The Impact of Spatial Resolution and Representation on Human Mobility Predictability

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Weicheng Qian

©Weicheng Qian, 2012. All rights reserved.

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Abstract

The study of human mobility patterns is important for both understanding human behaviour, a social phenomenon and to simulate infection transmission. Factors such as geometry representation, granularity, missing data and data noise affect the reliability, validity, and credibility of human mobility data, and any models drawn from this data.

This thesis discusses the impact of spatial representations of human mobility patterns through a series of analyses using entropy and trip-length distributions as evaluation criteria, Voronoi decomposition and square grid decomposition as alternative geometry representations. I further examine a spectrum of spatial granularity, from dimensions associated with social interaction, to city, and provincial scale, and toggle analysis between raw data and post-processed data to understand the impact of noisy data and missing data influence estimation. A dataset I was involved with collecting – SHED1 – featuring multi-sensor data collection over 5 weeks among 39 participants – has been used for the experiments.

An analysis of the results further strengthens the findings of Song et al., and demonstrates comparability in predictability of human mobility through geometric representation between Voronoi decomposition and square grid decompositions, suggesting a scale dependence of human mobility analysis, and demonstrating the value of using missing data analysis throughout the study.

Acknowledgements

My deepest gratitude goes to Dr. Kevin Stanley and Dr. Nathaniel Osgood, for their incredible support throughout my thesis. They also taught me how to think on my feet and survive in this fresh area. Dr. Stanley uncovered my interest into topics on human mobility pattern with his insights on sensor networks and ubiquitous computing. He also meticulously thoughtfully guided me to manage my thoughts and energy, my interests would not have veiled this concrete work without the improved self-management. Dr. Osgood firmed my foundation on mathematical thinking and software engineering, which enables me to analyze from mathematical hypothesis for customized computational utilization. He also mentored me in multi-discipline thinking and collaborating to apply my research in epidemic control. They have taught me different aspects required to approach a problem and improved my critical thinking and reasoning skills. I also appreciate the amount of effort they put in correcting my thesis.

Dr. Derek Eager and Dr. Michael Horsch have provided me their invaluable suggestions and feedback on my thesis. Dr. Derek Eager has also introduced me into computer system modelling, which boosted my study. Thanks also go to graduate secretary Janice Thompson, Gwen Lancaster and HPC technicians for their timely helpful whenever needed.

Dr. Christopher Dutchyn and the Problem Solving Club he founded with Dr. Daniel Nelson have brought me both fun and knowledge on efficient computing, which turned to be very useful in human mobility data processing.

My sincere thanks to each and every friend and colleague of mine accompanied me during those good and bad times.

To my parents, who offered me unconditional love and support throughout the course of this thesis.

Contents

Permission to Use **i**

Abstract **ii**

Acknowledgements **iii**

Contents **v**

List of Tables **vii**

List of Figures **viii**

1 Introduction **1**

2 Related Work **7**
2.1 Applications of Human Mobility 7
2.2 Methods of Data Collection 7
2.3 Mobility Models and Analysis 8

3 Theoretical Background **10**
3.1 Entropy and Predictability of Human Mobility 10
3.1.1 Information Entropy 10
3.1.2 Lempel-Ziv Estimation of Human Mobility Entropy 12
3.1.3 Human Mobility Predictability 13
3.2 Trip-Length Distribution 14
3.2.1 Implication for Human Mobility Characteristics 14
3.2.2 Implication for Data Collection and Resolution 15

4 Experimental Setup **16**
4.1 SHED1 Dataset 16
4.1.1 GPS Data Filtering 18
4.1.2 Average-Speed Filter 20
4.1.3 Bluetooth Records Assisted GPS Location Completion 21
4.1.4 Mapping to Indexed Location 21
4.2 Calculating Entropy and Predictability 23
4.2.1 Entropy 23
4.2.2 Predictability 24
4.2.3 Setting for Trip Length 25
4.2.4 Deployment to the Socrates Cluster 25

5 Results **26**
5.1 Entropy and Predictability 26
5.2 Trip-Length Distribution 31

6 Conclusion **34**
6.1 Contribution 34
6.1.1 Insensitivity to Geometry Representation 34
6.1.2 Sensitivity to Granularity 35
6.1.3 Insensitivity to Noise and Missing Data 35
6.1.4 Limitations 35

7 Summary **37**

References **38**

List of Tables

4.1 SHED1 Collected Data List 17
4.2 Available Location Records 19
4.3 GPS post-processing methods 20
4.4 Remaining Data Pairs Following Filtering 22

List of Figures

1.1 Aggegated Heatmap as a representation of the mobility pattern of a participant during the study. Red shows higher and blue shows lower number of samples within a given grid cell [16]. 2

4.1 Experimental Setup Flow showing data feeding, geometric representing and evaluating . . . 18
4.2 CCDF of GPS Displacement using Bluetooth as ground truth 21
4.3 Voronoi Decomposition fragment for much of southern Saskatchewan 23
4.4 Voronoi decomposition fragment for Urban areas of Saskatoon 24

5.1 Voronoi diagram of Saskatoon proximity area 27
5.2 Distributions of areas of Voronoi polygons in Saskatoon proximity Area (Shown in Figure 5.1) 28
5.3 Heat Map of Raw Cell Tower Voronoi 29
5.4 Heat Map of Raw WiFi Voronoi 30
5.5 Boxplot of entropy variation on geometry representation and granularity 31
5.6 Entropy vs bin width regression 32
5.7 BoxplotOfPredictability 33
5.8 Trip-length distribution 33

Chapter 1

Introduction

Prehistoric humans moved for survival, towards hunting and gathering grounds and away from catastrophe; modern humans move with inquiry – pondering why, how, and when are they going; technological humans move in more complex patterns, turning life to a sophisticated tangle of contacts, and shrinking the world into a flat village.

Just as agriculture contributed to the sedentarization and transition from a nomadic lifestyle, the transportation revolution – enabled by the industrial revolution – sped up urbanization and the division of labour. Throughout thousands of years, it appears that the pattern of our mobility evolved from urgency-driven to profit-driven, and then to comfort-driven. At no time in our history have we had the capacity as today to travel as rapidly, conveniently and safely on networks of roads, rails, causeways, and airlines. These benefits have been secured not only because the number of routes has significantly increased, but also because the topology of those routes are following the structure of network as demanded – in a word, human mobility is a governing factor in our infrastructure planning.

Recording time series of locations is a typical numerical approach to track human mobility. The pattern of human mobility underlies those time series, describing the relation between location-location and location-time such as concurrency, stationary, and frequency – just like chord, canon, and tempo.

Starting with smoke signals and drums, improved by telegraph and telephone, extended as radio and television, and finally revolutionized by computer networks and the Internet, telecommunications have been rooted in our system such that its performance shakes the entire system. Network protocols determine performance parameters, including efficiency, reliability, and power consumption. No universal protocol exists due to the large impact from dynamics of the transmitting environment and service requirements. People are eager to be serviced with telecommunications wherever they are, however it is impossible for service providers to deploy access points capable of serving in all physically possible service demands. Human mobility patterns could provide information like visiting schedules for access points, which could further used to estimate utility requirements for service dispatching. Knowledge of user mobility patterns becomes the secret ingredient of a network protocol at state of the art. A network protocols built and tested with a better mobility model is just like a student receives a simulated exam closer to the actual one, which largely increases their confidence in practical performance.

Human mobility underlies the transmission of contagious disease, as well as the footprint of alteration

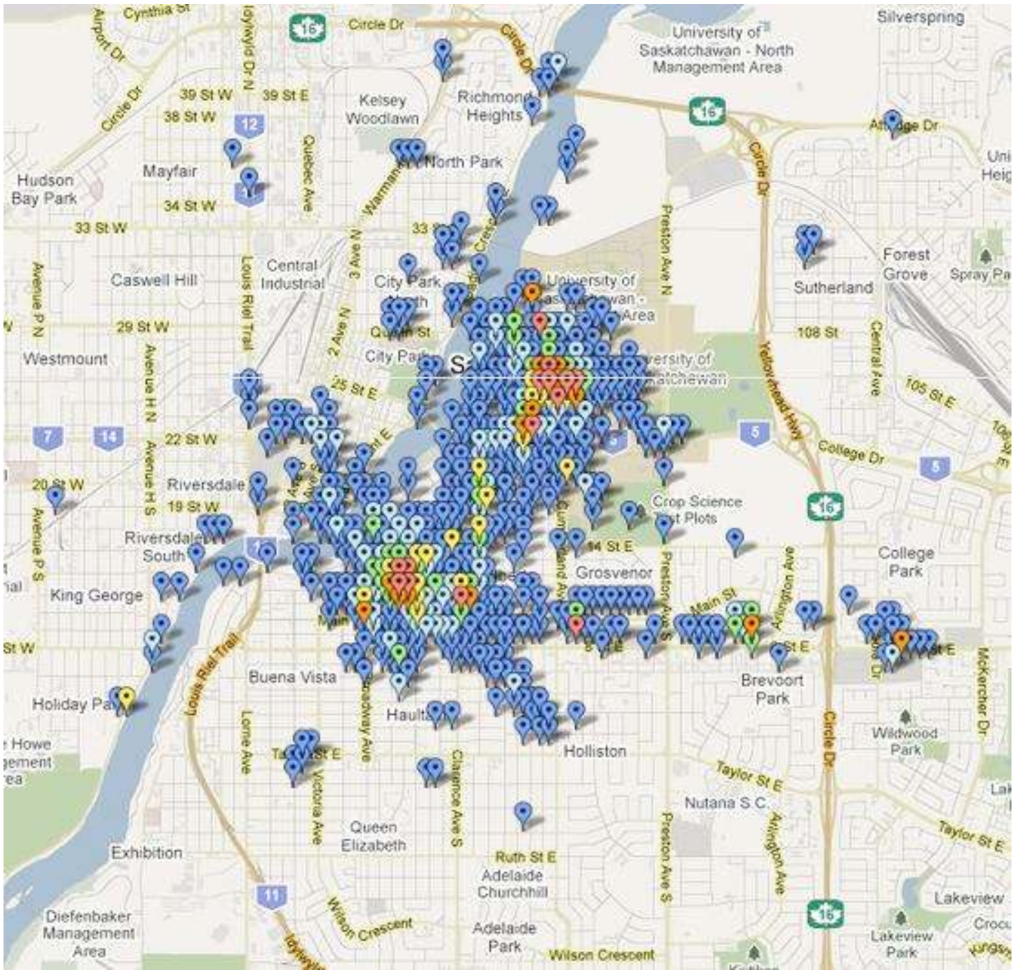

Figure 1.1: Aggegated Heatmap as a representation of the mobility pattern of a participant during the study. Red shows higher and blue shows lower number of samples within a given grid cell [16].

of our environment. People's mobility underlies contact patterns in between people-people and people-place, which drives the probability of infection transmission. Knowledge of these contact patterns between population members have a critical impact on the accuracy of simulations intended to help model disease transmission and epidemic control [44, 48]. Better mobility models would allow realistic synthetic behaviors for agents within Agent Based Models, which increases the fidelity of simulation and supports accurate simulation of intervention outcomes.

Pioneering work in building models for mobility patterns could be traced back to the beginning of the twentieth-century, when the Albert Einstein and Marian Smoluchowski brought their solution [10, 67] to Brownian motion, a term named after Robert Brown, referring to the random motion of small particles suspended in a gas or liquid. A Brownian motion could be abstracted as a continous stochastic process of a variable in N-dimensional real space with independent normally distributed increments. Standard Brownian motion has its start point at zero in coordination, which is also expressed as Wiener Process [47]. A Brownian motion with constraint on magnitude of increment forms a random walk [50] – a mathematical formalization of random motions usually taking a regular lattice as the discretized movement space. At this stage, mobility model assumes object moves wandering aimlessly.

A century later, Lévy walks were proposed as improved representations of human mobility. Lévy walk was named after mathematician Paul Pierre Lévy, who proved the existence of solutions to the N-step addition of Gaussians random variables other than still being Gaussians, although back in 1853, Augustine Cauchy was the first to realize this from pure mathematical point of view. Lévy's findings demonstrate that there could be no characteristic size for the random walk jumps, which results Lévy walk model with scale-invariant fractals. Rather than being based on a mathematical assumption like the Brownian motion model, the Lévy walk model was derived from empirical animal mobility patterns, drawn from mobility data analysis of creatures such as albatross, spider monkeys, bumblebees and deer [66, 51, 9].

An important characteristic of the Lévy walk model is that its trip length distribution is power-law distributed. A power-law distribution implies the heterogeneity of a population such as the Pareto-Zipf law (also known as Pareto-Mandelbrot law) showing that it is the sum of few high frequency events take big portion of occurrence rather than sum of many rare occurred events. The characteristic power-law distribution of Lévy walk model assumes that people tend to take trips with relatively short distance, while occasionally undertaking a long distance trip. In the opening decade of the twenty-first century, the Lévy walk model and related what is now called the continuous-time random walk (CTRW) [57] models of human mobility were created for application in different scenarios, such as Pocket Switched Network (PSN) [36, 19, 39], infrastructure location planning [6], and infectious disease simulation [65, 21, 42].

Although Lévy walk related models have allowed for significant improvements in caputuring human mobility patterns, there still exist the following questions:

- Credit for model generality: Rather than classic mechanics domain that have Newton's laws of motions as basis, which could give credence to other propositions through inference; lacking fundamental theories concerning the constitutive relationships linking different system components and collectively governing process evolution, we are unable to estimate the generality of those models built via current case-specified studies.

- Accuracy and domain of validity: The accuracy of prediction is one important criteria to assess theory validity, such as the domain of validity for Classic mechanics bounded by velocity (comparable to 3×10^8m/s) and size (near or less than 10^{-9}m). Currently we require an accuracy analysis to discern the domain of validity for mobility model to guide further research.

