emphasized on the fact that the reference group includes a large proportion of venues with effect size of 0. This clearly reduces the fraction of venues in the reference groups that have $d > 0$ leading to smaller bars for the reference group in Figure 4.3. A further examination of these cases shows that

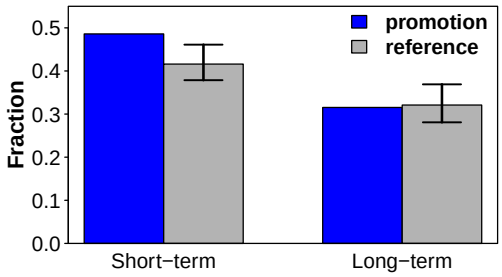
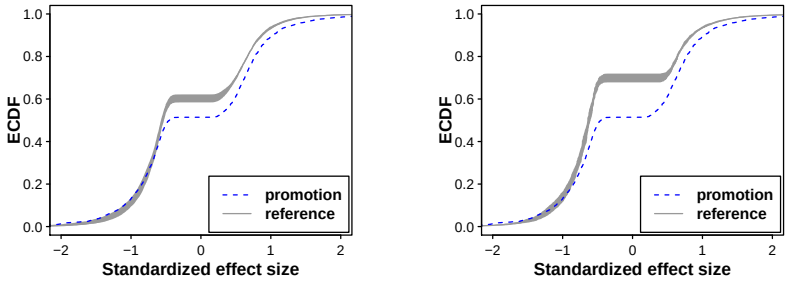


Figure 4.9: When considering venues with robust changes in their check-ins the effect of local promotions disappear.



(a) Short-term (b) Long-term

Figure 4.10: Small effect sizes do not provide robust observations based on our bootstrap tests (**daily check-ins**).

the vast majority of these venues exhibit 0 check-ins (and hence, 0 new customers as well) over the whole period. These data points do not represent real venues, but are venues that correspond to events such as extreme weather phenomena, traffic congestion, potentially spam venues etc. Hence, we can remove these venues from our reference groups. After doing so we are able to recover the results presented in Figure 4.9 further supporting the conditional independence between an increase in the mean number of check-ins per day and promotions. Note that our bootstrap tests for these venues are extremely underpowered (practically there is not any distribution since every observation is 0) and hence, are not included in the results presented in Figure 4.9. As we can further see from the plateau around $d_{c,*} = 0$ in Figure 4.10 that depicts the empirical CDF of Cohen's d for the venues used in Figure 4.9, small effect sizes do not constitute *robust* observations. Of course this can either be due to the

low power of the test to detect a small effect size, or due to the actual non-existence of any effect.

We perform exactly the same analysis for the daily new customers to venues. Even though $P(I_{p,d}|S,E) > P(I_{p,d}|E)$ this increase is not statistically significant as well, as we can see from Figure 4.11. The same is true for the long term impact of the campaign on the number of new daily customers. Finally, Figure 4.12 depicts the distribution of the corresponding effect sizes.

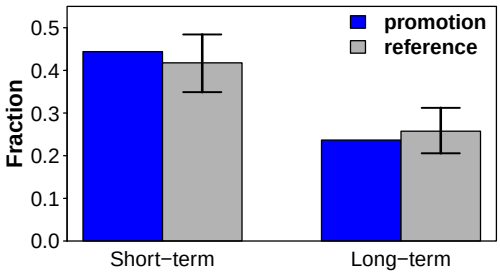
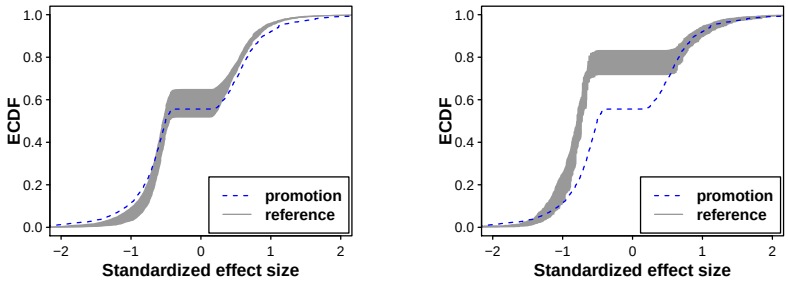


Figure 4.11: When considering venues with robust changes in their **daily unique customers** the effect of local promotions disappear.



(a) Short-term (b) Long-term

Figure 4.12: Small effect sizes do not provide robust observations based on our bootstrap tests (**daily unique customers**).

4.2.4 Anecdote Evaluation

As mentioned in the introduction there are different anecdote stories supporting the effectiveness of promotions through LBSNs. One of them is a burger joint in Philadelphia, which

we denote as v_P. At this part of our study we want to examine what our data imply for this specific venue and to verify whether our data and analysis are able to recover known ground truth. v_P publishes a special deal on the 37^{th} day of the data collection, which lasts until the end of the collection period. Therefore, we can only examine the short-term effectiveness. The standardized effect size for the daily check-ins is approximately 0.52 (0.41 for unique users), while our bootstrap test indicates that this increase is statistically significant. This is in complete agreement with reports about the specific venue [54]. Figures 4.13 and 4.14 further present the bootstrap distribution of $m^d_{c_v} - m^b_{c_v}$ and $m^d_{p_v} - m^b_{p_v}$ respectiveily under H_0 and H_1.

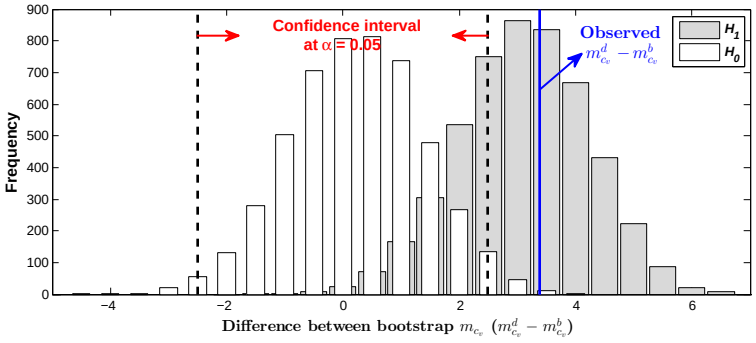


Figure 4.13: Our data support anecdote success stories for v_P.

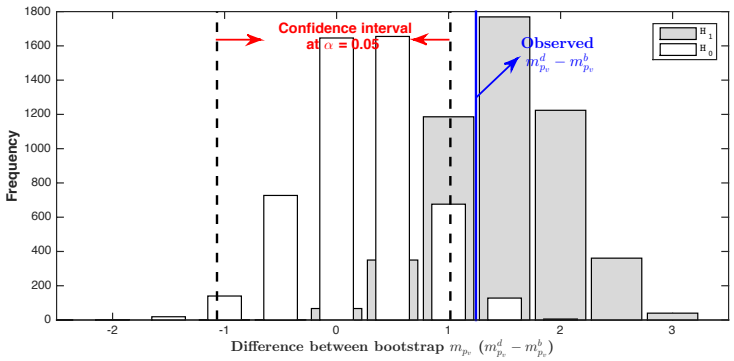


Figure 4.14: Our data support anecdote success stories for v_P (**for unique users**).

4.2.5 Difference-in-Differences

In this section, we analyze the promotion data using a method borrowed from the econometrics literature, namely difference-in-differences (DD) [7]. This is a quasi-experimental technique that aims in identifying the effect of an intervention using observational data. The reason for this analysis is to further support (or not) and hence, strengthen our conclusions from our statistical analysis using the bootstrap tests. An agreement between the two methods will also highlight another benefit of performing bootstrap tests. As we will elaborate on in the following, DD has a set of assumptions, with the strongest one being that of the parallel trend between treated and control subjects. This assumption does not always hold and hence, DD will not be applicable. Nevertheless, our approach presented in Section 4.2.3 does not rely on this assumption and hence, is applicable even in cases where DD is not.

DD requires observations obtained in different points in time, e.g., t_1 and t_2 ($t_1 < t_2$), for both the control (e.g., $y_{c,1}$ and $y_{c,2}$) and the treated (e.g., $y_{\tau,1}$ and $y_{\tau,2}$) subjects. The treated subject is exposed to the intervention only during t_2. The difference between $y_{\tau,2}$ and $y_{c,2}$ does not only includes the effect of the intervention but it also includes other "intrinsic" differences between the treatment and the control. The latter can be captured by their difference during time t_1, i.e., $y_{\tau,1} - y_{c,1}$, where the treated subject has not been exposed to the intervention. The DD estimate is then:

$$\delta_{\tau,c} = (y_{\tau,2} - y_{c,2}) - (y_{\tau,1} - y_{c,1}) \tag{4.6}$$

This removes any biases in the comparison during t_2 between the treatment and control that could be the result from (i) permanent differences between those points, as well as (ii) biases from comparisons over time in the treatment that could be the result of trends. Figure 4.15 further visualizes the process, where we see that the method takes advantage of the expected parallel trend between the treatment and the control if the intervention was not applied. In other words, this "parallel trend" assumption posits that the average change in the control group represents the counterfactual change in the treatment group if there were no treatment. This assumption is very important for the DD method to work and it

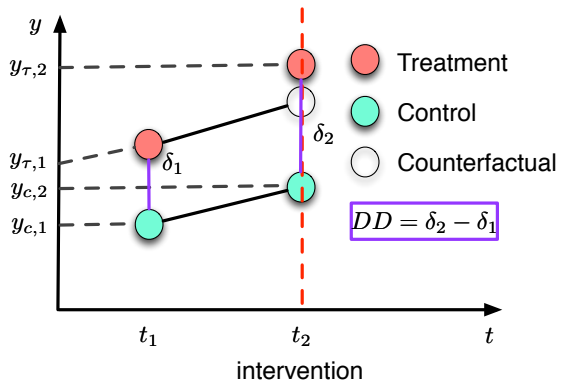
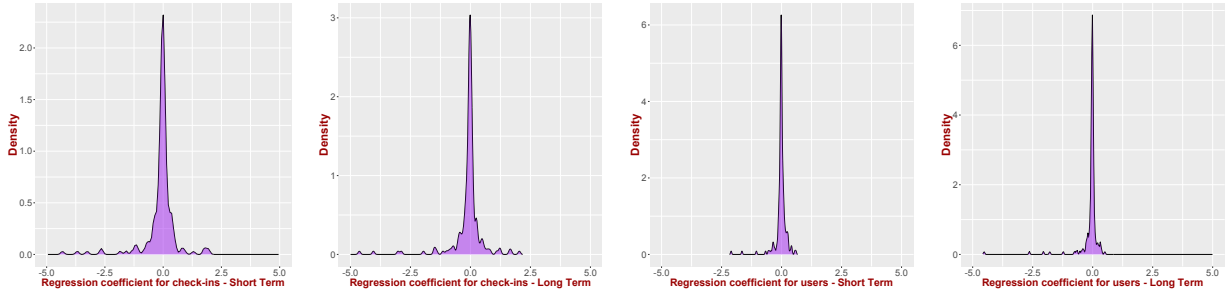


Figure 4.15: The difference-in-differences method.

is not always true and thus, not always applicable. The DD estimate can also be formally obtained through a linear regression that models the dependent variable y. More details are provided in Appendix C.

To reiterate, in our setting the intervention is the presence of a special offer. The observation/dependent variable y, is the average number of check-ins and the average number of new customers. In the case of analyzing the short-term effectiveness of a campaign, the time t_1 corresponds to the period prior to the special offer, while t_2 corresponds to the period during the promotion. Similarly, for the long-term study t_1 corresponds again to the period prior to the special offer, while t_2 corresponds to that after the promotion is seized. Every venue that has offered a promotion is considered a treatment, while the control venues for each of them are the same as in our analysis using the bootstrap tests. Then for every treated venue v_τ we compute the average difference-in-differences $\overline{\delta}_{v_\tau,v_c}$ with its matched venues v_c. With the dataset from all the treated venues, we can then compute the average difference-in-differences $\overline{\delta}_{\tau,c}$. If $\overline{\delta}_{\tau,c}$ is positive (at a predefined significance level α) then the promotions can be deemed successful. Figure 4.16 depicts the distribution of $\overline{\delta}_{\tau,c}$ for different dependent variable (i.e., check-ins and unique users) and for both the short and long term analysis. As we can see the observed values are concentrated around $\delta = 0$ for all the cases. In fact, the corresponding p-values for the statistical test $H_0 : \overline{\delta}_{\tau,c} = 0$, $H_1 : \overline{\delta}_{\tau,c} > 0$ is greater than 0.05 for all the 4 cases and hence, we cannot reject the null hypothesis, that is, the difference-in-differences is 0. This further means that on average there is not any impact

from the promotion campaigns. This result from difference-in-differences strengthens our conclusions from our bootstrap tests, that the impact of promotions through location-based social media is not as strong as anecdotal stories suggest.

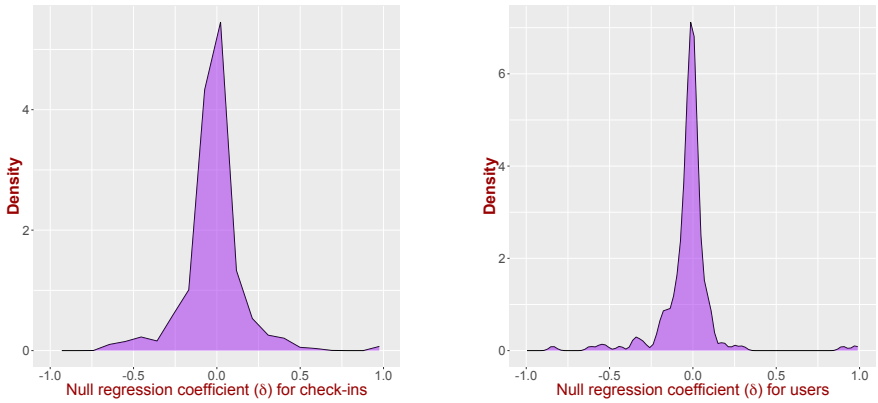


(a) Check-ins/Short-term (b) Check-ins/Long-term (c) Unique Users/Short-term (d) Unique Users/Long-term

Figure 4.16: The average difference-in-differences in all scenarios is statistically not different than 0!

