DEPARTMENT OF COMPUTER SCIENCE
SERIES OF PUBLICATIONS A
REPORT A-2015-4

Understanding Urban Human Mobility for Network Applications

Kai Zhao

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public examination in Auditorium A129, Chemicum building, Kumpula, Helsinki on November 6th, 2015 at 12 o'clock noon.

UNIVERSITY OF HELSINKI
FINLAND

Supervisor
Sasu Tarkoma, University of Helsinki, Finland

Pre-examiners
Eiko Yoneki, University of Cambridge, UK
Jukka Nurminen, Aalto University, Finland

Opponent
Vassilis Kostakos, University of Oulu, Finland

Custos
Sasu Tarkoma, University of Helsinki, Finland

Contact information

Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI-00014 University of Helsinki
Finland

Email address: info@cs.helsinki.fi
URL: http://cs.helsinki.fi/
Telephone: +358 2941 911, telefax: +358 9 876 4314

ISSN 1238-8645
ISBN 978-951-51-1692-5 (paperback)
ISBN 978-951-51-1693-2 (PDF)
Computing Reviews (1998) Classification: H.1.2, B.8.2
Helsinki 2015
Unigrafia

Understanding Urban Human Mobility for Network Applications

Kai Zhao

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
kai. zhao@cs.helsinki.fi
https://www.cs.helsinki.fi/en/people/kzhao

PhD Thesis, Series of Publications A, Report A-2015-4
Helsinki, October 2015, 62+46 pages
ISSN 1238-8645
ISBN 978-951-51-1692-5 (paperback)
ISBN 978-951-51-1693-2 (PDF)

Abstract

Understanding urban human mobility is crucial for various mobile and network applications. This thesis addresses two key challenges presented by mobile applications, namely urban mobility modeling and its applications in Delay Tolerant Networks (DTNs).

First, we model urban human mobility with transportation mode information. Our research is based on two real-life GPS datasets containing approximately 20 and 10 million GPS samples. Previous research has suggested that the trajectories in human mobility have statistically similar features as Lévy Walks. We attempt to explain the Lévy Walks behavior by decomposing them into different classes according to the different transportation modes, such as Walk/Run, Bike, Train/ Subway or Car/Taxi/Bus. We show that human mobility can be modelled as a mixture of different transportation modes, and that these single movement patterns can be approximated by a lognormal distribution rather than a power-law distribution. Then, we demonstrate that the mixture of the decomposed lognormal flight distributions associated with each modality is a power-law distribution, providing an explanation for the emergence of Lévy Walks patterns that characterize human mobility patterns.

Second, we find that urban human mobility exhibits strong spatial and temporal patterns. We leverage such human mobility patterns to derive an optimal routing algorithm that minimizes the hop count while maximizing the number of needed nodes in DTNs. We propose a solution framework, called Ameba, for timely data delivery in DTNs. Simulation results with experimental traces indicate that Ameba achieves a comparable delivery ratio to a Flooding-based algorithm, but with much lower overhead.

Third, we infer the functions of the sub-areas in three cities by analyzing urban mobility patterns. The analysis is based on three large taxi GPS datasets in Rome, San Francisco and Beijing containing 21, 11 and 17 million GPS points, respectively. We categorize the city regions into four categories, workplaces, entertainment places, residential places and other places. We show that the identification of these functional sub-areas can be utilized to increase the efficiency of urban DTN applications.

The three topics pertaining to urban mobility examined in the thesis support the design and implementation of network applications for urban environments.

Computing Reviews (1998) Categories and Subject Descriptors:
H.1.2 [User Machine Systems]: Human Factors
B.8.2 [Performance Analysis and Design Aids]

General Terms:
Urban Computing, Human Mobility, Mobile Computing, Mobile Social Networks

Additional Key Words and Phrases:
Urban Human Mobility, Delay Tolerant Networks, Mobile applications

Acknowledgements

First and foremost, I want to thank my supervisor Prof. Sasu Tarkoma for his continuous support of my Ph.D. study and research. He has been a tremendous mentor for me, with his inspiring and invaluable guidance, insights and advice over the past five years.