Novel data collection techniques could revolutionize our understanding of mobility patterns. Manual data collecting methods such as logging [9] and bank notes records [4] established the relationship between mobility patterns and the Lévy walk model. Wireless device assisted data collection approach such as using motes [17], cell phones [58], Wi-Fi [37, 25], RFID [49, 54, 28], GPS [52], and multi-sensor smartphones [8, 16], has improved the effectiveness of data collection and models for human mobility patterns.

Questions regarding the generality and existence of better human mobility models remained elusive until Barabási and his team confirmed the regularity and predictability of human mobility patterns. Taking human mobility data through a large population of cell tower routing records and auxiliary GPS data, gathered by a service provider, in [58] they calculated entropy to quantitatively evaluate the stochastics of human mobility patterns. They evaluated 3 forms of information content – random entropy, temporal-uncorrelated entropy and the actual entropy. Furthermore [58] summarized the results of entropy analysis by calculating the upper bound of predictability $\Pi^{\max}$ of temporal-spatially dimensioned human mobility, where predictability is reflected as the average of the probability of correctly predicting the next location based on all the previous location history. The highly bounded distribution peaked near 0.93 for $\Pi^{\max}$, effectively supported authors' conclusion that human mobility is regular and predictable.

The same data has been used to investigate scaling properties of human mobility, and three inconsistencies were reported based on number of distinct visited locations, visitation frequency and ultraslow diffusion of location between the entropy of real collected human mobility data and that implied by classical Continuous-time Random Walk models of human mobility [57].

The data collection of [58, 57] was relatively coarse. with an average accuracy of 3–4km over an irregularly shaped area, based on the spatial decomposition governed by cell towers' locations. Consequently, there is an absence of direct measurements of human mobility predictability with more regular spatial decomposition and at finer scales. Moreover, their data was biased towards oversampling intrinsically high-mobility people as only cellphones which recorded more than two hours per day of talk time were included in the analysis (approximately 0.5% of the total number of unique recorded devices in their data [57, 59]).

Injong Rhee's lab derived data from 24 hour GPS traces of participants and modeled human mobility pattern with four features, namely truncated power-law flights and pause-times; heterogeneously bounded mobility areas; truncated power-law inter-contact times; and fractal waypoints [16]. They have also created a model named Self-similar Least Action Walk (SLAW) as a result of the combination of fixed way points representing frequently visited locations and a Lévy Walk trajectory representing truncated power-law distributed trip-length for trajectories between way points.

At the same time, SLAW appears to be confined by weak considerations regarding temporal correlations observed in human mobility – instead of the absolute time that reflects a relatively firm schedule in human mobility bounded by societal structures and regulations, SLAW tackled the relative, focusing on the relative time between adjacent time stamps, thereby creating a memoryless process.

The idea of the Lévy walk of transitions among frequently visited locations in [16] has also brought an implicit classification of human mobility from intentional-flight to random-walk, which further brings up following issues:

- Will different geometric representation of location – namely, Voronoi decomposition in [16] and square grid decomposition in [58] – lead to an impact on predictability?

- Will finer grained location classification expose previously obscured random-walks and cause a large drop in predictability?

- Will sensor failure-caused data noise and missing data affect the reliability of answers offered to the previous two questions?

Despite the existence of qualitative and quantitative research on human mobility, there is currently lacking a sensitivity analysis study in this area to investigate the impact of various geometric representation of human location and the resulting implications for studies of human mobility. To be more specific, the sensitivity analysis should investigate the impact on the reliability of inferences regarding human mobility of geometric representation, of spatial granularity, and of noise and missing data.

Purely landmark-based localization methods return the coordinates of a known landmark, such as a cell phone tower, WiFi router, Bluetooth beacon, or an RFID tag. The possible location of an object can be characterized by a Voronoi cell centred on the landmark associated with that object (drawn from the collection of landmarks). All known landmarks divide the space into a set of cells that are collectively exhaustive and mutually exclusive, such that each cell contains one and only one landmark[1]. Thus, for all the locations within a given cell, the returned location will be the coordinate of the corresponding landmark. The area of each specific cell determines the error bounds associated with that cell, and the area of the cell is the accuracy of the location estimate corresponding to that cell.

Different GPS localization considerations apply when binning locations in a square grid. The level of accuracy may be determined not only by the GPS locating system itself, but also by the minimum grid size required by the end application. Computation can be saved if the application in question requires a significantly coarser resolution than that supplied by GPS, by binning GPS signals in larger geographic areas.

The granularity with which human mobility is represented affects the reliability of research results as well as the cost and the effort that one will have to invest to secure empirical evidence. For landmark-based studies, the granularity is directly determined by the number of landmarks one maintains in the experiment area. For gridded geometric representation, it is driven by the bin width of each square grid cell, and the minimum reliable resolution of the sensor used to obtain continuous location records.

It is inevitable that a man-made system will generate errors and might have no response at certain times due to service coverage lapses, energy exhaustion or participant opt-out. At the current time, studies on collecting human mobility data usually face a notable rate of missing data, especially for studies that last for months, and suffer from compliance fatigue. The location data available in [58]'s dataset comprises just 30 percent of the theoretical maximum that could be available and [15] has only 34 percent of the theoretically possible GPS data available.

In order to address the impact of geometric representation on spatial decomposition, I adopt the entropy evaluator used by [58] in cooperating with trip-length distribution used in [39, 29] as two criteria to investigate

[1] For the purpose of the thesis, I assume that all the points within a given Voronoi cell are capable of communicating with the landmark associated with that cell.

impact of Voronoi diagram represented location versus square grid represented location on human mobility data. Additionally, I address the influence of granularity, which is represented as landmark density for the Voronoi representation and bin width for a square grid representation.

The dataset – called Saskatchewan Human Ethology Dataset 1 (SHED1) – was used as dataset for the experiments. This dataset includes smartphone collected multi-sensor data with updating period down to 5 minutes. The analysis used the Global Positioning System (GPS) as the primary localizing approach, with WiFi, Bluetooth, and accelerometer as auxiliary data providers [16]. Detailed raw data enables simulation of the effects of coarser sampling than SHED1 collected; additionally, multi-sensor data has made data noise detecting, filtering and completing possible, so that I could investigate the impact of noisy data and missing data by comparison between different data feeds.

This analysis is based on experiments that compare entropy and trip-length distributions across different scenarios which are composed by cross combination of geometric representation (including Voronoi decomposition and square grid decomposition with different granularity) and data feed-in (raw data, filtered data, filtered completed data).

This study shows that geometric representation failed to cause human mobility predictability variation; instead, granularity is the primary factor that affects human mobility predictability. There is no acute increase of entropy detected with finer granularity, which further enhances the reliability of [58]. Finally, a comparison between raw data and filtered data indicates the accuracy of GPS system is enough to capture stable data.

The remainder of the thesis is organized as follows. Chapter 2 describes related research for each aspect of this work, including the application of human mobility, data collection approaches for human mobility patterns, and existing mobility models and analysis. Chapter 3 describes the theoretical background of the thesis, including entropy and trip-length distribution as evaluators to help analyze human mobility. Chapter 4 describes the experimental setup. Chapter 5 illustrated the results and analysis. Chapter 6 summarizes the thesis and outlines possible areas for future work.

Chapter 2

Related Work

2.1 Applications of Human Mobility

Intelligent routing decisions, whether based on infrastructure such as cell tower or WiFi router placement, or opportunistic peer-to-peer routing in delay tolerant networks, depend on reliable records, models and predictions of human mobility patterns. Significant research has attempted to elucidate the impact of employing user movement patterns to support transit opportunity estimation [56, 22, 18, 26], device resource management [24, 34], and battery savings [69, 38]. The same determinants of human mobility which inform the placement of cell towers and WiFi hotspots also help determine the placement of other services such as bus stops [20, 32].

Contact patterns show analysis including contact duration and intensity of population, it directly impact the calculation of critical parameters in health modelling, such as the basic reproductive number. The basic reproductive number of an infection is the number of cases one case generates on average over the course of its infectious period [11]. Contact patterns play a central role in infection transmission models and infection prevention and control decision-making and policies [44, 48]. Human mobility patterns underlie the contact patterns between both people and places, impacting the transmission of contagious diseases, attitudes and norms, access to services, and contribute to exposure to environmental risks such as toxins and pedestrian-unfriendly built environments, driving both environmentally mediated diseases such as asthma and socially mediated diseases such as obesity [60, 30, 31, 61, 63].

2.2 Methods of Data Collection

Because of the diverse and significant impacts of human mobility models on a variety of disciplines, researchers from several fields have attempted to leverage novel measurement techniques to infer, measure or constrain human mobility.

The cellular phone is a now ubiquitous item in modern society, and is undergoing rapid penetration into the developing world. Cell tower call and contact records form an attractive data source for studying human mobility because they are collected automatically, and cover thousands or millions of subscribers over prolonged sampling periods. For this reason, several notable contributions [58, 25, 13] have used traces that reflect connectivity to existing infrastructure (e.g., Access Points (APs) or cells) as their primary data source. However, the tower locations map to a Voronoi diagram with a spatial granularity dependent on geography and population density, which can reduce the overall accuracy, and position updates for call records are only available when a user initiates or receives a call or a text message [58].

The global positioning system (GPS) can provide near-continuous data with an accuracy down to a few meters. However when users move indoors, GPS suffers a significant reduction in accuracy due to fading and multi-path effects, up to the point of complete loss of signal within large facilities. Studies have employed GPS data sets with a temporal resolution of hours [58], down to minutes [16, 39, 1]; study durations from a day [16], to over a month [16, 1]; and study population from dozens [16], to hundreds [39, 1, 58] of people have been reported. However, these datasets are often characterized by poor participant compliance, ranging from 30% to 50% [58, 16]. Low participant compliance could cause risk of misjudging human mobility predictability and human mobility model creation.

To compensate for the poor resolution of cell tower based measures and the reduced coverage of GPS data, researchers have turned to shorter-distance radio-based devices such as WiFi [37], or RFID [49, 28, 54] systems, and have inferred position based on either the Voronoi diagram of beacons – as in the cell tower case – or used more sophisticated methods such as trilateration or fingerprinting.

2.3 Mobility Models and Analysis

Taking bank note tracing data as a secondary data to reflect the underlying human mobility, [4] brought human mobility research into a new phase by separating assumptions of human mobility research from analogies of physical models such as random walk and Lévy walks [64, 33]. Within their paper, an investigation regarding the distribution of traveling path lengths with relatively large scale led to the discovery of truncated power-law distributed trip length as one of the potential characteristics of human mobility. The truncated tails in the power-law distribution is mainly caused by the physical limitation of human beings and technology such as the speed of driving lying under 200km/h [2]. The researchers have also proposed a model deriving from lattice network assumptions that shows the ability to reproduce human tracks with truncated power-law distributed trip length. However, trip length as a restriction only focuses on memoryless or quasi-memoryless transitions, without considering the patterns of absolute location occupied by people. The memoryless processes have only relative location aspects recorded, and this could not used to represent the characteristic clustering of absolute locations.

The authors of [16] have developed an app for the Google Android operating system based smartphones. Multi-sensor collected data provides a potential opportunity for data fusion that could mitigate noise caused by a single sensor or complete missing data. However an analysis of collected data in terms of human mobility patterns is not provided.

The authors of [27] used both GPS and WiFi supported location information as feed-in data, using the same criteria as the Song and Barabási adopted [58]; their results advocate Song and Barabási's finding of a high entropy upper-bound [58]. Besides using the Lempel-Ziv estimator for actual entropy, they have also implied a first order Markov model with the transition probability estimated from the finite process and obtained the same result as did the Lempel-Ziv estimator.

Integrating human traces from six studies, [29] found that the distribution of inter-contact time possesses an invariant property: a characteristic elapsed time threshold – on the order of half a day – beyond which the distribution decays exponentially. Random Way Points method of generating synthetic human mobility records involves randomly generating points as persons' possible visiting locations, and randomly choosing destination from these points as the next trajectory. The Random Way Point method has the ability to aid in generating human trajectories while allowing truncated exponential decay of inter-contact time pairs in the simulated community.

This work is unique because it completed the post-processing of data collected from [16], and it have multi-sensor data that allow us to filter noisy data, and – in addition to previously used techniques [29, 27] – investigate the impact of and interaction between granularity and geometric representation.

Chapter 3

Theoretical Background

As an estimator for predictability, entropy originated in physics for describing how evenly energy is distributed in a system. Subsequently, this concept was introduced to information theory, and is now usually referred to as Shannon Entropy, which has been widely used in information processing, such as in search engines [45, 7], compression algorithms [35, 70], statistic tests [46], and parameter estimation [55, 5]. [58] used Shannon Entropy as a measurement of predictability of human mobility. Human mobility entropy, as far as I know, is the current best quantitative metric that is generally applicable to estimate the predictability of heterogeneous mobility records.

The trip-length distribution has been studied since the Brownian motion model era. The trip-length distribution is a practical analysis tool for a time-series of locations, because trips are one of the most common patterns within those time-series.

3.1 Entropy and Predictability of Human Mobility

The entropy and predictability of human mobility are statistical properties of the temporal and spatial characteristics of locations of a population of individuals. Such quantities are investigated in order to gain insight into the question of whether human mobility is regular enough to be predicted or simulated, rather than highly random without the potential to be accurately captured. The answer defines a general upper-bound on the best results that could be obtained by human mobility modeling.

3.1.1 Information Entropy

Information Entropy – also known as Shannon Entropy – is a measure of the uncertainty associated with a random variable [23]. The information entropy associated with a message specifies the amount of the information contained in a message [14].

The basic unit of information is a bit – or binary digit – which can have a value of either 1 or 0, and can be viewed as indicating that some proposition is either true or false. Series of bits represent information; however the length of the string of bits does not always indicate the amount of information contained within it. For example, within a string, if a long pattern was found repeated many times, a dictionary could be created, using a short index to refer to the long pattern and then replace all occurrences of that pattern in the string with that short index, thus the length of the string was reduced without any loss of information. This process is also called dictionary based compression, and its success proved that actually, that string did not contain as much information as it might appear.

