Understanding and Modeling Human Movement in Cities Using Phone Data

by

Fahad Alhasoun

Submitted to the Center for Computational Engineering (CCE) and the Department of Electrical Engineering and Computer Science (EECS)
in partial fulfillment of the requirements for the degree of
Master of Science in Compution for Design and Optimization and
Master of Science in Electrical Engineering and Computer Science
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
September 2016

Author ..... Signature redacted .....
Center for Computational Engineering (CCE) and the Department of Electrical Engineering and Computer Science (EECS)
August 18, 2016

Certified by.. Signature redacted .....
Marta C. González
Associate Professor, Civil and Environmental Engineering
Thesis Supervisor

Certified by... Signature redacted .....
Una-May O'Reilly
Principal Research Scientist, Computer Science and Artificial Intelligence Lab
Thesis Supervisor

Accepted by ... Signature redacted .....
Leslie Kolodziejski
Professor, Department of Electrical Engineering and Computer Science
Chair of the Committee on Graduate Students

Accepted by ... Signature redacted .....
Youssef Marzouk
Associate Professor, Department of Aeronautics and Astronautics
Co-Director, Computation for Design and Optimization

Understanding and Modeling Human Movement in Cities Using Phone Data

by

Fahad Alhasoun

Submitted to the Center for Computational Engineering (CCE) and the
Department of Electrical Engineering and Computer Science (EECS)
on August 18, 2016, in partial fulfillment of the
requirements for the degree of
Master of Science in Compution for Design and Optimization and Master of Science
in Electrical Engineering and Computer Science

Abstract

Cities today are strained by the exponential growth in population where they are homes to the majority of world's population. Understanding the complexities underlying the emerging behaviors of human travel patterns on the city level is essential toward making informed decision-making pertaining to urban transportation infrastructures This thesis includes several attempts towards modeling and understanding human mobility at the scales of individuals and the scale of aggregate population movement. The second chapter includes the development of a browser delivering visual insights of the aggregate behavior of populations in cities. The third chapter provides a computational framework for clustering regions in cities based on their attraction behavior and in doing so aids a predictive model in predicting inflows to newly developed regions. The fourth chapter investigates the patterns of individuals' movement at the city scale towards developing a predictive model for a persons' next visited location. The predictive accuracy is then increased by adding movement information of the population. The motivation behind the work of this thesis is derived from the demand of tools that provides fine-grained analysis of the complexity of human travel within cites. The approach takes advantage of the existing built infrastructures to sense the mobility of people eliminating the financial and temporal burdens of traditional methods. The outcomes of this work will assist both planners and the public in understanding the complexities of human mobility within their cities.

Thesis Supervisor: Marta C. González
Title: Associate Professor, Civil and Environmental Engineering

Thesis Supervisor: Una-May O'Reilly
Title: Principal Research Scientist, Computer Science and Artificial Intelligence Lab

Acknowledgments

To my parents, there will never be enough words to fully express my gratitude and love to you. You have always been behind every success over the years. To my sister and brothers, Yasser, Abeer, Ziad, Bader and Turki, you have always inspired me, believed in me and supported me to do my best. To my whole family, though we are thousands of kilometers apart, my heart will always be with you.

I would like to express my sincere gratitude to my advisor Marta C. González for her continuous support of my research. What she brings to the whole research group is not only domain expertise, but also support and enthusiasm. I would also like to thank my thesis reader Dr. Una-May O'Reilly for her support of my research.

This dissertation would not have been possible without the collaboration and the help from the following individuals: May Alhazzani and Zeyad Alawwad. I want to thank our collaborators from Saudi Arabia and namely the Saudi Telecom Company (STC) and Arriyadh Development Authority (ADA): Thank you for your support.

I would like to thank the Center for Computational Engineering (CCE) and the Electrical Engineering and Computer Science department (EECS) for providing such a delightful environment. All the courses I've taken and all the talks with the professors will give me a lifetime of benefits.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction and Overview

Cities today house over 50 percent of world's population, consuming 60-80 percent of global energy and emitting almost 75 percent of greenhouse gases [50]. Some have suggested that almost 70 percent of world's population will reside in cities by 2050 [50]. With the rapid urban population growth, cities' infrastructures are being strained to the point of becoming a major hindrance to socioeconomic activity. Left unaddressed, the problem threatens to weigh down the return on investment from public projects being constructed throughout cities and adversely affect the quality of life of all residents.

Understanding the complexities underlying the emerging behaviors of human travel patterns on the city level is essential toward making informed decision-making pertaining to urban transportation infrastructures [7]. Traditional methods of assessing the social demand on transportation are expensive and take longer periods of time to conduct [60, 54, 30]. Such assessments are usually in the form of surveys with considerably small sample sizes compared to the total population of a city. Furthermore, such methods lack the accuracy and resolution in time to provide fine-grained analysis of human travel with precise time resolution.

New road counter technologies such as pressure tubes, inductive loops and other traffic counting techniques allowed for counting travelers with a finer time resolution;

however, the drawback is the spatial resolution of such techniques. They are usually highly local and capture activity in a specific point in space that is miniscule with respect to the city as a whole [46]. Therefore, such techniques suffer from an inability to provide a holistic overview of the status of the system. In addition, deploying new traffic counting technologies can be extremely expensive when considering the mega-cities in the world.

An alternative approach toward capturing the social demand is by using data generated from mobile phones to model and understand the behavior of human mobility [60]. Data pertaining to mobile phone usage can be gathered at different levels within the GSM network. Telecom companies usually do not keep track of all the data traffic running across their networks; however, they store certain information for billing purposes and network development. The Call Detail Records, often referred to as CDRs, are one type of information telecom companies keep for billing purposes. Every time a user makes a phone call, sends a text message uses the Internet and even passively when the mobile communicates to the cellular network access points, the mobile network keeps a record of their usage information and location in the CDRs [33]. Therefore, such big data set can be utilized as a proxy to understand the social demand on transportation infrastructures.

The motivation behind developing the work of this thesis is derived from the demand of tools that provides fine-grained analysis of the complexity of human travel within cites. The approach takes advantage of the existing built infrastructures to sense the mobility of people eliminating the financial and temporal burdens of traditional methods. The outcomes of this work will assist both planners and the public in understanding the complexities of human mobility within their cities.

The nature of the code used in this thesis is divided into blocks that process data of varying aggregations formats such as trips in the Origin Destination matrix or set of inflows to a region or sequence of locations visited for a certain user under study. For example, a clear pipeline that digests chunks of the data from various formats is illustrated in chapter three figure 3-1. This thesis uses code blocks that were implemented in the Human Mobility and Networks Lab (HuMNet) such as the

code to mine the raw CDRs for the trips in a city. The code is available on the group's github account `github.com/Humnetlab`. Other codes in this thesis include developing predictive models for inflows (gaussian process), predicting a person's next location (dynamic bayesian network) or clustering attraction profiles (hierarchical agglomerative clustering). The code blocks are available upon request on the author's github account at `github.mit.edu/fha`.

This thesis includes text and experiments from a subset of my publications while at MIT [1, 2, 3]. The coauthors and collaborators of these publications deserve due credit and thanks: May Alhazzani for the work on the urban attractors in chapter three and the work on next place prediction on chapter four and Kael Greco for the work on the City Browser in chapter two.

1.2 Literation Review

1.2.1 Browsing City Mobility Data

Several research activities have been investigating approaches towards modeling and understanding mobility demand within cities. Traditional methods of demand modeling inferred the collective behavior of demand on transportation infrastructures through household or road surveys to gather information about user's behavior. Another approach has been to use theoretical models to estimate the number of trips and their directionality based on land use models. These approaches can be unreliable and can have financial and temporal costs. Today and with the emergence of pervasive technologies around the world, research started investigating human behavior through data gathered from mobile phones [26, 21, 32, 35]. Varying approaches have used the data as a proxy to better understand human mobility. The focus on human mobility ranges from decomposing the data onto the different dimensions to gain insights into behavioral patterns by applying algorithms and processes on models built on the data. Research investigating the dimensionality of the data includes work on utilizing the spatial decomposition of aggregate activity to understand the dynamics of cities and

universal patterns of human mobility [26, 10]. On the other hand, researchers have developed techniques to gain more insights from the data by creating algorithms capturing more of the hidden patterns [32, 17, 63]. For example, researchers have been modeling the social network based on the data captured from users' interactions to better understand whether the composition of social communities is correlating with the geographical constraints [49]. Another approach was to capture users' trips from the data set and aggregate trips to get insight on the flows of people around the city towards understanding the dynamics of flows of people [11, 60]. Such understanding can help identify flawed urban planning in cities [68].

1.2.2 Classification of Regions in a City

Multiple studies used human mobility behavior to classify urban areas. A recent study investigated the relationship between land use and mobility[36]. The authors showed that purposes of people's trips are strongly correlated with the land use of the trip's origin and destination. Recently, the availability of dynamic sources of data allowed for dynamic segmentation of the city according to human mobility behavior. Some studies combined human mobility with land use or POIs data to segment districts in urban areas according to their functions or use. The type of data used to capture human mobility behavior varies between individuals GPS traces [69, 23], taxi pick up/drop off locations as in [38, 45] , Call Detail Records (CDRs) as in [40, 59], social media check ins as in [66, 39, 6] , and bus smart card data as in [29]. However,to our knowledge, non of the previous studies quantified the attraction of places and used attraction profiles to segment the city.

Survey travel data has been used to detect the centers (significant places) of a city [71, 18]. A recent study proposed a method for measuring the centrality of locations that incorporates the number of people attracted to the location and the diversity of activities in which visitors engage [71]. The proposed method was tested on survey travel data in Singapore to identify the functional centers and track their significance over time. A similar approach focused on analyzing the aggregate behavior of the population to predicted highly attractive events such as the times square during

new years count down in new york [24]. Our method is based on validated origin destination matrices mined from cell phone data that captures human mobility. More significantly, our approach incorporate not just the amount of people a place attracts but also on where do they come from and the road distance they traveled.

Network analysis methods were used to detect hotspots based on flow patterns between locations[40, 62]. A recent paper [40] used Origin Destination (ODs) matrices extracted from cell phone data to identify the signature of mobility behavior as 4 main types of movements within the city: from hotspot to hotspot, to hotspots, originating at hotspots and the random flows. They showed how different cities have different mobility signatures. Additionally, a recent study used Taxi drop off/pick up GPS traces in Shanghai to create a network of flow between places. They applied community detection to extract sub regions and analyze the interaction between sub regions and within each sub region.

Researchers adapted modeling approaches from Natural Language Processing (NLP) in identifying functional zones in urban areas [65, 64]. One study applied a Latent Dirichlet Allocation (LDA) model on Foursquare check ins to detect local geographic topics that indicate the potential and intrinsic relations among the locations in accordance with users' trajectories. A recent study used LDA and POIs to detect functional zones[65].

1.2.3 Prediction of Individuals' Movement

Humans are creatures of habit where various aspects of human behavior exhibiting high periodicity. The nature of human movement is a combination of both periodic movement such as home-work daily trips as well as random explorations to attraction areas or social gatherings [12]. The random aspect of human mobility as well as the heterogeneity of individual preferences make the problem of predicting the next visited location of an individual a challenging one [27, 37, 70, 53]. Undoubtedly, the decision making process of humans is highly influenced by their social interactions [28]; statistically significant tests on the similarity of human mobility with respect to the existence of strong social ties suggests that social ties are an important influential

factor to the travel patterns of people [31]. This has provided an opportunity to improve human mobility models by incorporating the patterns of movement of friends [16].

Research towards predicting next locations of people have also shown to be very successful on data with varying resolution in space and time [41, 52, 20]. Targeting the locations of the social contacts showed to improve the prediction of the leisure locations of an ego when using GPS-geo-tagged Twitter or GPS type of data [51, 16, 15]. However, here we show that the same approaches fail greatly with Call Detailed Records (CDRs) from mobile phone data due to the sparsity of data on the temporal dimension significantly reducing the amount of observed mobility of an individual.

1.3 Dataset

The dataset consists of one full month of phone calling records for the entire country of Saudi Arabia, with 3 billion mobile activities to over 10 thousands unique cell towers, provided by a single carrier. Each record contains an anonymized user ID, the type of activity (i.e., SMS, MMS, call, data etc), the cell tower facilitating the service, duration if its a phone call, and time stamp of the activity. Each cell tower id is spatially mapped to its latitude and longitude. For privacy concerns, user id information were completely anonymized at the telecom operator side.

The focus of the thesis is on the capital of Saudi Arabia, Riyadh. The dataset for Riyadh consists of one month of Call Detail Records (CDRs) with around 1800 towers. The data provides CDRs for the duration of the month of December 2012 from a major telecom operator in the city. Phone activity for about 300 thousands users during the one month period makes about 109M records in the CDRs. Each cell tower ID is spatially mapped to its latitude and longitude where each voronoi cell in figure 1-1 correspond to a tower. The spatial granularity of a cell tower varies across the city; figure 1-1 shows the city of Riyadh and the voronoi cells of the towers in the city, areas that are closer to the center of the city have smaller cells while as we move away from the center, cells have larger sizes.

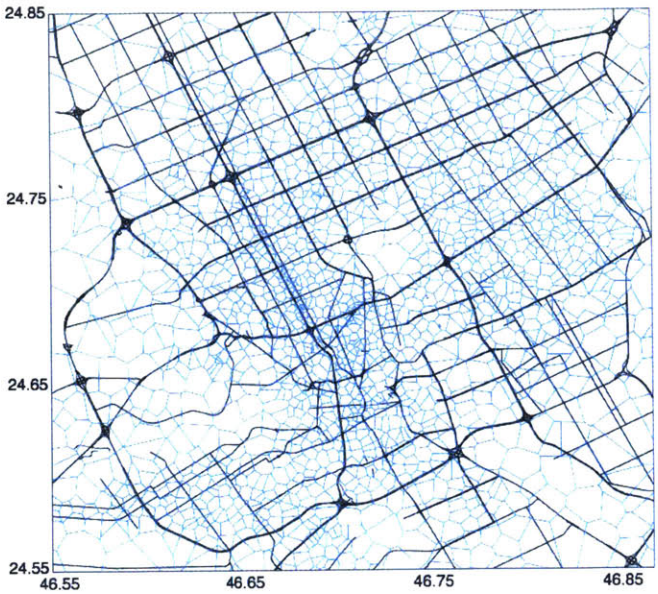

Figure 1-1: Riyadh city and the coverage of cell towers.

Previous studies [44, 26] have shown that human communication patterns are highly heterogeneous; where some users use their mobile phone much more frequently than others. The characteristics of the dynamics of individual communication activity obtained in Fig 1-2 supports such hypothesis.

The dataset was provided by the Saudi Telecom Company (STC) under a nondisclosure agreement for a project aimed at understanding Riyadh city dynamics and thus the data is not available for public access. However, the algorithms and procedures discussed in this thesis are applicable to any CDRs data due to them being a generic type of data telecom operators use for billing purposes.