Besides my advisor, I would like to thank Prof. Vassilis Kostakos for being my opponent, and Dr. Eiko Yoneki and Prof. Jukka Nurminen for their thorough pre-examination on my thesis.

I would like to thank my co-authors, Weixiong Rao, Mirco Musolesi, Pan Hui, Samuli Hemminki, Aaron Yi Ding, Petteri Nurmi, Yan Zhang and Eemil Lagerspetz, for their warm support, encouragement and insightful comments during my Ph.D. study in Finland. I would like to express my special appreciation and thanks to Prof. Weixiong Rao, who guided and led me during my first two years study.

I would like to express appreciation to the HECSE, FIGS and DoCS funding, for supporting my travels to the conference and visiting to other universities. My work would not have been possible without the funding from the Digile Internet of Things research program and ESENS project.

I would also like to thank all my friends in Finland. Thank you for accompanying me for the last few year and for the wonderful times we shared together.

Special thanks to my parents Yanjun Zhao and Guangmin Yuan, who bore me, raised me, supported me, taught me, and loved me. Finally, I am heartily thankful to my beloved wife, Zhe Chen, for her love, encouragement, inspiration and spiritually support throughout my life.

In Helsinki, October 20, 2015
Kai Zhao

Contents

List of Reprinted Publications

Research Paper I: Kai Zhao, Mirco Musolesi, Pan Hui, Weixiong Rao and Sasu Tarkoma. Explaining the Power-law Distribution of Human Mobility Through Transportation Modality Decomposition. In Nature Scientific Reports (NATURE SREP), Volume 5, Article number: 9136, Nature Publishing Group, March 2015.

Contributions: The present author Kai Zhao designed the research based on the initial idea by Kai Zhao and Sasu Tarkoma. The present author executed the experiments, performed statistical analyzes, and prepared the figures. Kai Zhao and Sasu Tarkoma wrote the manuscript. All the authors reviewed the manuscript.

Research Paper II: Weixiong Rao, Kai Zhao, Eemil Lagerspetz, Pan Hui and Sasu Tarkoma. Energy-Aware Keyword Search on Mobile Phones. In the Proceedings of ACM Special Interest Group on Data Communication, Mobile Cloud Computing workshop (SIGCOMM MCC), ACM, pages 59-64, August, 2012.

Contributions: The algorithmic development of the hybrid keyword search approach is a joint work with Weixiong Rao. The present author also contributed to building the mobile offloading system between mobile phones and a cloud server.

Research Paper III: Kai Zhao, Weixiong Rao, Yan Zhang, Pan Hui and Sasu Tarkoma. Towards Maximizing Timely Content Delivery in Delay Tolerant Networks. In IEEE Transactions on Mobile Computing (TMC), Volume 14, Number 4, pages 755-769, 2015.

Contributions: The present author Kai Zhao is the corresponding author of the journal paper. The original idea of the Ameba content-based DTNs algorithm is given by Weixiong Rao. The analysis of the Infocom06, MIT Reality and UCSD dataset, the development of the distributed Ameba and location-based Ameba algorithm are given by the present author. The author also contributed significantly to the writing of the article.

Research Paper IV: Kai Zhao, Mohan Prasath Chinnasamy, Sasu Tarkoma. Automatic City Region Analysis for Urban Routing. In the Proceedings of IEEE Conference on Data Mining, Mobility Analytics from Spatial and Social Data workshop (ICDM MASS), 7 pages, November 2015.

Contributions: The present author designed the research. The present author and Mohan Prasath Chinnasamy conducted the experiments, performed statistical analyzes, and prepared the figures. Kai Zhao and Sasu Tarkoma wrote the manuscript. All the authors reviewed the manuscript.