For this thesis, information regarding location is primarily dealt with. Consider a finite set L representing all the possible locations for a person. Then at each time interval for that person, there could be a particular element from the set L that would indicate the corresponding location the person is currently in. As the set is finite, a series of bits could be used to encode the corresponding element. In this way, individual's mobility pattern could be presented with certain level of accuracy in both the temporal and spatial dimension using an ordered series of bits. Similarly, to determine the mobility pattern of a person with the same level of accuracy on both spatial and temporal scales, our residual uncertainty regarding that mobility pattern depends on the amount of the information that is contained in that series.

With respect to the topic of human mobility, as Barabási described [58], there could be three types of entropy:

- The random entropy

  $$S_i^{\text{rand}} = \log_2 N_i \tag{3.1}$$

  Within equation 3.1, N_i is the number of distinct locations visited by user i. This entropy only focuses on the number of unique locations that a person has ever visited during the experiment, and is independent of history.

- The temporally-uncorrelated entropy

  $$S_i^{\text{unc}} = -\sum_{j=1}^{N_i} \mathrm{p}_i(j) \log_2 \mathrm{p}_i(j) \tag{3.2}$$

  Within equation 3.2, $\mathrm{p}_i(j)$ is the fraction of visits to locations by user i, that were to location j. This entropy not only captures the number of unique locations that person visits, but also the potential heterogeneity in the frequency with which a person visits all of those possible locations.

- The actual entropy S

  $$S = \lim_{n\to\infty} \frac{1}{n} \mathrm{S}(X_1, X_2, \ldots, X_n) \tag{3.3}$$

  The actual entropy is intended to capture both the order of visiting and heterogeneity in frequency of visiting to locations that a person will visit. With regular observation intervals for the location information of a person, actual entropy also captures the stationary property of human mobility, which could reflect the physical inter-contact time of the human community once the observations schedules for each entity are synchronized (i.e. a common temporal scale is defined).

Within the three types of entropy above, the first two types of entropy could be evaluated directly from collected location records, with only a very modest amount of processing. However, the third type of entropy cannot be evaluated directly, because it requires finding those definitions (which might not be unique) of dictionaries that could minimize the real entropy of the compressed string. Hence a method is required in order to estimate the actual entropy.

Data compression is the process of reducing the redundancy within an ordered series of bits – and thus reducing its size – so as to benefit transmission and storage. The compression ratio limit that the data compression algorithm can reach on a specified ordered series of bits is determined by the information or the entropy the series contains [35]. To estimate the entropy of a series of characters, it is possible to simply run a universal compression algorithm (as not all such compression algorithms, for instances, will lead to a ratio that approaches the entropy), and then calculate the ratio of compression [35]. This ratio approaches the actual entropy of the string when the length of the string tends to infinity [58, 59].

3.1.2 Lempel-Ziv Estimation of Human Mobility Entropy

If the human mobility records are defined during duration T with a sampling interval (temporal resolution) Δt, and a spatial resolution β such that all the possible locations that the observed group of people could have is reflected in the discretization of geographical location mapping into a set L, each $x_i \in L$ represents a non-overlapping sub-area containing one and only one landmark, and generally the scale of x_i is on the order β. For example, a string like:

$$x_1, x_2, x_3, \ldots, x_n$$

where $n = \lfloor T/\Delta t \rfloor$ and $i \in [1, n]$, could be used to represent a time series of locations with temporal resolution Δt and spatial resolution β , which represents a person's mobility records with specific resolution.

Lempel-Ziv estimation of actual entropy can be calculated with the formula:

$$S^{\text{est}} = \left(\frac{1}{n} \sum_{i=1}^{n} \Lambda_i \right)^{-1} \ln n \tag{3.4}$$

Within equation 3.4, Λ_i is the length of the shortest substring starting at position i which doesn't previously appear from position 1 to $i-1$, and n is the length of all samples for which records is available. The estimation S^{est} converges to the actual entropy S when the string length n (corresponding to total recorded time $n \cdot \Delta t$) tends to infinity.

In realistic scenarios, the records that have been exhibited a limited n, and within the limited records there might be missing data on a person's location information at specific times. Such missing data can be treated as being represented by a distinguished character (without loss of generality, a question mark "?"). Even for those times in which the location information $x_i \in L$ is available, there might be noise associated with the location measurement, where – for example – the location of a person at time t is x_i, while as a result of the limitation of the locating method and related equipment, the recorded location is $x_j, i \neq j$.

Such noisy location records could, for example, flip among the cells adjacent to the actual location. This effect determines the noise floor (the measure of the signal created from the sum of all the noise sources and unwanted signals within a measurement system) for a particular human mobility pattern analysis. Sensor failures which produce large relative displacements – such as a GPS restarting and reporting (latitude 0, longitude 0) as the coordinate – can create outliers with significant impact on the entropy calculation.

In the supplementary material of [58], Barabási et al. have investigated the impact of missing data on entropy estimation with an experiment that randomly assigns a fraction of known locations to unknown for a dataset that they collected every hour over eight days for two typical users. This forms a subset of a dataset that collects the nearest cell tower coordinates as personal location. From their result, the estimated entropy will slightly (less than 12%) overestimate the actual entropy when there is more than 30% of location information remaining. By contrast, when more than 70% of the location information is missing, the system tends to underestimate the actual entropy (by up to 50%). This suggests that as long as less than 70% of the total data is missing, Estimated entropy could still be used to bound the predictability in a conservative fashion (in which the entropy might be overestimated – and thus underestimate the predictability – but not the reverse).

Using a question mark "?" to represent unknown location information will tend to underestimate the entropy of human mobility patterns, as once there is unknown location information, the "hole" will be patched by the same location "?" throughout the records, such that when the missing rate is over 50%, one could actually find a person over half of the time in the "unknown" bin. Secondly, in performing experiments in which known locations were replaced with unknown locations, the newly formed pattern underplays the clustering of the missing locations. Based on our experience, within the data that has been collected, the missing data tends to be clustered rather than evenly distributed between known location information. The preceding two points might have impact on the validity of sensitivity analyses to be presented.

3.1.3 Human Mobility Predictability

Predictability [58] is intended to quantify the ability to predict a person's location given information regarding all previous location information for that person. This corresponds to the probability that the location of a person could be correctly guessed, if their history of previous locations were token into account.

The average rates of correct guesses of a person's location might move and down as the algorithm exploited additional past records; however, there exists a upper bound on predictability for any algorithm that could be adopted. The upper bound of predictability over some period $\Pi^{\max}$ is defined implicitly by – and solved by reference to – a set of equations considering both entropy S and number of unique locations N that a person has visited during that period:

$$\begin{aligned} S &= \mathrm{H}(\Pi^{\max}) + (1-\Pi^{\max})\log_2(N-1) \\ \mathrm{H}(\Pi^{\max}) &= -\Pi^{\max}\log_2(\Pi^{\max}) - (1-\Pi^{\max})\log_2(1-\Pi^{\max}) \end{aligned} \tag{3.5}$$

Within equation 3.5, N must be no less than 2. In solving these equations for $\Pi^{\max}$, I make use of the value of S calculated using the Lempel-Ziv Entropy Estimation algorithm noted above as well as the value of N. $\Pi^{\max}$ provides an upper limit for the performance of any algorithm which utilizes the complete past for a person to guess at what specific location that person lies at a specified time [27].

Within [27], the authors note that there is a cluster of the upper-bound of predictability at 93% across the 45,000 people in their dataset. The clustering also shows the lack of variability in predictability. This upper-bound does not consider potential difficulty and cost in devising estimates that approach this limit.

3.2 Trip-Length Distribution

Trip-length distributions are intended to describe the character of human movement scales. Currently, human movement is detected via location change between scheduled observations. The accuracy of the trip-length distribution is mainly determined by the accuracy of the locating method and the sampling regime.

There is a trade-off between the accuracy of trip-length information that could be gathered and the cost required to achieve it. Considerations for this cost include the storage capacity, complexity of sensor deployment, power consumption, and (in a multiplicative fashion) the size of population that could be monitored.

3.2.1 Implication for Human Mobility Characteristics

The truncated power-law distribution [29] reflects the heterogeneity in the distance traveled by humans. At the same time, it also reflects limits in human energy [2]. This limitation could be further summarized as the result of physical bounds and psychological bounds. Physical bounds emphasize on the limit of human body tolerance and limitations of transportation devices and regulatory guidelines that shape human mobility, such as the fact that an acceleration of 3 times of gravity is usually the limit that normal people can bear and the fact that the upper-bound of freeway speed is usually no more than 140 miles per hour (about 225 km/h). The speed could also been restricted by psychological factors such as caution regarding the road condition, cognitive mapping, or personal schedules. For example, a new settler of a city or campus might choose a common path instead of a shortcut, but after some time familiarity with the path and its surround might lead them to switch to shortcuts; at the same time, people might sacrifice the convenience of shortcut if they were attracted by a scenic view. In a word, social and spatial context could change the behavior of human mobility and cause a truncated tail as the distribution of trip-length.

In short, the rapid convergence of trip-length reflects existing bounds that people can't overcome – it could be the limit of the available time that people could have in a day, also with several physical boundaries associated with the speed of displacement.

A trip-length distribution could not only determine the next location that one could transit to, but could also to some extent reflect the specific habits of human beings, as the trip-length could related to two typical classes of human mobility: flight and random walk [16].

Flight is defined as a long distance trip with only minor direction changes. This type of mobility is primarily related to intentional human movement, and could be tagged with important events that have a long term plan that might transport resources, immigrate individuals and communicate between groups with relatively little interactions. As flight is more related to intensional trajectories, it can also be related to scheduled behavior. Thus, it would inherently provide a high degree of predictability [16].

By contrast, a random walk focuses on small scale wandering without obvious direction or bounding region. This kind of mobility is mostly related to instant decision–making without a long-term forecast or plan. The relatively small scale of unscheduled movement could cause difficulty in determining the exact location of a person.

3.2.2 Implication for Data Collection and Resolution

Modern localization approaches typically rely on WiFi, Cell tower, GPS systems, but also specific indoor localization using RFID tags [49, 28, 54]. The localization could be based on reference to known landmarks, either directly duplicating the known location as the estimate, or using time-of-flight distance measurement or trilateration to infer an intermediate location from a set of several such landmarks.

Cell tower or WiFi-based call and contact records form an attractive data source for studying human mobility because they are collected automatically, and cover thousands or millions of subscribers over prolonged sampling periods. For this reason, several notable contributions [58, 25, 13] have used traces that reflect connectivity to existing infrastructure (e.g., Access Points (APs) or cells) as their primary data source. However, the tower locations map to a Voronoi diagram with a spatial extent dependent on geography and population density, which can reduce the overall accuracy. Moreover, because position updates for call records are only available when a user initiates or receives a call or a text message [58], the records can over-sample client who primarily rely on mobile phones such as real estate agents and cab drivers and (by extension) high mobility individuals.

To compensate for the poor resolution and selection bias of cell tower-based measures and the reduced coverage of GPS data, researchers have turned to shorter-distance radio-based devices such as WiFi [37], or RFID [49, 28, 54] systems, and have inferred position using either the Voronoi diagram of beacons [58] – as in the cell tower case – or used more sophisticated methods such as trilateration or fingerprinting [68, 12, 43, 62, 3] to interpolate a cartesian position from multiple records.

Chapter 4

Experimental Setup

To address the impact of geometric representation on spatial decomposition, the influence of granularity and noisy incomplete data, and adopting both Entropy-Predictability and the distribution of trip lengths as the metrics, the experiment is designed to compare the variety of human mobility pattern as a composite of different human mobility records, varying sensor type (GPS or WiFi), data quality (raw data or filtered data), geometric representation (Voronoi or squared grid) and location record granularity. The SHED1 dataset is used, which includes multi-sensor data records for 39 participants over 5 weeks.

To reflect the variation in location record granularity, for Voronoi decomposition, Cell Towers are chosen as low-density landmarks and WiFi access points as high-density landmarks. For square grid decomposition, Five different granularities were used to represent the location, with bin width ranging from 15.626 m to 4 km with a multiplicative step factor of 2 (15.625 m, 31.25 m, 62.5 m, 125 m, 250 m, 500 m, 1 km, 2 km, 4 km).

Throughout the experiment, I obtained 52 boxplots for entropy evaluation and 1 triplength distribution for further analysis.

4.1 SHED1 Dataset

Central to our analysis is the Saskatchewan Health Ethology Dataset (SHED1) collected over 5 weeks in April and May of 2011 [16]. This dataset contains telemetry recorded from HTC Magic Android smartphones and includes direct (through GPS) and indirect (through WiFi and Bluetooth) information on participant geographic location. This data was crossed with estimated recorded positions of WiFi routers and cell towers in the city of Saskatoon [40]. Collected data was parsed and filtered prior to performing entropy calculations on the Socrates cluster at the University of Saskatchewan.

After receiving approval from our Research Ethics Board employing the software described in [15], we deployed the data collection system for 5 weeks during April and May of 2011 using Android Dev Phone 2s running a custom version of the Android 2.1 operating system. Forty participants were recruited from the Computer Science department, consisting of graduate students from several laboratories, and technical and administrative staff. Participants met one-on-one with at least one study organizer, were walked through the experimental protocols and use of the phone, filled out consent forms, and had the opportunity to ask questions. Participants were requested to carry the phones with them at all times during the day, unless the phone was low on batteries, in which case they were requested to plug it into a computer near them. Participants were also requested to take the phone home with them at night, and to initiate charging just prior to going to sleep. One participant withdrew within the first week, leaving 39 participants who completed the entire study. Of those, one participant experienced a hardware failure on the GPS receiver on their phone, leaving 38 participants with reliable location data. Results are presented here for those 38 participants.