1.4 Thesis Outline

The rest of the thesis will show the applications of CDRs and other digital traces in different aspects of mobility modeling. Chapter 2 discusses the development of the City Browser, a tool used to browse through City Mobility data. The tool mines the CDRs for insights inferred from the movement of people as well as their daily home and work locations. The tool is composed of several components that interact with a visual interface for the delivery of information. Chapter 3 goes into further

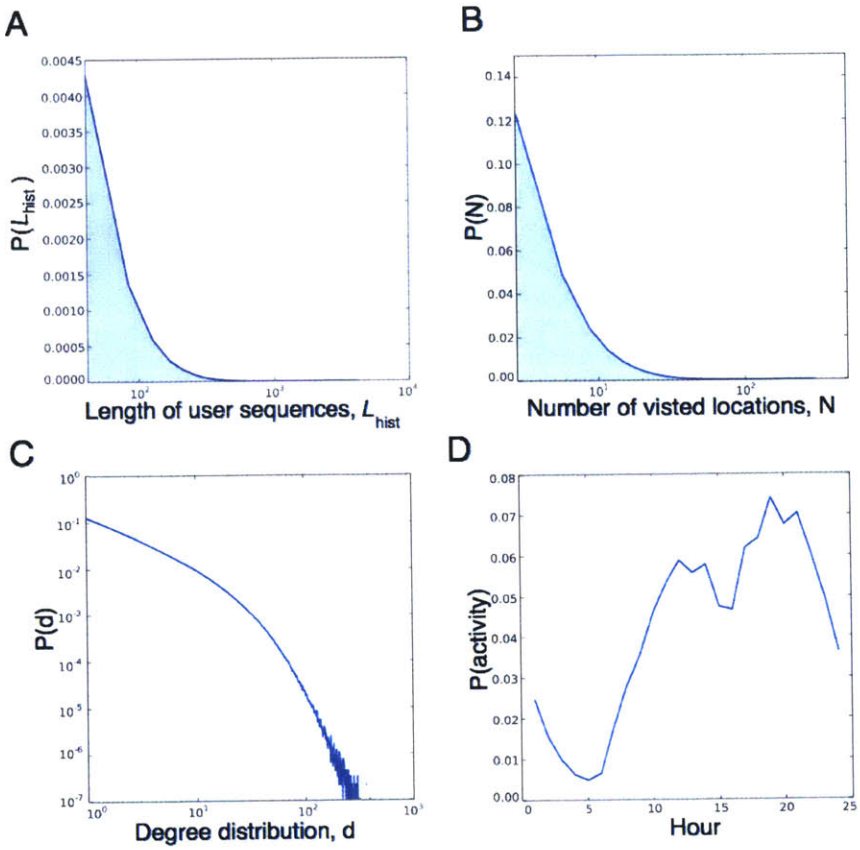

Figure 1-2: The distributions of (a) Length of sequences of locations L_{hist} and (b) Number of visited locations N and (c) Degree distribution of the social network and (d) Total activity of phone usage around a typical day.

detail in terms of mobility on the aggregate scale of the population. The chapter goes into further detail in classifying regions with the city depending on how they attract people. The chapter then moves to developing a predictive model for the estimation of incoming flows for newly developed regions. The developed method is then shown to perform better than a version of the gravity model chosen as the baseline model for our study. Chapter 3 investigates mobility at the individual scale. The chapter investigates how one can predict a person's next location given their movement history as CDRs records. The chapter investigates how one can increase the accuracy of the prediction by observing the mobility of others, this includes people of similar mobility patterns. The developed Dynamic Bayesian Network (DBN) achieves accuracy better than that of the a baseline of a Markov Chain.

Chapter 2

The City Browser: Visualizing City Mobility Dynamics through Massive Call Data

2.1 Introduction

The motivation behind developing the browser is derived from the demand of a tool that provides fine-grained analysis of the complexity of human travel within cites. The approach takes advantage of the existing built infrastructures to sense the mobility of people eliminating the financial and temporal burdens of traditional methods. The outcomes of the tool will assist both planners and the public in understanding the complexities of human mobility within their cities.

In this chapter we will present the City Mobility Browser, a tool that facilitates a simplified understanding of human mobility across a city. The chapter is divided into three sections: Section 2.2 presents the methodology of the browser, Section 2.3 describes the general architecture of the system, Section 2.4 describes each component of the tool in detail, and Section 2.6 presents results of the case study of city of Riyadh in Saudi Arabia. The contributions of this chapter can be summarized into the following two points:

- We propose an architecture that combines several known techniques for data collection, storage and analysis in one framework in a meaningful context to develop the "City Browser", that can aid in simplifying the complexity of human mobility across a city.

- We examine the usefulness of the system through a case study of Riyadh, Saudi Arabia. The case study contained 100 million real mobile phone activity and demonstrates the process of analyzing massive amount of data and through visualization, distilling the bits into actionable insights.

2.2 Methodology

The objective of the browser is to provide an understanding of the complexity underlying human mobility within a city. The browser will capture the dynamics of the distribution of the population to investigate aspects pertaining to flows of people as well as the structure of the community. Investigating population localization dynamics provides information pertaining to emerging zones with higher population densities; certain dense zones emerge on daily basis like commercial areas on weekdays while others emerge as consequence of events that are not of periodic nature. The browser will investigate whether the formation of periodic dense zones has an influence on the segregating of the population of the city into communities. On the other hand, it will provide information about how the city interacts with events in terms of population commuting flows.

The approach towards simplifying the complexity of human mobility is staged into four steps. Starting with step 1, the browser decomposes population distribution across the spatial dimension on a time resolution of a day capturing the emergence of dense zones (see Subsection 2.2.1). Step 2 then analyzes each individual in the CDRs to capture their home/work locations (see Subsection 2.2.2). Step 3 as explained in subsection 2.2.3 investigates the formation of communities within cities as a result of their home/work choices. Step 4 estimates people flows within the city within a day time scale (see Subsection 2.2.4).

2.2.1 Spatial-Temporal Decomposition

The first phase of the methodology decomposes the population over the spatial dimension of the city on the day scale; it will capture time series information of densities of people at every zone with time granularity in minutes. The technique quantifies the magnitudes of mobile user activities within the defined time window, generating time series data for user activity densities for each zone covered by a cell tower. Observing densities with such fine time granularity provides fine grained detail on the emergence of such populated zones by identifying when, where and how fast different dense zones emerge.

2.2.2 Home/Work Places Capturing

The second phase takes a larger time granularity spanning weeks to capture residential and business areas. The approach towards that is by identifying locations where users spend most of their time during day and night (i.e. home/work locations) across a sufficient time interval. Aggregating the number of users spending most of their times over a particular location captures zones that are emerging as a result of daily routine activities like regular business areas and schools.

2.2.3 Community Detection

To better understand the influence of where people live and work, this phase investigates the formation of segregated communities based on their home and work locations. The formation of a mobility community within the population indicates that there is a subset of the population traveling within confined bounds of the city and tend not to cross those bounds (i.e. a neighborhood or group of neighborhoods). Such analysis can provide insights on the level of heterogeneity of trips' sources and destinations.

2.2.4 Flows Estimation

To better understand daily commuting within a city, this phase captures flows within the city through the origin destination estimation algorithm. The algorithm captures trips generated by users around the day and then aggregates the flows of people on a specified time window. The results of the origin destination estimation algorithm will provide information about how dense zones emerge in terms of the source of the population visiting those zones.

2.3 General Architecture

The general architecture of the browser is composed of three major components; data warehouse, modules and algorithms, and the visualization interface. The data warehouse contains the needed data for the modules and algorithms to produce insights and information visualized through the visualization interface. The general architecture is shown in the figure 2-1. The data warehouse contains data pertaining to human mobile phone usage as well as GIS information of the city and traffic counts. There are four major modules residing within the modules and algorithms component that are spatial-temporal decomposition module, home/work capturing module, community detection module and flows estimation module. Finally the visualization interface takes the results produced by the modules and algorithms together with GIS information of the city to provide a comprehensive dynamic view of human mobility within a city.

2.4 Components

The City Browser is decomposed into components following the general architecture described in section 4. This section will provide the details of each component. The breakdown of the browser into components is to allow for a more scalable, modular and simpler architecture for development. Each of the components is describes below.

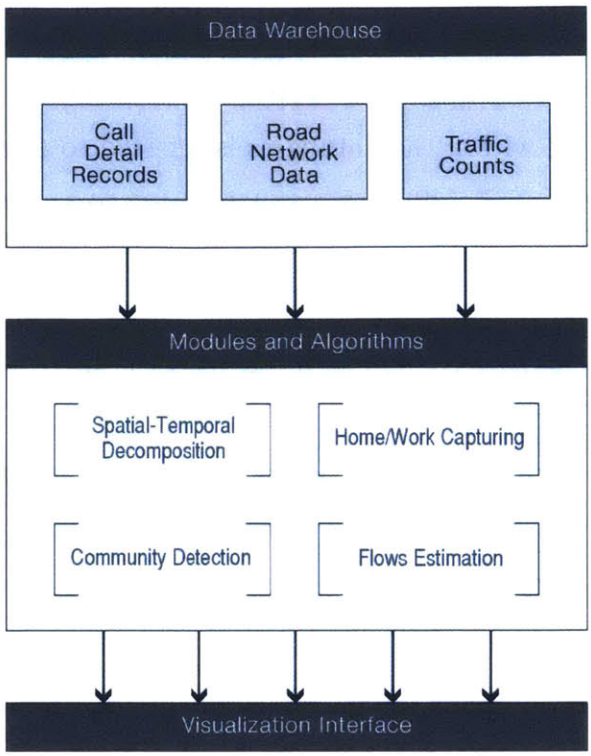

Figure 2-1: City Browser general architecture

2.4.1 Data Warehouse

The data warehouse houses several datasets containing information of the structure of the city as well as the dynamics of it. It contains a geospatial database of the city including the lookup table of the locations of the cell towers for the purpose of mapping mobile phone activity to locations. In addition, it contains information of the time series mobile phone usage data as well as traffic counts.

The major part of the data warehouse is mobile phone billing information, also known as Call Detail Records (CDRs), which are records that telecom companies usually keep for the purpose of generating bills for customers. The CDRs are generated by mobile switching centers (MSCs) within GSM networks and go through several processing methods to be usable by telecom providers. The CDRs are finally structured in a table-like format, withholding information about phone activity details. Each entry in the CDRs table is a record representing an activity generated by a user. Every time a user makes a phone call, sends a text message or accesses the Internet, the CDRs keeps a record of the cell tower that was used to facilitate activ-

ity. In addition, the data warehouse contains a lookup table for cell tower geospatial information where each cell tower is mapped to its coordinates (i.e. latitude and longitude). Each record within the databases is referred to as an activity and is described by time t, user u and cell tower c and represented as $a(t, c, u)$. For each user, the dataset contains a series of activities captured and are represented as:

$$A_u = \{a_0, a_1, a_2, ...a_n | u = u_{a_0} = u_{a_1} = u_{a_2} = ... = u_{a_n}\}$$

where a_0 is an activity record and u_{a_0} is the user generating activity a_0. The data warehouse also contains traffic volume counts at specific points on the road network. Traffic counts are usually taken for a defined period of time where pressure tubes are placed on certain links to count the number of times vehicles pass across them. Furthermore, information about the geometry of the road network is housed within the data warehouse as a spatial database. The road network spatial database contains information about the geometry of roads such as number of lanes, category, length and speed limit.

2.4.2 Modules and Algorithms

The Modules and Algorithms component is composed of four components: spatial-temporal decomposition module, home/work capturing module, flow estimation module, and community detection module. Each of the components is described below.

Spatial-Temporal Decomposition Module

The first step toward understanding the dynamics of a city on the day scale is to look at the dynamics of population densities across the city through aggregate user activities for each cell tower. This module breaks down the total activities of users on both the spatial and temporal dimensions. A similar approach was developed in [11]. For each cell tower within the city, the module generates a time series data for activity levels for a specified time granularity Δt. To capture the collective behavior of the population across the city, the module captures the aggregate activity level of users

at every cell tower c_i within the city. The aggregate phone activity level denoted $AL(c_i, \Delta t)$ at cell tower c_i for a time window Δt is computed as follows from the dataset.

$$AL(c_i, \Delta t) = \sum_{c \in c_i, t \in \Delta t} a(c, t, u)$$

Where $a(c, t, u)$ is an activity generated through cell tower c at time t. Each time series data for every location c_i gives insights on the nature of the zone where the cell tower resides in terms of its use. For example, work areas within cities are expected to have a higher density of activity during work hours compared to residential areas. The module also provides insight into collective population behavioral characteristics showing when the city becomes alive in the morning. It also captures information on how users are interacting with events in terms of localization or behavior of service usage. The objective of developing this module is to provide a holistic overview of the change in population densities across space and time.

Capturing Home/Work Places Module

Expanding the time interval of the analysis, this module captures work zones as well as residentail zones. This is essentially capturing places where the majority of daytime calls are as a proxy to work locations. First, we segregate activity records on two time windows to capture most visited zones at daytime versus nighttime for a particular user u. Activities that would hold potential work locations are separated in a set as:

$$\begin{aligned} day_u = &\{a_0, a_1, a_2, ...a_n | u = u_{a_0} = u_{a_1} = u_{a_2} = ... = u_{a_n} \\ &\wedge t_{a_i} \in daytime\} \end{aligned}$$

Where a_0 is an activity record, u_{a_0} is the user generating activity a_0 and t_{a_i} is time tag of activity a_i. Similarly, $night_u$ is obtained with the same logic for nighttime activity. Then, $work_u$ location for user u is chosen to be the most occuring location in day_u and the same applies to $home_u$ as it is chosen to be the most occuring location in

$night_u$.

After determining the $work_u$ and $home_u$ for each user. The aggregation of the resulting zones where users spend most of their times during the day and night identifies dense zones that pertain to business/residential areas since the module considers larger time granularity for the analysis. Thus, this module quantifies the extent to which a zone is considered as residentail/business zone.

Community Detection Module

Following on the output of section 2.4.2, this module will investigate whether there are groups within the population forming communities that have similar home and work locations. The module begins with the city-wide network of connected zones $G(N, E)$ where N is the set of cell towers within the city representing the zones and E is composed of weighted directed edges defined as the number of users who have a particular home/work pair, respectively, in the zones corresponding to the starting and terminating nodes. The adjacency matrix A of the discussed network is as follows:

$$A = \begin{pmatrix} w_{0,0} & w_{0,1} & \cdots \\ w_{1,0} & w_{1,1} & \cdots \\ \vdots & \vdots & \cdots \\ w_{m,1} & w_{m,2} & \ddots \end{pmatrix}$$

Where $w_{0,1}$ is the number of users having their $home_u$ as c_0 and $work_u$ as c_1. The algorithm then uses a modularity optimization scheme, such that sets of nodes are clustered in a way that minimizes internal arc disruption [13, 43]. Each resulting community represents an area where a large fraction of users are mostly located during the day and night.

Modularity is a standard objective function in the field of community detection; it measures how well a partition of network nodes into communities reflects the characteristics of the underlying network (in our case the commuting flow among zones). The rationale behind modularity is that a group of nodes with connections mostly directed towards its own members represent a community with higher modularity

while a set of nodes with intra-community connections is what we would expect by randomly rewiring all the links.

Communities resulting from modularity optimization of telecommunication data have been empirically shown to be representative of the actual social and administrative boundaries at the level of whole countries [17].

In the case of a city, we went further and studied communities at the level of the neighborhood. The interesting results we obtained are discussed in Section 2.6.

Flows Estimation Module

To capture the directionality and mobility of the population across the city, the browser houses an algorithm that provides information about the collective behavior of human mobility through mining mobile phone activity. The module of estimating the aggregate flows of people across the city from the CDRs is a three step algorithm that has the CDRs as inputs and the aggregation of flows of people between locations at every time window Δt as its result (i.e. Origin Destination matrix). A similar approach was developed in [11]. The module starts by arranging data on a user level and considering each of their displacements as a potential trip. After that, the resulting potential trips go through a filtration process that filters out noise in the data from the potential trips generated. Finally the last step aggregates the resulting trips on both the spatial and temporal dimensions to generate an origin-destination matrix based on the provided time slice of interest.

The first step in the algorithm looks at phone activities on a user level and gathers all activities generated for each user sorted in time as follows.

$$\begin{aligned} A_u = &\{a_0, a_1, a_2, ...a_n | u = u_{a_0} = u_{a_1} = u_{a_2} = ... = u_{a_n} \\ &\wedge t_{a_0} < t_{a_1} < t_{a_2}....t_{a_n}\} \end{aligned}$$

Where A_u is the set of all activities generated by user u, u_{a_i} is the user generating the activity a_i and t_{a_i} is the time tag of activity a_i. Every consecutive records belonging to the same user are merged into pairs of location records with their associated

times representing a potential displacement of the user. The set of displacements of a user are represented as given by:

$$D_u = \{(c_{a_i}, c_{a_{i+1}}, t_{a_i}, t_{a_{i+2}}) \mid a_0, a_1, \in A_u\}$$

Where D_u is the set of all potential displacements of user u, c_{a_i} is the cell tower facilitating the activity a_i, t_{a_i} is the time tag of activity a_i and u_{a_i} is the user generating activity a_i. The set of potential displacement considers each successive user activity a potential trip though this includes noisy data such as users who did not change their locations between the successive activities but where nevertheless served by different nearby cell towers, a phenomena referred to as localization error. In order to capture user trips in which a displacement actually occurs, we apply further filtering on the set of potential displacements D_u. The goal of the filtering process is to eliminate all captured pairs of location records that are considered as noise in terms of trip-capturing. The filtration process eliminates all records that are considered as localization error, have very long time intervals or no movement detected. Entries in the data that corresponds to localization error are filtered out by eliminating all trips that are less than a specified distance of the maximum distance between any neighboring cell towers within an urban setting. Given any two neighboring cell towers that c_{a_i} and c_{a_j} , each element within D_u must satisfy the below predicate.

$$distance(c_{a_i}, c_{a_{i+1}}) > max[distance(c_{a_i}, c_{a_j})]$$

Where $disance(c_{a_i}, c_{a_{i+1}})$ is the distance between the towers c_{a_i} and $c_{a_{i+1}}$. The filter eliminates potential displacements having a distance larger than that of the maximum distance between any two neighboring towers in the city. In addition, each pair of records satisfy $t_{a_{i+1}} - t_{a_i} > \alpha$, where t_{a_i} is the time tag of activity a_i. That is a time difference between consecutive activity records being more than a threshold is filtered out of the set of displacements D_u for the purpose of reducing the uncertainty in capturing the actual departure and arrival times for trips.