Chapter 1

Introduction

Understanding human mobility is crucial for epidemic control [75, 5, 8, 26], urban planning [125, 114], traffic forecasting systems [58, 38] and, more recently, various mobile and network applications [46, 121, 124, 51, 30]. Nowadays, a variety of urban human mobility data have been gathered and published. The pervasive GPS data can be collected by mobile phones. A mobile operator can track people's movement in cities based on their cellular network location. This urban human mobility data contains rich knowledge about locations and can help in addressing many urban challenges such as traffic congestion or air pollution problems. This thesis aims to utilize the knowledge of urban human mobility patterns to improve the performance of urban network applications. Urban human mobility patterns [6, 103, 36, 45, 77, 64, 59] pertain to how people move in cities, for example, walking, biking, driving and utilizing public transportation.

Delay Tolerant Networking (DTNs) [119, 91, 105, 57, 32, 63, 112] is an enabler for the urban network applications. It provides intermittent communication for humans with mobile devices (vehicles, mobile phones, etc.), by exchanging data through short-range communications such as Bluetooth or WiFi direct, which can significantly reduce mobile data traffic of cellular networks. The data transferred in DTNs is delay tolerant, such as weather forecasts, football score or regional information. Since humans carry their mobile devices everywhere everyday, understanding and utilizing urban human mobility can help in delivering the data in DTNs more efficiently [19, 88, 106, 87]. Urban DTNs provide complementary caching and offloading abilities for the congested cellular networks in a city. They can also provide the basic network support during disasters, such as earthquakes or sudden power failures. In addition, urban DTNs also benefit from the huge number of people living inside a city, e.g., tens of thousands of people gather together for a football match. These large numbers of people increase the network density with their mobile devices allowing the network to operate faster and with more messages.

This thesis addresses the two key challenges presented by mobile applications, namely urban mobility modeling and its network applications especially DTN applications [83]. First, we present three findings concerning urban human mobility:

- Urban human mobility follows different log-normal distributions with different transportation modes, such as car, subway or bike.
- Urban human mobility exhibits strong spatial and temporal patterns.
- Urban taxi traffic correlates with the functions of city areas.

Then we show that we can utilize our knowledge of these urban human mobility patterns to improve the performance of urban applications that require delay tolerant operation.

1.1 Motivation

Nowadays, a variety of urban human mobility data have been gathered and published due to the significant growth of sensing technologies and large-scale computing infrastructures. This urban human mobility data contains rich knowledge about locations and can help in addressing many urban challenges. For example, understanding human movements inside a city can help forecasting of the traffic. Another example is that we can identify the functions of locations by the means of the transitions between these locations, e.g., people usually go to work during daytime on weekdays, and visit shopping centers after work.

Currently mobile and cellular networks are heavily utilized by mobile applications. As an example, we can consider a photo post in Facebook[1]. One picture taken by a smartphone is typically between 2 MB and 3 MB. Considering the huge number of users who post millions of photos on Facebook everyday, the total traffic over mobile networks is high. According to AT& T, its network has seen a 5,000 percent surge of mobile data transfer[2]. Thus, it is important to design and deploy techniques that alleviate network bandwidth issues. DTNs provide the basis for a promising optimization technique that offloads cellular network traffic to the opportunistic DTNs. DTNs use complementary network communication technologies (WiFi direct, Bluetooth) for delivering the data that is originating from the cellular network and destined for the cellular network subscribers.

Motived by the availability of large-scale urban mobility data and the current urban cellular network congestion problem, this thesis aims to utilize the

[1] `www.facebook.com`

[2] `http://www.att.com/gen/press-room?pid=17961&cdvn=news&newsarticleid=30838`

knowledge of urban human mobility patterns to improve the performance of urban DTNs. The network applications of the urban DTNs include offloading of delay-tolerant traffic from the cellular network, content delivery in public events where traditional networks become congested, environmental sensing, and providing data delivery for environments that do not have ubiquitous cellular connectivity.