The phone was programmed to collect data in bursts with a 5 minute duty cycle to manage data size and battery life. Every duty cycle, the phone logged 1 minute of accelerometer records, 1 minute of Bluetooth contacts, 3 seconds of WiFi contacts and 10 records of battery state. GPS records were collected for 2 minutes, but given that the GPS required significant time to acquire satellites to achieve position lock, the first approximately 90 seconds often did not contain data. The information recorded by each sensor is summarized in Table 4.1. Values in the ALL row correspond to common fields.

Table 4.1: SHED1 Collected Data List

Parameter	Variables recorded
ALL	Participant ID, time stamp
GPS	Latitude, longitude, velocity, accuracy
Acceleration	Acceleration in x, y, z
Bluetooth	MAC address, signal strength
WiFi	BSSID (MAC address), SSID (Network Name), signal strength, frequency, security protocol
Battery	Battery level, plugged status, battery status

Data was opportunistically uploaded by phones when participants were in contact with the university's secure wireless network. Data on the server was accumulated in flat files and parsed at regular intervals and inserted into an MS SQL Server database. Participants with low compliance and those whose reported data dropped significantly were notified through email. At the conclusion of the study, participants returned their phones and filled out a questionnaire, which contained basic demographic information, information about perceived compliance and lab/office affiliations.

Time series of location information were used to reformat those records of mobility. As SHED1 collected data has several locations within the same duty-cycle, I summarized the location of a person in a duty cycle. Because SHED1's GPS data was in units of degrees of latitude and longitude, a way need to be adopted to convert between GPS degree pairs and kilometers specified in two dimensional coordinates.

A space-time log is an ordered series of location indices representing the mobility of a participant during the experimental period; each location index within the series representing the specific location that such participant is in at a particular duty cycle. For those duty cycle when there is no location data for a

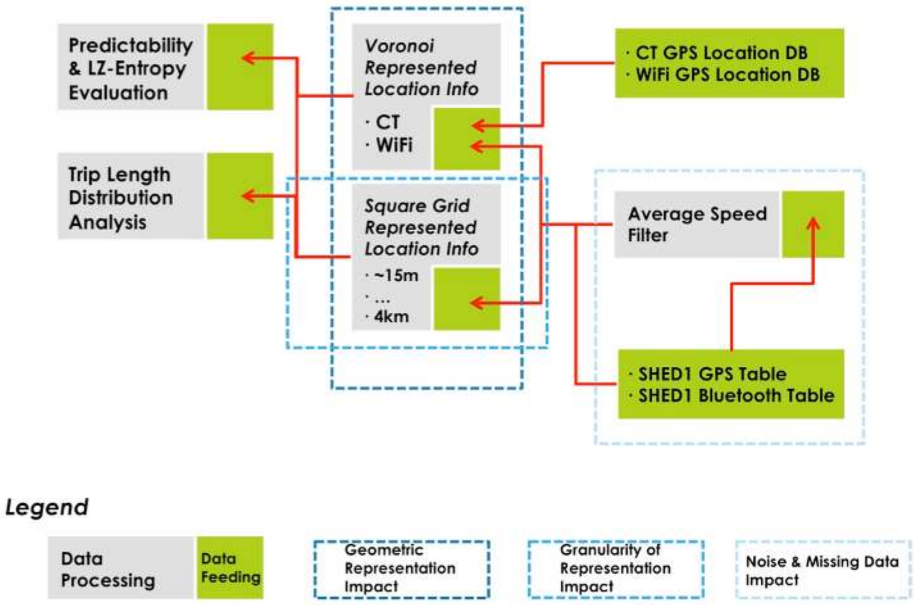
Predictability & LZ-Entropy Evaluation
Trip Length Distribution Analysis
Voronoi Represented Location Info
· CT
· WiFi
Square Grid Represented Location Info
· ~15m
· ...
· 4km
· CT GPS Location DB
· WiFi GPS Location DB
Average Speed Filter
· SHED1 GPS Table
· SHED1 Bluetooth Table
Legend
Data Processing
Data Feeding
Geometric Representation Impact
Granularity of Representation Impact
Noise & Missing Data Impact

Figure 4.1: Experimental Setup Flow showing data feeding, geometric representing and evaluating

participant, the special location index "-1" was assigned as a placeholder (analogous to the "?" described in section 3). Therefore, we had 38 records of 9793 duty cycles corresponding to participant locations for every duty cycle in the study.

For both geometric representations, the location index is used to indicate in which cell the participant was currently located based on GPS coordinates. For the Voronoi geometric represented location, Cell tower ID was used directly as the location index; for the square gridded geometric representation, I numbered the bins from top to bottom with each row from left to right continuously; and for the WiFi location, I used the MAC address of each access point as the location index.

To convert from GPS degree coordinates to two dimensions, I used the Spherical Law of Cosines to calculate distance between two degree coordinates with the formula: Let $p_1(lat_1, lon_1), p2(lat_2, lon_2)$ be two points whose GPS degree coordinates is already known,

$$\text{Dist}(p_1, p_2) = \sin^{-1}\Big(\sin(lat_1)\sin(lat_2) + \cos(lat_1)\cos(lat_2)\cdot\cos(lon_2 - lon_1)\Big)\cdot R \tag{4.1}$$

Within equation 4.1, R is the earth's radius (mean radius 6,371km). This method typically gives errors of up to 0.3%[41].

4.1.1 GPS Data Filtering

GPS data filtering enables us to inspect the impact of missing data and noisy GPS data on perceptions of human mobility by removing data based on considerations of physically plausible velocities and detected co-location with other nodes as measured by Bluetooth.

Raw GPS data typically exhibits two major sorts of problems: by recording noisy data, and by missing data which should have been recorded. Noisy data could, for example, be caused by an inappropriate report of a default location while restarting the GPS sensor, or by presence indoors causing unreliable position estimates from the GPS satellites. Missing data could be due to the phone running out of power, the phone being in a location where no satellite signals are available or by the participant temporarily disabling recording.

Table 4.2: Available Location Records

Representation	Available Records	Missing	Meaning
Grid Raw	126,789	65.93%	Fed with raw GPS data, square grid decomposition
Grid ASF	112,268	69.83%	Fed with average speed filtered GPS data,square grid decomposition
Cell Tower Raw	126,788	65.93%	Fed with raw GPS data, Voronoi decomposition with cell tower as landmark
Cell Tower ASF	112,268	69.83%	Fed with average speed filtered GPS data, Voronoi decomposition with cell tower as landmark
WiFi Raw	192,976	48.14%	Fed with WiFi access point scanning data using the access point with highest RSSI level each duty cycle per participant, Voronoi decomposition with WiFi access points as landmarks

As shown in the Table 4.2, SHED1 collected raw GPS data has a missing rate of 65.93% which indicates that only 34% of the potentially GPS data is available over all participants. This is primarily due to participant compliance, which averaged a little over 50% [15]. Within this 34% of data that is available, I also evaluated the reliability of these data via cross referencing with Bluetooth records.

Theoretically, the Bluetooth sensor on a smartphone will record the MAC address of all the devices having Bluetooth radio in “discoverable” mode within approximately 10 meters. Bluetooth records were aggregated into a table, listing any device that has ever been spotted for all participants [15] in the study for every duty cycle. Because Bluetooth radio range is generally roughly 10 meters, one should expect that the GPS locations of a pair of Bluetooth records should also exhibit displacement between their GPS location no more than 10 meters (much less 100 meters). However, of those records containing both Bluetooth and GPS data, 14.7% exhibited a relative participant-participant displacement of over 1 km.

4.1.2 Average-Speed Filter

The average speed filter is primarily intended to filter GPS data associated with physically impossible duty cycle – duty cycle transitions. Note that average-speed filtering is not an absolute threshold filter such as low-pass filter, high-pass filter, or band-pass filter in signal processing that will only working with signal values without considering the variation of signal values; instead, this filter is based on the physical limits of human speed.

A simply implemented average-speed filter could just use common speed (such as walking speed or driving speed) as a threshold. But a more complex one could consider the acceleration that governs the potential speed change during selected session. For instance, if average speed is found to be around 100km/h, then it is very likely that the person is on a highway, and at this time, the person's trajectory should exhibit properties closer to flight, and has a higher chance to last many sessions.

The following CCDF plot shows the distribution of displacement of GPS coordinates for a pair of participants when Bluetooth records show that at least one of them has detected the other at that duty cycle. Different colors denote distinct different data feeds, with details as shown in Table 4.3.

Table 4.3: GPS post-processing methods

Data Field Name	Processing Method
filteredMean	Calculate mean latitude and longitude on all remaining records of each duty cycle after filtering those records that have Euclidian distance more than 1.5km from the median latitude and longitude of the duty cycle
MaxAccuracy	Use the GPS records that have the maximum accuracy among all records within each duty cycle
RawMean	Calculate mean latitude and longitude on all records of each duty cycle
AvgSpdX	Calculate mean latitude and longitude on all remaining records of each duty cycle after filtering those records that have Euclidian distance more than X km from the median latitude and longitude of the duty cycle, and then discounting all duty cycle records that have their average speed from last duty cycle larger than Xkm/duty cycle

With the average speed filter, I found that the noise spotted by using the Bluetooth table as the underlying truth has been reduced from 15% down to 7%. That is, half of the mis-estimated positions were due to erroneous reports of "teleportation" of one of the nodes likely due to a GPS reset.

An average-speed filter is insensitive to incompleteness. Even if there are several duty cycles with missing data separating two tested sessions, the average-speed bound is still defined – as here "average-speed" also indicates the minimum average-speed between two tested duty cycles, with the assumption that in the intermediate missing-data sessions, I assume travel in a straight line with this average speed serving as a

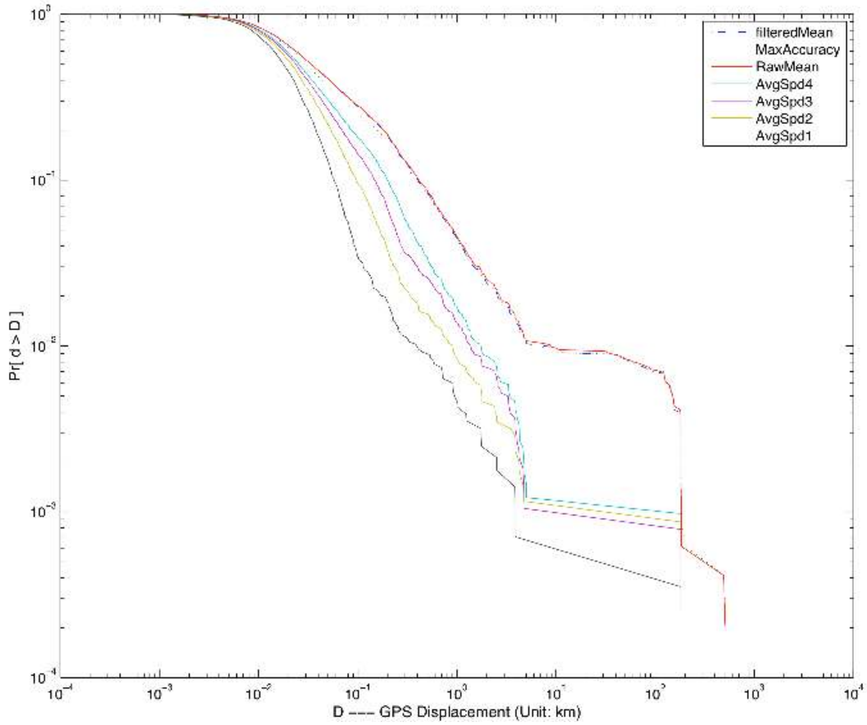

Figure 4.2: CCDF of GPS Displacement using Bluetooth as ground truth

lower bound.

4.1.3 Bluetooth Records Assisted GPS Location Completion

A bluetooth device in the phone could serve as a near-field contact detector able to detect connections in the range of tens of meters.

Since there is a notable number of 51,113 duty cycle level Bluetooth connections recorded in SHED1 across 9,792 duty cycles, Bluetooth spot connections could be a tool to estimate location data for pairs of nodes in Bluetooth contact where one node has GPS reading while its connected pair does not. This selection criteria was used to enhance the number of GPS records after pruning unrealizable positions using the average speed filter.

As we can see from table 4.4, using Bluetooth (BT) pairs to complete missing GPS records could increase the single GPS Bluetooth pair raises the useful GPS data by 12%.

4.1.4 Mapping to Indexed Location

To calculate the entropy of human mobility, I have to map location into a symbol, and use a series of these symbols to represent the location variation for a person over time.

For a square grid geometric representation, I first set the bin width and then extended if it is necessary to each side equally from the boundary of both latitude and longitude, such that the final frame would have

Table 4.4: Remaining Data Pairs Following Filtering

	FullGPS BtPair	SingleGPS BtPair
RawData	4,842	19,836
Median	4,842	19,836
FilteredMean	4,842	19,836
AvgSpd4	(48km/h)	4,099
18,054	AvgSpd3	3,802
17,221	AvgSpd2	3,455
15,977	AvgSpd1	2,829

integer count of bins spanning latitude and longitude.

Mapping to Synthetic Cell Tower and Wifi Location Index

The website [40] provides the location of all 14,000 cell sites across Canada, including all 32 national and regional carriers, including Rogers, Bell, Telus, and Wind, giving us locations for all cell towers, which would been likely contacted during the study.

The following two figures reflect both large and small scale representations of the area associated with the coordinate of a cell towers. I acquired the location of cell towers from [40], and each cell tower is treated as a landmark used in Voronoi decomposition. As I did not have the access to the routing table used in some previous contributions [58], I assumed a person would secure services through the nearest tower.