However, as alluded to above, in order for the results of the difference-in-differences method to provide robust results the parallel trend assumption needs to be satisfied. In order to test whether this assumption is satisfied in our dataset, we can compute the difference-in-differences between the treated and the control venues for earlier time periods that do not include the presence of a promotion. If the computed difference-in-differences is insignificant, i.e., $\delta = 0$ for all statistical purposes, then the parallel trend assumption holds [10]. In our case we use the one-month period prior to the special promotion. We consider a "pseudo-intervention" at the middle of this period, i.e., two weeks, and we compute the difference-in-differences coefficient between the two null time-periods. The results are presented in Figure 4.17. As we can see the estimated null DD coefficient is distributed around 0. In fact, the corresponding t-tests for the daily check-ins and unique users fail to reject the null hypothesis (p-value > 0.15), i.e., that they are equal to 0. Thus, we can conclude that the parallel trend assumption holds.

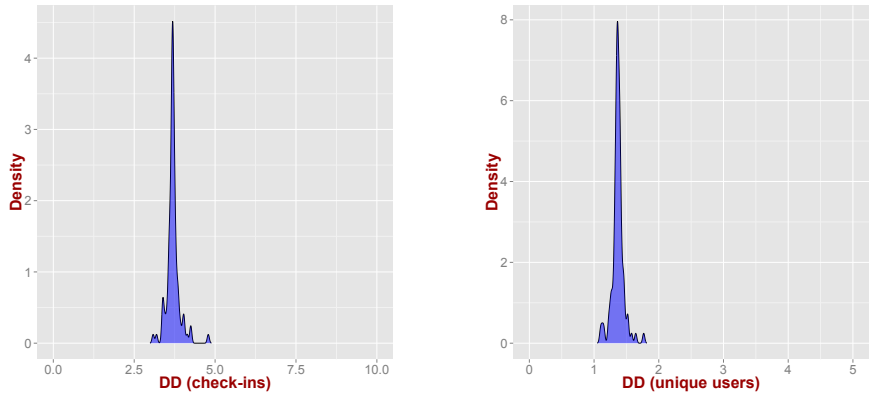
Anecdote evaluation: We further present the results for the difference-in-differences analysis for venue v_P. In particular, we compute the distribution of the difference-in-differences $f(\delta_{v_P,c})$ and the venues matched with v_P, for both the number of check-ins as



(a) Check-ins (b) Unique Users

Figure 4.17: The parallel trend assumption is satisfied in our dataset for both the daily check-ins and the daily new users.

well as the number of unique users visiting v_P. The results are presented in Figure 4.18 (recall that we can only compute the effectiveness for the short-term). As we can see both of the distributions are in the positive side of the horizontal axis and hence, further verify the anecdote stories that the specific venue has benefited from the social media campaign.



(a) Check-ins (b) Unique Users

Figure 4.18: The average difference-in-differences for venue v_P is 3.68 (p-value < 0.01) for the check-ins and 1.36 (p-value < 0.01) for the unique users.

4.2.6 Summary of Analysis

To summarize, we have analyzed the impact of promotions through LBSNs on the mean daily number of check-ins to a venue as well as on the mean daily number of its new customers. Our results indicate that the presence of a promotion does not alter the probability of observing an increase in either of these daily means. Consequently there are not any significant evidence in support of Hypotheses 4.1 and 4.2. For our analysis, we relied on quasi-experimental techniques using bootstrap hypothesis testing. We further verified our results using the difference-in-differences method increasing the confidence in our results.

Before describing our prediction models we want to explore the connection between these two metrics. In particular, we compute the conditional probability of observing a statistically significant increase in the mean daily number of check-ins given a statistically significant increase in the new customers and vice versa. Our results are presented in Table 4.3. As we can see these conditional probabilities are very high revealing the high correlation between the two metrics. For this reason, in the following section we will focus on building models based on the mean daily number of check-ins.

Table 4.3: The two metrics we used to evaluate the effect of LBSN promotions are correlated.

Probability	short-term		long-term	
	treatment	control	treatment	control
$\Pr(I_p \mid I_c)$	0.887	[0.804 0.827]	0.764	[0.684 0.722]
$\Pr(I_c \mid I_p)$	0.931	[0.893 0.933]	0.898	[0.907 0.948]

4.3 MODELS FOR LOCAL PROMOTIONS

In this section our goal is to examine whether there are specific attributes that contribute to the success of a promotion. For this we build models that can provide an educated decision

on whether a special deal will "succeed" or not considering the short and the long term cases separately. We will treat these as two separate binary classification problems (one for the short term effectiveness and one for the long term). Based on the bootstrap tests the positive class includes the offers that exhibit statistically significant increase in $m^d_{c_v}$ ($m^a_{c_v}$), while the negative class includes the special deals with a statistically significant decrease or a failure to reject the null hypothesis with a powerful test ($\pi \geq 0.8$). We begin by extracting three different types of features. Note that some of these features are specific to promotions, while others aim at capturing more general factors that can affect the popularity of a venue. For instance, the urban form of the neighborhood of a venue, that is, the composition of the environment nearby with respect to venue types can be crucial as we explain later. We then evaluate the predictive power of each individual feature using a simple unsupervised learning classifier. We further build a supervised learning classifier to predict the effect of a special deal using the extracted features.

The above classification problem provides a binary response, i.e., whether the promotion will succeed or not. However, it does not inform about the extent of this success. For this reason we also built a classic linear regression model using the same set of features, where the dependent variable is the change in the check-ins/unique users both in short and long term.

4.3.1 Feature Extraction

4.3.1.1 Venue-based features ($\mathcal{F}_v$) The set $\mathcal{F}_v$ includes features related with the properties of the venue publishing the special deal. The intuition behind extracting such features lays on the fact that the effectiveness of the special offer can be connected to the characteristics of the venue itself. For instance, a special deal might not help at all a really unpopular venue but it might be a great boost for a venue with medium levels of popularity. In particular, the features in $\mathcal{F}_v$ include:

Venue type: This is the top-level type $T(v)$ of venue v. Table 4.4 depicts the fraction of special deals offered from different types of venues that are associated with a statistically significant increase in the daily number of check-ins, i.e., the conditional probability

$P(I|T(v))$.

Table 4.4: Probability for the positive class conditioned on the type of the venue.

Category		Nightlife	Food	Shops	Arts	College	Outdoors	Travel	Residence	Professional
% Positive class	*short-term*	62.07%	57.74%	42.90%	52.87%	56.25%	58.33%	66.84%	54.54%	61.86%
	long-term	50.00%	41.51%	28.22%	43.75%	37.04%	25.00%	53.80%	14.29%	39.68%

Popularity: For the venue popularity we use two separate features; (i) the mean number of check-ins per day at the venue for the period before the special offer starts, $m^b_{c_v}$ and (ii) the cumulative number of check-ins in v just before the beginning of the special offer, $c_{av}[t_{s-1}]$.

Loyalty: We define the loyalty λ of users in venue v as:

$$\lambda_v[t_{s-1}] = \frac{c_{av}[t_{s-1}]}{p_{av}[t_{s-1}]} \tag{4.7}$$

where $p_{av}[t_{s-1}]$ is the accumulated number of unique users that have checked-in to venue v at time t_{s-1}. At a high-level λ indicates the average return (check-in) rate of users in v.

Rating score: Each venue in Foursquare has a rating score ranging from 0 to 10. Foursquare calculates this rating based on a number of signals [59] such as the number of positive/negative reviews that the venue has received from Foursquare users. We use the rating $\gamma_v[t_{s-1}]$ of venue v at time t_{s-1} as another feature.

Likes: Foursquare allows users to like or dislike a venue. We will use the accumulated number of likes $\iota_v[t_{s-1}]$ a venue has received (at time t_{s-1}) as a feature for our classifiers.

Tips: Foursquare allows users to leave short reviews for the venues. We use the total number of such reviews (tips in Foursquare's terminology) $\mathtt{N}_{\mathtt{t}v}[t_{s-1}]$ for venue v up to time t_{s-1} as a feature for our classifiers.

4.3.1.2 Promotion-based features ($\mathcal{F}_p$)

The set $\mathcal{F}_p$ includes features related to the details of the special offer(s) that exist during the promotion period. The details of the deal(s) might be important on whether the promotion will succeed or not. For instance, a

short-lived offer might have no impact because people did not have a chance to learn about it. The features in $\mathcal{F}_p$ include:

Duration: The duration `D` is the promotion period length. Intuitively, a longer duration allows users to learn and "spread the word" about the promotion, which consequently will attract more customers to check-in to the venue.

Type: There are 7 types of special deals that can be offered from a Foursquare venue during the promotion period. Each type provides different kind of benefits but has also different unlocking constrains. Table 4.5 shows the probability distribution of the positive class conditioned on the different types of special offers that are part of the promotion.

Table 4.5: Probability distribution of the positive class conditioned on the different types of special offers.

Type		Newbie	Flash	Frequency	Friends	Mayor	Loyalty	Swarm	Multi-type
% Positive class	*short-term*	62.24%	60.00%	45.56%	84.62%	67.74%	50.50%	57.14%	60.60%
	long-term	59.32%	62.50%	30.07%	43.75%	54.84%	50.00%	0.00%	44.23%

If a venue publishes two (or more) different types of deals we refer to this as "Multi-type" offer. In order to be able to easily distinguish between different combinations of offers in this "Multi-type" deals, we encode this categorical feature in a binary vector $\boldsymbol{\xi}_s \in \{0,1\}^7$, where each element represents a special type. "Multi-type" promotions will have multiple non-zero elements.

Count: Count $\mathtt{N_s}$ is the average number of special deals per day associated with a promotion period. $\mathtt{N_s}$ captures how frequently a venue published specials during a specific promotion period. Note that $\boldsymbol{\xi}_s$ is a binary vector and hence, if a venue is offering two deals of the same type this can only be captured through $\mathtt{N_s}$.

4.3.1.3 Geographical features ($\mathcal{F}_g$) The effectiveness of a promotion can be also related to the urban business environment in the proximity of the venue. The latter can be captured through the spatial distribution of venues. For example, an isolated restaurant might not benefit from a special deal promotion, simply because people do not explore the

specific area for other attractions. For our analysis, we consider the neighborhood $\mathcal{N}(v, r)$ of a venue v to be the set of venues within distance r miles from v (we use $r = 0.5$)[2]. The features in $\mathcal{F}_g$ include:

Density: We denote the number of neighboring venues around v as the density ρ_v of $\mathcal{N}(v, r)$. Hence,

$$\rho_v = |\mathcal{N}(v, r)| \tag{4.8}$$

Area popularity: The density ρ_v captures a static aspect of v's neighborhood. To capture the dynamic aspect of the overall popularity of the area, we extract the total number of check-ins observed in the neighborhood at time t_{s-1}:

$$\phi_v = \sum_{v' \in \mathcal{N}(v,r)} c_{av'}[t_{s-1}] \tag{4.9}$$

Intuitively, a more popular area could imply higher likelihood for Foursquare users and potential customers to be in the area, learn about the promotion and be influenced to visit the venue.

Competitiveness: A venue v of type $T(v)$, will compete for customers only with neighboring venues of the same type. Hence, we calculate the proportion of neighboring venues that belong to the same type $T(v)$:

$$\kappa_v = \frac{|v' \in \mathcal{N}(v, r) \wedge T(v') = T(v)|}{\rho_v} \tag{4.10}$$

Neighborhood entropy: Apart from the business density of the area around v, the diversity of the local venues might be important as well. To capture diversity we typically rely on the concept of information entropy. In our setting we calculate the entropy of the distribution of the venue types in $\mathcal{N}(v, r)$. With f_T being the fraction of venues in $\mathcal{N}(v, r)$

[2]We have also used $r = 0.3$ and $r = 0.8$ and we obtained similar results.

of type T the entropy of the neighborhood around v is:

$$\varepsilon_v = -\sum_{T \in \mathcal{T}} f_T \cdot log(f_T) \tag{4.11}$$

where, $\mathcal{T}$ is the set of all (top-level) venue types.

4.3.2 Predictive Power of Individual Features

We now examine the predictive ability of each of the numerical features described above in isolation. We will compare descriptive statistics of the distribution of each feature (in particular the median) for the two classes. We will then compute the ROC curve for each feature considering a simple, threshold-based, unsupervised classification system.

Mann-Whitney U test for each feature's median: A specific numerical feature X can be thought of as being strongly discriminative for a classification problem, if the distributions of X for the positive and negative instances are "significantly" different. To that end we examine the sample median of these distributions by performing the two-sided Mann-Whitney U test for the median values in the positive and negative classes for each of the features. The p-values of these tests are presented in Table 4.6.