The result of the filtering process is the set of displacements $\bar{D}_u$ containing all

pairs of locations where movement was detected and reasonable time duration for the trip was captured. After that, the final step towards the generation of OD matrices is to aggregate the trips according to the specified time slice into the origin destination matrix given by:

$$OD(\Delta t) = \begin{pmatrix} 0 & T_{0,1} & T_{0,2} & \cdots \\ T_{1,0} & 0 & T_{1,2} & \cdots \\ T_{2,0} & T_{2,1} & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix}$$

Where each element $T_{i,j}$ gives the number of trips captured between c_i to c_j during the time slice Δt. The value of $T_{i,j}$ is computed by:

$$T_{i,j}(\Delta t) = \sum \bar{D}_u(c_{a_n}, c_{a_{n+1}}, t_{a_n}, t_{a_{n+1}})$$

Where $c_{a_n} \in i$, $c_{a_{n+1}} \in j$ and $t_{a_{n+1}} - t_{a_n} \in \Delta t$. Thus, $T_{i,j}$ quantifies the flows from zone i to zone j during the time window Δt.

2.5 Visualization Interface

The visualization component shows the results of the modules and algorithms on two time scales depending on the nature of their outputs. It will visualize population density distribution and major flows of people across the city dynamically over the span of a day while on longer time scales it will show a static map of the communities forming around the analysis of dense zones.

The visualization will start by showing the spatial-temporal decomposition of the population over the scale of a day. A dynamic visualization with time granularity of 15-minutes will capture population density variations across the day and night. The browser shows mobile activity over a dynamic period of time broken up into 15-minute intervals as shown in figure 3. This visualization presents a rotatable, scalable map onto which a shifting, three-dimensional grid is superimposed to show locational agglomerations of cellphone activity. Grid sectors will rise and fall, and brighten and

fade as people move across the city using their mobile devices.

On the same scale of a day, the visualization components shows the directionality of human mobility through the output of flow estimation module as well as the car counts stored in the data set. Major flows within the city showing the aggregate behavior of commuting around the day are visualized with a time window of 15-minutes. The component visualizes the generation of trips on each time slice by as an arc that rose from originating to terminating cell tower. As shown in figure 6, each arc embodies a variable number of trips, and to illustrate this we altered its thickness and height in correspondence to the intensity of activity along that route (on a logarithmic scale). The arcs are drawn over the same city base geography, on top of the social interaction mesh from above, in an effort to reveal unseen connections between the two results. In addition, car counts were built into the visualization as half-spheres placed at their respective intersections. Each sphere changes shape and color at an hourly rhythm in line with the measured volume.

On the longer time scale and towards visualizing the output of the community detection module, the visualization interface overlays the community network over the spatial dimension of the city to show if there are correlations between the formation of communities and the urban fabric of the city. Nodes represent zones of the city and arcs represent groups of people spending most of their times across the day/night between connected nodes. The community detection module provides the set of nodes that belong to the same community. To visualize the output of the community detection algorithm, nodes belonging to the same community are colored with the same color as shown in figure 5. Thus, areas where sub communities spend most of their time during the day and night are bounded within zones of the same color.

2.6 Case Study

Over the past decade, Saudi Arabia has taken strong steps towards developing a diversified economy. Specifically on enhancing its Information and Communication

Technology (ICT) infrastructure [4]. Today, Saudi Arabia has one of the highest Internet penetration percentages in the gulf area with current penetration at 14.7 million. It is ranked among the highest countries worldwide in mobile penetration rates with 188% of the population possess mobile phones [14]. The high penetration rate of mobile in Saudi Arabia make it an ideal candidate for utilizing the Call Detail Records (CDRs) as in situ sensors for human mobility.

The City Browser was implemented for the Urban Transportation System (UTS), a system developed to provide city planners with insights with regards to the mobility of the population. The project started with gathering information related to the structure of the city as well as the dynamics of the population. The data gathered includes Records CDRs spanning a period of the month of December, a spatial database of the road network of Riyadh city and traffic counts data on different points within the city. Currently, the data is housed within the data warehouse where several modules and algorithms are using it to generate insights on the dynamics of the city.

2.6.1 City Spatial-temporal Decomposition

The first step towards understanding the data in the city of Riyadh is to decompose cellular activity on the spatial and temporal dimensions. The visualization in figure 2-2 shows cellular activity through color, transparency, and height (in logarithmic scale) gridded across the metropolitan expanse of Riyadh. As opposed to seeing the cell towers as discrete points in the city, we show network traffic interpolated over a 100 by 100 grid. In this sense, each grid cell is assigned an intensity based on its distance to surrounding antennas and their activity levels using a Gaussian smoothing function. The temporal activity is interpolated in a similar manner, showing smooth transitions between each time-slice in the dataset.The city's downtown core quickly becomes clouded in smog of network activity early in the morning that hangs over region for the entire day. Clear sub centers emerge that follow construction density, and these sub-centers appear to be partitioned by the roadway network itself.

The city's shifting activity profile also highlights a rich temporal signature of

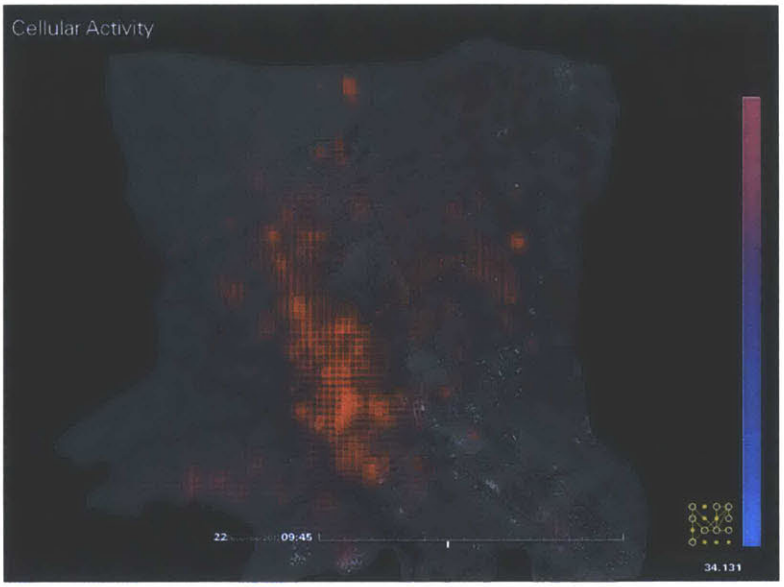

Figure 2-2: Spatial-Temporal Decomposition out for a single time slice. The figure demonstrates the time-cumulative spatial mobile activity conducted between 9:45am to 10:00am.

communication that is all Riyadh's own. Watching the oscillations of the activity landscape, we see that Riyadh comes alive at around 6:15am. We also see strong regional delineation: the residential neighborhoods to the southwest and northeast of the downtown core come alive well before the rest of the city, and experience the strongest inter-hour fluctuations throughout the course of the day. Finally, we see some peculiar discontinuities in aggregate talk throughout the day almost as if all phone traffic was suddenly halved at strange intervals.

2.6.2 Capturing Home/Work Places

A fundamental quality of mobility behavior is to analyze the emergence of zones with higher densities along a wider time granularity to understand the distribution of residential and business zones. Expanding our time intervals to capture broader day and night variation we can begin to differentiate dense business areas and schools versus dense residential neighborhoods.

The map in Figure 2-3 highlights the discrepancy between the purely day zones shifting towards the red color and the purely night dense zones shifting towards blue

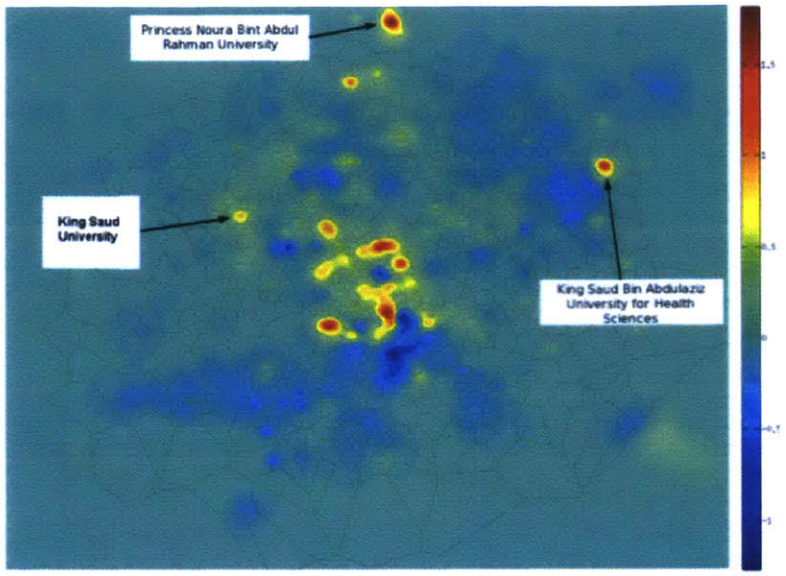

Figure 2-3: Dense work zones during the day versus home locations during the night. We observe high day-densities at the periphery where major universities are located.

color, showing some mono-centrically clustered day hotspots that follow the overall spatial logic of the city. At the periphery we also see a number of universities show up strongly as day locations. Lastly, we see high agglomerations of residences to both the south and east of the city, with smaller pockets scattered throughout.

2.6.3 Detecting Mobility Communities

The work/home dense zones visualizations shown in section 2.6.2 point to an organizational logic of the city. Conceptualizing the totality of day/night commutes as a city-wide mobility network, we can conceivably break this network into sub-communities by applying a regional delineation algorithm.

By overlaying the results of the community detection module on geography of the city (see Figure 2-4), a number of interesting relationships are revealed between the detected communities and the built form of the city. Most strikingly, the resulting clusters closely correlate to the main arterials of city's roadway infrastructure. Mobility communities seem to be partitioned by the street network itself, underscoring the city's dependence on highway infrastructure, while also supporting the commonly held belief that heavily trafficked streets, on many levels, are instruments of segregation

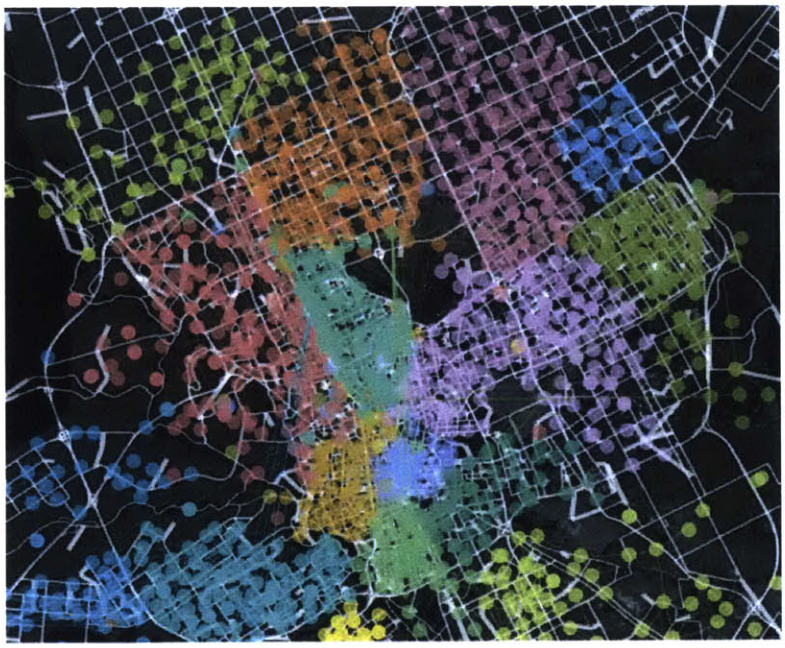

Figure 2-4: Community Detection Module results plotted by Latitude and Longitude on the map of Riyadh. We find support to the commonly held belief that heavily trafficked streets, on many levels, are instruments of segregation and control.

and control, or, perhaps more optimistically: good streets make good neighbors.

2.6.4 Flow Estimation

The approach toward understanding flows that contribute dense-zone emergence on smaller time granularity unveils rich information pertaining to the sources of dense zones as well as the distribution of flow over time. By collecting and filtering each user's mobile activity as sequence of cell tower locations and then aggregating collective users' trips, we are able to estimate flows in terms of origins and destinations of trips. We've observed that these estimated flows contributed to the emergence of high density zones in the city of Riyadh; however this approach includes the added benefit of capturing travel demand at highly dynamic time slices ranging from seasonal variations to hourly fluctuations. Such a high temporal resolution has the potential to transform our understanding of urban mobility[56].

The resulting dynamic maps held a striking similarity to the local intuition of vehicular flows across the city (see Figure 2-5). Overall flows correspond quite closely

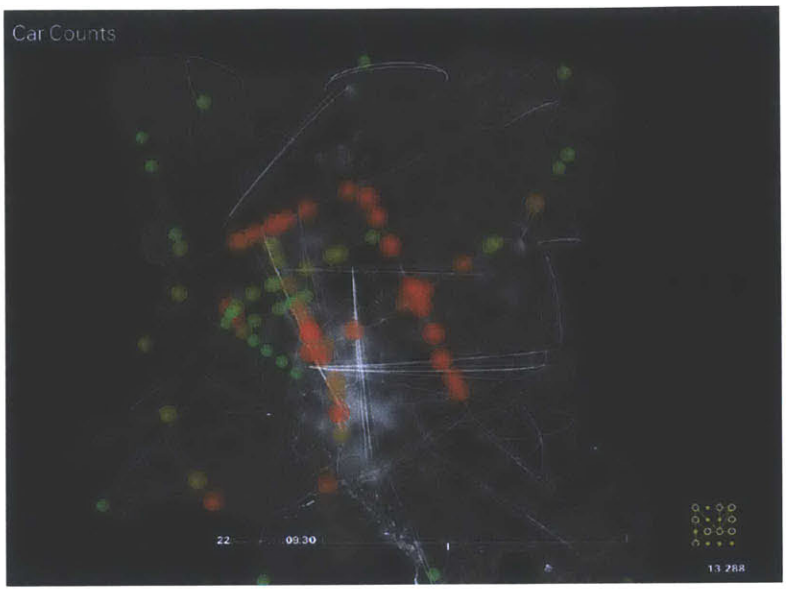

Figure 2-5: The extracted Origin Destination (OD) matrix across Riyadh at the time slice of 9:30-9:45am. The height of the line corresponds to the number of trips between a specific OD.

to the underlying street network. Most notably, Figure 2-5 shows intense activity along the city's main arterials; King Fahd Road and the Northern and Eastern Ring roads. This agrees with the local community's subjective understanding of commute patterns across the city. But to further validate our results, we compared them against the best ground-truth measurements of roadway activity: car count volumes captured by pressure-tube sensors placed at multiple intersection across the city.

2.7 Discusssion

In this chapter, we have presented a tool addressing the complexity of city human mobility and showed its application to the city of Riyadh the capital of Saudi Arabia. The browser is built to work with historical data and thus would provide an after-the-fact analysis and does not allow for the parsing and analysis of the data in real time. A potential future work would be investigating the possibility of enabling the browser to parse such big data in real time through establishing a live connection of data feed with GSM network operators.The city mobility browser synthesizes and extends existing algorithms to provide a holistic decomposition of the complexity of mobility across

multiple dimensions. Although the browser captures the dynamics of the demand on transportation, it does not map the demand over the road network of the city.The visualizations provided by the tool give a dynamic qualitative understanding of the spatial attributes of the city as well as its population directionality across different times of the day. Future work could also enable the visualization interface to provide quantitative analysis and a better understanding of emerging patterns.