The larger region shown in Figure 4.3 showed the set of cell towers that has been referenced in over 99% of records; the remaining 0.1% not shown primarily correspond to towers in the Toronto area. While trips further east and west were made, they were made by air, and contain limited path data, as participants had to turn off their smartphones on an aircraft in keeping with air traffic safety regulations, and are omitted for figure clarity.

The smaller area in Figure 4.4 represents the greater Saskatoon area, covering an area of approximately 570 km2, where 95% of records were concentrated. The Voronoi regions corresponding to cell towers are also shown in these figures to provide a sense of scale for cell tower division of space.

The phone recorded all available Access Points during each duty cycle and the signal strength to each router. Picking the access point with the strongest signal strength at each duty cycle, I obtained the MAC address of the theoretically closest Access Point of the person's current position (ignoring potential obstacles such as walls and other signal interference). Approximate GPS coordinates for WiFi routers were obtained by cross referencing WiFi traces with the GPS table. The location of a WiFi router was estimated to be the GPS location of the participant who observed the WiFi router with the greatest signal strength. 26% of WiFi routers were localized in this manner.

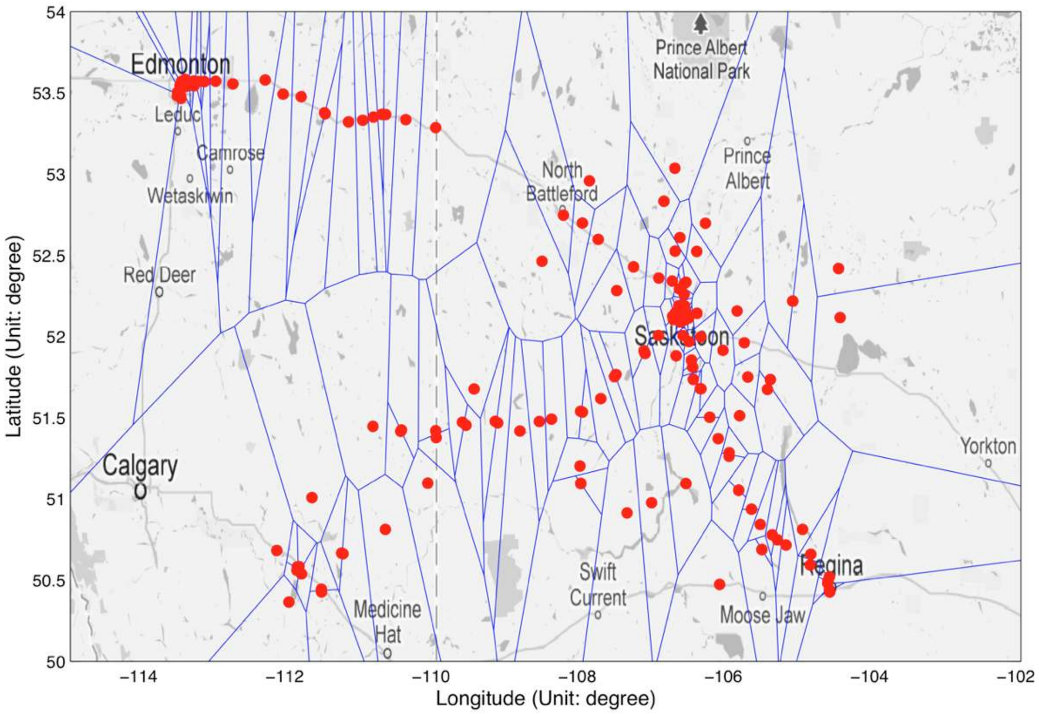

Figure 4.3: Voronoi Decomposition fragment for much of southern Saskatchewan

4.2 Calculating Entropy and Predictability

4.2.1 Entropy

Within the computation of Lempel-Ziv Entropy estimation, I calculate Λ, which is defined as the length of the shortest string that started at the current point that has not yet occurred in the previous part of the string. For example, for the string "ABCAB", if one currently stands at the first "B", then Λ will be equal to 1, as "C" has not previously occurred. This could raise a challenge, because if the current position is beyond the midpoint of the string, there could be a case where I could not find the correct length of the shortest string that has not occurred previously. Still taking string "ABCAB" as an example, if I currently stand at the second "A" and want to calculate Λ, I cannot be sure if Λ is 3 or more – because information after the second "B" are missing. If the character following the second "B" is "D", then Λ will be 3; if it is "C", then Λ will be at least 4. To deal with the issue, there are two approaches.

- Cycling the whole string
- Calculating the entropy on the first half of the string

For my calculation, I choose the later one, as personally I think introducing complete replication of human mobility could lead to more dangerous deviation of result rather than does the risk of not having a long enough length of location records to let the Lempel-Ziv estimator converge to the actual entropy, due

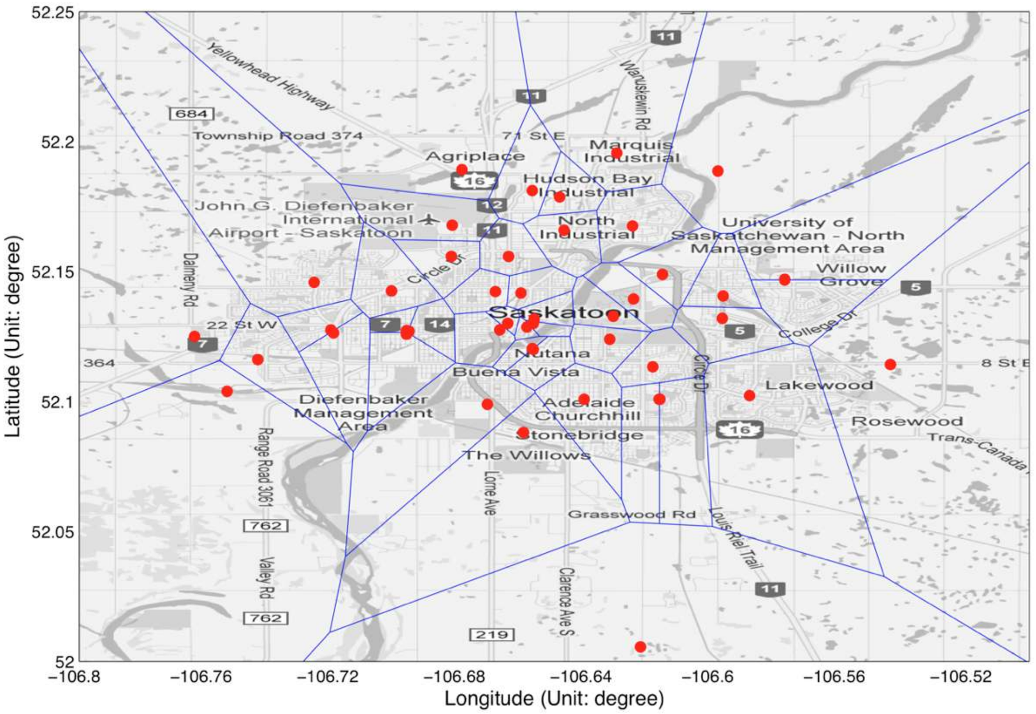

Figure 4.4: Voronoi decomposition fragment for Urban areas of Saskatoon

to the proven quick convergence characteristic of Lempel-Ziv Entropy estimator [35].

When calculating entropy, a method for treating missing location data is required. There are typically two approaches, one is ignore the missing part, the other is to use a place holder (denoted "?") to mark the missing location data. I computed the entropy and predictability with both methods to try to evaluate bounds on their actual values.

4.2.2 Predictability

As the upper-bound of predictability is defined by equations involving actual entropy and number of unique locations, and the equation contains non-linear components, I have employed the non-linear equation solver "fsolve" function in Matlab to get the approximate solutions of the equation with an accuracy of at least 10^{-5}. Because equations could have multiple solutions, to ensure solutions returned are meaningful, I set the initial point of "fsolve" function to be 0.5 (which is in the middle of the meaningful range $[0, 1]$, thus unless the equation could have two solutions within the range of $[0, 1]$ – it will occur no matter which solver is adopted), and also applied a post-check to ensure no solution out of range $[0, 1]$ is returned by the solver function.

4.2.3 Setting for Trip Length

I define path or trip length as the displacement by a participant with minimal stops. The distribution of path lengths could be used to describe the characteristic of trajectories [29]. It can also be used to distinguish random walk and intentional trips between common waypoints [39, 53] on the one hand and Levy walks on the other.

To calculate trip length given a representation of space, one must first define what is meant by a trip. Given that all of my representations are inherently rasterized, I chose to represent a trip as any sequence of records for an individual where the location changes between each record. The length of the trip was either summed over cell sizes for the grid representation, or calculated from transmitter to transmitter in the cell tower and WiFi cases. For example, if there were 5 adjacent cells (A, B, C, D, and E), then sequential records across duty cycles containing ABCD or ABD would correspond to a trip of length 4, while ABBC would correspond to two trips of length two. Making a conservative assumption, I treated missing data in a duty cycle as ending a trip, so the sequence AB?C corresponded to a single trip of length 2.

4.2.4 Deployment to the Socrates Cluster

To calculate the entropy, I need to calculate Λ, which require computation that requires on the order of 10^{11} string comparisons for each person under each scenario. Hence, to parallelize the computation, 1 node of the Socrates cluster (2x Quad core Intel Xeon L5420 at 2.5GHz, and 2GB memory) was assigned to calculate the entropy for all participants for a single spatial representation.

Chapter 5

Results

My results could be categorized into two groups according to the evaluator used to execute comparisons on scenarios. In section 5.1, we focus on the results of entropy and predictability analysis, where I enhanced the regularity of human mobility pattern brought up by [58], demonstrate the similarity between Voronoi decomposed versus square grid decomposed geometric representation, and support the impact of granularity on human mobility predictability. In section 5.2, the trip-length distribution is used to demonstrate how data granularity impacts the ability to capture trips in human mobility records, which further supports findings advanced in 5.1.

5.1 Entropy and Predictability

Within this section, I used heatmaps of locations and entropy boxplots to illustrate the impact of issues such as geometric representation, granularity on human mobility predictability.

Heatmaps in this section can be used to show the empirical distribution of a person's location amongst cells, either Voronoi or grid. The size of a cell is determined by the granularity of location information; for reference-based geometric representations such as a cell tower Voronoi map, the size of a cell is directly affected by the density of landmarks in an area; for coordinate based geometric representations such as square grid maps, the size of a cell is dictated by the accuracy of the measured displacement between target position and reference position (the origin). As shown in the following figure, the area was calculated of Voronoi cells associated with Cell Towers or WiFi as landmarks in Saskatoon proximity area:

Figure 5.1 shows the Saskatoon proximity area, which contains over 97% of GPS data. Focusing our Voronoi convex hull area distribution calculation in Saskatoon proximity area allows us to convert all GPS location into UTM coordinates, so that Voronoi convex hull area in unit of square meters could be acquired.

To avoid overestimating cell size, the analysis excluded cells that reached the margin, because the area of those cells will also be affected by landmarks that haven't been included in the figure. For a cell tower Voronoi map within the overall area, over 90% of cells are larger than 1 km2. The WiFi Voronoi map is more fine-grained than cell tower, as overall, only 2.5% of cells are larger than 1 km2. Nearly 5% of the cells of the WiFi Voronoi map exhibit area less than 10 m2, which is mainly because locations of WiFi Access Points in different floors from the same building might become close when they are projected on to a plane.

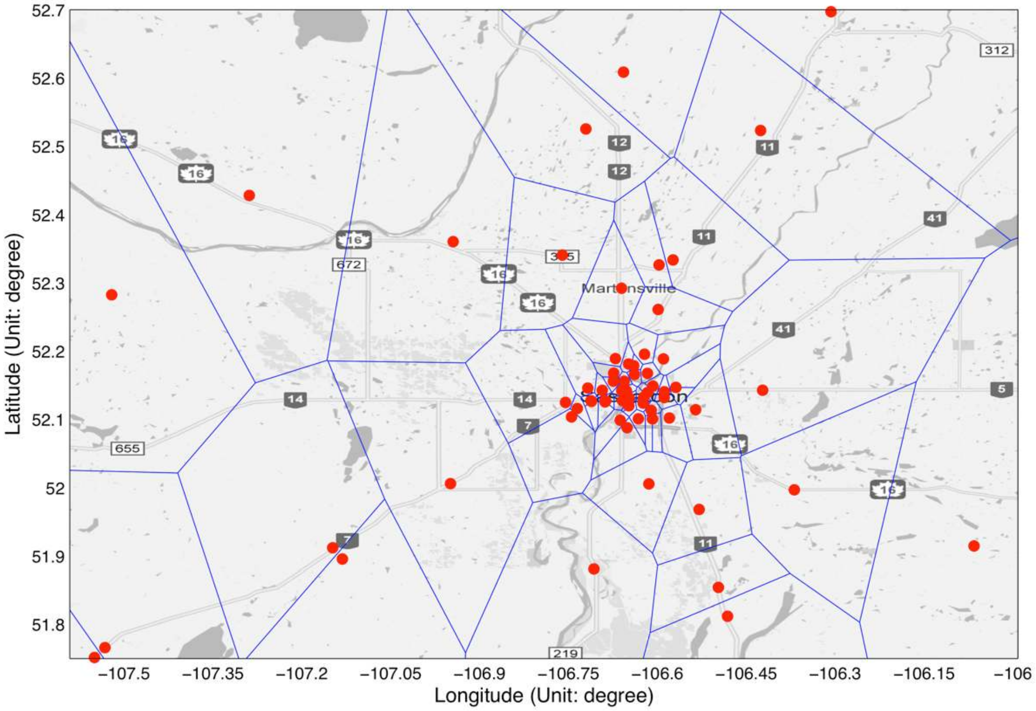

Figure 5.1: Voronoi diagram of Saskatoon proximity area

The area of the cell reflects the spatial error; the larger the cell is, the lower the accuracy that will result from mapping locations to these cells. If there are relatively few times that person's location lies within a larger cell and most of the time it lies within smaller area cells, then Voronoi diagrams could be an efficient tool in representing peoples' locations. However, as shown in the following heatmap figure, there are large-scale cells that exhibit frequent participant occupancy, which indicates that there might be a risk of lower accuracy, due to larger mean cell size.