ROC curves for individual features: We now compute the ROC curve for each feature based on a simple unsupervised classifier. The latter considers each feature X in isolation and sets a threshold value for X that is used to decide the class of every instance in our dataset. For each value of this threshold we obtain a true-positive and false-positive rate. Using all the true-positive, false-positive rate points we finally obtain the ROC curve for X. Our results for the short-term are presented in Figure 4.19 for both short and long-term predictions. As we can see these curves are fairly close to the line $y = x$, which corresponds to the performance of a random classifier! We further calculate the area under the ROC curve (AUC). Interestingly, there is a connection between the Mann-Whitney U test and the AUC given by [35]:

$$AUC = \frac{U}{n_p \cdot n_n} \tag{4.12}$$

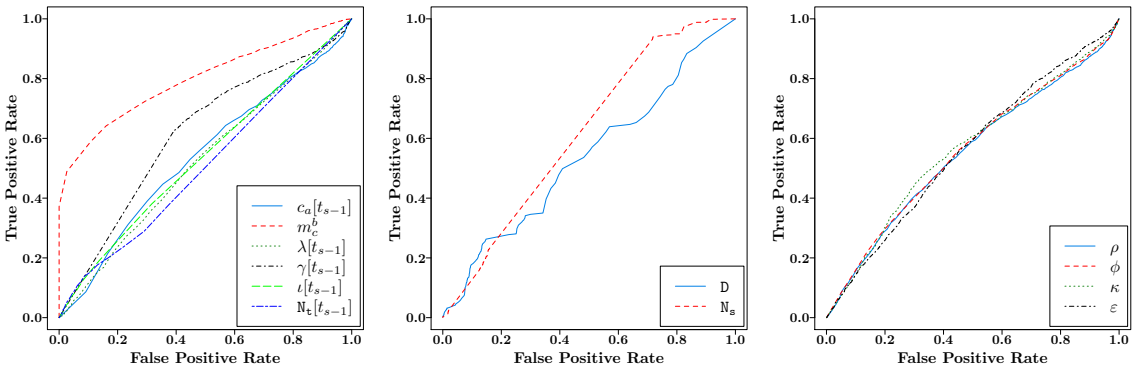
Table 4.6: While the median of the features for the two classes are significantly different, the actual distribution appear to not be discriminative (low AUC)

Features		short-term		long-term	
		AUC	p-value	AUC	p-value
$\mathcal{F}_v$	$c_{av}[t_{s-1}]$	0.537	10^{-6}	0.519	0.047
	$m^b_{c_v}$	**0.799**	0	**0.702**	0
	$\lambda_v[t_{s-1}]$	0.526	10^{-4}	0.535	10^{-4}
	$\gamma_v[t_{s-1}]$	**0.614**	0	0.580	0
	$\iota_v[t_{s-1}]$	0.537	10^{-9}	0.557	0
	$\mathtt{N_{tv}}[t_{s-1}]$	0.510	0.178	0.546	10^{-7}
$\mathcal{F}_p$	$\mathtt{D}$	0.539	10^{-7}	0.580	0
	$\mathtt{N_s}$	**0.617**	0	**0.609**	0
$\mathcal{F}_g$	ρ_v	0.551	0	0.551	10^{-8}
	ϕ_v	0.558	0	0.558	10^{-9}
	κ_v	0.565	0	0.557	10^{-9}
	ε_v	0.559	0	0.574	0

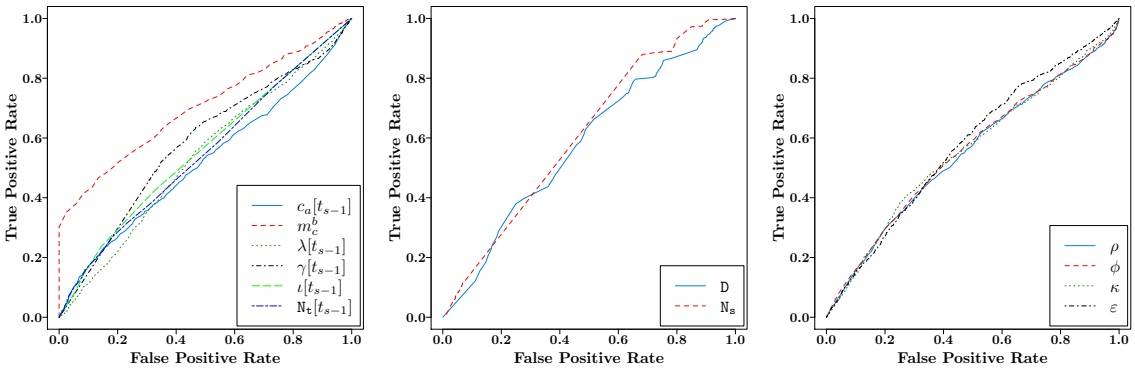
where U is the value of the Mann-Whitney U test statistic, n_p is the number of positive instances and n_n is the number of negative instances. Table 4.6 presents the values for AUC. As we observe while there are some features that deliver a good performance (e.g., $m^b_{c_v}$ and $\mathtt{N_s}$) most of the features give a performance close to the random baseline of 0.5. Hence, each feature individually does not appear to be a good predictor for the effect of special offers through LBSNs. However, in the following section we will examine a supervised learning approach utilizing combinations of the different types of features extracted.

4.3.3 Supervised Learning Classifiers

In this section we turn our attention to supervised learning models and we combine the extracted features to improve the performance achieved by each one of them individually. We evaluate various combinations of the three types of features for the binary classification problem. Our performance metrics include accuracy, precision, recall, F-measure and AUC for each classification model, and we also report the magnitude and significance of coefficients for the logistic regression model. For the classification we examine two different models, a



(a) $\mathcal{F}_v$ (b) $\mathcal{F}_p$ (c) $\mathcal{F}_g$



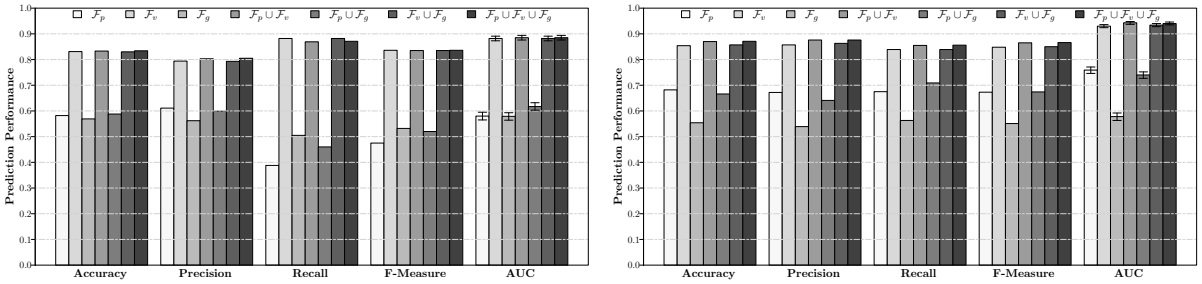
(d) $\mathcal{F}_v$ (e) $\mathcal{F}_p$ (f) $\mathcal{F}_g$

Figure 4.19: ROC curve of individual feature evaluation for the *short-term* (top row) and *long-term* (bottom row) prediction.

linear one (i.e., logistic regression) and a more complex based on ensemble learning (i.e., random forest).

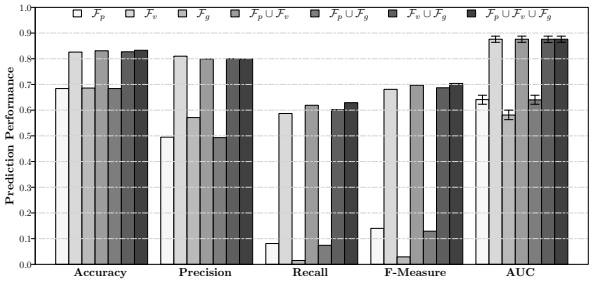
We begin by evaluating our models through 10-fold cross validation on our labeled promotion dataset. The results for the different combinations of features and for the different classifiers are shown in Figure 4.20. As the results indicate, even when we use simple linear models the performance is significantly improved compared to unsupervised models. It is also interesting to note that the most important type of features appears to be the venue-based features $\mathcal{F}_v$. The promotion-based as well as the geographic features while improving the classification performance when added, do not provide very large improvements.

The above models were built and evaluated on the data points identified through the bootstrap statistical tests in an effort to keep the false positives/negatives of the labels low.

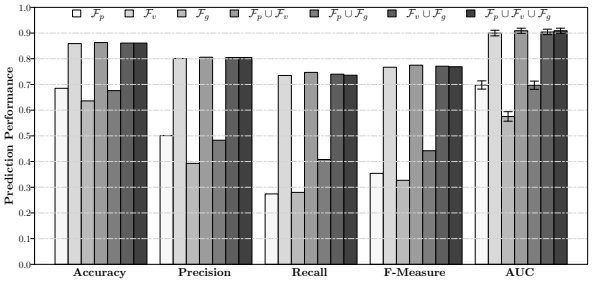


(a) Short-term Prediction with Logistic Regression

(b) Short-term Prediction with Random Forest



(c) Long-term Prediction with Logistic Regression

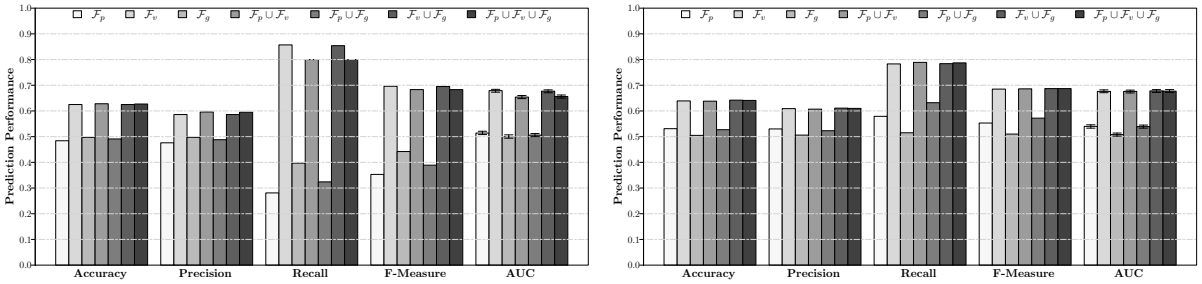


(d) Long-term Prediction with Random Forest

Figure 4.20: Using supervised learning models improves the performance over unsupervised learning methods.

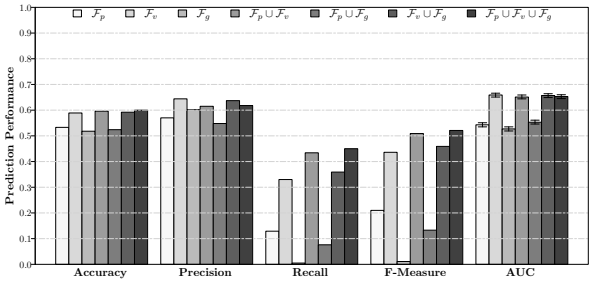
However, while this is important for building a robust model, in a real-world application the model will need to output predictions for cases that might not provide statistically significant results a posteriori. After all, a venue owner is interested in what he observes, and not whether this was a false positive/negative (i.e., an increase/decrease that happened by chance). Hence, we test the performance of our models on the data points in the promotion group for which we were not able to identify a statistically significant change ($\alpha = 0.05$) in the average number of check-ins per day. A positive observed value of d corresponds to the positive class. Note that we do not use these points for training. This resembles an out-of-sample evaluation of our models, testing their generalizability to less robust observations. Our results are presented in Figure 4.21. While as one might have expected the performance is degraded compared to the cross-validation setting, it is still good.

Finally we focus on the results from logistic regression, which has a genuine probabilistic interpretation. In particular, the accuracy performance when using the set of features $\mathcal{F}_v \cup \mathcal{F}_g$ and $\mathcal{F}_p \cup \mathcal{F}_v \cup \mathcal{F}_g$ is very similar. We compute the actual outcome of the model, i.e., before

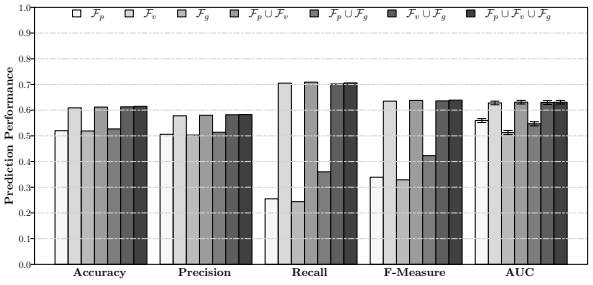


(a) Short-term Prediction with Logistic Regression

(b) Short-term Prediction with Random Forest



(c) Long-term Prediction with Logistic Regression



(d) Long-term Prediction with Random Forest

Figure 4.21: Our supervised models deliver good performance on out-of-sample evaluation on the less robust observations.

applying the classification threshold, which is the probability of observing an increase in the mean daily check-ins of the corresponding venue. Hence, the outcome of the two models provide the probabilities $P(I|\mathcal{F}_v, \mathcal{F}_g)$ and $P(I|\mathcal{F}_v, \mathcal{F}_g, \mathcal{F}_p)$ respectively. Table 4.7 presents the root mean square differences between these probabilities for the various cases examined, which is small for all the scenarios. Since features $\mathcal{F}_v$ and $\mathcal{F}_g$ capture various (environmental) externalities, while the set $\mathcal{F}_p$ captures attributes related with the promotion itself, these results further support our findings from our statistical analysis in Section 4.2. Of course these features do not capture all the externalities, and thus the actual probabilities might differ, even though the classification outcome is very accurate.

Finally, we examine the logistic regression coefficients of the various features used. The results are presented in Table 4.8. As we can see the majority of the features provide statistically significant information for the success of a promotion. However, the most important feature appears to be the popularity of a venue as captured by the number of check-ins prior to the promotion. The direction of the effect of this feature is negative, i.e., a venue with

Table 4.7: The root mean square distance of the logistic regression output for the features $\mathcal{F}_v \cup \mathcal{F}_g$ and $\mathcal{F}_p \cup \mathcal{F}_v \cup \mathcal{F}_g$ further supports our statistical analysis.

Cross-validation		Out-of-sample	
short-term	long-term	short-term	long-term
0.081	0.067	0.072	0.074

smaller popularity appears to be more probable to benefit from a promotion as compared to a more popular venue (keeping all the other independent variables constant). At a hindsight this might be expected since an already popular venue might have already saturated the nearby clientele and hence, it will be extremely hard to benefit from such promotions. On the contrary, venues with lower popularity (e.g., newer venues) might be able to attract more of the nearby customer base.

4.4 DISCUSSION AND IMPLICATIONS

Our results suggest that the benefits from local promotions through LBSNs (and to be more specific through the platform examined in our study, which is currently the largest LBSN) are much more limited than what anecdotal stories suggest. However, we acknowledge that the time-series of daily check-ins and unique users serve only as a proxy for the actual revenue generated. Nevertheless, we believe that these proxies can indirectly drive revenue, by increasing the *visibility* of a venue. In addition, customers attracted by LBSN campaigns are arguably more motivated to check-in than others. In fact, as we discussed in Section 4.1 LBSN campaigns *require* users to check-in in order to claim their badges, discounts or other types of rewards. Therefore, a revenue increase due to the influx of such customers should be reflected in these proxies.