Chapter 3

Urban Attractors: Discovering Patterns of Regions Inflows in Cities

3.1 Introduction

Understanding how different places in the city influence human mobility is significant for urban planning tasks. A challenging one for large, complex and congested cities is maintaining a robust transportation infrastructure. Understanding the patterns by which places in the city attract visitors is essential for planning and modifying the transportation system. Specifically, identifying the major regions that play a major role in road congestions help identify the regions requiring higher accessibility in the planning process. This issue is of a particular significance to the city of Riyadh, Saudi Arabia where the largest metro project is being developed and promised to be running in 2019 [5, 40]. Moreover, understanding how different types of places affects the flow of trips in the city differently is essential in making decisions and policies related to placing and modifying services in the metro system undergoing construction. For instance, where to place an industrial area and how would it influence the flow of trips during different times of the day. Researchers investigated a similar question about where to place new business stores for higher profitability [34].

Today and with the ubiquity and pervasiveness of technology, data generated from mobile phones enabled researchers to better understand the behavior of individuals

across many dimensions [67, 8]. Most of the previous work on categorizing urban areas aimed to classify districts by their functionality and land use (i.e. commercial, educational, ...etc.). Some papers considered the human mobility aspects to classify districts. A recent paper [65] proposed a topic modeling approach to classify districts into functional; zones according to peoples socioeconomic activities mined from taxi and public transport traces and points of interests (POIs) data. Another paper [45] proposed a land use classification approach based on the social functions of districts analyzed from GPS taxi traces where districts witness change of land use class dynamically. While the work in [59] analyzed cell phone data to measure spatiotemporal changes in population and classified land use based on similar cell phone activity patterns. However, classifying urban places based on how attractive they are to visits within cities has not been explored.

In this work, we present a computational framework for classifying urban districts by their attraction patterns. we define attraction profiles in terms of statistical and contextual features of regions incoming trips on a specified temporal window. Different places in the city have different patterns of attracting visitors. Some places like universities and hospitals attract a large amount of visitors who come from all over the city and they travel long distance to visit those types of Attractors. On the other hand, some districts in the city that provide local services such as restaurants, schools, and small clinics only attract few people from nearby areas. We aim to automatically identify pattern of attraction based on three main dimensions: how many visitors a TAZ receives, how spatially spread the origins of the trips traveled to that TAZ, and the shape of the distribution of the distance traveled by all visitors to that TAZ. Additionally, we aim to understand what makes a district have a certain attraction behavior. To do that we used statistical significance testing to automatically relate the decomposition of POI types (services) in a district to its attraction pattern. We know how each type of services is related to attraction behavior, that can be useful for planning and locating services around the city while considering the attraction that type of place is expected to cause.

The contribution to the literature can be summarized in the following points:

- We propose measures of regions attractiveness by quantifying the spatial dispersion of where visitors come from and by using the distribution of distances traveled by visitors on the road network.

- We propose a computational framework for detecting attraction patterns and rigorously relating each POI type to each detected Attractor behavior.

- We provide a technique for predicting the inflows to newly developed regions using a Gaussian Process.

3.2 Related work

Multiple studies used human mobility behavior to classify urban areas. A recent study investigated the relationship between land use and mobility[36]. The authors showed that purposes of people's trips are strongly correlated with the land use of the trip's origin and destination. Recently, the availability of dynamic sources of data allowed for dynamic segmentation of the city according to human mobility behavior. Some studies combined human mobility with land use or POIs data to segment districts in urban areas according to their functions or use. The type of data used to capture human mobility behavior varies between individuals GPS traces [69, 23], taxi pick up/drop off locations as in [38, 45] , Call Detail Records (CDRs) as in [40, 59], social media check ins as in [66, 39, 6] , and bus smart card data as in [29]. However,to our knowledge, non of the previous studies quantified the attraction of places and used attraction profiles to segment the city.

Survey travel data has been used to detect the centers (significant places) of a city [71, 18]. A recent study proposed a method for measuring the centrality of locations that incorporates the number of people attracted to the location and the diversity of activities in which visitors engage [71]. The proposed method was tested on survey travel data in Singapore to identify the functional centers and track their significance over time. A similar approach focused on analyzing the aggregate behavior of the population to predicted highly attractive events such as the times square during

new years count down in new york [24]. Our method is based on validated origin destination matrices mined from cell phone data that captures human mobility. More significantly, our approach incorporate not just the amount of people a place attracts but also on where do they come from and the road distance they traveled.

Network analysis methods were used to detect hotspots based on flow patterns between locations[40, 62]. A recent paper [40] used Origin Destination (ODs) matrices extracted from cell phone data to identify the signature of mobility behavior as 4 main types of movements within the city: from hotspot to hotspot, to hotspots, originating at hotspots and the random flows. They showed how different cities have different mobility signatures. Additionally, a recent study used Taxi drop off/pick up GPS traces in Shanghai to create a network of flow between places. They applied community detection to extract sub regions and analyze the interaction between sub regions and within each sub region.

Researchers adapted modeling approaches from Natural Language Processing (NLP) in identifying functional zones in urban areas [65, 64]. One study applied a Latent Dirichlet Allocation (LDA) model on Foursquare check ins to detect local geographic topics that indicate the potential and intrinsic relations among the locations in accordance with users' trajectories. A recent study used LDA and POIs to detect functional zones[65]. Our work is different where we aim to analyze the attraction behavior of a place using measures that has not been used in the previous work.

3.3 Urban Attractors Framework

In this paper, we propose a framework to categorize districts in the city according to their attraction patterns. Additionally, the framework relates type of services in the city to different attraction behavior. Figure 3-1 shows the general structure of the process of analyzing attraction patterns in cities with the input datasets and the outputs. The first step in the process is to extract trips information from Call Detail Records (CDRs) of cell phones using the validated origin destination extraction algorithm implemented in [58]. We use the ODs as a data source for estimating human

mobility, where it provides the amount of trip from each pair of origin and destination. From the ODs, we extract three statistical features that quantify how attractive a place is: the number of trips a place receives, the spatial dispersion of the origins of all incoming trips, and the distance distribution visitors traveled to visit the place on the road network. We use these attraction features to classify all districts in the city according to their attraction behavior. Finally, using a statistical significance testing approach we relate the types of places that are significantly concentrated in each types of attractors identified. In the following sections we explain each process and its output in details.

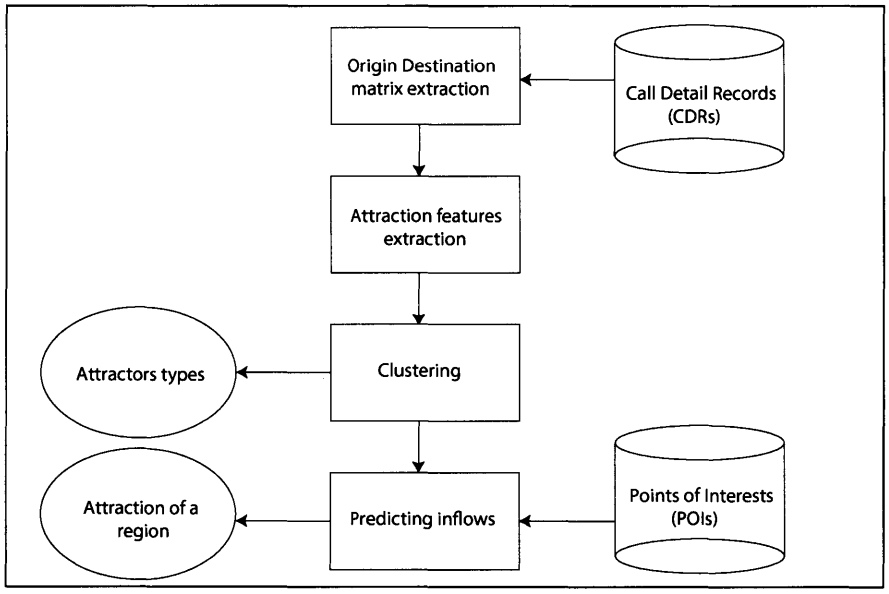

Figure 3-1: Discovering Urban Attractors Models and Data

3.4 Origin Destination Matrix Extraction

The aim of this process is to extract the number of trips between each pair of locations in the city. Our primary source of data is one month (December 2012) of CDRs of anonymous mobile phone users in The city of Riyadh, Saudi Arabia. Within the CDRs, each record contains an anonymized user ID of the caller and receiver, the type of communication (i.e., SMS, MMS, call, data etc), the cell tower ID facilitating

the service, the duration, and a time stamp of the phone activity. Each cell tower ID is spatially mapped to its latitude and longitude where each Voronoi cell in figure 3-2 correspond to a tower. The CDRs provide a proxy for tracking human mobility behavior in the city. Computational steps are needed to extract clean trajectories from the CDRs.

Methods of estimating validated ODs around the day range from very traditional methods more modern ones. Traditional methods include running surveys within cities and estimating the flows between locations of the city from the feedback of those surveys. Such methods consume longer periods of time rather than being inaccurate at times. In addition, traditional surveying methods usually span smaller population sample sizes and thus are more prone to biases. Recent research in the domain of ubiquitous computing provided alternative methodologies for estimating ODs faster and more accurate. The methods proposed in [58, 57] uses mobile phone location traces (i.e. CDRs) to estimate the flows of people between areas in the city. The large amounts of phone data provide more sample sizes and more accurate information compared to traditional methods. I use state of the art methods of extracting OD matrices for the city of Riyadh between each pair of TAZes as shown in figure 3-2.

The output of this process is the OD matrix that provides information of the number of individuals traveling from location i to j at the cell value T_{ij}. The spatial scale we used is based on Traffic Analysis Zones (TAZes), which is the official segmentation used in transportation planning. Conventionally segmenting the city into TAZes are based on census block information such as population, where zones tend to be smaller in denser areas and larger in areas of low density.

3.5 Attraction Features

We aim to quantify how attractive a place is through statistical features of the inflow for that place. The first feature is the total amount inflow a place receives. As the more visitors a place receives, the more attractive that place is. Additionally, a place is more attractive if it attracts people from various places in the city. Some

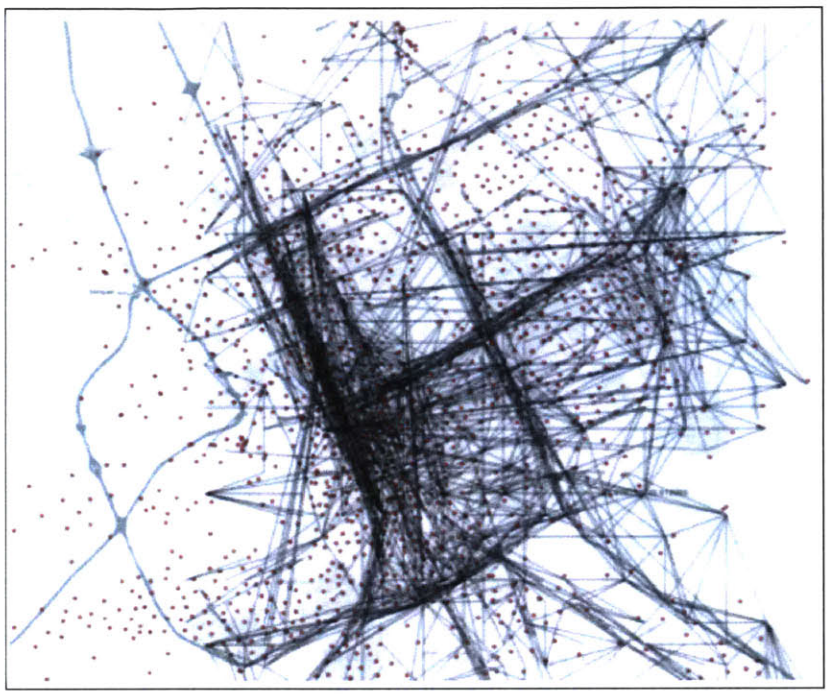

Figure 3-2: The ODs in Riaydh during the morning period. Each line represents a trip from a source to a destination.

places only attract people nearby which makes them local in terms of from where they attract people. On the other hand, some place attracts people from all over the city such as universities and hospitals. Thus, we quantify how spatially dispersed the original location of visitors using a spatial standard dispersion measure. Moreover, we quantify the distance by which visitors traveled to visit the place on the road network. If visitors are willing to travel long distances to visit a place, it makes that place more attractive. In the following sections, we explain each characteristic we used to describe the attraction behaviors of destinations.

Inflow magnitude

The amount of visitors a place receives is a strong indicator of how attractive the place is. This feature measures the attraction force of a location where locations that have high inflow (number of visitors) are major attractors in the city. Figure 3-3 shows the distribution of the number of TAZes according to their inflow amount. The majority of TAZes have small to moderate inflow. However, there are few TAZes

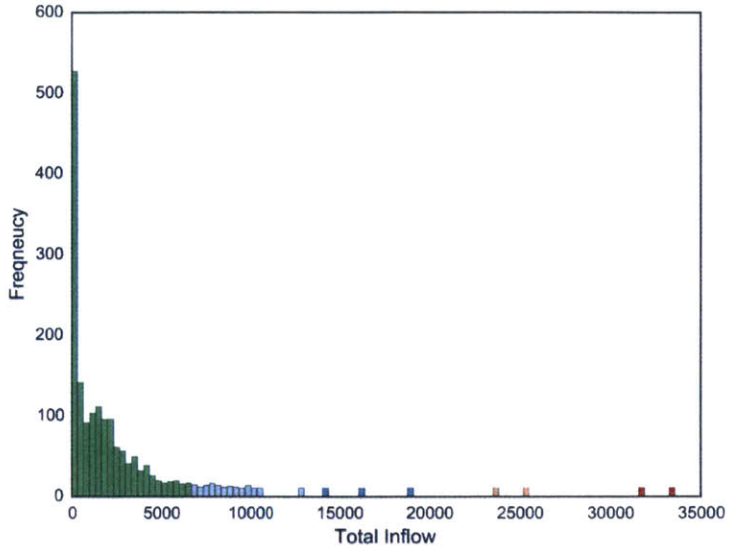

Figure 3-3: The distribution of total inflow received by TAZ's in Riaydh. the majority of places receive small to medium inflow amount (number of visitors). Few places receive very high inflow (colored red),which makes them highly attractive.

that have a very large inflow (colored red), which makes them extreme outliers. The inflow magnitude of a location i is calculated from the OD matrix as follows:

$$Inflow_i = \sum_{j=1}^{n} T_{ji} \tag{3.1}$$

Spatial dispersion

An important and novel feature to analyze the attraction behavior of a place is to measure the spatial dispersion of where visitors come from. The spatial dispersion quantifies how spread out the locations of the origins of trips are to the center of mass of where visitors come from. A place is more attractive if it attracts visors from various and spread out places in the city. Major attractors tend to attract people from all over the city (large spatial dispersion), while insignificant attractors only attract people nearby (small spatial dispersion).