As shown in Figure 5.3, those cells colored black include over 10k counts of reported participant-dutycycle records, which means that the location of the associated cell tower is at least 10 times more frequently referred to than for the other lighter cells. However the size of those cells is relatively big, which could span over 1 square kilometer. Location information recorded within a cell with area 1km2 makes it difficult to detect the movement under 1km, because it will often be referred to the same landmark; at the same time it could overestimate the fraction of trip trajectories of approximately 1km in length for those points wandering across the cell edge.

Figure 5.4 shows Voronoi decomposition using WiFi access points as landmarks. The density of WiFi access points is much higher than for cell towers, but the shape of the WiFi access point distribution is similar to the cell tower distribution, in that both of them are clustered in urban area but sparse at the rural side. The dispersal of participant-duty cycles occurring in each cell is also similar to that for the cell tower Voronoi heatmap, which includes a predominance of participant occurrence in small size cells, but still exhibits a considerable amount of participant occurrence on whose area is on the order of square kilometers.

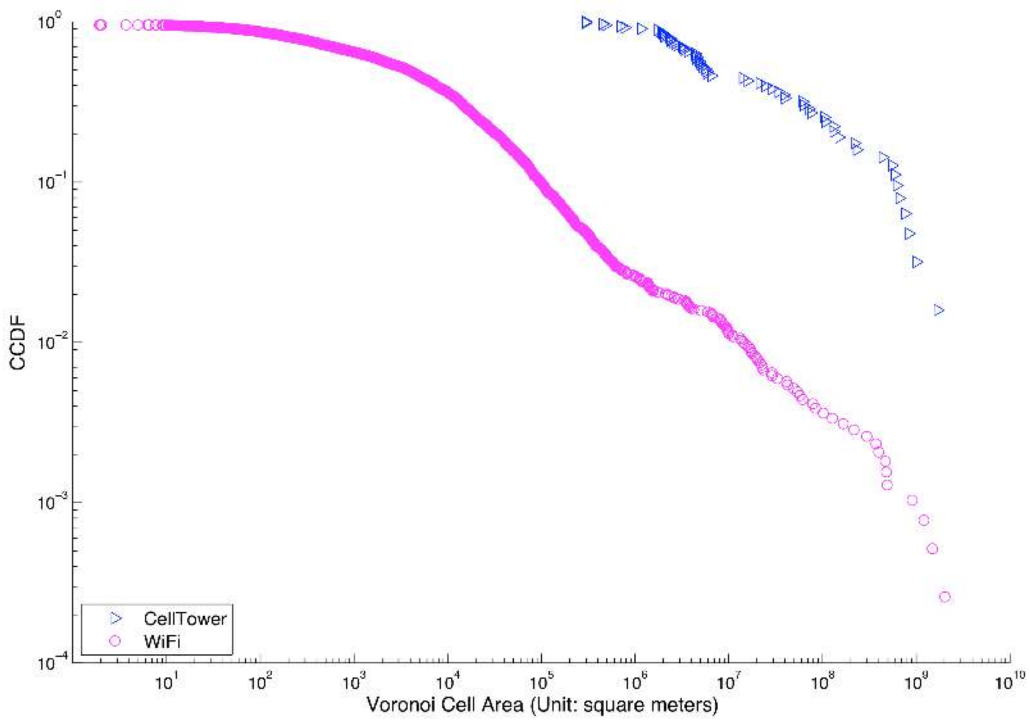

Figure 5.2: Distributions of areas of Voronoi polygons in Saskatoon proximity Area (Shown in Figure 5.1)

Interestingly, I also found that even in those extremely clustered urban areas, with a very fine grained cells, the dissemination of human location exhibits considerable heterogenity. This is another piece of evidence supporting the regularity of human mobility, which is not hard to image, because our mobility in the crowded area is significantly shaped by the path formed by amenities such as corridors, stairs and elevators.

One of the central questions I sought to investigate in this thesis is the degree to which spatial representation impacts the predictability of human motion. If spatial resolution and representation have little impact, it would further support Song et al.'s characterization of Western mobility patterns as inherently predictable, potentially allowing for use of much coarser measurement tools – such as cell tower records – to investigate mobility patterns for planning at all scales. If, however, [16] is correct, then there should be a pronounced resolution effect on entropy, potentially to the point where a phase change occurs at small scales. The entropy was calculated for all representations described here. The results are presented in Figure 5.5 as boxplots, where each distribution is computed across participants. A boxplot of predictability is shown in Figure 12. The boxplot is a descriptive approach to demonstrate the distribution of the population over a single dimension. Within Figure 5.5, there are boxplot showing the predictability distribution over 38 people under 13 different scenarios. The boxes extend from the 25th to 75th percentiles of participant mobility entropy and the maximum whisker length is 1.5 times the 3rd to 1st inter-quartile difference, which corresponds to approximately 99.3% coverage if the data are normally distributed. The line in the center of the box represents the mean.

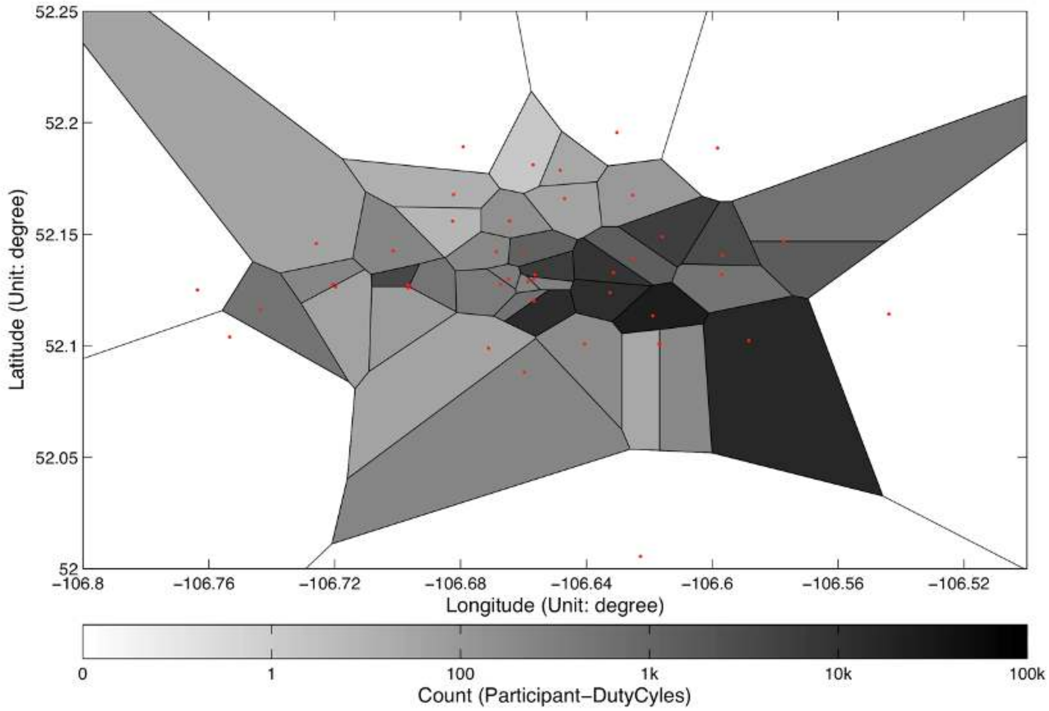

Figure 5.3: Heat Map of Raw Cell Tower Voronoi

It is readily apparent from Figure 5.5 that while the pruning for the average speed filter removed a significant amount of data, it did little to change the entropy. This effect lends confidence that the measurements represent a viable sample of participant mobility entropies, as selectively removing those points least likely to be correct had limited impact. If removing points had a significant impact, one would have to conclude that the calculations were strongly biased by the data quality.

As expected – but in contrast to some previous findings – there were significant resolution impacts, with larger bins – either Voronoi or grid – producing smaller entropies. Indeed, looking only at the means, it appears that quadrupling the area approximately halves the entropy, demonstrating that the entropy increases in step with the resolution, albeit at a slower rate. There appeared to be little sensitivity to the geometry of the representation, as cell tower entropy closely resembles that for the 4 km bin width, approximately the average scale of the Voronoi cell size considered in the Greater Saskatoon area in which most of the records fell, and the entropy for the WiFi Voronoi lies between that for the 62.5m and 250m grid; the latter is a typical nominal range for WiFi routers.

The large variance in the entropies for the smallest grids (15.625 m and 62.5 m bin width, respectively) are likely indicative of both greater mobility at small scales and the increasing impact of sensor noise on the distribution. It is worth noting that the 15.625 m grid is the only case where a zero entropy is not within the 99% confidence interval. I suspect – but cannot conclude – that this is primarily due to the bin size being near the sensor noise floor. At any given duty cycle, the signal could then wander randomly about its neighbors, giving rise to an artificially higher entropy. In fact, the relationship between bin area and entropy

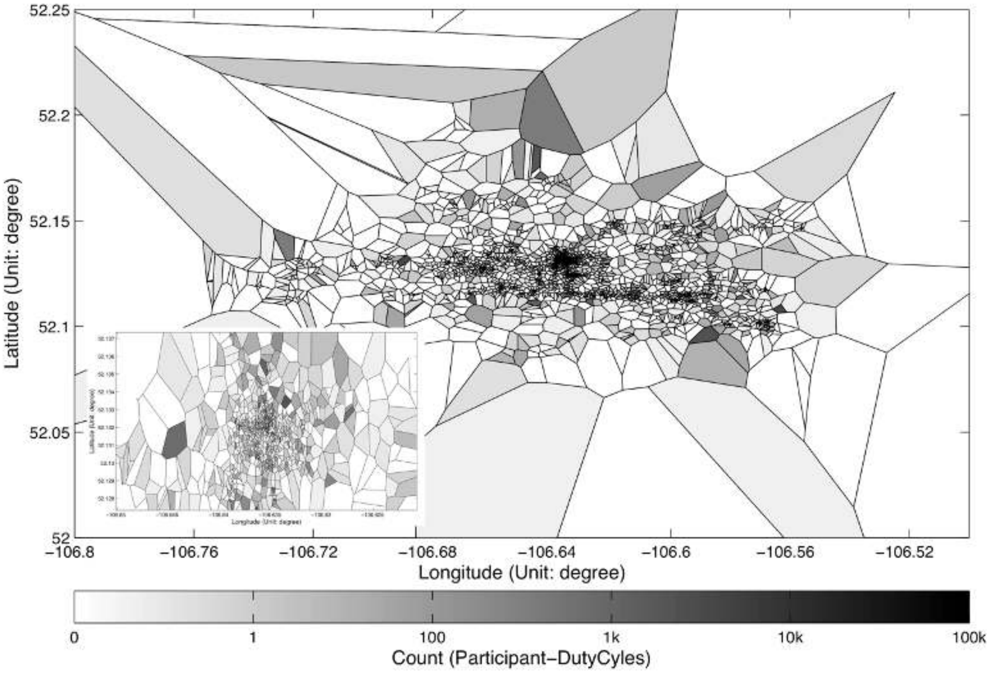

Figure 5.4: Heat Map of Raw WiFi Voronoi

seems to follow an inverse power relationship. For every quadrupling of bin width (with a resulting 16-fold increase in bin area) the number of possible states decreases by a factor of 16; but the entropy only decreases by a factor of 2. To investigate this further, the mean entropies from Figure 5.5 have been regressed against the inverse of the square root of bin dimension, as shown in Figure 5.6.

As is clear from Figure 5.6, there is a strong inverse power relationship in evidence. Another way of considering this relationship is to note that every decrease in cell area of 16 times increases the number of possible states by 16, but participants would only double the number of states that they occupied. The enumeration of the scaling of mobility entropy with area is one of the significant contributions of this work.

To facilitate comparison with other work, the upper-bound of predictability ($\Pi^{\max}$) has also been plotted, which denotes the upper bound for the predictability of an event based on the entropy. A boxplot of predictability is shown in Figure 5.7.

As shown in Figure 5.7 the consistency between predictability data represented in Raw Bin and ASF Bin further underscores the reliability of findings in human mobility [58], and that there is regularity in human mobility pattern even though SHED1 data collection exhibits limitations. At the same time, granularity does impact the predictability of human mobility – the predictability of Wi-Fi has a notable drop compared with CT-Raw and CT-ASF in the "Voronoi" region, and for column "Raw Bin" and column "ASF Bin", with the four-fold reduction of bin area, there is relatively slowly accelerating predictability drop.