Even though our study suggests the limited potential of such campaigns, we choose to use

Table 4.8: Coefficients for logistic regression

(a) Short-term

		Est.	Signif.
	(Intercept)	0.62	
$\mathcal{F}_v$	$c_{av}[t_{s-1}]$	0.003	***
	$m^b_{c_v}$	**-3.104**	***
	$\lambda_v[t_{s-1}]$	-0.014	.
	$\gamma_v[t_{s-1}]$	-0.139	***
	$\iota_v[t_{s-1}]$	-0.003	
	$\mathtt{N_{tv}}[t_{s-1}]$	0.018	.
$\mathcal{F}_p$	D	0.007	***
	$\mathtt{N_s}$	0.410	**
$\mathcal{F}_g$	ρ_v	-0.001	*
	ϕ_v	0.000	**
	κ_v	-0.574	
	ε_v	0.354	**
	not verified	*Reference level*	
	verified	-0.026	
Category	Arts	*Reference level*	
	College	-0.777	
	Food	0.069	
	Nightlife	0.227	
	Outdoors	0.700	
	Residence	-1.845	*
	Shops	-0.529	
	Travel	0.735	.
	Work	-0.511	
Type	Newbie	*Reference level*	
	Flash	0.726	
	Frequency	-0.501	**
	Friends	1.261	*
	Mayor	0.979	**
	Regular	0.221	
	Swarm	0.269	
	Multitype	-0.640	**

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1

(b) Long-term

		Est.	Signif.
	(Intercept)	0.771	
$\mathcal{F}_v$	$c_{av}[t_{s-1}]$	0.004	***
	$m^b_{c_v}$	**-3.782**	***
	$\lambda_v[t_{s-1}]$	-0.028	
	$\gamma_v[t_{s-1}]$	-0.074	***
	$\iota_v[t_{s-1}]$	0.163	***
	$\mathtt{N_{tv}}[t_{s-1}]$	-0.018	
$\mathcal{F}_p$	D	-0.010	***
	$\mathtt{N_s}$	0.202	
$\mathcal{F}_g$	ρ_v	-0.001	**
	ϕ_v	0.000	***
	κ_v	-0.387	
	ε_v	0.271	.
	not verified	*Reference level*	
	verified	-0.589	
Category	Arts	*Reference level*	
	College	-1.161	
	Food	0.922	*
	Nightlife	1.369	.
	Outdoors	1.122	
	Residence	-2.138	
	Shops	0.633	
	Travel	1.769	***
	Work	0.801	
Type	Newbie	*Reference level*	
	Flash	1.909	*
	Frequency	-0.976	**
	Friends	-1.117	
	Mayor	0.260	
	Regular	-0.466	
	Swarm	-13.153	
	Multitype	-0.982	*

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1

these findings as motivation for improving the design of similar advertising efforts. In this direction, the promising results of the predictive models that we introduced in Section 4.3 suggest the usefulness of such methods for the purpose of estimating the effectiveness of

alternative campaign designs and choosing the best possible option for each setting.

Further, recent relevant work has exposed reasons why people check-in to venues [101]. Furthermore, design flaws that can explain some of the shortcomings of current LBSN campaigns have also been identified. For example, Cramer et al. [36] revealed possible reasons that lead people to check-in to a location long after they arrive. It is likely that these users might have not have used the LBSN to discover nearby venues during their visit and would, thus, be oblivious to any location-based campaigns. This suggests the need for more active communication channels, such as geo-fenced push notifications. In fact, Fang *et al.* [52] showed through randomized experiments that **active notifications** for location-aware mobile promotions can generate 12 times more sales as compared to conventional notifications. Moreover, the way that a promotion is redeemed might also play a role. For example, a large fraction of the promotions on Foursquare limit their scope to users that have a particular credit card (e.g., American Express). While such constraints are typically motivated by agreements with credit card companies or with other 3rd-party vendors, further research is required to verify whether the benefits outweigh the cost of eliminating a significant part of the customer base.

4.5 RELATED WORK

In this section we discuss relevant to our work literature and differentiate our study. In brief, there is a large volume of research in the area of online and social-media advertising. Nevertheless, to the best of our knowledge, our study is the first to analyze at scale promotions through location-based social networks. These platforms bring together both the location component as well as the social media.

Effects of Promotions: Studies in the management science have examined the impact of promotions on marketing. For example, [20] found that temporary discounting substantially increases short term brand sales. However, its long term effects tend to be much weaker. This pattern was further quantified by Pauwels *et al.* [119] who found that the significant short time promotion effects on customer purchases die out in subsequent weeks or

months. Furthermore, Srinivasan *et al.* [147] quantified the price promotion impact on two targeted variables, namely, revenues and total profits, by using vector autoregressive modeling. The authors found that the price promotion has a positive impact on manufacture revenues, but for retailers it depends on multiple factors such as brand and promotion frequency. Finally, in [90] Kopalle *et al.* proposed a descriptive dynamic model which suggests that the higher-share brands tend to over-promote (i.e., offer promotions very frequently), while the lower-share brands do not promote frequently enough.

Online Deals and Advertising: Online promotions have gained a lot of attention in recent literature. Such promotions have been a popular strategy for local merchants to increase revenues and/or raise the awareness of potential customers. A detailed business model analysis on Groupon was first presented in [4], while in [43] the authors surveyed businesses that provide Groupon deals to determine their satisfaction. Edelman *et al.* [46] considered the benefits and drawbacks from a merchant's point of view on using Groupon and provided a model that captures the interplay between advertising and price discrimination effects and the potential benefits to merchants. Byers *et al.* [24] designed a predictive model for the Groupon deal size by combining features of the offer with information drawn from social media. They further examined the effect of Groupon deals on Yelp rating scores and similar to our study they identified that Groupon deals do not offer a sustainable means of advertisement; venues offering Groupon deals see a reduction in their Yelp ratings after the promotion. Finally, Adamopoulos and Todri [2] examined the long-term effect of promotions through social media platforms (in particular Twitter) and report abnormal returns for the participating brands in terms of expanding the firm's social media fan base.

Tangential to our work is also literature on web advertising and its efficiency. In this space, Fulgoni *et al.* [62] present data for the positive impact of online display advertising on search lift and sale lift, while Goldfarb *et al.* [68] further examined the effect of different properties of display advertising on its success through traditional user surveys. Papadimitriou *et al.* [118] study the impact of online display advertising on user search behavior using a controlled experiment, while CARESOME [17] was designed in order to assess the ability of social media to acquire and retain customers.

Mobile Marketing and Social Media: Mobile marketing serves as a promising

strategy for retail businesses to attract, maintain and enhance the connection with their customers. Sliwinski [143] built a prototype application that utilizes customer spatial point pattern analysis to target potential new customers, while Luhur and Widjaja [103] describe a mobile application that can facilitate location-based search for restaurants and promotions. Furthermore, Banerjee *et al.* [15] studied the effectiveness of mobile advertising. Their findings indicate that the actual location of the participant as well as the context of that location, significantly influence the potential effectiveness of these advertising strategies. Recently, there have also been efforts to quantify through models [11] the financial value of location data, which are in the center of mobile marketing operations.

In another direction, location-based social media have gained a lot of attention. Data collected from such platforms can drive novel business analysis. Qu and Zhang [127] proposed a framework that extends traditional trade area analysis and incorporates location data of mobile users. As another example, Karamshuk *et al.* [89] proposed a machine learning framework to predict the optimal placement for retail stores, where they extracted two types of features from a Foursquare check-in dataset. Furthermore, these platforms can serve as mobile "yellow pages" with business reviews that can influence customer choices. For instance, Luca [102] has identified a causal impact of Yelp ratings on restaurant demand using the regression discontinuity framework.

Overall, our study examines the performance of promotions through location-based social media. It significantly contributes to the empirical literature in the area, since it is the first study at scale on the specific problem. While the analysis and conclusions are tight to the platform where data were collected from, our work points to clear actionable directions for both the social media provider and the participating businesses as we discuss in detail in Section 4.4.

4.6 SUMMARY

In this chapter, I formally evaluates both the long-term and short-term effects of LBSN campaigns for participating businesses to attract the visits of customers. Our main result

indicates that the positive effects of special offers through the LBSN platform examined are significantly more limited than what anecdotal success stories seem to suggest. I validate our findings by adopting two alternative methods for statistical testing, which lead to the same conclusions. In addition, in order to gain a deeper understanding of our results and increase the practical value of our methodology, I design and implement a model by extracting three types of features for predicting the popularity of a venue during and after a campaign. Our findings can be used to inform strategies for improving campaign effectiveness.

In the next chapter, I further present our study on the impact of urban environmental factors, e.g., street fairs, on local economy and the underlying human movement.

5.0 IMPACT OF URBAN EVENTS ON LOCAL ECONOMY

A healthy local business sector is important for the prosperity of the surrounding community. City governments design policies and community organizations take actions that aim in boosting the growth of such businesses. This growth can have rippling positive externalities, such as, reducing local unemployment rates, keeping the local economy alive[1] and facilitating regional resilience to name just a few. These are even more important during periods of economic crises and recession, similar to the recent one in 2008 that US is just getting itself out of.

However, these efforts might not have the results expected. For example, many local governments during the "Small Business Saturday" (last Saturday of November) offer free curb parking. The rationale behind this policy is to give incentives to city dwellers (i.e., reduced trip cost to the business) to shop locally. However, the outcome is in many cases radically different. The underpricing of curb parking creates latent incentives for drivers to keep their cars parked for longer than normal periods of times. This leads to low turnover per parking spot and hence, ultimately to fewer number of customers in the local stores [138]. Therefore it is crucial to evaluate the efficiency of similar *interventions*. Knowing what boosts the local economy and what does not, can allow the involved parties to make educated decisions for their future actions and ultimately lead to *urban intelligence* through data-driven decisions and policy making. In this study we are interested in a specific question and in particular, we are studying a research hypothesis related with the **impact of street fairs on neighboring local businesses.**

The golden standard for evaluating public policies is randomized experiments. However,

[1] As per the New Economics Foundation "local purchases are twice as efficient in terms of keeping the local economy alive".

in many cases designing and running the experiment is impossible from a practical point of view. Hence, quasi-experimental techniques [137] have been developed to analyze observational data in such a way that resembles a field experiment. To complicate things more with respect to our specific research hypothesis, evaluating the economic impact of street fairs requires access to the appropriate revenue data. While a city government office can obtain access to information such as sales tax revenue, local business advocates and citizens organizations will certainly face obstacles in obtaining such kind of data. This type of information is not part of the Open Data released by local governments and are accessible (if at all) in a very limited form through pay-per-request APIs (e.g., http://zip-tax.com/pricing). This lack of transparency can be compensated to a certain extend by utilizing information from social networks and social media. While similar types of data can potentially suffer from well-documented biases (e.g., demographic biases), they form an open platform that can be easily accessed and analyzed by citizens themselves to facilitate further investigation of issues, leading to a grassroots approach to urban governance.

In our case, given that we do not have actual revenue data for the businesses in the area of Pittsburgh as aforementioned, we collect Foursquare *check-ins* from the city of Pittsburgh over a three-month period (June-August 2015) and evaluate the effect of summer street fairs on local economy. The check-in information can serve as a proxy - even though not perfect - for the revenue ρ generated [166]. We would like to emphasize here that, our study aims in evaluating the impact of street fairs on the brick-and-mortar stores that are adjacent to the event location and not that on the participating entities – which is expected to be positive in order for them to participate.

In order to analyze our data we rely on two quasi-experimental techniques. First, an increase in the check-ins for the venues near the street fair does not necessarily mean that this was due to the event. One or more control areas need to be used for comparison. However, our data are not generated through a randomized experiment but they are purely observational. For our analysis, this essentially means that we cannot assume that the area hosting a street fair event is chosen at random. Consequently, we cannot assume that the areas that do not host street fairs exhibit the same characteristics with respect to unobserved confounding features and hence, we cannot compare the revenue in the treated area with any untreated

area. For overcoming this problem, we rely on quasi-experimental design techniques that identify appropriate control areas. In particular, we rely on propensity score matching [131], adopted in our setting by utilizing expert domain knowledge, in order to pick a set of *matched* areas $\mathcal{A}_m$ with the treated area α that will serve as our control subjects. Second, once the matched areas for comparison are chosen, we adopt the difference-in-differences method [7] in our setting in order to quantify the impact of the street fairs on local businesses. In a nutshell, the difference-in-differences is a regression model that examines the average change of the treatment group once the treatment has been applied and compares it with the control group. The implicit assumption is that this difference would be zero if the treatment had not been applied. We elaborate further on these two methods in the following section.

The main contributions of this chapter can be summarized as follows:

- We provide quantifiable evidence that support the positive impact of street fairs on local businesses.
- We show how social media data - despite their potential biases - can be useful to public policy makers and local governments since they are transparent, accessible and are able to provide good evidence when analyzed properly.

Scope of this chapter: While in the current study we are focusing on the effect of street fairs on local businesses the method can be applied in a variety of scenarios that include an external event/stimulant. For example, one can use our framework to quantify the effect of short-term road closures and/or constructions on the local economy. This is especially important during the bidding phase of a construction project since these effects should be included in the calculation of liquidated damages [67]. However, they are not currently included since there is not a framework to estimate this effect.