We measure the spatial dispersion of visitors by calculating the weighted standard distance deviation, which is a standard method used to measure the statistical

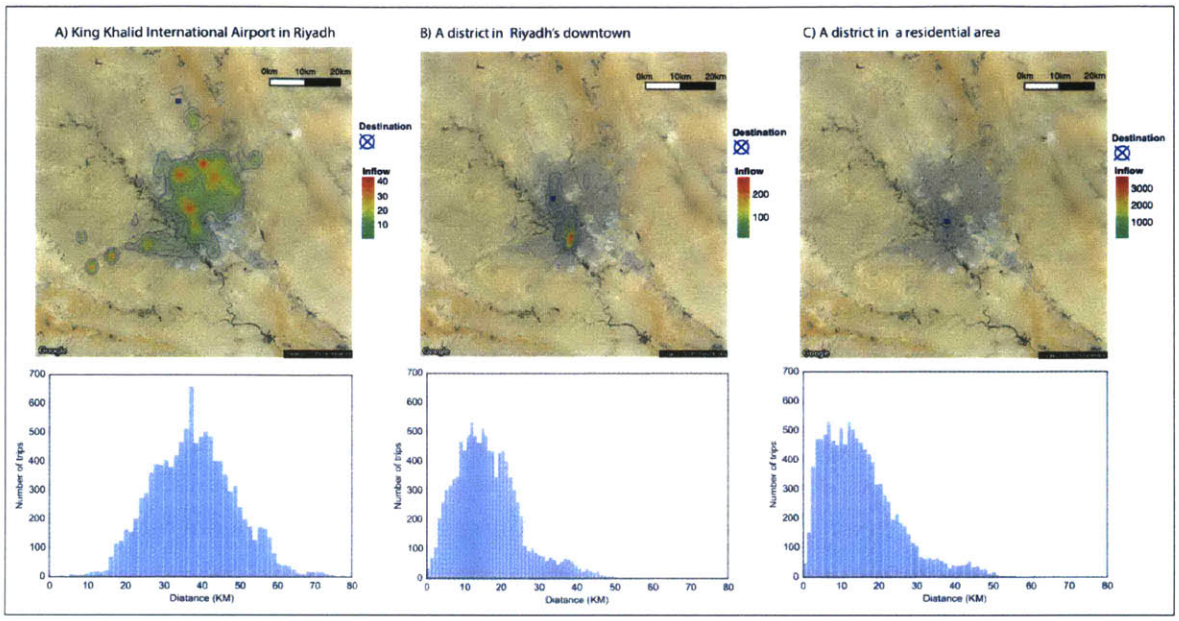

Figure 3-4: Spatial dispersion and the corresponding distance distributions of three different examples of attractors. Example A is the international airport in Riaydh. Example B is a place in the downtown area. Example C is a place in a residential area. The top row shows the heatmaps of the origins of the inflow, where the heat color corresponds to the amount of trips that orientated from that place. The bottom row is the corresponding distance distribution of each example. The distance distribution represents the distances traveled by all visitors, where each type of place has different distribution signal.

dispersion of spatial data [42]. Mathematically, the weighted spatial dispersion (SD) for a TAZ i is defined as follows:

$$SD_i = \frac{\sqrt{\sum_{i=1}^{n} w_i(X_i - X_c)^2 + \sum_{i=1}^{n} w_i(Y_i - Y_c)^2}}{\sum_{i=1}^{n} w_i} \tag{3.2}$$

Where n is the number of source TAZes from where trips originated. X_i and Y_i are the spatial coordinates of the origin of a trip i. w_i is the amount of inflow from source TAZ i. X_c,Y_c are the coordinates of the spatial center of mass of all origins of all the incoming flow calculated as follows:

$$X_c = \frac{\sum_{i=1}^{n} w_i.X_i}{\sum_{i=1}^{n} w_i}, Y_c = \frac{\sum_{i=1}^{n} w_i.Y_i}{\sum_{i=1}^{n} w_i} \tag{3.3}$$

Figure 3-4 shows three examples of places with different attraction behavior. The

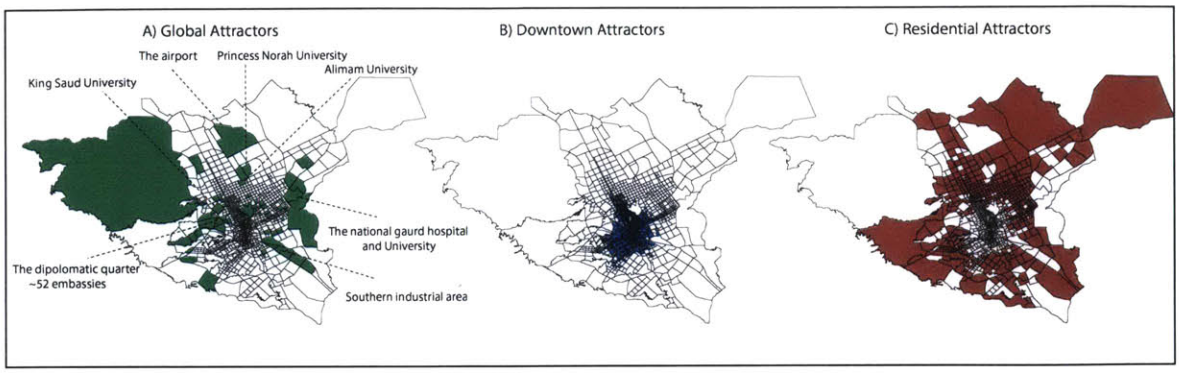

Figure 3-5: The three types of attractors detected

top row shows the heat maps of the inflow sources and their concentration. The destination TAZ is labeled with a target sign on the maps. Example A shows the heat map of the international airport in Riyadh city, where the heat is spread all over the city, which indicates strong attraction. Example B is a TAZ in the downtown area, where the inflow sources are moderately spread. Example C is in a TAZ in a residential area, where it only attract visitors nearby with small spatial dispersion.

Distance distribution

One main property to describe the attraction pattern is the distribution of distances traveled by all the trips that attractor received. The trip distance from each source to the centroid of the attractor were calculated on the road network of Riyadh by using the Dijkstra shortest path algorithm [19] to find the optimal routes between all of the origin-destination pairs. This provides a more accurate estimation than the Euclidean or Manhattan distance metrics as it accounts for the variation in the geometry of the roads.

The bottom row in Figure 3-4 shows the distance distribution of all trips a TAZ received. In example A the distance distribution for the airport is very unique. The mean distance is very high (around 40 KM) and there is a shift in the distribution due to the distant location of the airport in the very far north of the city. On the other hand, the distance distributions are different in example B and C. In B the mean is moderate and there is a tail to the distribution that corresponds to the long distance

traveled by visitors to visit that downtown place. For example C, we notice that the largest amount of trips traveled shorter distances on average. We conclude that for each attraction behavior the distance distribution signal differ. Thus, we capture the mean and the standard deviation of the distribution which are the most critical features to describe the signal of the distribution to distinguish attraction behaviors.

3.6 Clustering

To discover common patterns of inflow within cities, regions are clustered using the attraction features of their incoming flows as discussed in the previous section. We used a Hierarchical Agglomerative Clustering (HAC) approach to categorize TAZes based on their attraction features. HAC classifies objects, where each object is represented as a vector of features that describe that object, based on specified similarity metrics. Here, a vector x_i represent the attraction features that describe TAZ i as follows:

$$x_i = [inflow_i, SD_i, \mu_i, \sigma_i] \tag{3.4}$$

Where $inflow_i$ is the inflow magnitude, SD_i is the spatial dispersion of the inflow sources, μ_i is the mean of the distances traveled to TAZ i, and σ_i is the standard divination of the traveled distance distribution.

HAC starts by assigning each single object to a separate cluster, and sequentially merge the most similar clusters until it results in one cluster. Thus, HAC requires defining how to merge clusters and how to measure the distance between them.For merging clusters, We used complete-linkage algorithm, which merges two clusters based on their most dissimilar objects a follows:

$$D(X,Y) = \max_{x \in X, y \in Y} d(x,y) \tag{3.5}$$

Where $d(x,y)$ is the distance between two objects $x \in X$ and $y \in Y$, and X and Y are the 2 sets of clusters. Complete-linkage algorithm is conservative when merging clusters, thus it tends to find very compact clusters, which fits our objective in

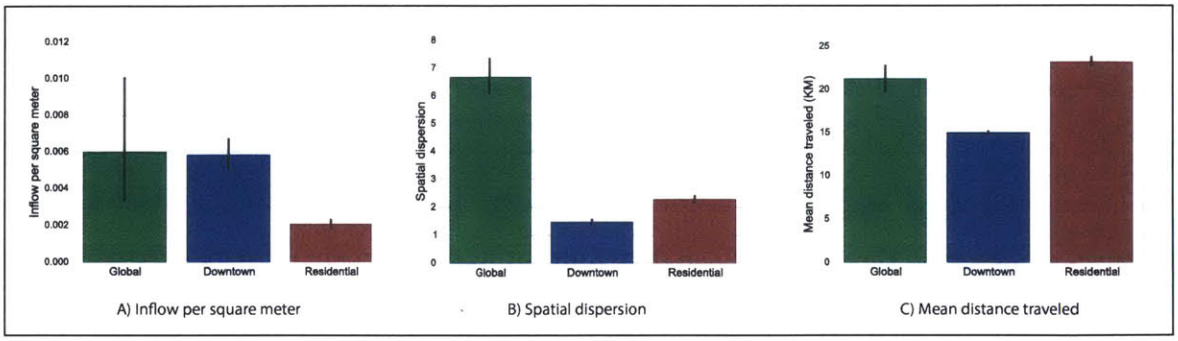

Figure 3-6: The attraction features of the three clusters

in finding closely related attraction patterns. For measuring the distance between clusters' objects $d(x, y)$, we use correlation distance metric defined as follows:

$$d(x,y) = 1 - \frac{(x-\overline{x}).(y-\overline{y})}{\|(x-\overline{x})\|_2 \|(y-\overline{y})\|_2} \tag{3.6}$$

Where $\overline{x}$ and $\overline{y}$ are the mean of the elements of vector x and y correspondingly, and $(x-\overline{x}).(y-\overline{y})$ is the dot product of the vectors $(x-\overline{x})$ and $(y-\overline{y})$.

Correlation distance works well for finding unbalanced clusters sizes as we expect to have small number of places behaving very uniquely as strong attractors and larger number of places that are not as attractive. Additionally, the correlation score can correct for any scaling within a feature, while the final score is still being tabulated. Thus, different features that use different scales can still be used.

HAC provides a hierarchy structure of the classified regions in the city. To determine the number of clusters k that best divide the data, we calculate the total within-cluster variation for each possible k from 1 to 20. The varience ratio drops as k increase until it doest decrease significantly. We select the k that correspond to the point where the variance stops decreasing significantly. The method is known as the elbow curve method. The classification process all TAZes in the city of Riaydh produce three types of attractors that have distinct features. The following section extends the findings on the behavior of region clusters in the city.

3.6.1 Attractor Types

Districts in Riyadh city are classified into three main types of attractors based on distinguishable attraction behavior. Figure 3-5 shows the 3 types of attractor classes detected, where the colored polygons are the TAZes that belong to the labeled type of attractors in Riaydh. The global attractors are the ones that have significant influence on the whole city, hence the name. Unlike the remaining clusters, the locations of these places seem to be random around the city. The second detected type is the downtown attractors, which play a significant influence ,after the global, on human mobility. They are mostly clustered in the downtown area of Riyadh. Lastly, the residential attractors, are the least influential attractors in the morning period of typical weekdays. They are mostly located on the outer places of Riyadh city. In the following sections we describe each type of attractors and their behavior.

Global Attractors

The most significant type is the global attractors, colored green in Figure 3-5. These types of attractors are not common and show outliers behavior in terms of attraction characteristics. The most distinguishable feature of global attractors is the extreme spatial dispersion of the incoming flows, as visitors come from all over the city to visit these places, as shown in Figure 3-6 B. Additionally, the amount of visitors they attract range from high to extremely high as shown in A. We use inflow per squared meter due to the unbalanced sizes of the TAZes. Additionally, the mean distances traveled by visitors to these locations is extremely high, which makes these places highly attractive and unique. We call these places global attractors because they strongly influence human mobility over the whole city. Global attractors always offer some unique *services* that makes them distinguished from other regions where visitors only find such services in those regions. Significant places in the city like the airport, major universities, and hospitals ,that occupy whole TAZes by their own and are easy to identify from the map.Figure 3-5 A shows annotation of these major places in the city of Riyadh.

Downtown Attractors

The second type of attractors is the downtown attractors, colored blue in Figure 3-5.It contains places that are mostly clustered in the center of Riyadh city. These are TAZs that have relatively high inflow. However, because of their central location in the city, visitors from allover the city have short routs to access these places. They have smaller average distances compared to the other types as shown in Figure 3-6. Due to the same reason, the dispersion of the origins of inflows is relatively small. The major feature of these places is that they attract a lot of visitors but because of their ideal location they are very accessible.

Residential Attractors

Residential attractors, colored red in Figure 3-5, are the least significant in terms of attraction power. They attract a very small number of visitors. However, as they are located in the outer sides of the city far from where most of the population resides, the few visitors these places attract travel long distance on average to visit them. Also, it is why the dispersion of the small number of visitors is higher than the downtown attractors as shown in Figure 3-6.

3.7 Prediction of Inflows

The goal here is to fit a model that can predict incoming flows into a region in a city to aid the process of city planning. I used the validated estimates of flows discussed earlier in this chapter to develop a probabilistic model that can predict those flows given the flows of other regions of similar attractions. Then, I compare the results of my approach to the gravity model which is a common method to estimate trips in the domain of transportation engineering.

Given people's flow data in the city, I went through various attempts at modeling the problem starting with a Dirichlet-Multinominal model where the Dirichelet distribution has a simplex of the types of places (attractions) and the distribution models the variability of regions in terms of the places that are in them, the multinomial

would then model the flows. The model didn't fit the problem quite well where I then moved to an implementation of a Gaussian process for the problem. My initial attempts suffered accuracy issues and computational challenges that didn't end as I wished. Then I moved on to developing a spatial Gaussian process model that learns the pattern of inflow of those regions that have similar attraction profile and use the model for prediction. The intuition behind the modeling approach is that inflows of people are usually driven by the places in a destination region and similar regions exhibit similar spatial inflow signatures [3, 64]. For example, to model the inflow for a university in a city, we fit a model on the other existing universities in the city and use the model to predict the inflow. Therefore, we utilize our prior knowledge about the visitors of universities where they usually come from similar locations.

3.7.1 Gaussian Process Model (GP)

The Gaussian Process model parametrizes the incoming flows to regions with similar attraction profiles by their geographical coordinates. We define a set k as the set of regions with similar places of interest. In the universities example, the set k represents the set of regions having universities in them. A sample of the data is included in table 1. Each row in the data has the form $[lat_i, lon_i, l_{ik}]$ meaning that we have an inflow of l_{ik} people from the geographical coordinate $[lat_i, lon_i]$ where i is the index of the region that is the source of the flow and k is an index of regions with similar attraction places.

i	lon_i	lat_i	l_{ik}
1	46.5766	24.7173	45
2	46.5194	24.7461	68
3	46.5920	24.7166	6

Table 3.1: sample of inflow data for regions set k from region i

The first step towards this problem is to define a Gaussian process for the input data $\{x_1, x_2\} \subset X$ where x_i represents a row of the parameters discussed above that is $[lat_i, lon_i]$ and l_{ik} is the flow from that point. The Gaussian process is parameterized by a mean function $m(x)$ and a covariance function or kernel $k(x, x')$ where we get a

finite set of functions equal to the number of points we have. The Gaussian process is then given by

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{3.7}$$

Where the mean function and kernel function are given by

$$m(x) = 0, \qquad k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2l}\right)$$

To implement the model, I used the `Matlab GPML toolbox` developed by Carl Edward Rasmussen and Hannes Nickisch [48] which is an implementation of the topics covered in their book [47]. The method allows for various choices of mean functions, covariance functions, likelihood functions and inference methods. For the purpose of this project, I used a Gaussian likelihood and used exact inference for the parameters. Exact inference is computationally feasible in this model as the number of points is ≈ 1500 which is the number of regions in the city of Riyadh shown in Figure 1. My choice of the hyper parameters σ and l was based on experimentation of the output of the model where I found $\sigma = 1, l = 0.01$ to be performing reasonably well.

3.7.2 Results

This section includes the results of the implementation of the model on the city of Riyadh in Saudi Arabia. I will be estimating the flows into King Saud University shown in red in the figure below. To do that, I will develop a GP for spatial inflow to the regions in the set k defined as the regions with universities in them and shown in the green color in the figure below.