1.2 Problem Statement

Human mobility has been studied for a very long time. In 1885, the publication of The Laws of Migration [84] in the Journal of the Royal Statistical Society can be considered as the first modern attempt to understand human mobility. Due to the significant growth of mobile phones, the study of human mobility has significantly changed. Mobile phones utilize cell tower information and the Global Positioning System (GPS) for fine-grained location tracking. Billions of people carry their phone every day, which provides a large amount of data on human movement. In 2008, one of the first large scale human mobility studies based on mobile phones was published in Nature [39]. By studying cell phone user's locations it was shown that trajectories in human mobility have statistically similar features to Lévy Walks [39]. According to the this model, human movement contains many short flights and some long flights, and these flights follow a power-law distribution. Similar results have been published in an earlier work in Nature in 2006 by studying the tracing of bank notes [15]. Rhee et al. published their study of human mobility at the IEEE Infocom conference in 2008 [87]. They also demonstrated that human walk patterns closely follow Lévy Walks patterns based on approximately one thousand hours of GPS trace studies in various outdoor settings including two different college campuses, a metropolitan area, a theme park and a state fair [88]. Later studies [110, 44, 56, 118, 90, 93] also identified the Lévy Walks patterns of human mobility and the researchers propose their explanations of the reasons behind them. The first explanation for the Lévy Walks patterns was given by Marta et al. [39] and later Yan et al. [110] provided additional insight by examining the individual mobility patterns. Our work helps to understand the formation of the pattern by decomposing it into transportation mode specific segments. The relationship of the transportation modes and the emergence of the Lévy Walks pattern was not studied before our work presented in this thesis.

DTNs [119, 91, 105, 57, 32, 63, 112] provide intermittent communication for humans with mobile devices (vehicles, mobile phones, etc.), by exchanging content through short-range communications such as Bluetooth or WiFi direct. The impact of human mobility on network applications, especially on DTNs, has been studied recently [62, 47, 111, 22, 52]. Many DTN routing algorithms utilize the power of human mobility patterns to disseminate the content in the DTN more effectively and efficiently.

Vahdat et al. proposed the first Flooding-based algorithm in DTNs called Epidemic routing [100], and many later human-mobility-based routing algorithms are based on this algorithm. The key idea of Epidemic routing is to utilize random pair-wise message exchanges among mobile devices for achieving eventual message delivery. The protocol aims for the maximization of the message delivery rate, minimization of the message latency, and minimization of the total resources spent in the message delivery process.

To utilize the mobility property of DTNs, the ProPHET protocol [67] leveraged history information for calculating the delivery probability for encountered nodes and used this information for selecting carriers that maximize the delivery rate. The Spray and Wait algorithm first sprays the data to a random set of carriers, and then each data carrier waits for encountering the data destinations [95]. The paper [13] presented a Markovian model for developing a utility function for data dissemination in DTNs. In terms of location-aware dissemination in DTNs, Fan et al. [33] analyzed mobile users' movement and observed that mobile users usually visit several locations regularly rather than moving randomly. They formulated the data dissemination problem in terms of a superuser that broadcasts data to other users in the network. Based on user movement data, they proposed an efficient algorithm for constructing superuser trajectories that either minimize total duration or maximize dissemination ratio.

We observe that DTNs have been an active research topic for the last decade; however, urban human mobility has not yet been extensively studied in this context. This thesis focuses on the following research questions:

- RQ1. How can we model urban human mobility?
- RQ2. How to model and optimize mobile phone energy usage with applications supporting both local and remote processing?
- RQ3. How can human mobility be used to improve network application efficiency?
- RQ4. To what extent does an urban human mobility model improve network application efficiency?

1.3 Methodology

Table 1.1 shows an overview of the research methodology used in this thesis. In PI, we build a human mobility model based on two real-life GPS datasets containing approximately 20 and 10 million GPS samples with transportation mode

Research Questions	Methodology	Publications
RQ1. How can we model urban human mobility?	Statistical data analysis and stochastic modeling	PI: Explaining the Power-law Distribution of Human Mobility Through Transportation Modality Decomposition. (Nature SREP 15)
RQ2. How to model and optimize mobile phone energy usage with applications supporting both local and remote processing?	Creating a prototype for modeling networking and energy.	PII: Energy-Aware Keyword Search on Mobile Phones. (Sigcomm MCC 12)
RQ3. How can human mobility be used to improve network application efficiency?	Routing algorithm and simulation	PIII: Towards Maximizing Timely Content Delivery in Delay Tolerant Networks. (TMC 14)
RQ4. To what extent does an urban human mobility model improve network application efficiency?	Large-scale data analysis	PIV: Automatic City Region Analysis for Urban Routing. (ICDM MASS 15)