5.1 QUASI-EXPERIMENTAL ANALYSIS METHODS

Let us denote the total volume of revenue within area α at day t with $\rho_{t,\alpha}$. Furthermore, $\mathcal{T}_\alpha$ is the set of days that a street fair took place within area α. The trending of $\rho_{t,\alpha}$ by

itself cannot reveal anything with respect to the contribution of the street fair at the revenue generated in area α. Hence, in order to account for various confounding factors and other externalities we will need to get a "baseline" for comparison. When experimental design and implementation is possible this happens with random assignment of the treatment (in our case the street fair) to the experimental subjects. However, in our case this is not possible and hence, we rely on matching techniques and more specifically we use propensity score matching. Matching techniques provide us with the ability to analyze observational data in a way that mimics some of the particular characteristics of a randomized trial. In particular, we choose a matched, with area α, neighborhood, say, α_m, to analyze and compare the corresponding revenues generated.

Our analysis is inspired by the difference-in-differences method [7]. In brief, we compare the daily revenue differences between the area with the street fair and the corresponding matched area(s) both during the period of the street fairs as well as during the period without any street fair. The comparison with the matched area(s) - that are exposed to the same externalities - accounts for various confounding factors that can affect revenues, and hence, any observed difference can be attributed to the treatment, i.e., the street fairs in our case. In what follows, I describe in detail the building blocks of our analysis, i.e., propensity score matching and difference-in-differences.

5.1.1 Propensity Score Matching

Propensity score matching can be used to reduce (or even eliminate) the effect of confounding variables on the analysis of observational data. To reiterate propensity score matching allows an analysis in a way that mimics a randomized trial. In our own context, the ***treatment*** of interest is whether or not there is a street fair in neighborhood i. The propensity score of each (untreated) instance (i.e., every untreated neighborhood) represents the probability of this instance to be treated, conditional on a set of confounding variables. In a real randomized experiment, the instances are randomly assigned to the treatment and control groups. This ensures (given sufficiently large number of instances) that on average the two groups will only differ with respect to the reception of the treatment. In the case of observational data,

the treatment is not randomly assigned but usually the "treated" instances are chosen due to some specific characteristics (i.e., the confounding factors). Therefore, in order to identify an appropriate control group we need to calculate the probability of the untreated instances obtaining the treatment.

In order to calculate the propensity scores, i.e., the conditional probabilities of the instances receiving the treatment, we employ a logistic regression model similar to [5]. In particular, given a feature vector $\boldsymbol{Z}$ that is formed by a set of neighborhood characteristics (i.e., the confounding factors) we estimate the following conditional probability:

$$\Pr(b_i = 1|\boldsymbol{Z}_i) = \frac{\exp(w_i^T \cdot \boldsymbol{Z}_i)}{1 + \exp(w_i^T \cdot \boldsymbol{Z}_i)} \tag{5.1}$$

where b_i is a binary indicator variable, which takes the value 1 if area i is treated and 0 otherwise. In our case, $\boldsymbol{Z}_i$ includes **three types of features** for every type of establishment T that exists in neighborhood i that captures **(a)** the fraction of type T venues in i, as well as, **(b)** the fraction of the revenue (check-ins in our case) within α that was generated by venues of type T. Finally, for every business venue type, we use **(c)** the "stickiness" of the users in this type as an additional feature. The "stickiness" is defined as the ratio between the total number of check-ins in the corresponding category over the number of unique users that generated these check-ins.

After training the aforementioned logistic regression model, we estimate the probability from Equation (5.1) for all neighborhood instances $i \in \mathcal{N}$ (both treated and untreated), where $\mathcal{N}$ is the set of areas/neighborhoods. Then we match the treated neighborhood α, with:

$$\alpha_m = \min_{i \in \mathcal{N} \setminus \{\alpha\}} |\Pr(b_i = 1|\boldsymbol{Z}_i) - \Pr(b_\alpha = 1|\boldsymbol{Z}_\alpha)| \tag{5.2}$$

Essentially, this means that area α_m is the one that has the closest probability of hosting a street fair to that of area α, under the assumption that the only features that affect the decision are the ones captured by the observable confounding variable vector $\boldsymbol{Z}$.

In many scenarios (such as in our case study) we might only have one treated area α,

i.e., only one area has hosted a street fair. In this case, evaluating Equation (5.2) is *trivial*, since, the minimum is observed for the area i for which the vector distance $d(\boldsymbol{Z}_i, \boldsymbol{Z}_\alpha)$ is minimized. Simply put, the matched area α_m is the one whose feature vector $\boldsymbol{Z}_{\alpha_m}$ is closer to that of the treated area $\boldsymbol{Z}_\alpha$. We would like to emphasize here that, there might be other, unobserved, factors that lead to the choice of an area for a street fair. This is a limitation of the quasi-experimental techniques in general and propensity score matching can only account for observable confounders $\boldsymbol{Z}$.

One way we propose to use in order to alleviate some of the potential problems associated with the aforementioned limitation is to initialize the matching process with expert knowledge. In particular, the matched area α_m can be chosen using expert knowledge (e.g., urban planners in our case). The benefit of this approach is that the domain expert is - implicitly or explicitly - considering various (potentially unobserved) confounders simultaneously. We can then use the expert matching as a "seed" for matching more than one neighborhoods to α using the propensity scores.

In particular, with $\pi_{m,e}$ being the propensity score of the (domain expert) matched area $\alpha_{m,e}$, we can pick the following set of matched areas:

$$\mathcal{A}_m = \{\alpha_{m_j} : \ |\pi_{m_j} - \pi_\alpha| < |\pi_{m,e} - \pi_\alpha| + \epsilon\} \tag{5.3}$$

Essentially, as per Equation (5.3), the set $\mathcal{A}_m$ includes neighborhoods that have propensity scores that are closer to the score of the treated area (within a tolerance factor ϵ) as compared to the expert matched area. Once set $\mathcal{A}_m$ is obtained we can analyze the corresponding revenues generated using the difference-in-differences method described in what follows.

5.1.2 Difference-in-Differences

We apply again the difference-in-differences (introduced in Section 4.2.5) to quantify the impact of the street fairs on surrounding businesses. The control and treatment subjects in our setting are urban neighborhoods, and the treated subjects includes neighborhoods that host street fairs.

Figure 5.1: The treated neighborhood with street fairs and a matched area selected with domain knowledge.

5.2 DATASET AND ANALYSIS SETUP

In this section we will present the dataset we collected, as well as, the hypothesis and the experiment setup for our analysis.

5.2.1 Dataset

For the purposes of our study we collected time-series data using Foursquare's venue public API. We queried daily all Foursquare venues in Pittsburgh for the three-month period between **06/01/2015 - 08/30/2015**. This period includes six street fairs/events[2] that took place at a specific neighborhood in the city of Pittsburgh (see the street marked with red in Figure 5.1).

Our time-series data include information with respect to the number of check-ins $c_v[t]$ that have been generated in venue v during day t. To reiterate, given the fact that we do not have actual revenue data for the businesses in Pittsburgh we rely on the check-in information

[2] http://thinkshadyside.com/events/

as a proxy for the corresponding revenue of venue v, $\rho_v[t]$. This information will allow us to build the aggregate volume daily check-ins c_α within area α, i.e., $c_\alpha[t] = \sum_{v \in \alpha} c_v[t]$. Every area is defined as a circle of radius r centered at the centroid of the neighborhood under consideration. In our experiments, we examine various values for r in order to explore the spatial distribution of the impact.

We have also collected meta-data information. In particular, Foursquare associates each venue v with a type/category T (e.g., restaurant, school etc.). This classification is hierarchical and at the top level of the hierarchy there were 9 categories at the time of data collection. In order to obtain the feature vector $\boldsymbol{Z}$, we use the top-level categories and hence $\boldsymbol{Z}$ includes 21 features (2 for each category and 3 for the stickiness of each type of business venue). Our final dataset includes 27,263 venues in the city of Pittsburgh, where 21.53% (5,869) are business venues (i.e., *Nightlife Spots*, *Food* and *Shops & Services*). There are in total 32,501 check-ins in our dataset, among which 44.46% were generated in business venues.

5.2.2 Hypothesis Development

In this chapter, we will examine the following two hypotheses.

Hypothesis 5.1 (Street fairs impact on people's movement to local businesses)**.** *Street fair events lead to an increase in customer visitations for nearby business venues.*

Hypothesis 5.2 (Spatial impact of street fairs)**.** *The impact of street fairs on the customer visitations is geographically contained in a very small area.*

In order to support or reject Hypotheses 5.1 and 5.2 we will rely on data we collected from Foursquare described in the next section, utilizing the difference-in-differences method described in Section 4.2.5. We will further examine contextual dependencies, i.e., whether specific types of business venues benefit more than others.

5.2.3 Experimental Setup

In our study we consider a single area α that has hosted street fairs during our data collection period. This area is a small business center, with a number of restaurants, cafes, retail stores

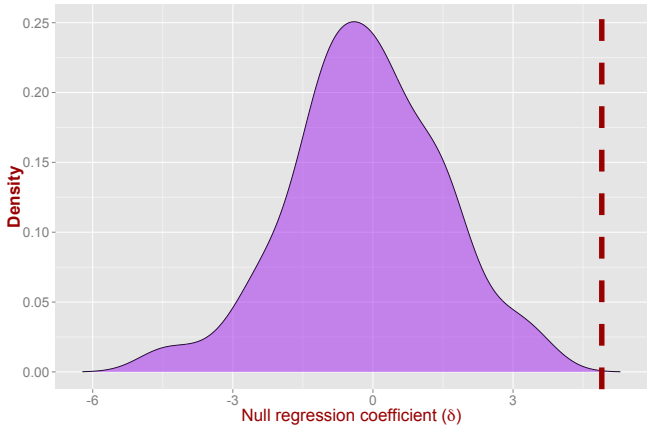


Figure 5.2: The null difference-in-differences coefficient is practically equal to 0, hence, allowing us to apply the model with high confidence.

(e.g., clothing stores, galleries etc.) and services (e.g., bank branches). The *treated* area is also accessible through public transportation, Pittsburgh's shared bike system as well as through private vehicle with parking facilities nearby. We (initially) perform the matching process based on the *expertise*[3] of local urban planners. Based on their recommendations we choose another small business area, with a similar urban form and accessibility patterns not very far from the treated area (approximately 2 miles away - green area in Figure 5.1). We have further used Equation (5.3) to build a set of matched areas. More specifically, we first pick 2,000 random points in the city of Pittsburgh and create a neighborhood of radius 0.3 miles around this point. We further eliminate areas with less than 60 venues. We consequently obtain the matched area set $\mathcal{A}_m$ using Equation (5.3) with $\epsilon = 0$ and we filter out overlapping matched neighborhoods, in order to remove possible dependencies in our datasets originating from the overlapping regions. In particular, when k matched areas overlap we only keep the final matched set the area with a propensity score matching closest to the treated area. We would like to emphasize here that we have examined different values for the radius of the control neighborhood area selection and the tolerance factor ϵ and the results obtained were very similar.

[3]We have consulted with urban planners familiar with the city of Pittsburgh.

5.3 ECONOMIC IMPACT OF STREET FAIRS

The metric of interest for our analysis is the mean number of daily check-ins in area α, denoted with y_α. For every area α we compute the average number of daily check-ins during the treatment period, $y_{\alpha,\mathcal{T}_\alpha}$, as well as, during the days with no street fair, $y_{\alpha,\mathcal{T}_\alpha^c}$, where $\mathcal{T}_\alpha^c$, represents the complement of $\mathcal{T}_\alpha$, i.e., the set of days in our dataset where no street fair took place in α. With this setting the difference-in-differences coefficient is equal to **4.95 (p-value $<$ 0.001)**. Simply put, there are 5 more check-ins every day with a fair in area α on average. This corresponds to an almost 100% increase in the check-ins in the area, since the average daily check-ins for the days with no event is 5.3.

As mentioned in Section 4.2.5 one of the crucial assumptions for the difference-in-differences to provide robust results is the parallel trend assumption. Typically the way that has been followed in the literature for verifying this assumption is to calculate the difference-in-differences coefficient for periods that the treatment has not been applied [108, 121]. Hence, for the days that in reality no street fair occurred we randomly assign pseudo-treatments in order to calculate a null coefficient δ. Figure 5.2 depicts the distribution of the corresponding coefficients obtained from 100 randomizations. As we can see the mass of the distribution is concentrated around $\delta = 0$, while the 95% confidence interval is $[-0.42, 0.37]$. Hence, we cannot reject the hypothesis that the null coefficient δ is actually 0, hence, verifying the parallel trend assumption needed for the difference-in-differences method.

We also want to examine the spatial extent of this impact, i.e., how the impact decays with space. For this, we compute the difference-in-differences coefficient for zones of different radius around the treated area making sure that there is not any overlap with control areas. In particular, we examine zones of [0, 0.1], [0.1, 0.3], [0.3, 0.6] miles. Our results are depicted in Figure 5.3 where as we can see there is a clear decreasing trend of the impact. In fact, the coefficient for the range [0.1, 0.3] miles is much smaller, and equal to 0.89 (p-value ¡ 0.1), while going further away from the area of the event (i.e., [0.3, 0.6] miles) the effect is practically eliminated ($\delta_{[0.3,0.6]} = 0.33$, p-value $= 0.61$). These results indicate - as one might have expected - that the impact of a street fair event is highly localized within a very small

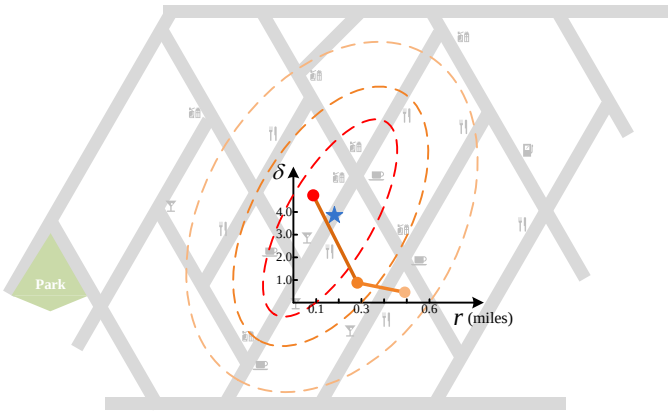


Figure 5.3: The impact of street fairs on local businesses rapidly decays with the spatial distance from the event.

area around the epicenter of the event.