Figure 2-a shows the mean predictive after training the GP on the inflows to regions in the set k. The model captures the main sources of inflows to universities. We can see that there is a major inflow to universities from the south-western region of the city which is known to be highly residential. The significant inflows in general significantly overlap with highly residential regions. Figure 2-b shows the variance around the mean predictive of of inflow. The variance is relatively low inside the

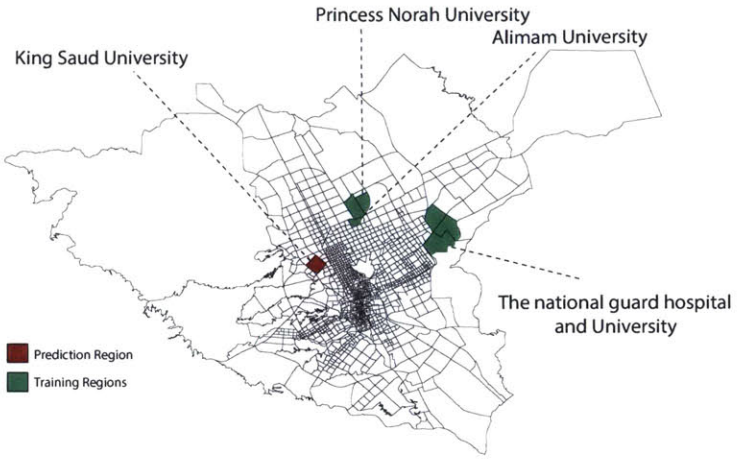

Figure 3-7: Training regions versus prediction region, training regions represent the set k

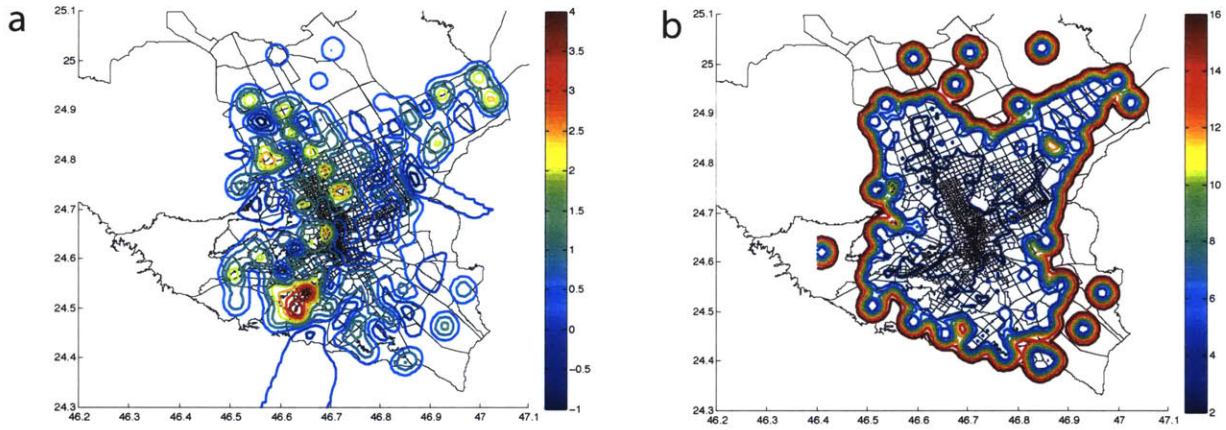

Figure 3-8: The figure shows (a) log of predictive mean and (b) predictive variance for the inflow to regions in the set k

boundaries of the city where we have inflow data and is highest outside the city where no inflow data is available. I will use the value of the mean predictive when predicting inflows to King Saud University.

The metrics I used in quantifying the accuracy of the model are the Root Mean Square Error (RMSE) and the Mean Error (ME) given by

$$\text{RMSE}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(L_{ij} - l_{ij})^2} \quad , \qquad \text{ME}_j = \frac{1}{n}\sum_{i=1}^{n}|L_{ij} - l_{ij}|$$

where L_{ij} is the predicted flow to location j from location i and l_{ij} is the actual flow.

The ME is more interpretable for me to evaluate the model in terms of the average error of flow quantities but I also used RMSE to evaluate how far the predicted flows are from the actual values.

3.7.3 Baseline model

For the purpose of evaluating the performance of the model compared to existing methods, I compare the results of the model to the gravity model as the baseline model used in the domain of transportation engineering [22]. The model is given by

$$L_{ij} = \frac{O_i T_j}{d_{ij}^{\alpha}}$$

where O_i is the total outflow from a location i and T_j is the total inflow into location j and d_{ij} is the street distance between the i and j regions and α is a calibration parameter to be estimated. There are many versions of the gravity model, variations are always in the number of calibration parameters. This is a result of the lack of generalization of the model between cities which is a drawback of gravity models resulting in the disadvantage of overfitting the data. In our example, I chose a moderately complex version where I have one calibration parameter $\alpha = 4.8$ for the city of Riyadh.

Figure 3 shows plots of the predicted flows using the Gaussian process in (a) and using the gravity model in (b) versus the actual inflow values to King Saud University. The figure shows that the predicted inflows using the GP are closer to the $y = x$ line than that of the gravity model. Table 2 shows the performance of the Guassian process compared to the gravity model in terms of ME and RMSE where we find that GP has a ME of 10.5 people and a RMSE of 54.58 compared to that of the gravity model have a ME of 212.05 and a RMSE of 900.52.

The modeled GP estimates inflows more accurately than a gravity model fitted on the city of Riyadh. The initial results found in this report suggests that probabilistic models enabled by the abundance of phone data can sometimes provide better decision tools for urban planners than existing models in the literature of transportation

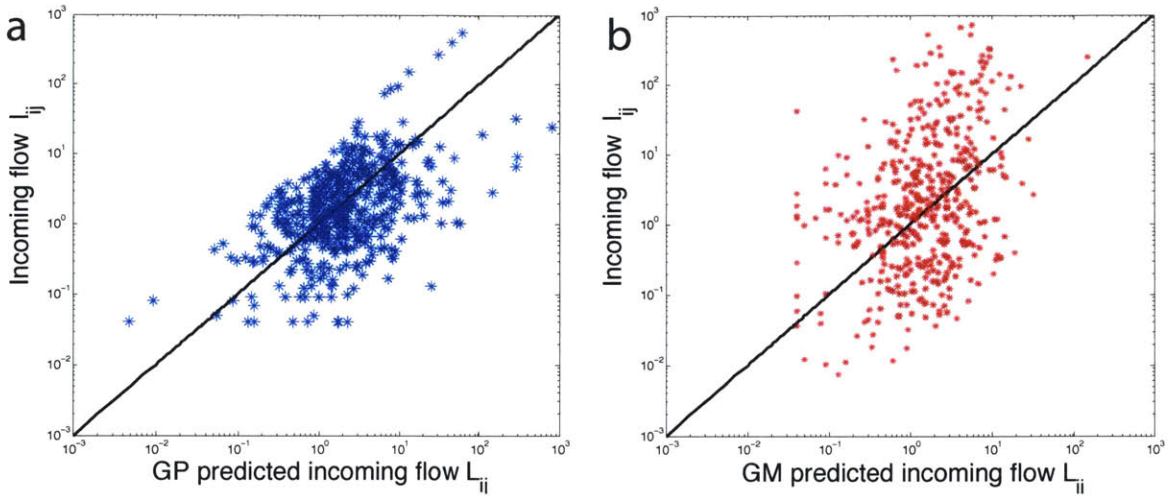

Figure 3-9: Performance of (a) Gaussian process (GP) versus (b) gravity model (GM) for predicting the inflow to King Saud University

method	ME	RMSE
Gravity model	212.05	900.52
Gaussian process	10.5	54.58

Table 3.2: ME and RMSE for GP and gravity model in predicting inflows to King Saudi University

engineering. I chose this project to experiment with the potentials of such models in city planning problems compared to traditional methods. I think it is worthwhile to extend the work to other cities and regions of other functionalities for further investigation of how GPs compare to gravity models.

3.8 Discussion

In this chapter, we presented a computational framework to discover different attraction patterns of districts in cities and to predict the inflow to newly developed regions. We proposed 3 dimensions to define attraction of a district : total number of trips the district receives, the spatial dispersion of the origins of trips, and the shape of the distribution of all distances traveled by visitors to reach that district. Additionally, we presented a method for understanding the relationship between the decomposition of the types of POIs in a spatial zoena and its attraction behavior. We applied the

methodology and discussed the results on the data set from the city of Riyadh the capital of Saudi Arabia.

The chapter shows the results of implementing the discussed modules that mined data from mobile phones to provide a coherent understanding of the dynamics of the interaction between the flows of people to a district and types of services (POIs) that are located in that district. We detected three attraction patterns in the city of Riyadh according to the morning mobility dynamics.

The chapter then proposes a probabilistic modeling approach for predicting the flows between regions in a city using phone calling data. The modeled GP estimates inflows more accurately than a gravity model fitted on the city of Riyadh. The initial results found in this chapter suggests that probabilistic models enabled by the abundance of phone data could provide better decision tools for urban planners than existing models in the literature of transportation engineering. Such predictive models can help decision makers estimate the inflow to a location prior to making decisions on the functionality of a region or the residential population capacity.

Chapter 4

City Scale Next Place Prediction from Sparse Data through Similar Strangers

4.1 Introduction

Research towards predicting next locations of people have shown to be very successful on data with varying resolution in space and time [41, 52, 20]. Targeting the locations of the social contacts showed to improve the prediction of the leisure locations of an ego when using GPS-geo-tagged Twitter or GPS type of data [51, 16, 15]. However, here we show that the same approaches fail greatly with Call Detailed Records (CDRs) from mobile phone data due to the sparsity of data on the temporal dimension significantly reducing the amount of observed mobility of an individual.

We propose a new model that tackles the problem of sparsity in the data in order to improve the accuracy of next location prediction in highly sparse datasets in general, and motivated by improving the usability of CDRs in particular. In sparse phone calling data, despite being massive, we observe that a mutual availability of locations logs from social contacts is very rare since users have few records. While coupling friends records have increased the accuracy in predicting the next location

of GPS-tagged records, here we show that this is not the case with lower resolution datasets like phone activity records. This chapter presents an alternative mechanism of developing a Dynamic Bayesian Network in a way that reduces effects of the sparsity in the data and improves the accuracy of predicting the next location in $5-6\%$ over the baseline case. This framework is targeted for human mobility prediction on very sparse temporal datasets with spatial accuracy of cell towers. The chapter will compare the performance of the proposed model to a Markov Chain (MC) as a baseline where it was suggested as one way of approach the theoretical predictably limit of human movement [41, 25]. I also compare the results of the model to the ones of [51], which incorporates social contacts information and has time as an observed node on the model. Finally, we relate the model's results to the results of a theoretical upper bound of predicting human movement proposed in [55] in the evaluation section. The contributions of this chapter can be summarized in the following points:

- Investigate approaches towards limiting the negative effects of sparsity on next location prediction in CDRs. The chapter propose several human mobility similarity metrics used to identify other users with similar mobility characteristics (i.e. *similar strangers*).

- Model human mobility as a Markovian process and propose a Dynamic Bayesian Network model that incorporates the mobility patterns of similar strangers towards better predicting next locations given the whereabouts of *similar strangers*.

- Provide a case study of the model on sparse mobile phone calling logs on the city scale in the city of Riyadh, Saudi Arabia. The case study shows an improve of 5-6 % in prediction accuracy compared to the baseline model.

4.1.1 Dataset Sparsity

The data is ordered as sequences of locations $l_1, l_2, l_3, l_4, l_5, l_6 ... l_n$ for each user i, where we have the corresponding times for those locations as $t_1, t_2, t_3, t_4, t_5, t_6 ... t_n$. We also denote the length of the sequence as L_{hist} and the length of the unique locations a user

has been to as L_i. For example, a users with a visit sequence of $(100, 200, 100, 200)$ will have $L_{hist} = 4$ and $L_i = 2$. In the context of sparse datasets, individuals will have missing location information for most of the time as discussed in the following section.

Previous studies have shown that human communication patterns are highly heterogeneous [27], with some users using their mobile phone much more frequently than others. Figure 1-2 (a) shows the distribution of the number of records for the population of users where the majority of users have a few number of activity records. Thus, the majority of the users in the data have a small number records introducing sparsity in the data that we will use to learn and predict their mobility patterns. The sparsity extends to the availability of the data given an hour of the day and day of the week for a user. Thus the data suffers from low granularity in terms of the number of records as well as low granularity in location data given time of the day and week. The number of visited locations shown in figure 1-2 (b) where the majority of users are seen in a few number of places and this effect is partially due to the few number of records in addition to the fact that users are mostly seen at home. Figure 1-2 (d) shows the levels of phone calling activity along the day with peaks during the day time and a minimum during the night where there is less data to model movement. Figure 1-2 (c) shows the degree distribution of the social network resulting from reciprocal phone communications showing that the majority of users have fewer contacts.

The aforementioned characteristics of the dataset introduce challenges when using the social circle of an individual in predicting their next visited locations. This is due to that fact that the probability of two users being observed (i.e. have their location traces logged) at the same time for users i and j is $\frac{L_i * L_j}{(hours\ in\ a\ month)^2}$ which is very small given the small empirical values of L_i and L_j shown in figure 1-2. This implies that it is harder to learn from the mobility of such users in ways similar to methods of adding friends' movement [51]. Such methods will be severely limited by the availability of contacts mobility data to learn from. In addition, the dependency of such models on the time of day and day of the week adds more dimensions to the data which is not in the favor of reducing the sparsity of the data.

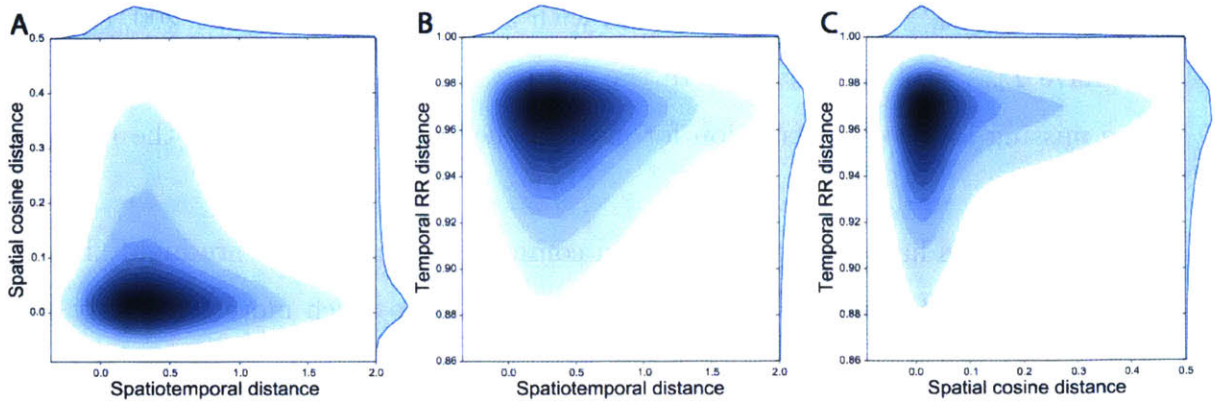

Figure 4-1: Kernel density estimation of the joint distributions of ϕ distance against the RussleRao (RR) distance in (a), ϕ against RR distance in (b) and the Cosine distance against RR in (c). The distances are calculated for every pair of users in the dataset in all three cases.

4.2 Temporal and Spatial Similarity

To overcome the sparsity of the data, we propose to add information of users who are not necessarily within the social circle of the user we are modeling. Our approach is to employ similarity measures on both the temporal and spatial dimensions of the data to find users who will potentially help in predicting the location of a given user. To that end, we couple users who have the maximum temporal, spatial and spatiotemporal similarity scores; Such individuals most similar to the target user are identified as the *similar strangers* in the data. We discuss each of these similarity metrics in the following sections.

4.2.1 Temporal Closeness

To gain more insight about the mobility patterns of an individual with respect to other patterns in the population, we need to have as much information as possible about the mobility patterns of the population surrounding an individual. However, the limitation is that it is infrequent to find location data for multiple users on the same hour. In order to overcome this problem, we quantify how two sequences of location data overlap by a temporal closeness measure. We model the location data as two boolean vectors of length n, where n is the length of the time period in

hours. Each vector cell is 1 if location data for that particular hour is available and 0 otherwise. Then, we employ Russell Rao (RR) distance measure that calculates the distance between the two vectors as follows:

$$RR\,distance = \frac{n - c_{11}}{n} \tag{4.1}$$

where c_{11} is the number of times where we have location information for two sequences and n is the number of hours in the training time interval. This distance measure will be very small (i.e. close to zero) if both users have location data all the time and will have a value of one if they have location data that don't overlap on the time dimension. Thus, by picking the lowest RR between two users, we are maximizing the availability of location information between two users. Figure 4-1 B and C show the distribution of the temporal RR distance between individuals in the CDRs population, the plot suggests a vast majority of very high distances between individuals in terms of mutual data availability which is essentially the limiting factor of observing the location of an individual with respect to the social contacts or any selected set from the population.