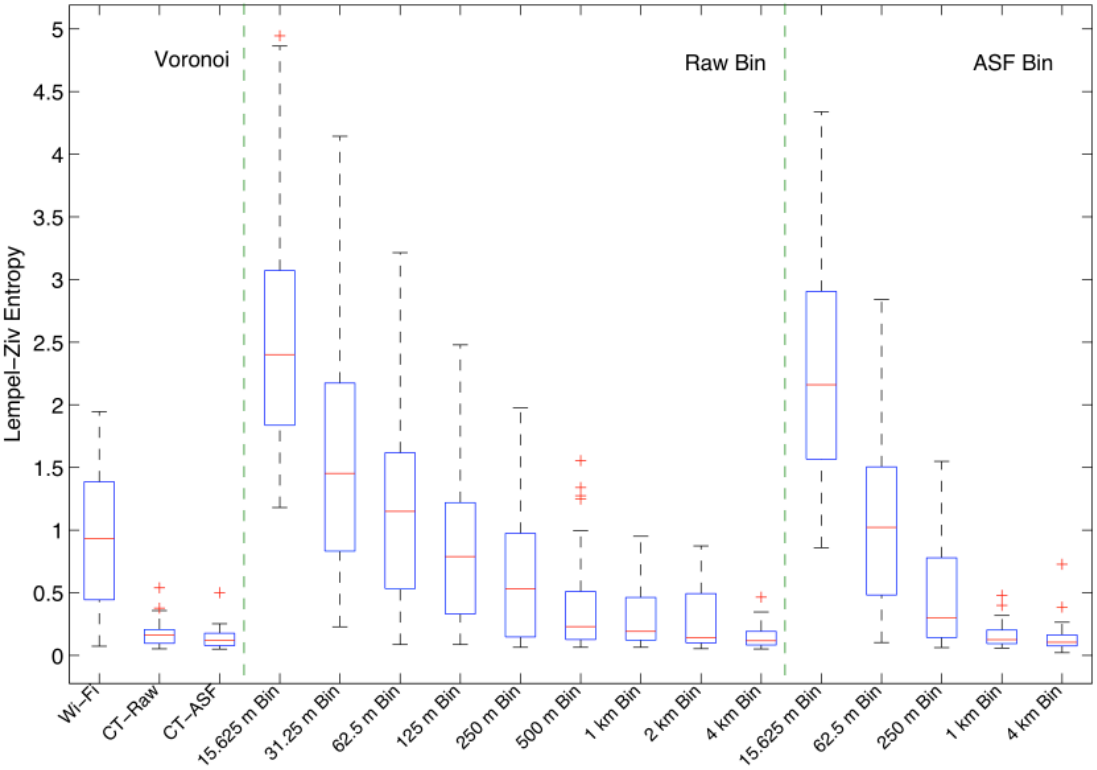

Figure 5.5: Boxplot of entropy variation on geometry representation and granularity

5.2 Trip-Length Distribution

The impact of spatial representation on trip length can provide some insight into the potential impacts of resolution on entropy. If trip length distributions tend to zero density beyond a minimum scale, then little change in entropy would be expected for increasing resolution beyond that point. However, if the minimum trip length corresponds to the resolution of the spatial representation, personally I cannot assume that entropy will saturate with resolution. Trip length complementary cumulative distributions (CCDFs) for 15.625 m, 62.5 m, 250 m and cell tower grids are shown in Figure 5.8.

Most of the curves are characterized by a power law relationship over the 0.1 to 1 km distance interval, with the exception of the cell tower trip length which is characterized by a much steeper decay in the vicinity of 1 km. This is clearly a spatial sampling effect, as the underlying data is the same in all cases. Because of the large mean size of the cell towers Voronoi cells, fewer trips occur, because cell changes in sequential duty cycles are much less common as shown in Figure 13, and because when transitions do occur, they result in much larger increments of distance. Further evidence of the resolution effect can be seen by comparing the three grid resolutions, where fewer trips are reported for ever larger grid bin widths.

The saturation effect evident in all cases at around 2 km bin width is likely an artifact of my sample, composed primarily of graduate students with a dependency on public transport, and the modest size of the

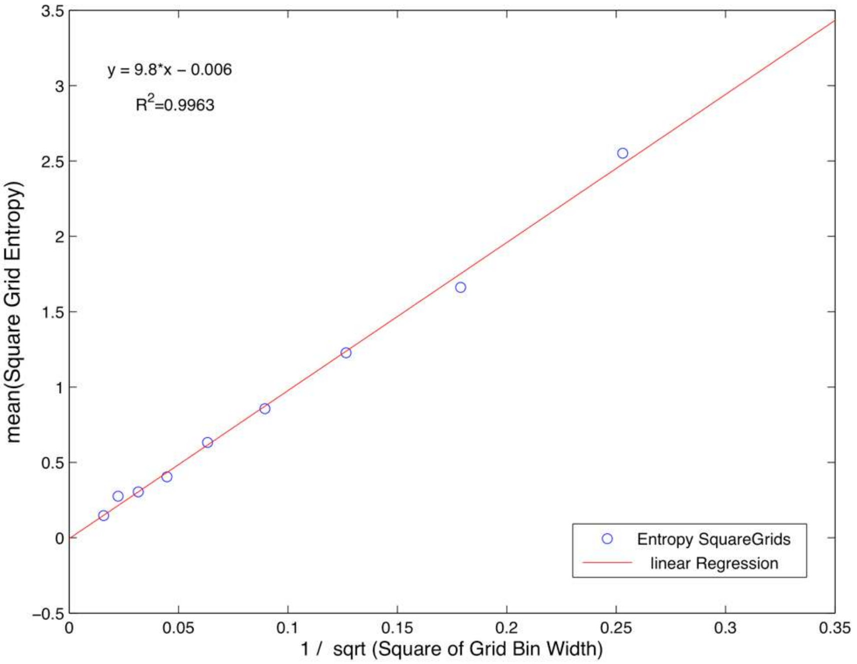

Figure 5.6: Entropy vs bin width regression

city, which can easily be traversed by bus in less than 45 minutes. My strong trip definition – which split trips if either a repeated cell was found or a duty cycle without GPS was found – also likely split some of the longer trips into several smaller trips. Relaxing this assumption would likely bend the tail back towards a power law distribution.

The two-piece power law evident in the 15.625 m grid case is likely due to two effects – an increase in the probability of shorter trips as noted in [16], and noise contributed from the GPS sensor providing the illusion of many short trips. In the 15.625 m and 62.5 m and 250 m grid resolutions, almost 90% of the probability exists between the minimum grid size and 300 m, again confirming our intuition that there should be a significant increase in entropy when considering smaller spatio-temporal granularity (higher spatio-temporal resolution).

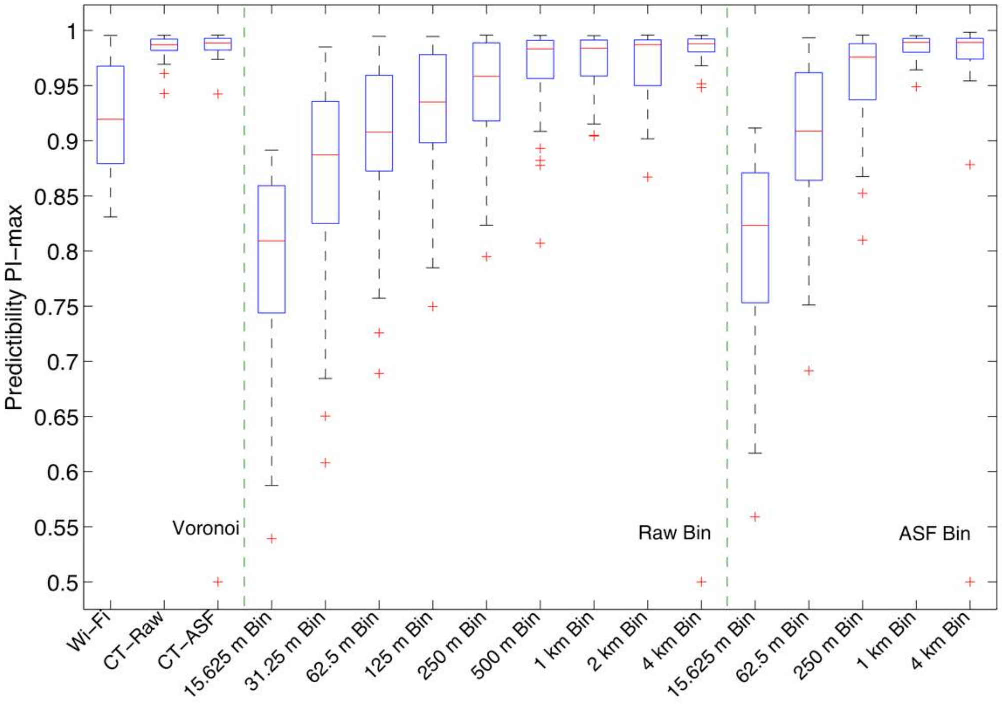

Figure 5.7: BoxplotOfPredictability

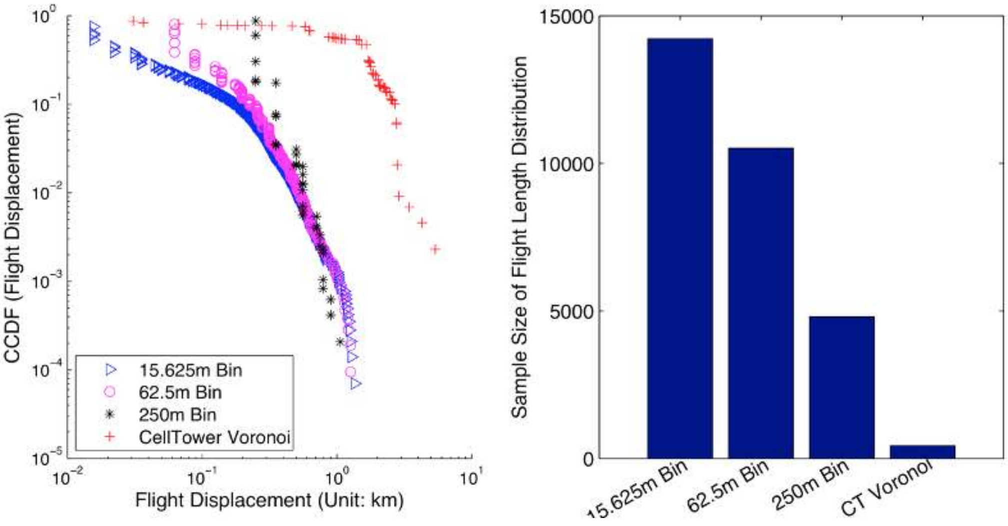

Figure 5.8: Trip-length distribution

Chapter 6

Conclusion

6.1 Contribution

As one looks at human motion in greater detail, more information will he require to capture that mobility. There is no entropy saturation point observable in SHED1 data – suggesting that, from a purely analytic perspective, there is no scale within the resolutions examined at which no additional data is required to represent motion as scale decreases. However, It could be noted that the entropy grows more slowly than resolution, indicating that, for SHED1 data at least, there is a declining return for incremental investment in resolution, as the additional variability captured will grow more slowly than the resolution – and therefore sensing effort – employed. This in turn indicates that the resolution of choice for a particular study should be application – and not computationally – dependent; sufficient resolution for the task should be selected, with the recognition that predictions and conclusions may not apply at finer scales.

There are barely notable interaction between scale size and spatial representation. Voronoi regions generated similar entropies for their scale as rectangular bins. This finding is important, as it suggests that experimental designers can safely choose the most appropriate spatial partition based on the available infrastructure.

While most of the trips recorded with higher resolution representations were found on sub-100 m scales, the difference in entropy and scale size remained largely consistent, echoing the findings at larger scales [58]. While larger cells obscure the shorter trips, the inherent predictability does not increase faster than the number of available states. While it is more difficult to encode the trajectory of one of SHED1's participants at a smaller scale, this is in line with number of states, and not indicative of a phase change in mobility patterns, where entropy would jump significantly.

6.1.1 Insensitivity to Geometry Representation

The most fundamental finding – the fact that entropy scales with resolution (but at a slower rate) – provides guidance to intelligent system design based on the detection of human patterns such as those described in [22, 44]. My analysis suggests that the complexity of representing mobility should scale slower than the resolution, implying that the memory and computational requirements will also scale in a relatively favorable manner, and designers should not be inhibited from employing higher-fidelity data to meet their system's needs.

6.1.2 Sensitivity to Granularity

It could be concluded that there was little difference in predictability due to spatial representation, and that the dominant parameter was the size of a cell – not the shape of a cell – at least between the two most common representations: grid and Voronoi. Given the relative costs in terms of hardware and battery life, and the extremely dense WiFi networks common in even moderately sized cities, personally I believe future researchers should preferentially leverage WiFi for urban mobility data, supplementing with GPS only when in rural areas lacking adequate WiFi coverage.

6.1.3 Insensitivity to Noise and Missing Data

Participant-intensive research in the wild will inevitably produce data with gaps and errors due to participant neglect and sensor failure. My analysis of the data has shown that additionally pruning data based on speed filters did little to change the calculated entropy. This suggests that datasets with smaller participant size taken over a sufficiently large timeframe may be able to provide sufficient resolution to capture the underlying mobility patterns, although this hypothesis will require further data and analysis to confirm.

6.1.4 Limitations

While this research has provided significant insight into the role of spatial resolution and representation on the apparent entropy of human mobility, it suffers from two distinct shortcomings related to the structure of the underlying SHED1 dataset. First, the dataset is predominantly comprised of graduate students, who have unstructured lives and may have a bias towards a higher mobility entropy than average citizens – effects moderated by their low income and associated reliance upon public transit. However, even with this highly flexible population, a lower entropy was found than Song et al. [58] whose reliance on heavy mobile phone users was built into their sampling protocol. Secondly, SHED1 sample size, while comprising a significant proportion of graduate students in the Department of Computer Science, is insufficient to provide broadly generalizable empirical conclusions. However, for the population under study, these findings are reasonable, providing strong motivation for studies of larger and broader populations using similar techniques to more fully elucidate the predictability of human motion. The third shortcoming of the dataset relates to the compliance of the participants. SHED1 dataset captured slightly more than half the available data that would have been obtained with perfect compliance, and only one third of the total possible data had viable GPS measurements. However, I still obtained more than 1,000,000 GPS records at multi-minute resolution, corresponding to a significant sample of the participants' lives.

The use of GPS proximity to cell towers as a measure of cell tower based localization – rather than call records from cell towers themselves, as reported in [58] – is also a limitation. However, Song et al. were forced to use – and accept the associated biases of – call records as their baseline because they did not have access to measurements of participant positions. In a sense, my representations is in an idealized case of what Song et al. were hoping to achieve, and represents an upper bound on the fidelity of data that could be obtained from cellular call records. However, calls placed from inside a building would register with the cell tower, whereas GPS might not. The resolution and scope of the WiFi records provides us with additional confidence that my methods were sufficient to capture the overall trends in mobility patterns.