We further examine the impact of each event individually, i.e., we consider a single day treatment. Table 5.1 presents our results. As we can see every event contributes to the overall local business sector a positive increase to the check-ins, which can further be translated to increase foot traffic and revenue. The only exception is the Vintage GP Car show. Compared to the other events, this attracts a very specific part of the population - i.e., car-lovers - and this might have affected its overall impact.

Table 5.1: All events - except the Vintage GP Car Show - exhibit a statistically significant and positive coefficient δ.

Event	Difference-in-differences coefficient δ
Jam On Walnut 1	9.7^{***}
Vintage GP Car Show	-2.01^{***}
Jam on Walnut 2	5.45^{***}
Jam on Walnut 3	6.64^{***}
Arts Festival on Walnut 1	4.45^{***}
Arts Festival on Walnut 2	5.53^{***}

Significance codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1

Our analysis until now has considered all of the business venues together regardless of their type. This essentially captures the aggregate impact of the street fair in the neighbor-

hood. However, we would like to decompose this effect in order to understand better what type of establishments benefit from the fairs. In particular, we compute the difference-in-differences regression coefficient for the three different types of business venues our dataset contains. Figure 5.4 depicts our results, where the 95% confidence interval of the estimated coefficients is also presented. As we can see shopping venues are the ones that benefit the most from the street fairs, while nightlife and food establishment exhibit a much (but significant and positive) lower coefficient δ. However, one crucial point here is that the coefficient provides the cumulative - additional to the counterfactual - check-ins recorded in all venues of the specific type. Hence, if a specific venue type is overrepresented in the area the estimated DD coefficient might be *inflated* [4]. In order to avoid similar issues, we can normalize the obtained coefficients from the regression model by the number of venues for every establishment type. In particular, the number of shop, nightlife and food venues in the treated area are 60, 13 and 25 respectively. Therefore, the normalized coefficients for the shop and nightlife are practically equal (0.066 and 0.061 respectively). However, the food venues still have a much smaller normalized coefficient, that is, 0.014.

Overall, we can say that our results support the two research hypotheses put forth in Section 5.2.2. In particular, street fairs have a positive impact on nearby businesses as captured by the check-ins on Foursquare and the difference-in-differences method. Furthermore, this impact is highly concentrated in the areas around the street fair (i.e., 0.1, 0.2 miles) and drops extremely fast as we move further away.

5.4 DISCUSSION AND IMPLICATIONS

As discussed in Section 1.2.3, one of the main critics that studies relying on social media get is that of the potential demographic biases that the data include. This is certainly true and is one of our study's limitation as well. Nevertheless, location-based social media is a very

[4]Note here that, this is not an issue when we applied the difference-in-differences at the level of a neighborhood. In that case, we were interested in the total additional check-ins in the neighborhood as compared to the counterfactual. Hence, if a control area had a different number of venues this would not impact the results.

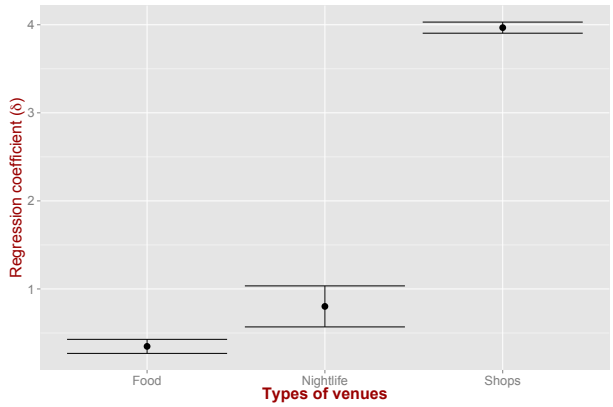


Figure 5.4: The shopping businesses appear to have the largest benefit from the street fairs among the local establishments around the area.

good, and accessible, proxy for the economic activities in urban areas. Certainly there will be noise in the obtained signal, but this information is valuable for providing supporting (or not) evidence in a variety of research hypotheses similar to ours. For example, similar datasets have been used to study urban gentrification, deprivation, emotions in a city [79, 154, 63] etc.

In our difference-in-differences regression model we included fixed time and location effects. One might argue that we should also control for the day of the week. However, this is not necessary since the null regression model essentially shows us that the different days of the week will exhibit the same "trending" on average (of course the absolute values of the check-ins will be different). To verify this we run the regression model by adding an independent variable that captures the day of the week. Our results for the various zones around the treated neighborhood are presented in Table 5.2.

As we can see even when controlling for the day of the week the impact is strong and significant. In fact, when controlling for the day of the week the impact appears to be significant even for distances beyond the 0.1 miles. Nevertheless, the impact itself is weak (i.e., the coefficient is small). Furthermore, even though it appears that the further zone has a stronger effect, the 95% confidence intervals for the two coefficients (i.e., for the ranges [0.1, 0.3] and [0.3, 0.6]) overlap, and hence, we cannot confidently support the presence of a trend.

Table 5.2: Even when controlling for the day of the week, the impact of the street fair remains.

Radius r	Difference-in-differences coefficient δ
[0, 0.1]	[3.71, 4.1]
[0.1, 0.3]	[0.17, 0.85]
[0.3, 0.6]	[0.51, 1.61]

5.5 RELATED WORK

In this section we briefly discuss related methodological literature as well as literature relevant to the specific application domain.

Quasi-experimental methodologies: The gold standard for evaluating the impact of a policy is a field experiment. However, when it comes to public policy many times this is not possible for a variety of reasons. In this case we need to rely on quasi-experimental techniques [137] in order to quantify the potential impact. Quasi-experimental designs allows to control the assignment to the treatment condition, but using some criterion different than random assignment as in field experiments.

There are various techniques that can be used depending on the type of observational data one has. For example, the difference-in-differences method [7] compares the average change over time in the outcome variable for the treatment group to the average change over time for the control group. One of the major problems when applying this method is the parallel trend assumption, that is, that the two groups exhibit the same temporal trend on their averages without the treatment. Regression discontinuity [81] is another technique that can be used to quantify the effects of treatments that are assigned by a threshold. The key idea is that observations lying very closely on either side of the threshold while differing in the reception of the treatment, they are *equal* for all practical purposes. Hence, their

treatment assignment mimics that of a randomized control trial. It should be clear that not all quasi-experimental designs are applicable in all scenarios (for example regression discontinuity cannot be applied in our setting), while there can be settings were no method is applicable. A nice survey of various quasi-experimental techniques can be found in [74].

Local businesses and urban economy: Small shops and businesses are the backbone of local economy and quantifying the effect of external events and policies on their prosperity is of utmost importance. Given the absence of large scale data, most of the existing studies have been based on survey data. For instance, a survey research conducted by Lee *et al.* [96] during the 2002 World Cup identified that the event-related tourists yielded much higher expenditure as compared to *regular* tourists, indicating that such mega-events could have a positive economic impact for local businesses. As another example, a report from a Toronto-based think tank has identified the positive impact that bike lanes have on the revenue of local businesses despite the fact that business owners systematically underestimate it [6]. In a similar direction, based on merchant and pedestrian surveys in Toronto's Annex Neighborhood, the "Clean Air Partnership" [33] recommended reallocating a curb parking lane to bike lanes, since this is likely to increase commercial activity. A recent study further showed that the installation of shared bike system can lead to an increase of the housing property values [121]. Moreover, in a briefing paper DeShazo *et al.* [42] using a survey conducted over a small sample of businesses quantified the effect of CicLAvia on local businesses. CicLAvia[5] is a car-free event that happens once every year in various areas in Los Angeles. Furthermore, anecdotal hard evidence from Seattle [41] show that increasing the price of curb parking can be beneficial to restaurants and local businesses mainly due to the increased turnover of each parking spot [138].

During the last years, and driven by the proliferation and availability of geo-tagged social media data, there has been a surge of studies on business analytics. For instance, Qu and Zhang [127] proposed a framework that extends traditional trade area analysis and incorporates location data of mobile users. Their framework can answer crucial questions in retail management such as "where are the customers of a business coming from?". As another example, Karamshuk *et al.* [89] proposed a machine learning framework to predict

[5] http://www.ciclavia.org

the optimal placement for retail stores, where they extracted two types of features from a Foursquare check-in dataset. Furthermore, these platforms can serve as mobile "yellow pages" with business reviews that can influence customer choices and business revenue. For example, Luca [102] has identified a causal impact of Yelp ratings on restaurant demand using the regression discontinuity framework. Closer to our study, Georgiev *et al.* [64] using data collected from Foursquare study the impact of the 2012 Olympic Games on the businesses in London, while Zhang *et al.* [166] quantify the effectiveness of special deals offered through location-based services as an affordable advertisement for local businesses.

To the best of our knowledge no one has examined the impact of street fairs on the adjacent businesses, even though local authorities expect this policy to have a positive outcome for businesses[6]. Studies that examine the economic effects of special events/festivals exist (e.g., [29]) but their focus is slightly different, focusing on the participating entities/kiosks themselves. On the contrary, our study is focused on the "network" effects a street fair can have for the nearby businesses.

5.6 SUMMARY

In this chapter, I apply two quasi-experimental techniques, i.e., propensity score matching and difference-in-differences, to quantify the impact of street fairs on nearby businesses. I take the number of check-ins at business venues as a proxy of revenues in local businesses. Our findings indicate that such urban event as street fairs can significantly increase the frequency of human movement to surrounding business venues, but the effect decays fast with the spatial distance from the event center. Also the impact is contextually dependent on the type of businesses. Our analysis framework is general and can be applied to evaluate the effect of many other policy making and urban external events.

[6] E.g., http://tinyurl.com/zdved39

6.0 CONCLUSION AND FUTURE DIRECTIONS

6.1 CONCLUSION

In this dissertation, human urban movement is studied within the social, economic and urban contexts they emerge in as captured through location-based social networks. I design statistical analysis and modeling frameworks using randomization and quasi-experimental techniques to investigate and quantify the effects of various contextual factors on human movement in urban space. The scenarios I examine include social interactions, local business advertising strategy and urban events initiated by local government and neighborhood communities. Based on statistical analysis on the three scenarios, we can see the data generated in location-based social networks, though have potential biases as discussed in Section 1.2.3, can capture a much richer contexts where urban movement emerge and thus provide an unprecedented opportunity for a better and deeper understanding on how people move and act in urban space.

Our randomization experiments in Chapter 3 indicate that human movement to local places exhibit significant levels of homophily via social ties. While the similarity of people's geo-trails at the geographically global scale cannot be attributed to peer influence, the latter can explain a significant proportion of localized similarity between friends. The level of influence contextually depends on the type of places. Due to the "network value" of peers [44], understanding peer influence with regard to their movements in real-world places can potentially help improve targeting customers in local marketing. Also social connections tend to be stimulated by non-trivial similarity captured by places with special network characteristics, thus mining features from individuals' mobility data can facilitate social recommendation [135], which can further help build a virtuous circle of the local ecosystem.

In Chapter 4, I investigate the effectiveness of a direct local advertising mechanism, i.e., "special offer", for attracting the visits of customers to local businesses in both short-term and long-term period. I utilize quasi-experimental techniques with various confounding factors considered and design two statistical hypothesis testing frameworks. Both the two frameworks reveal that online promotions in LBSNs are not as effective as anecdotal stories might suggest in attracting customers. I build a supervised learning model and evaluate the effects of three types of features, venue-based, promotion-based and geographical, on the popularity of the promoted businesses during and after the promotion. The promotion-based features actually help only little with the prediction task, which further confirm our conclusion in previous statistical analysis. Our studies are envisioned to provided educated support for the design and cost-benefit analysis of campaigns in LBSNs. This findings of this work have been featured in multiple media press, e.g., Pittsburgh Post-Gazette [1].

As present in Chapter 5, check-ins in LBSNs can serve a good, though not perfect, proxy for human economic activities in urban neighborhood. With such observational data available and the applied quasi-experimental techniques, we provide an educated guidance on how urban events and local government decisions impact human visits to nearby businesses, thus overall the local economy. In this dissertation, I take street fairs as a study case, but the analysis framework is rather general and can be applied to evaluate various policy making.

6.2 FUTURE DIRECTIONS

Overall, in this dissertation I design statistical analysis frameworks and provide a general viewpoints on how the social, economic and urban environmental factors interplay with human movement across urban space. The outcome of this dissertation would provide guidance for a better understanding of human urban movement and further foster applications in sociology, local economy and urban planning. To be further noted, human movement behaviors are rather complex and can be attributed to many other factors, which might not be fully captured by our current statistical analysis methods. Below I provide a brief list showing some possible future directions of our work.

- In Section 3.3.2, we design a popularity-based reference model by assuming people are more likely to go to globally popular places. This assumption might not hold for every user since people usually have different location preferences inherently. The next step would be to further control such factors for a more appropriate randomized model. This can help reduce the bias of estimation of peer influence.
- Furthermore, in Section 4.2.2 and 5.1.1, given only the observational data, I apply matching techniques to select reference venues or neighborhoods to get a baseline for estimating treatment effects. An accurate matching in geographical space is rather difficult since every place or neighborhood could be quite different even with a limited number of confounding factors considered. Novel techniques, such as causality analysis of confounding factors, for location matching would be helpful to give a more robust estimation of intervention.
- Overall in this dissertation, we design and conduct statistical experiments separately to examine the effects of different factors on urban movement. One potential direction is the design of a unified model that can quantify the level of influence of different factors on human mobility explained. There are previous similar ideas [151, 150] using graphical model to quantify the level of social influence in time-varying geo-social networks, but currently no external factors were combined into modeling.

6.3 OUTLOOK

With the rapid urbanization, cities are becoming a more and more complex system. Human mobility in urban space has been important part that indicate how people interact with the urban environment they live, such as emergence events, transportation infrastructure and local government policies. The advent of online and mobile social webs and applications enable recording the digital footprints of human at an unprecedented large geographical scale, spatial and temporal granularities, population size, and more importantly, in a much richer contents. Every piece of information regarding users' online behaviors can now be geo-tagged. The big data of human footprints in urban space allow researchers and scientists to

answer important research questions, and enable urban planners and local governments to understand the way people act and behave through their movements in our cities, in order to design cities that can deliver a livable, resilient and sustainable urban environment that is relevant to the city dwellers needs and finally toward the goal of smart cities.