4.2.2 Spatial Closeness

This metric measures closeness between individuals' mobility by measuring the similarity of individuals' spatial distributions. For each user, we construct a vector that is of the length of possible locations visited in the city denoted $k \approx 1800$ and each cell in the vector holds the probability of a person being observed in that location. Then we employ the cosine distance measure to quantify the similarity of the locations visited by two users with vectors u and v, as follows:

$$Cosine\,distance = 1 - \frac{u.v}{||u||_2.||v||_2} \tag{4.2}$$

The cosine distance gives an indication of the similarity of places where individuals stay. Figure 4-1 C shows the kernel density estimation of the joint distribution of

the temporal distance against the spatial distance of the visited locations. The figure shows that it is very rare to find users that have high mutual availability of location data while it is still common to find people having similar spatial distribution of their visited places. In general, individuals that have very low spatial scores are individuals that share the same visited locations with relatively similar weights. For example, people living within the same vicinity as well as people working in the same area are expected to have scores that are on the lower scale of the spatial distance. Next, we introduce a measure that combines the temporal and spatial closeness together.

4.2.3 Spatiotemporal Closeness

This metric selects an informative subset of the population by combining the spatial and temporal aspects of location traces through measuring the association of observations between two individuals. Given two users a and b and their corresponding locations logs denoted as l_t^a and l_t^b corresponding to user a and b at time t, respectivly. We construct a contingency table of locations; each cell in the contingency table denoted as $C_{i,j}$ has a count of the number of times users a and b have been in the associated locations given the same hour.

$$C_{i,j} = \sum_{t=1}^{T} \mathbb{1}_{l_t^a=i} \times \mathbb{1}_{l_t^b=j}$$

Given the contingency table, we measure the degree of association between users a and b using a chi-squared test on the table C. Then, we calculate the ϕ distance of association as follows:

$$\phi = 1 - \sqrt{\frac{\chi^2}{n}} \tag{4.3}$$

Where χ^2 is the chi squared test value of the contingency table C and n is the sample of data in the table. This measure will have a lower distance if users a and b move in a synchronous manner and have mutually available data. Hence, the measure combines metrics discussed above. For example, people with synchronous daily home-work trips will have a lower spatiotemporal distance given the data to capture

that is available. Figure 4-1 A and B show the joint distributions of the spatiotemporal distances, the ϕ association measure captures the level of predictability of a user from another utilizing both the spatial and temporal information of their sequences. The model proposed handles human mobility as a Markov process where people transition between locations with associated probabilities called *transition probabilities.* Each individual's transition between locations is intuitively related to how the rest of the population is moving in an urban setting. The model aims at utilizing such information to better predict human movement in the CDRs. This section will describe a baseline model and our improved methodology using a Dynamic Bayesian Network (DBN) incorporating the mobility information of users with similar behavior or *similar strangers* that we quantify using the closeness measures we defined in the previous section.

4.2.4 Dynamic Bayesian Networks

We approach the problem by proposing a DBN with T slices where each slice corresponds an hour of the day. Figure 4-2 shows the schematics of the model. The nodes l_i depend on the last visited location a user was observed as well as the location of the closest sequences depending on the distance measure used. The model has no dependency on the time of day and day of the week to maximize availability of data per hour as will be discussed in the evaluation section.

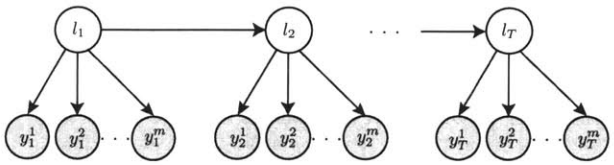

Figure 4-2: Dynamic Bayesian Network where l_i is location of a person at time i and y_i^j represent location of the jth coupled person at time i.

The location sequence $y_1^j, y_2^j ....y_T^j$ correspond to the j^{th} closest individual in distance spatially, temporally or spatiotemporally as discussed earlier. The DBN allows for the coupling of more users as desired.

Learning

The DBN will use the data of the user that we intend to model in addition to the data of the coupled users to model a user. We consider having the observations $l_t \in \{1, 2...k_l\}$ in the set of k_l cell towers visited by an individual at time t. In addition we consider the observations $y_t^j \in \{1, 2...k_j\}$ in the set of k_j cell towers visited by the j^{th} closest user. Then the joint probability distribution of the model is of this form:

$$p(l_{1:T}, y_{1:T}^1, ...y_{1:T}^m) = p(l_{1:T}) \prod_{i=1}^{m} p(y_{1:T}^i | l_{1:T})$$
$$= \prod_{t=2}^{T} p(l_t | l_t - 1) \prod_{t=1}^{T} \prod_{i=1}^{m} p(y_t^i | l_t)$$

Where the model structure is shown in figure 4-2. The parameters of the model correspond to the transition probabilities of an individual captured by $p(l_t|l_t - 1)$ in the model as well as the the observational probabilities of other individuals captured by $p(y_t^i|l_t)$ in the model.

The DBN uses a maximum likelihood estimator to learn the transitions of a person and the observational probabilities of the locations of others. The parameter corresponding to the transition probabilities is a matrix of size $k_l \times k_l$ where each cell corresponds to the transition probability between two places.

$$Tr(a, b) = p(l_t = a | l_{t-1} = b)$$

Where Tr is the transition probability matrix of size $k_l \times k_l$.
In addition, the DBN learns the conditional probabilities of the location of a user given their similar strangers using a maximum likelihood estimator as well. The model will learn m observational matrices each corresponding to a couple of the user and similar stranger i:

$$Ob^i(a, k) = p(y_t^i = a | l_t = k) \; for \; i = 1...m$$

Where Ob^i is a $k_j \times k_l$ matrix.

Inference

After the DBN learns the transitions and observational probabilities, we then can infer the most probable next locations. To get the most probable sequence of locations visited given the locations of the most similar strangers at different times of the day, we use a MAP estimator to find $l^*_{1:t}$ by finding the argmax in the following:

$$l^*_{1:t} = \arg\max_{l_{1:t}} p(l_{1:t}|y^1_{1:t}, ...y^m_{1:t})$$

Where the predicted locations depend on the parameter $Tr(a, b)$ of transition probabilities between locations a to b and the parameters for the observational probabilities Ob^i for each similar stranger.

4.2.5 Markov Chains (Baseline Model)

The model developed by Lu et all shows that a Markov Chain approaches the limits of predictability in human mobility using mobile phone data of a user alone, we will use the model as the baseline model. The Markov Chain is similar to the employed DBN where we consider having the observations $l_t \in \{1, 2...k_l\}$ in the set of k_l cell towers visited by an individual at time t. However, the approach doesn't make use of social contacts or similar strangers mobility patterns. Figure 4-3 shows an example user with three locations and transition probabilities as shown on the edges. The Markov Chain only learns from individuals historical records and thus doesn't couple information of others. The joint probability of the model is given by:

$$p(l_{1:T}) = \prod_{t=2}^{T} p(l_t|l_{t-1})$$

The Markov Chain learns the transition probability matrix of size $k_l \times k_l$ where each cell corresponds to the transition probability between places:

$$p(l_t = a | l_{t-1} = b) = Tr(a, b)$$

Similar to the DBN model, Tr is a transition probability matrix of size $k_l \times k_l$ estimated using a maximum likelihood estimator.

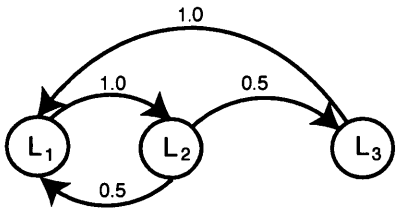

Figure 4-3: Example of a person moving between three locations

Using the learned parameters, we use a MAP estimator to estimate the most probable locations by looking for the sequence that has the highest probability $l^*_{1:t}$.

$$l^*_{1:t} = \arg\max_{l_{1:t}} p(l_{1:t}) = \arg\max_{l_{1:t}} \prod_{i=2}^{t} p(l_t | l_t - 1)$$

Hence, the baseline model only uses transition probabilities to estimate the most probable sequence of locations $l^*_{1:t}$.

4.3 Evaluation

In order to evaluate the proposed methodology, we compare the accuracy with existing methods of coupling with social contacts [51] as well as the Markov Chain baseline model described above. The accuracy of the model is given by:

$$accuracy = \frac{\sum_{t=1}^{T} \mathbb{1}_{l_t = l^*_t}}{T}$$

Where l_t is the true location of a the predicted user at time t and l_t^* is the predicted value. We split the data into three training periods (the first one, two and three weeks of the data) and then testing with the remaining data. Then we predict the location sequence for the remaining time in the period where we testing data that is three, two and one week respectivly. Table 4.1 shows that coupling with social contacts location observations doesn't improve the accuracy in sparse data compared to other data sources [51]. Furthermore, our method of coupling with non-social contacts that are closest spatially, temporally or spatio-temporally improves the accuracy of the DBN. The differences in the change in accuracy between the the distance measures is minimal as shown in the table. Note that we present the results of coupling with a

Algorithm	1w	2w	3w
Markov Chain (baseline)	53.46%	53.48 %	54.12%
Sadilek et all.	52.80%	54.19 %	55.42%
DBN (RR distance)	57.11%	58.66%	59.88%
DBN (Cosine distance)	57.14%	58.61%	59.82%
DBN (ϕ distance)	56.82%	58.63%	60.03%

Table 4.1: Accuracy achieved with different learning periods of first week (1w), first two weeks (2w) and first three weeks (3w). In addition, the table shows how the proposed approach compare to existing methods in the literature.

single closest similar stranger and still get significant improvements compared to the methodology of [51] that couples all of the social contacts of a person. While coupling with the social contacts is intuitive, it does suffers greatly when the data is sparse and doesn't perform well with CDRs.

4.3.1 Accuracy by time of day

The average accuracy of predicting the location of users varies according to the time of day as shown in figure 4-4. Both algorithms have higher accuracies where people are usually at home relative to the remaining hours of the day where the average accuracy drops, indicating a less predictable movement. The coupling of mobility

information of closest person on the spatio-temporal ϕ distance allows for the DBN to increase the predictability during the day times, while not influencing the accuracy significantly during the late nighttime (i.e. 12am to 8am). In addition, the average accuracy of predictions fluctuates in varying degrees around the day; one reason is that phone calling is lower during nighttime compared to the rest of the day. As also shown in Fig. 4-4, the scarcity in the data during the nighttime causes the average prediction to have the larger confidence intervals.

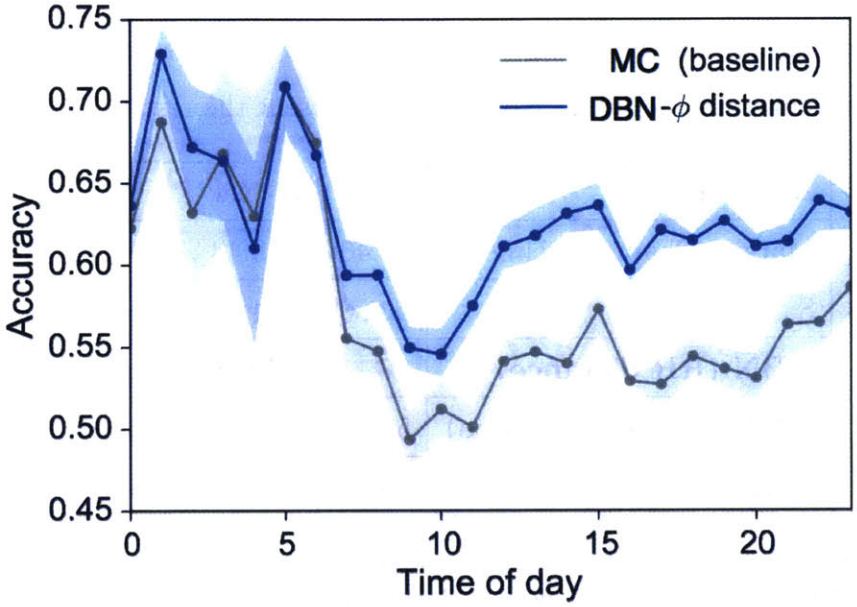

Figure 4-4: The average accuracy of the DBN-ϕ distance compared to the HMM (baseline) at different hours of the day with 95% confidence interval.

4.3.2 Accuracy and mobility entropy

The population of users doesn't move in a similar fashion. There are individuals that visited way more locations relative to others. In order to capture the randomness of the location sequences of people, we calculate the entropy of the observed location sequences. Given a sequence of mobility observations, its entropy S^{real} is defined as:

$$S_i^{real} = -\sum_{x_i' \subset x_i} P(x_i') \log_2 P(x_i') \tag{4.4}$$

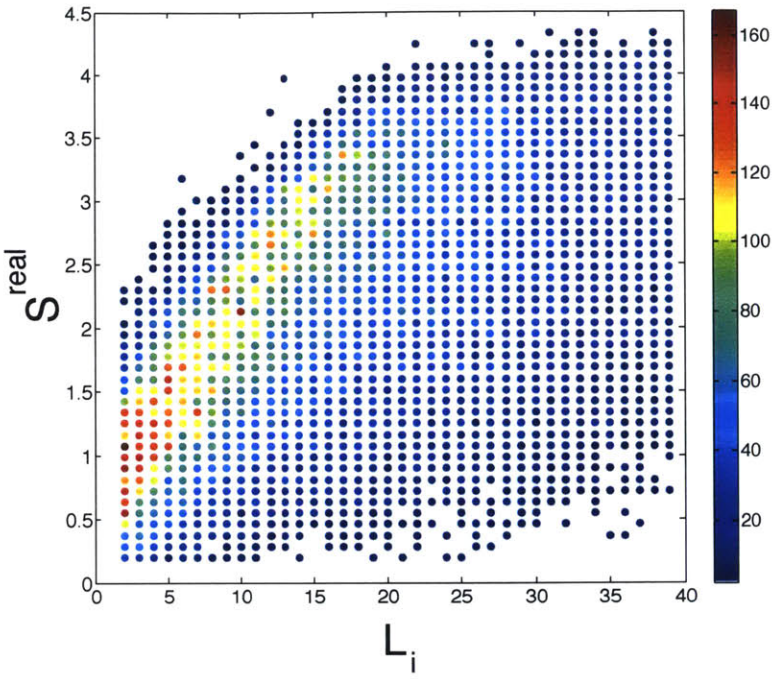

Figure 4-5: The Entropy and the number of uniquely visited location. The entropy of a location sequence (i.e. user's mobility data) increases as the user explores more locations. The figure shows that the majority of the user population visit few locations and have relatively lower entropy.

Where x_i is an observed location sequence for user i, x_i' is the sequences of lengths $1, 2, ...n$ in X_i. The entropy accounts for the observed sequences of movement within the traces of a person; enabling us to account for the spatial and temporal patterns of individual movements[55]. Figure 4-5 shows that there exists a positive correlation between the number of places in which a person was seen and their respective S^{real} value. It also shows that the majority of users visit very few locations (i.e. $L_i < 5$), which makes their entropy lower and therefore more predictable.

The greater the randomness in a location sequence, the lower the predictability of the user that generates it. Figure 4-6 shows the negative correlation of the prediction accuracy and the entropy of the user's data S^{real}. Taken together these results demonstrate that reporting accuracy of a method as a single number for the entire population is not very informative given the high variability within the population's behavior.

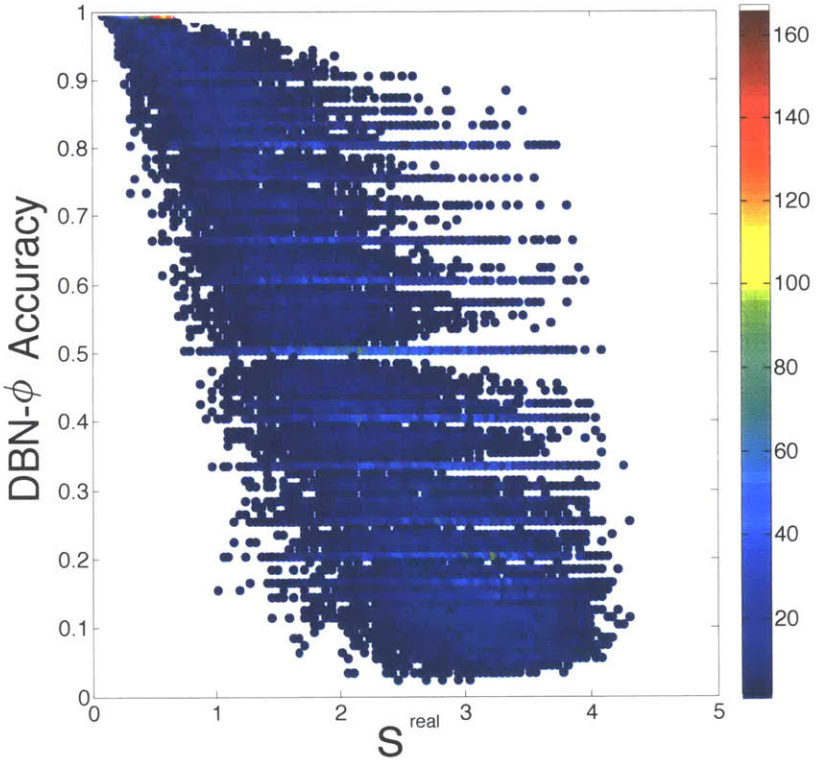

Figure 4-6: The accuracy of DBN-ϕ predictions versus S^{real}. As users explore more location, their entropy increases and they become harder to predict.