Finally, it is worth noting that all automated studies cited in this work – including this one – focus exclusively on Western mobility patterns. I do not contend that the mobility patterns of a tribesman in Papua New Guinea would follow similar trends, although differences between culturally and environmentally mediated mobility patterns would make for interesting future anthropological research.

Chapter 7

Summary

This thesis focused on analyzing impact of spatial resolution on human mobility pattern by trying to answer following three questions: will different geometric representation of location (Voronoi and square grid decomposition) lead to an impact on predictability; will finer grained location classification expose previously obscured random-walks and cause a large drop in predictability; will sensor failure-caused data noise and missing data affect the reliability of answers offered to the previous two questions.

In order to answer proposed questions, I designed the experiment using entropy and trip-length distribution as criteria to evaluating on groups mixing by the Cartesian product of Voronoi decomposition, square grid decomposition, (Sparse CellTower landmarking, finer WiFi landmarking as location classification for Voronoi decomposition), (15.625 m to 4 km bin width with scale of 2 as location classification for square grid decomposition, and Raw data, Filtered data, Completed data. I also calculated the trip-length distribution of both geometric decomposition under representative grained location classification.

Experimental results show the absence of obvious differences between Voronoi decomposition and square grid decomposition in representing human mobility pattern. Granularity of location classification has an influence on predictability of human mobility but the drop of predictability is relatively slow when area (Voronoi convex hull area for Voronoi decomposition and grid area for square grid decomposition) belongs to a unique landmark arise. Missing data and noisy data did not jeopardize the reliability of my analysis.

References

[1] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6):643–659, December 2011.

[2] L.A.N. Amaral, A. Scala, M. Barthélémy, and H.E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, 97(21):11149, 2000.

[3] M. Azizyan, I. Constandache, and R. Roy Choudhury. Surroundsense: mobile phone localization via ambience fingerprinting. In *Proceedings of the 15th annual international conference on Mobile computing and networking*, pages 261–272. ACM, 2009.

[4] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–5, January 2006.

[5] R.C.H. Cheng and N.A.K. Amin. Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 394–403, 1983.

[6] Z. Cheng, J. Caverlee, K. Lee, and D.Z. Sui. Exploring millions of footprints in location sharing services. *AAAI ICWSM*, 2011.

[7] Z. Dou, R. Song, and J.R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th international conference on World Wide Web*, pages 581–590. ACM, 2007.

[8] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[9] A. M. Edwards, R. Phillips, N. W. Watkins, M. P. Freeman, E. J. Murphy, V. Afanasyev, S. V. Buldyrev, M. G. E. da Luz, E. P. Raposo, H. E. Stanley, and G. M. Viswanathan. Revisiting lévy flight search patterns of wandering albatrosses, bumblebees and deer. *Nature*, 449(7165):1044–8, October 2007.

[10] A. Einsten. Über die von der molekularkinetischen theorie der warme geforderten bewegung von ruhenden flüssigkeiten suspendierten teilchen. *Ann Phys*, 17:549–560, 1905.

[11] C. Fraser, C.A. Donnelly, S. Cauchemez, W.P. Hanage, M.D. Van Kerkhove, T.D. Hollingsworth, J. Griffin, R.F. Baggaley, H.E. Jenkins, E.J. Lyons, et al. Pandemic potential of a strain of influenza a (h1n1): early findings. *Science*, 324(5934):1557–1561, 2009.

[12] Q. Fu and G. Retscher. Active rfid trilateration and location fingerprinting based on rssi for pedestrian navigation. *Journal of Navigation*, 62(02):323–340, 2009.

[13] M. C. González, C. Hidalgo, and A. L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–82, June 2008.

[14] R.M. Gray. *Entropy and information theory.* Springer Verlag, 2010.

[15] M. Hashemian, D. Knowles, J. Calver, W. Qian, M.C. Bullock, S. Bell, R.L. Mandryk, N.D. Osgood, N.D., and K.G. Stanley. iepi: an end to end solution for collecting, conditioning and utilizing epidemiologically relevant data. In *Proceedings of the 2nd ACM international workshop on Pervasive Wireless Healthcare*, pages 3–8. ACM, 2012.

[16] M.S. Hashemian, K.G. Stanley, D.L. Knowles, J. Calver, and N.D. Osgood. Human network data collection in the wild: the epidemiological utility of micro-contact and location data. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 255–264. ACM, 2012.

[17] M.S. Hashemian, K.G. Stanley, and N. Osgood. Flunet: Automated tracking of contacts during flu season. In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2010 Proceedings of the 8th International Symposium on*, pages 348–353. IEEE, 2010.

[18] S. Havlin and D. Ben-Avraham. Diffusion in disordered media. *Advances in Physics*, 36(6):695–798, January 1987.

[19] S. Hong, I. Rhee, S.J. Kim, K. Lee, and S. Chong. Routing performance analysis of human-driven delay tolerant networks using the truncated levy walk model. In *Proceedings of the 1st ACM SIGMOBILE workshop on Mobility models*, pages 25–32. ACM, 2008.

[20] M. W. Horner and M. E. O'Kelly. Embedding economies of scale concepts for hub network design. *Journal of Transport Geography*, 9(4):255–265, December 2001.

[21] Y. Hu, D. Luo, X. Xu, Z. Han, and Z. Di. Effects of levy flights mobility pattern on epidemic spreading under limited energy constraint. *arXiv preprint arXiv:1002.1332*, 02 2010.

[22] P. Hui, J. Crowcroft, and E. Yoneki. Bubble rap: social-based forwarding in delay tolerant networks. In *Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing*, pages 241–250. ACM, 2008.

[23] S. Ihara. *Information theory for continuous systems*, volume 2. World Scientific Pub Co Inc, 1993.

[24] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva. Directed diffusion for wireless sensor networking. *Networking, IEEE/ACM Transactions on*, 11(1):2–16, 2003.

[25] S. Isaacman and R. Becker. Human mobility modeling at metropolitan scales. *Proceedings of the 10th*, pages 239–251, 2012.

[26] S. Jain, K. Fall, and R. Patra. Routing in a delay tolerant network. *Proceedings of the 2004 conference on Applications, technologies, architectures, and protocols for computer communications - SIGCOMM '04*, page 145, 2004.

[27] B.S. Jensen, J. Larsen, L.K. Hansen, J.E. Larsen, and K. Jensen. Predictability of mobile phone associations. In *Inter. Workshop on Mining Ubiquitous and Social Environments*, 2010.

[28] X. Jiang, Y. Liu, and X. Wang. An enhanced approach of indoor location sensing using active rfid. In *Information Engineering, 2009. ICIE'09. WASE International Conference on*, volume 1, pages 169–172. IEEE, 2009.

[29] T. Karagiannis, J.Y. Le Boudec, and M. Vojnovic. Power law and exponential decay of intercontact times between mobile devices. *Mobile Computing, IEEE Transactions on*, 9(10):1377–1390, 2010.

[30] M. Keeling. The implications of network structure for epidemic dynamics. *Theoretical Population Biology*, 67(1):1–8, 2005.

[31] M.J. Keeling. The effects of local spatial structure on epidemiological invasions. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 266(1421):859–867, 1999.

[32] R. Kitamura, C. Chen, R.M. Pendyala, and R. Narayanan. Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation*, 27(1):25–51, 2000.

[33] J. Klafter, M.F. Shlesinger, and G. Zumofen. Beyond brownian motion. *Physics Today*, 49(2):33–39, 1996.

[34] C. Konstantopoulos, A. Mpitziopoulos, D. Gavalas, and G. Pantziou. Effective determination of mobile agent itineraries for data aggregation on sensor networks. *Knowledge and Data Engineering, IEEE Transactions on*, 22(12):1679–1693, 2010.

[35] I. Kontoyiannis, P.H. Algoet, Y.M. Suhov, and AJ Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to english text. *Information Theory, IEEE Transactions on*, 44(3):1319–1327, 1998.

[36] S. Kosta, A. Mei, and J. Stefa. Small world in motion (swim): Modeling communities in ad-hoc mobile networking. In *Sensor Mesh and Ad Hoc Communications and Networks (SECON), 2010 7th Annual IEEE Communications Society Conference on*, pages 1–9. IEEE, 2010.

[37] D. Kotz and K. Essien. Analysis of a campus-wide wireless network. *Wireless Networks*, 11(1-2):115–133, 2005.

[38] J.S. Lee. Performance evaluation of ieee 802.15. 4 for low-rate wireless personal area networks. *Consumer Electronics, IEEE Transactions on*, 52(3):742–749, 2006.

[39] K. Lee, S. Hong, S.J. Kim, and I. Rhee. Slaw: A new mobility model for human walks. *INFOCOM 2009, IEEE*, pages 855–863, 2009.

[40] Loxcel. Canadian cell tower map. `http://www.loxcel.com/celltower`.

[41] Movable Type Ltd. Calculate distance, bearing and more between latitude/longitude points. `http://www.movable-type.co.uk/scripts/latlong.html`.

[42] A. Madan, M. Cebrian, D. Lazer, and A. Pentland. Social sensing for epidemiological behavior change. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 291–300. ACM, 2010.

[43] D.E. Manolakis. Efficient solution and performance analysis of 3-d position estimation by trilateration. *Aerospace and Electronic Systems, IEEE Transactions on*, 32(4):1239–1248, 1996.

[44] R. M. May. Network structure and the biology of populations. *Trends in Ecology & Evolution*, 21(7):394–399, 2006.

[45] Q. Mei and K. Church. Entropy of search logs: how hard is search? with personalization? with backoff? *Urbana*, 51:61801, 2008.

[46] P.A.P. Moran. The random division of an interval. *Supplement to the Journal of the Royal Statistical Society*, 9(1):92–98, 1947.

[47] P. Mörters and Y. Peres. *Brownian motion*, volume 30. Cambridge University Press, 2010.

[48] M.E.J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(1):016128, 2002.

[49] L.M. Ni, Y. Liu, Y.C. Lau, and A.P. Patil. Landmarc: indoor location sensing using active rfid. *Wireless networks*, 10(6):701–710, 2004.

[50] K. Pearson. The problem of the random walk. *Nature*, 72(1865):294, 1905.

[51] G. Ramos-Fernandez, J.L. Mateos, O. Miramontes, G. Cocho, H. Larralde, and B. Ayala-Orozco. Lévy walk patterns in the foraging movements of spider monkeys (ateles geoffroyi). *Behavioral Ecology and Sociobiology*, 55(3):223–230, 2004.

[52] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong. On the levy-walk nature of human mobility. *IEEE/ACM Trans. Netw.*, 19(3):630–643, 2011.

[53] J. Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of statistics*, 1983.

[54] S.S. Saad and Z.S. Nakad. A standalone rfid indoor positioning system using passive tags. *Industrial Electronics, IEEE Transactions on*, 58(5):1961–1970, 2011.

[55] Y. Shao and M.G. Hahn. Limit theorems for the logarithm of sample spacings. *Statistics & probability letters*, 24(2):121–132, 1995.

[56] A. Smith, H. Balakrishnan, M. Goraczko, and N. Priyantha. Tracking moving devices with the cricket location system. In *Proceedings of the 2nd international conference on Mobile systems, applications, and services*, pages 190–202. ACM, 2004.

[57] C. Song, T. Koren, P. Wang, and A. L. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, September 2010.

[58] C. Song, Z. Qu, N. Blumm, and A.L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[59] C. Song, Z. Qu, N. Blumm, and A.L. Barabási. (supporting online material of) limits of predictability in human mobility, 2010.

[60] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Régis, J.F. Pinton, N. Khanafer, W. Van den Broeck, et al. Simulation of an seir infectious disease model on the dynamic contact network of conference attendees. *BMC medicine*, 9(1):87, 2011.

[61] P.D. Stroud, S.J. Sydoriak, J.M. Riese, J.P. Smith, S.M. Mniszewski, and P.R. Romero. Semi-empirical power-law scaling of new infection rate to model epidemic dynamics with inhomogeneous mixing. *Mathematical biosciences*, 203(2):301–318, 2006.

[62] F. Thomas and L. Ros. Revisiting trilateration for robot localization. *Robotics, IEEE Transactions on*, 21(1):93–101, 2005.

[63] A.R. Tuite, A.L. Greer, M. Whelan, A.L. Winter, B. Lee, P. Yan, J. Wu, S. Moghadas, D. Buckeridge, B. Pourbohloul, et al. Estimated epidemiologic parameters and morbidity associated with pandemic h1n1 influenza. *Canadian Medical Association Journal*, 182(2):131–136, 2010.

[64] G.E. Uhlenbeck and L.S. Ornstein. On the theory of the brownian motion. *Physical Review*, 36(5):823, 1930.

[65] N. Valler, B. Prakash, H. Tong, M. Faloutsos, and C. Faloutsos. Epidemic spread in mobile ad hoc networks: Determining the tipping point. *NETWORKING 2011*, pages 266–280, 2011.

[66] G.M. Viswanathan, V. Afanasyev, and S.V. Buldyrev. Lévy flight search patterns of wandering albatrosses. *Nature*, 1996.

[67] M. von Smoluchowski. Zur kinetischen theorie der brownschen molekularbewegung und der suspensionen. *Annalen der Physik*, 326(14):756–780, 1906.

[68] T. Wei and S. Bell. Indoor localization method comparison: Fingerprinting and trilateration algorithm. In Renee Sieber, editor, *Proceedings of the 2011 Spatial Knowledge and Information Canada Conference*, volume 1, pages 66–68, 2011.

[69] W. Ye, J. Heidemann, and D. Estrin. Medium access control with coordinated adaptive sleeping for wireless sensor networks. *Networking, IEEE/ACM Transactions on*, 12(3):493–506, 2004.

[70] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *Information Theory, IEEE Transactions on*, 24(5):530–536, 1978.