In this dissertation, I have attempted to take a step forward a better understand human urban mobility in the contexts of social interaction, economic incentives and urban events, by developing systematic statistical analysis frameworks to explain the underlying processes. I envision our methodologies to be extended to other scenarios in terms of data sources and applications. Also I hope our findings can inspire researchers in various disciplines from academia, industry and local government to participate in this area, design novel models and build new applications.

APPENDIX A

ERDŐS-RÈNYI RANDOM GRAPHS

In a random graph model some properties of the network are fixed, while others are generated randomly. In the Erdős-Rènyi random graph model, denoted as $G(n, m)$, we fix the number of nodes to n and the number of edges to m. $G(n, m)$ is then a probability distribution $P(G)$ over all the possible networks G such that if G has n vertices and m edges and Ψ is the number of such (simple) graphs, then $P(G) = 1/\Psi$; otherwise $P(G) = 0$. A slightly different and more tractable model, is the $G(n, p)$. In this case the number of nodes is still fixed (n) but now instead of fixing the actual number of edges in the network, we fix the probability of an edge between any two nodes to be equal to p. More details on random graph models can be found in [48].

APPENDIX B

STATISTICAL SIGNIFICANCE RESULTS

While we observe in Figures (3.8)-(3.10) that the average degree, clustering coefficient and entropy of the common venues for the friend's dataset are different from that of the reference group of user pairs, we further delve into the statistical significance of these results. In particular, every point in Figure 3.8 corresponds to the mean value for the degree of the common venues of friends residing in distance d, $\mu^f_{deg}(d)$ (bottom line) or of the reference pairs of users $\mu^r_{deg}(d)$ (top line). In order to examine whether the difference observed in the Figure is statistically significant we perform a one-tailed t-test on the mean values for each distance-bin d. In particular, the hypothesis test is:

$$H_0: \quad \mu^f_{deg}(d) = \mu^r_{deg}(d) \tag{B.1}$$

$$H_1: \quad \mu^f_{deg}(d) < \mu^r_{deg}(d) \tag{B.2}$$

The p-values indicate that for all distances d the null hypothesis can be rejected at the 95% significance level, while for the majority of the cases (and in particular for small d) it can also be rejected at the 99% significance level. Similar results are also obtained for the differences observed at the clustering coefficient and the entropy of the venues. Given that the t-test makes the assumption of normality in the data, we also performed the Mann-Whitney U test for the median. The latter does not have the normality assumption and lead us to similar conclusions with respect to the statistical significance of the differences in the network characteristics for the two sets.

APPENDIX C

REGRESSION FOR DIFFERENCE-IN-DIFFERENCES

The exact same estimate from Equation (4.6) for the DD can be formally derived through a linear regression that models the dependent variable y. In particular, we have the following model:

$$y_{ilt} = \gamma_0 + \gamma_1 \cdot \alpha_l + \gamma_2 \cdot \beta_t + \delta \cdot D_{lt} + \epsilon_{ilt} \tag{C.1}$$

where y_{ilt} is the dependent variable for instance i (at time t and location l), α_l and β_t are binary variables that capture the fixed effects of location and time respectively, D_{lt} is a dummy variable that represents the treatment status (i.e., $D_{lt} = \alpha_l \cdot \beta_t$) and ϵ_{ilt} is the associated error term. The coefficient δ captures the effect of the intervention on the dependent variable y. It is then straightforward to show that the DD estimate $\hat{\delta}$ is exactly Equation (4.6). In particular, if $\overline{y}_{lt}$ is the sample mean of y_{ilt} and $\overline{\epsilon}_{lt}$ is the sample mean of ϵ_{ilt}, and using Equation (C.1) we have:

$$(\overline{y}_{11} - \overline{y}_{01}) - (\overline{y}_{10} - \overline{y}_{00}) = \quad \delta(D_{11} - D_{01}) - \delta(D_{10} - D_{00}) \quad + \overline{\epsilon}_{11} - \overline{\epsilon}_{01} + \overline{\epsilon}_{00} - \overline{\epsilon}_{10}$$

Taking expectations and considering the i.i.d. assumptions for the errors for the ordinary least squares we further get:

$$\mathrm{E}[(\overline{y}_{11} - \overline{y}_{01}) - (\overline{y}_{10} - \overline{y}_{00})] = \delta(D_{11} - D_{01}) - \delta(D_{10} - D_{00}) \tag{C.2}$$

Given that the dummy variable D is equal to 1 only when $l = 1$ and $t = 1$ (i.e., for the treatment group after the intervention), we finally get for the DD estimator:

$$\hat{\delta} = (\overline{y}_{11} - \overline{y}_{01}) - (\overline{y}_{10} - \overline{y}_{00}) \tag{C.3}$$

which is essentially the same as Equation (4.6). Therefore, one can estimate the DD using either of the Equations (4.6) or (C.1).

BIBLIOGRAPHY

[1] Pitt study finds location-based marketing a mixed bag. *Pittsburgh Post-Gazette*, 2015. http://www.post-gazette.com/business/tech-news/2015/08/30/Pitt-study-finds-location-based-marketing-a-mixed-bag/stories/201508300063.

[2] P. Adamopoulos and V. Todri. The effectiveness of marketing strategies in social media: Evidence from promotional events. In *ACM SIGKDD*, 2015.

[3] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *ACM KDD*, 2008.

[4] Ahmadali Arabshahi. Undressing groupon: An analysis of the groupon business model, 2010.

[5] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[6] D. Arancibia. Cycling economies: Economic impact of bike lanes. *Report. Toronto Cycling; Think and Do Tank*, 2012.

[7] Orley Ashenfelter and David Card. Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, 67(4):648–60, 1985.

[8] Kalidas Ashok and Moshe E Ben-Akiva. Dynamic origin-destination matrix estimation and prediction for real-time traffic management systems. In *International Symposium on the Theory of Traffic Flow and Transportation (12th: 1993: Berkeley, Calif.). Transportation and traffic theory*, 1993.

[9] Kalidas Ashok and Moshe E Ben-Akiva. Alternative approaches for real-time estimation and prediction of time-dependent origin–destination flows. *Transportation Science*, 34(1):21–36, 2000.

[10] David H Autor. Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing. *Journal of labor economics*, 21(1):1–42, 2003.

[11] F. Baccelli and J. Bolot. Modeling the economic value of location and preference data of mobile users. In *IEEE INFOCOM*, 2011.

[12] James P Bagrow and Yu-Ru Lin. Mesoscopic structure and social aspects of human mobility. *PloS one*, 7(5):e37676, 2012.

[13] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.

[14] Eytan Bakshy, Brian Karrer, and Lada A Adamic. Social influence and the diffusion of user-created content. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 325–334. ACM, 2009.

[15] Syagnik Banerjee and Ruby Dholakia. Mobile advertising: does location based advertising work? *International Journal of Mobile Marketing*, 2008.

[16] Richard A Becker, Ramon Caceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):18–26, 2011.

[17] J. Bernab-Moreno, A. Tejeda-Lorente, C. Porcel, H. Fujita, and E. Herrera-Viedma. Caresome: A system to enrich marketing customers acquisition and retention campaigns using social media information. *Knowledge-Based Systems*, 80:163 – 179, 2015.

[18] Brian Joe Lobley Berry and Quentin Gillard. *The changing shape of metropolitan America: Commuting patterns, urban fields, and decentralization processes, 1960-1970*. Ballinger Publishing Company, 1977.

[19] Adam Blake. Economic impact of the london 2012 olympics. 2005.

[20] Robert C Blattberg, Richard Briesch, and Edward J Fox. How promotions work. *Marketing Science*, 14(3 supplement):G122–G132, 1995.

[21] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

[22] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[23] John W Byers, Michael Mitzenmacher, and Georgios Zervas. The daily deals marketplace: Empirical observations and managerial implications. In *SIGecom Exchanges, Vol. 11, No. 2*. ACM, 2012.

[24] John W Byers, Michael Mitzenmacher, and Georgios Zervas. Daily deals: Prediction, social diffusion, and reputational ramifications. In *ACM WSDM*, 2012.

[25] John W Byers, Michael Mitzenmacher, and Georgios Zervas. The groupon effect on yelp ratings: A root cause analysis. In *ACM EC*, 2012.

[26] Julián Candia, Marta C González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[27] Captical bikeshare. https://www.capitalbikeshare.com/system-data/.

[28] Gerald AP Carrothers. An historical bedew of the gravity and potential concepts of human interaction. *Journal of the American Institute of Planners*, 22(2):94–102, 1956.

[29] Rachael D. Carter and Jeannie W. Zieren. Festivals that say cha-ching!measuring the economic impact of festivalst. In *Main Street Now*, 2012.

[30] Christy MK Cheung, Matthew KO Lee, and Neil Rabjohn. The impact of electronic word-of-mouth: The adoption of online opinions in online customer communities. *Internet Research*, 18(3):229–247, 2008.

[31] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[32] Yohan Chon, Hyojeong Shin, Elmurod Talipov, and Hojung Cha. Evaluating mobility models for temporal prediction with high-granularity mobility data. In *Pervasive computing and communications (percom), 2012 ieee international conference on*, pages 206–212. IEEE, 2012.

[33] CleanAir-Partnership. Bike lanes, on-street parking and business. *Report*, 2009.

[34] Freek Colombijn. *Patches of Padang: The history of an Indonesian town in the twentieth century and the use of urban space*. Number 19. Research School CNWS, 1994.

[35] C. Cortes and M. Mohri. Auc optimization vs. error rate minimization. *NIPS*, 2003.

[36] H. Cramer, M. Rost, and L.E. Holmquist. Performing a check-in: Emerging practices, norms and 'conflicts' in location-sharing using foursquare. *ACM MobileHCI*, 2011.

[37] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *ACK KDD*, 2008.

[38] David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.

[39] Justin Cranshaw, Raz Schwartz, Jason I Hong, and Norman M Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *ICWSM*, 2012.

[40] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 119–128. ACM, 2010.

[41] E. de Place. Are parking meters boosting business?

[42] J.R DeShazo, C. Callahan, M. Brozen, and B. Heimsath. Economic impacts of ciclavia: Study finds gains to local businesses. In *Briefing Paper - UCLA Luscin School of Public Fairs*, 2013.

[43] Utpal M Dholakia. How effective are groupon promotions for businesses. *Social Science Research Network*, 2010.

[44] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM, 2001.

[45] L. Drell. 6 successful foursquare marketing campaigns to learn from. http://mashable.com/2011/07/13/foursquare-marketing-campaigns/#VwgnuIw6.qqn (Last accessed: February 23, 2016).

[46] Benjamin Edelman, Sonia Jaffe, and Scott Kominers. To groupon or not to groupon: The profitability of deep discounts. *Harvard Business School NOM Unit Working Paper*, (11-063), 2011.

[47] B. Efron and R.J. Tibishirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993.

[48] P. Erdős and A. Rènyi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5: 17-61*, 1960.

[49] Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis: theory and extensions*, volume 3. Vsp, 1990.

[50] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *ACM KDD*, 1996.

[51] Facebook places. https://www.facebook.com/places/.

[52] Zheng Fang, Bin Gu, Xueming Luo, and Yunjie Xu. Contemporaneous and delayed sales impact of location-based mobile promotions. *Information Systems Research*, 26(3):552–564, 2015.

[53] Foursquare. https://foursquare.com/.

[54] Foursquare success stories. http://sproutsocial.com/insights/foursquare-success-stories.

[55] James H Fowler and Nicholas A Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *Bmj*, 337:a2338, 2008.

[56] Lauretta Conklin Frederking. A cross-national study of culture, organization and entrepreneurship in three neighbourhoods. *Entrepreneurship & Regional Development*, 16(3):197–215, 2004.

[57] Jordan Frith. Turning life into a game: Foursquare, gamification, and personal mobility. *Mobile Media & Communication*, 1(2):248–262, 2013.

[58] Jon Froehlich, Joachim Neumann, and Nuria Oliver. Measuring the pulse of the city through shared bicycle programs. *Proc. of UrbanSense08*, pages 16–20, 2008.

[59] Foursquare place ratings. http://support.foursquare.com/entries/21942938-Place-Ratings-.

[60] Foursquare special types, 2011. http://www.slideshare.net/opt4digital/an-introduction-to-foursquare-specials.

[61] Discover how bars, restaurants, shops, and more have used foursquare to promote their business. http://business.foursquare.com/discover.

[62] Gian M Fulgoni and M Morn. How online advertising works: Whither the click. *comScore. com Whitepaper*, 2008.

[63] Luciano Gallegos, Kristina Lerman, Arthur Huang, and David Garcia. Geography of emotion: Where in a city are people happier? In *WWW*, 2016.

[64] Petko Georgiev, Anastasios Noulas, and Cecilia Mascolo. Where businesses thrive: Predicting the impact of the olympic games on local retailers through location-based services data. *arXiv preprint arXiv:1403.7654*, 2014.

[65] Donald Getz et al. *Festivals, special events, and tourism.* Van Nostrand Reinhold, 1991.

[66] S. Goel, D.J. Watts, and D.G. Goldstein. The structure of online diffusion networks. In *ACM EC*, 2012.

[67] Charles J. Goetz and Robert E. Scott. Liquidated damages, penalties and the just compensation principle: Some notes on an enforcement model and a theory of efficient breach. *Columbia Law Review*, 77(4):pp. 554–594, 1977.

[68] Avi Goldfarb and Catherine Tucker. Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3):389–404, 2011.

[69] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[70] Chris Gratton, Simon Shibli, and Richard Coleman. Sport and economic regeneration in cities. *Urban studies*, 42(5-6):985–999, 2005.