4.4 Discussion

This chapter proposes a Dynamic Bayesian Network for predicting human mobility using CDRs as a target case. The model utilizes information from similar strangers that behave similarly in space and time but do not necessarily share a social link. The results show that humans exhibit higher mobility randomness during the day compared to the night; the proposed DBN-ϕ model performs better than the baseline during the later parts of the day when human mobility exhibits more randomness and the coupling with similar strangers proves to be more useful at those times.

While we use CDRs as a proxy to traces of locations, we have shown here that without sampling these sources are very sparse and aren't optimal to recover human mobility of the entire dataset. Furthermore, phone communication is also an imperfect indicator of a social link, failing in some cases to indicate homophily or a type of relation resembling social media connections. On the other hand, these datasets spans over huge sample of the population worldwide.

Chapter 5

Conclusion

The increasing availability of massive passive data opens several venues to study and research the behavior of humans in cities. In this thesis, the second chapter of this thesis investigates how visualizing massive passive data provides insights on patterns of mobility in cities. Chapter two included the application to the city of Riyadh the capital of Saudi Arabia through the UTS project. The project developed the Mobility Browser for the city of Riyadh. It included implementing several modules that mined data generated from mobile phones for insights about dynamics of the interaction between its social structure and transportation infrastructures. The Riyadh city mobility browser synthesizes and extends existing algorithms to provide a decomposition of the complexity of mobility across multiple dimensions.

The visualizations provided by the tool give a dynamic qualitative understanding of the spatial attributes of the city as well as its population directionality across different times of the day. The city mobility browser is envisioned to be a tool that can provide planners, engineers and the public with an easy to understand analysis while capturing fine grained details about the city. Future work could also enable the visualization interface to provide quantitative analysis and a better understanding of emerging patterns.

Chapter three presented a computational framework to discover different attraction patterns of districts in cities and to predict the inflow to newly developed regions. We proposed 3 dimensions to define attraction of a district : total number of trips

the district receives, the spatial dispersion of the origins of trips, and the shape of the distribution of all distances traveled by visitors to reach that district. Additionally, we presented a method for understanding the relationship between the decomposition of the types of POIs in a spatial zone and its attraction behavior. We applied the methodology and discussed the results on the data set from the city of Riyadh the capital of Saudi Arabia.

The chapter shows the results of implementing the discussed modules that mined data from mobile phones to provide a coherent understanding of the dynamics of the interaction between the flows of people to a district and types of services (POIs) that are located in that district. We detected three attraction patterns in the city of Riyadh according to the morning mobility dynamics. Most interesting, the Global attractors, which attract a large portion of the visiotrs traveling relatively high distances and coming from all over the city. These attractors have places of interest that are the destination of large student bodies, factory workers, hospital associates, and embassies. The second type of attraction behavior is that of the downtown area in Riyadh, which attract high inflow of people but with moderate distance and spatial dispersion due to its central location in the city that makes it accessible. The most significant POI types located in the downtown attractors are business based places like firms, and shopping and services places. The least significant attractors behavior is that of the residential areas in the morning hours, where the amount of inflow is very minimal. Residential areas attractors contain non-unique POIs that serve the local people in the neighborhood like apartments, mosques, and schools.
The chapter then proposes a probabilistic modeling approach for predicting the flows between regions in a city using phone calling data. The modeled GP estimates inflows more accurately than a gravity model fitted on the city of Riyadh. The initial results found in this report suggests that probabilistic models enabled by the abundance of phone data could provide better decision tools for urban planners than existing models in the literature of transportation engineering. Such predictive models can help decision makers estimate the inflow to a location prior to making decisions on the functionality of a region or the residential population capacity.

Chapter four proposes a Dynamic Bayesian Network for predicting human mobility using CDRs as a target case. The model utilizes information from similar strangers that behave similarly in space and time but do not necessarily share a social link. We include three ways to calculate closeness measures between individuals as a proxy to finding similar strangers; the closeness measures depend on the spatial and temporal aspects of locations sequences. Coupling users with their most similar stranger achieves a prediction accuracy of 60.03% compared to the addition of the whole social contacts of a person that gives an accuracy of 55.42%. Our results show that humans exhibit higher mobility randomness during the day compared to the night; the proposed DBN-ϕ model performs better than the baseline during the later parts of the day when human mobility exhibits more randomness and the coupling with similar strangers proves to be more useful at those times.

The approach discussed here help predict and understand the mobility patterns of humans at the individual scale. Achieving better prediction accuracy impacts developing recommender systems [9] and generate more accurate human mobility flow models [57, 61]. City flow model are essential for planning the future of transportation in a city [61].

While we use CDRs as a proxy to traces of locations, we have shown here that without sampling these sources are very sparse and aren't optimal to recover human mobility of the entire dataset. Furthermore, phone communication is also an imperfect indicator of a social link, failing in some cases to indicate homophily or a type of relation resembling social media connections. On the other hand, these datasets spans over huge sample of the population worldwide. While data from GPS traces or other technologies with higher resolutions become more ubiquitous, exploring methods to enhance the usability of CDR data for technological applications deserves attention.

Bibliography

[1] Fahad Alhasoun and May Alhazzani. City scale next place prediction from sparse data through similar strangers. *Under Review at Data Mining and Knowledge Discovery*, 2016.

[2] Fahad Alhasoun, Abdullah Almaatouq, Kael Greco, Riccardo Campari, Anas Alfaris, and Carlo Ratti. The city browser: Utilizing massive call data to infer city mobility dynamics. In *3rd International Workshop on Urban Computing (UrbComp 2014). UrbComp: New York, NY*, 2014.

[3] May Alhazzani, Fahad Alhasoun, Zeyad Alawad, and Marta Gonzalez. Urban attractors: Discovering patterns of regions attraction in cities. *Under Review at IEEE International Conference on Data Mining (ICDM)*, 2016.

[4] Abdullah Almaatouq, Fahad Alhasoun, Riccardo Campari, and Anas Alfaris. The influence of social norms on synchronous versus asynchronous communication technologies. In *Proceedings of the 1st ACM International Workshop on Personal Data Meets Distributed Multimedia*, PDM '13, pages 39–42, New York, NY, USA, 2013. ACM.

[5] Arriyadh Development Authority. King abdulaziz project for riyadh public transport. `http://www.ada.gov.sa/ADA_e/DocumentShow_e/?url=/res/ADA/En/Projects/RiyadhMetro/index.html`, 2016. Accessed: 2016-03-30.

[6] Aleix Bassolas, Maxime Lenormand, Antònia Tugores, Bruno Gonçalves, and José J Ramasco. Touristic site attractiveness seen through twitter. *EPJ Data Science*, 5(1):1–9, 2016.

[7] M. Batty, K.W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali. Smart cities of the future. *The European Physical Journal Special Topics*, 214(1):481–518, 2012.

[8] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *arXiv preprint arXiv:1502.03406*, 2015.

[9] Joan Borras, Antonio Moreno, and Aida Valls. Intelligent tourism recommender systems: A survey. *Expert Systems with Applications*, 41(16):7370–7389, 2014.

[10] D Brockmann, L Hufnagel, and T Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, January 2006.

[11] Francesco Calabrese, Massimo Colonna, Piero Lovisolo, Dario Parata, and Carlo Ratti. Real-time urban monitoring using cell phones: A case study in rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):141–151, 2011.

[12] 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.

[13] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.

[14] Communications and Information Commission. Ict indicators, q2-2012, 2012.

[15] Manoranjan Dash, Kee Kiat Koo, João Bártolo Gomes, Shonali Priyadarsini Krishnaswamy, Daniel Rugeles, and Amy Shi-Nash. Next place prediction by understanding mobility patterns. In *Pervasive Computing and Communication Workshops (PerCom Workshops), 2015 IEEE International Conference on*, pages 469–474. IEEE, 2015.

[16] Manlio De Domenico, Antonio Lima, and Mirco Musolesi. Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing*, 9(6):798–807, 2013.

[17] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, March 2013.

[18] Marco De Nadai, Jacopo Staiano, Roberto Larcher, Nicu Sebe, Daniele Quercia, and Bruno Lepri. The death and life of great italian cities: A mobile phone data perspective. *arXiv preprint arXiv:1603.04012*, 2016.

[19] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.

[20] Nathan Eagle, Aaron Clauset, and John A Quinn. Location segmentation, inference and prediction for anticipatory computing. In *AAAI Spring Symposium: Technosocial Predictive Analytics*, pages 20–25, 2009.

[21] Nathan Eagle and Alex (Sandy) Pentland. Reality mining: Sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, March 2006.

[22] Sven Erlander and Neil F Stewart. The gravity model in transportation analysis: theory and extensions. 3:18–19, 1990.

[23] Zipei Fan, Xuan Song, and Ryosuke Shibasaki. Cityspectrum: A non-negative tensor factorization approach. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 213–223. ACM, 2014.

[24] Zipei Fan, Xuan Song, Ryosuke Shibasaki, and Ryutaro Adachi. Citymomentum: an online approach for crowd behavior prediction at a citywide level. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 559–569. ACM, 2015.

[25] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, page 3. ACM, 2012.

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

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

[28] Mark Granovetter. Economic action and social structure: the problem of embeddedness. *American journal of sociology*, pages 481–510, 1985.

[29] Haoying Han, Xiang Yu, and Ying Long. Discovering functional zones using bus smart card data and points of interest in beijing. *arXiv preprint arXiv:1503.03131*, 2015.

[30] J. Herrera, D. Work, R. Herring, X. Ban, Q. Jacobson, and A. Bayen. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. *Transportation Research Part C*, 18(4):568–583, August 2010.

[31] CM Schneider Jameson L. Toole, C Herrera-YagÃije and Marta C. González. Coupled mobility of social ties. *Journal of the Royal Society Interface*, 2015.

[32] S Jiang, J Ferreira Jr, and M.C. González. Discovering urban spatial-temporal structure from human activity patterns. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, pages 95–102, 2012.

[33] Shan Jiang, Gaston A. Fiore, Yingxiang Yang, Joseph Ferreira, Jr., Emilio Frazzoli, and Marta C. González. A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. In *Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing*, UrbComp '13, pages 2:1–2:9, New York, NY, USA, 2013. ACM.

[34] 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.

[35] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. Activity recognition using cell phone accelerometers. *SIGKDD Explor. Newsl.*, 12(2):74–82, March 2011.

[36] Minjin Lee and Petter Holme. Relating land use and human intra-city mobility. *PloS one*, 10(10):e0140152, 2015.

[37] Lin Liao, Donald J Patterson, Dieter Fox, and Henry Kautz. Learning and inferring transportation routines. *Artificial Intelligence*, 171(5):311–331, 2007.

[38] Xi Liu, Li Gong, Yongxi Gong, and Yu Liu. Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography*, 43:78–90, 2015.

[39] Xuelian Long, Lei Jin, and James Joshi. Exploring trajectory-driven local geographic topics in foursquare. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 927–934. ACM, 2012.

[40] Thomas Louail, Maxime Lenormand, Miguel Picornell, Oliva García Cantú, Ricardo Herranz, Enrique Frias-Martinez, José J Ramasco, and Marc Barthelemy. Uncovering the spatial structure of mobility networks. *Nature Communications*, 6, 2015.

[41] Xin Lu, Erik Wetter, Nita Bharti, Andrew J Tatem, and Linus Bengtsson. Approaching the limit of predictability in human mobility. *Scientific reports*, 3, 2013.

[42] Andy Mitchell. The esri guide to gis analysis, volume 2: Spatial measurements and statistics. redlands, 2005.

[43] M E Newman. Modularity and community structure in networks. *Proc Natl Acad Sci U S A*, 103(23):8577–8582, June 2006.

[44] J.P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi. Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. USA*, 104(18):7332–7336, 2007.

[45] Gang Pan, Guande Qi, Zhaohui Wu, Daqing Zhang, and Shijian Li. Land-use classification using taxi gps traces. *Intelligent Transportation Systems, IEEE Transactions on*, 14(1):113–123, 2013.

[46] Srinivas Peeta and Pengcheng and Zhang. Counting device selection and reliability: Synthesis study. Technical report, Purdue University.

[47] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[48] Carl Edward Rasmussen and Hannes Nickisch. Gaussian processes for machine learning (gpml) toolbox. *The Journal of Machine Learning Research*, 11:3011–3015, 2010.

[49] Carlo Ratti, Stanislav Sobolevsky, Francesco Calabrese, Clio Andris, Jonathan Reades, Mauro Martino, Rob Claxton, and Steven H. Strogatz. Redrawing the map of great britain from a network of human interactions. *PLoS ONE*, 5(12):e14248, 12 2010.

[50] Philipp Rode and Ricky Burdett. Cities: investing in energy and resource efficiency. In *United Nations Envrionment Programme, (corp. ed.) Towards a Green Economy: Pathways to Sustainable Development and Poverty Eradication*, pages 453–492. United Nations Environment Programme, 2011.

[51] 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.

[52] 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.

[53] Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.

[54] W. Shen and L. Wynter. Real-time traffic prediction using GPS data with low sampling rates: A hybrid approach. In *91st Transportation Research Board Annual Meeting*, number 12-1692, Washington, D.C., January 2012.

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

[56] Jameson L Toole, Serdar Colak, Fahad Alhasoun, Alexandre Evsukoff, and Marta C Gonzalez. The path most travelled: Mining road usage patterns from massive call data. *arXiv preprint arXiv:1403.0636*, 2014.

[57] Jameson L Toole, Serdar Colak, Fahad Alhasoun, Alexandre Evsukoff, and Marta C González. The path most travelled: mining road usage patterns from massive call data. *arXiv preprint arXiv:1403.0636*, 2014.

[58] Jameson L Toole, Serdar Colak, Bradley Sturt, Lauren P Alexander, Alexandre Evsukoff, and Marta C González. The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, 58:162–177, 2015.

[59] Jameson L Toole, Michael Ulm, Marta C González, and Dietmar Bauer. Inferring land use from mobile phone activity. In *Proceedings of the ACM SIGKDD international workshop on urban computing*, pages 1–8. ACM, 2012.

[60] Pu Wang, Timothy Hunter, Alexandre M. Bayen, Katja Schechtner, and Marta C. González. Understanding Road Usage Patterns in Urban Areas. *Scientific Reports*, 2, December 2012.

[61] Pu Wang, Timothy Hunter, Alexandre M Bayen, Katja Schechtner, and Marta C González. Understanding road usage patterns in urban areas. *Scientific reports*, 2, 2012.

[62] Lihua Wu, Henry Leung, Hao Jiang, Hong Zheng, and Li Ma. Incorporating human movement behavior into the analysis of spatially distributed infrastructure. *PloS one*, 11(1), 2016.

[63] 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.

[64] 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.

[65] Nicholas Jing Yuan, Yu Zheng, Xing Xie, Yingzi Wang, Kai Zheng, and Hui Xiong. Discovering urban functional zones using latent activity trajectories. *Knowledge and Data Engineering, IEEE Transactions on*, 27(3):712–725, 2015.

[66] Xianyuan Zhan, Satish V Ukkusuri, and Feng Zhu. Inferring urban land use using large-scale social media check-in data. *Networks and Spatial Economics*, 14(3-4):647–667, 2014.

[67] 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.

[68] Yu Zheng, Yanchi Liu, Jing Yuan, and Xing Xie. Urban computing with taxicabs. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 89–98. ACM, 2011.

[69] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800. ACM, 2009.

[70] Yu Zheng and Xiaofang Zhou. *Computing with spatial trajectories*. Springer Science & Business Media, 2011.

[71] Chen Zhong, Markus Schläpfer, Stefan Müller Arisona, Michael Batty, Carlo Ratti, and Gerhard Schmitt. Revealing centrality in the spatial structure of cities from human activity patterns. *Urban Studies*, page 0042098015601599, 2015.