[71] Michael J Greenwood. Research on internal migration in the united states: a survey. *Journal of Economic Literature*, pages 397–433, 1975.

[72] Nir Grinberg, Mor Naaman, Blake Shaw, and Gilad Lotan. Extracting diurnal patterns of real world activity from social media. In *ICWSM*, 2013.

[73] Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

[74] Anthony D Harris, Jessina C McGregor, Eli N Perencevich, Jon P Furuno, Jingkun Zhu, Dan E Peterson, and Joseph Finkelstein. The use and interpretation of quasi-experimental studies in medical informatics. *Journal of the American Medical Informatics Association*, 13(1):16–23, 2006.

[75] Kingsley E Haynes, Dudley L Poston Jr, and Paul Schnirring. Intermetropolitan migration in high and low opportunity areas: Indirect tests of the distance and intervening opportunities hypotheses. *Economic Geography*, 49(1):68–73, 1973.

[76] Martin L Hazelton. Inference for origin–destination matrices: estimation, prediction and reconstruction. *Transportation Research Part B: Methodological*, 35(7):667–676, 2001.

[77] Wenbo He, Xue Liu, and Mai Ren. Location cheating: A security challenge to location-based social network services. In *Distributed Computing Systems (ICDCS), 2011 31st International Conference on*, pages 740–749. IEEE, 2011.

[78] Petter Holme and Mark EJ Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E*, 74(5):056108, 2006.

[79] D. Hristova, M.J. Williams, M. Musolesi, P. Panzarasa, and C. Mascolo. Measuring urban social diversity using interconnected geo-social networks. In *ACM WWW*, 2016.

[80] GM Hyman et al. The calibration of trip distribution models. *Environment and Planning*, 1(3):105–112, 1969.

[81] Guido Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. Working Paper 13039, National Bureau of Economic Research, April 2007.

[82] Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.

[83] Pablo Jensen. Network-based predictions of retail store commercial categories and optimal locations. *Phys. Rev. E*, 74:035101, Sep 2006.

[84] Pablo Jensen. Analyzing the localization of retail stores with complex systems tools. In *Advances in Intelligent Data Analysis VIII*, volume 5772 of *Lecture Notes in Computer Science*, pages 10–20. Springer Berlin Heidelberg, 2009.

[85] Lei Jin, Ke Zhang, Jianfeng Lu, and Yu-Ru Lin. Towards understanding the gamification upon users scores in a location-based social network. *Multimedia Tools and Applications*, pages 1–25, 2014.

[86] Michael A Jones, David L Mothersbaugh, and Sharon E Beatty. The effects of locational convenience on customer repurchase intentions across service types. *Journal of Services Marketing*, 17(7):701–712, 2003.

[87] Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. Understanding human mobility from twitter. *PloS one*, 10(7):e0131469, 2015.

[88] Andreas Kaltenbrunner, Rodrigo Meza, Jens Grivolla, Joan Codina, and Rafael Banchs. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4):455–466, 2010.

[89] Dmytro Karamshuk, Anastasios Noulas, Salvatore Scellato, Vincenzo Nicosia, and Cecilia Mascolo. Geo-spotting: Mining online location-based services for optimal retail store placement. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 793–801. ACM, 2013.

[90] Praveen K Kopalle, Carl F Mela, and Lawrence Marsh. The dynamic effect of discounting on sales: Empirical analysis and normative pricing implications. *Marketing Science*, 18(3):317–332, 1999.

[91] G. Kossinets and D.J. Watts. Empirical analysis of an evolving social network. In *Science 311:88-90*, pages 405–450, 2006.

[92] G. Kossinets and D.J. Watts. Origins of homophily in an evolving social network. In *American Journal of Sociology, Volume 115, Number 2*, pages 405–450, 2009.

[93] Gautier Krings, Francesco Calabrese, Carlo Ratti, and Vincent D Blondel. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):L07003, 2009.

[94] Kevin S Kung, Kael Greco, Stanislav Sobolevsky, and Carlo Ratti. Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one*, 9(6):e96180, 2014.

[95] H.R. Künsch. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3):1217–1241, 1989.

[96] Choong-Ki Lee and Tracy Taylor. Critical reflections on the economic impact assessment of a mega-event: the case of 2002 fifa world cup. *Tourism Management*, 26(4):595–603, 2005.

[97] Everett S Lee. A theory of migration. *Demography*, 3(1):47–57, 1966.

[98] K. Lewis, M. Gonzalez, and J. Kaufman. Social selection and peer influence in an online social network. In *Proceedings of the National Academy of Sciences, Volume 109, Number 1*, pages 405–450, 2012.

[99] Xiao Liang, Jichang Zhao, Li Dong, and Ke Xu. Unraveling the origin of exponential law in intra-urban human mobility. *Scientific reports*, 3, 2013.

[100] Yu Liu, Chaogui Kang, Song Gao, Yu Xiao, and Yuan Tian. Understanding intra-urban trip patterns from taxi trajectory data. *Journal of geographical systems*, 14(4):463–483, 2012.

[101] Pin Luarn, Jen-Chieh Yang, and Yu-Ping Chiu. Why people check in to social network sites. *International Journal of Electronic Commerce*, 19(4):21–46, 2015.

[102] Michael Luca. Reviews, reputation, and revenue: The case of yelp. com. *Com (September 16, 2011). Harvard Business School NOM Unit Working Paper*, (12-016), 2011.

[103] Luhur, H.S. and Widjaja, N.D. Location-based social networking media for restaurant promotion and food review using mobile application. *EPJ Web of Conferences*, 68, 2014.

[104] A Craig MacKinlay. Event studies in economics and finance. *Journal of economic literature*, pages 13–39, 1997.

[105] Vijay Mahajan, Eitan Muller, and Frank M Bass. New product diffusion models in marketing: A review and directions for research. In *Diffusion of technologies and social behavior*, pages 125–177. Springer, 1991.

[106] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[107] Edward Miller. A note on the role of distance in migration: costs of mobility versus intervening opportunities. *Journal of Regional Science*, 12(3):475–478, 1972.

[108] Ricardo Mora and Iliana Reggio. Treatment effect identification using alternative parallel assumptions. 2012.

[109] S.A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *ACM KDD*, 2012.

[110] M.E.J. Newman. Mixing patterns in networks. In *arXiv:cond-mat/0209450v2 [cond-mat.stat-mech]*, 2002.

[111] Anastasios Noulas, Cecilia Mascolo, and Enrique Frias-Martinez. Exploiting foursquare and cellular data to infer user activity in urban environments. In *2013 IEEE 14th International Conference on Mobile Data Management*, volume 1, pages 167–176. IEEE, 2013.

[112] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5):e37027, 2012.

[113] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. Mining user mobility features for next place prediction in location-based services. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1038–1043. IEEE, 2012.

[114] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. A random walk around the city: New venue recommendation in location-based social networks. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 144–153. IEEE, 2012.

[115] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare. *ICwSM*, 11:70–573, 2011.

[116] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. *The Social Mobile Web*, 11, 2011.

[117] Nyc opendata. https://nycopendata.socrata.com/.

[118] Panagiotis Papadimitriou, Hector Garcia-Molina, Prabhakar Krishnamurthy, Randall A Lewis, and David H Reiley. Display advertising impact: Search lift and social influence. In *ACM SIGKDD*, 2011.

[119] Koen Pauwels, Dominique M Hanssens, and Sivaramakrishnan Siddarth. The long-term effects of price promotions on category incidence, brand choice, and purchase quantity. *Journal of marketing research*, pages 421–439, 2002.

[120] K. Pelechrinis and P. Krishnamurthy. Location affiliation networks: Bonding social and spatial information. *ECML/PKDD*, 2012.

[121] Konstantinos Pelechrinis, Marios Kokkodis, and Theodoros Lappas. On the value of shared bike systems in urban environments: Evidence from the real estate market. *Available at SSRN*, 2015.

[122] Pittsburgh healthyride. https://healthyridepgh.com/data/.

[123] Yaniv Poria, Richard Butler, and David Airey. The core of heritage tourism. *Annals of tourism research*, 30(1):238–254, 2003.

[124] Sergio Porta, Vito Latora, Fahui Wang, Salvador Rueda, Emanuele Strano, Salvatore Scellato, Alessio Cardillo, Eugenio Belli, Francisco Crdenas, Berta Cormenzana, and Laura Latora. Street centrality and the location of economic activities in barcelona. *Urban Studies*, 49(7):1471–1488, 2012.

[125] Michael E Porter. Location, competition, and economic development: Local clusters in a global economy. *Economic development quarterly*, 14(1):15–34, 2000.

[126] Krishna PN Puttaswamy and Ben Y Zhao. Preserving privacy in location-based mobile social applications. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*, pages 1–6. ACM, 2010.

[127] Yan Qu and Jun Zhang. Trade area analysis using user generated mobile location data. In *ACM WWW*, 2013.

[128] Ernest George Ravenstein. The laws of migration. *Journal of the Statistical Society of London*, 48(2):167–235, 1885.

[129] Jonathan Reades, Francesco Calabrese, and Carlo Ratti. Eigenplaces: analysing cities using the space–time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836, 2009.

[130] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186. ACM, 1994.

[131] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[132] Adam Sadilek, Henry Kautz, and Jeffrey P Bigham. Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 723–732. ACM, 2012.

[133] Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Andrew T Campbell. Nextplace: a spatio-temporal prediction framework for pervasive systems. In *Pervasive Computing*, pages 152–169. Springer, 2011.

[134] Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo. Socio-spatial properties of online location-based social networks. *ICWSM*, 11:329–336, 2011.

[135] Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1046–1054. ACM, 2011.

[136] Jerry Scripps, Pang-Ning Tan, and Abdol-Hossein Esfahanian. Measuring the effects of preprocessing decisions and network forces in dynamic network analysis. In *Proceedings*

of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 747–756. ACM, 2009.

[137] William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Cengage Learning, 2001.

[138] Donald Shoup. *The High Cost of Free Parking*. American Planning Association, 2011.

[139] Vibhooti Shukla and Paul 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*, 21(2):225–253, 1991.

[140] Filippo Simini, Marta C González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

[141] Parag Singla and Matthew Richardson. Yes, there is a correlation:-from social networks to personal behavior on the web. In *Proceedings of the 17th international conference on World Wide Web*, pages 655–664. ACM, 2008.

[142] Larry A Sjaastad. The costs and returns of human migration. In *Regional Economics*, pages 115–133. Springer, 1970.

[143] Adam Sliwinski. Spatial point pattern analysis for targeting prospective new customers: bringing gis functionality into direct marketing. *Journal of Geographic Information and Decision Analysis*, 6(1):31–48, 2002.

[144] Stephen LJ Smith. Restaurants and dining out: geography of a tourism business. *Annals of Tourism Research*, 10(4):515–549, 1983.

[145] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[146] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, and Ryosuke Shibasaki. Prediction of human emergency behavior and their mobility following large-scale disaster. pages 5–14, 2014.

[147] Shuba Srinivasan, Koen Pauwels, Dominique M Hanssens, and Marnik G Dekimpe. Do promotions benefit manufacturers, retailers, or both? *Management Science*, 50(5):617–629, 2004.

[148] Josef Steindl. *Random processes and the growth of firms: A study of the Pareto law*. Griffin London, 1965.

[149] Samuel A Stouffer. Intervening opportunities: a theory relating mobility and distance. *American sociological review*, 5(6):845–867, 1940.

[150] Chenhao Tan, Jie Tang, Jimeng Sun, Quan Lin, and Fengjiao Wang. Social action tracking via noise tolerant time-varying factor graphs. In *Proceedings of the 16th ACM*

SIGKDD international conference on Knowledge discovery and data mining, pages 1049–1058. ACM, 2010.

[151] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *ACM KDD*, 2009.

[152] J. Tang, S. Wu, and J. Sun. Confluence: Conformity influence in large social networks. In *ACM KDD*, 2013.

[153] The mathematics of gamification. http://engineering.foursquare.com/2014/01/03/the-mathematics-of-gamification/.

[154] Alessandro Venerandi, Giovanni Quattrone, Licia Capra, Daniele Quercia, and Diego Saez-Trumper. Measuring urban deprivation from user generated content. In *ACM CSCW*, pages 254–264, 2015.

[155] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. ACM, 2011.

[156] Qi Wang and John E Taylor. Resilience of human mobility under the influence of typhoons. *Procedia Engineering*, 118:942–949, 2015.

[157] Alan G Wilson. A statistical theory of spatial distribution models. *Transportation research*, 1(3):253–269, 1967.

[158] Lun Wu, Ye Zhi, Zhengwei Sui, and Yu Liu. Intra-urban human mobility and activity transition: Evidence from social media check-in data. *PloS one*, 9(5):e97010, 2014.

[159] Mao Ye, Peifeng Yin, and Wang-Chien Lee. Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 458–461. ACM, 2010.

[160] Yelp. http://www.yelp.com.

[161] Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 186–194. ACM, 2012.

[162] Ke Zhang, Wei Jeng, Francis Fofie, Konstantinos Pelechrinis, and Prashant Krishnamurthy. Towards reliable spatial information in lbsns. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 950–955. ACM, 2012.

[163] Ke Zhang, Qiuye Jin, Konstantinos Pelechrinis, and Theodoros Lappas. On the importance of temporal dynamics in modeling urban activity. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, page 7. ACM, 2013.

[164] Ke Zhang, Yu-Ru Lin, and Konstantinos Pelechrinis. Eigentransitions with hypothesis testing: The anatomy of urban mobility. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

[165] Ke Zhang, Konstantinos Pelechrinis, and Prashant Krishnamurthy. Acm hotmobile 2013 poster: detecting fake check-ins in location-based social networks through honeypot venues. *ACM SIGMOBILE Mobile Computing and Communications Review*, 17(3):29–30, 2013.

[166] Ke Zhang, Konstantinos Pelechrinis, and Theodoros Lappas. Analyzing and modeling special offer campaigns in location-based social networks. In *ICWSM*, 2015.

[167] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):38, 2014.

[168] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 312–321. ACM, 2008.