Table 1.1: The methodology of this thesis contains empirical measurement and statistical analysis.

information. Previous research has suggested that the trajectories in human mobility have statistically features similar to Lévy Walks. We explain the Lévy Walks behavior by decomposing them into different classes according to the different transportation modes, such as Walk/Run, Bike, Train/ Subway or Car/Taxi/Bus. Our transportation-decomposed Lévy Walks model deepens our understanding of the human mobility in the city environment with different transportation modes. This answers RQ1, how can we model urban human mobility?

In PII, we model the mobile phone energy usage with applications supporting both local and remote processing by building a keyword search prototype for mobile phones. We show that a hybrid computation task division approach between local device and remote server can reduce the energy usage of mobile phones significantly. This answers RQ2, how to model and optimize mobile phone energy usage with applications supporting both local and remote processing? This result indicates that hybrid solutions that combine local and remote processing are suitable for energy efficient keyword searching. This result is useful for creating hybrid DTN applications that can leverage remote servers for reducing energy consumption in low power situations.

DTN routing protocols aim for the delivery of a message from a source node to the destination through a series of opportunistic pair-wise encounters. Thus DTNs do not rely on a complete end-to-end path in data delivery, but rather the path is formed at runtime by encountering other nodes. Thus the DTN routing protocols adopt the ”store and forward” approach that provides delay tolerance through message buffering. The key metrics for optimizing DTN protocols include energy consumption, number of routing steps (or hops), and the maximization of the message delivery rate.

The overall DTN design consists of two subproblems, namely the target-set problem [42] and the routing problem [83]. The former pertains to the selection of the initial data carrier nodes and the latter relates to the runtime routing behaviour and its characteristics. Typically both problems are solved in order to minimize the store and forward steps while maximizing the number of devices that receive the desired data. The thesis addresses these two problems for environments with urban human mobility.

In PIII, we develop a routing algorithm, Ameba, for solving the routing problem in DTNs. We find that human mobility exhibits strong spatial and temporal patterns. In Ameba, we leverage human mobility patterns for deriving an optimal routing hop count for each message in order to maximize the number of carrier nodes. Simulation results with experimental traces indicate that Ameba achieves a delivery ratio comparable to a Flooding-based algorithm, but with only 3% energy cost. This result answers RQ3, how can human mobility be used to improve network application efficiency?

In the DTN target-set problem [42], the protocol selects a subset of devices that are carriers of the data. The devices in the subset then further distribute the data to other devices through opportunistic DTN communications. How can we choose the initial target-set for maximizing the number of devices that further receive the desired data is called the target-set problem in DTNs [42]. In PIV, we analyzed the temporal taxi mobility patterns and inferred the functions [37, 43, 69, 108] of the sub-areas in three cities. We then showed that the identification of these functional sub-areas can be utilized to solve the target-set problem. This result answers RQ4, to what extent does an urban human mobility model improve network application efficiency?

1.4 Thesis Contributions

This thesis consists of four scientific articles presented in Table 1.1. In this section, we summarize the contributions of the four articles.

The contribution of study PI is twofold. First, we built an urban mobility model and extracted the distribution function of displacement with different transportation modes. This is important for many applications that model and predict urban movement [46]. Our result deepens the understanding of urban human mobility with different transportation modes. The transportation mode information can also help us enhancing the prediction of the next place the user will visit, which can also improve the DTN routing efficiency. Second, we demonstrate that the mixture of different transportation modes can be approximated with a truncated Lévy Walks. This result is a step towards explaining the emergence of Lévy Walks patterns in human mobility.

In PII, we built an energy model for keyword search in the mobile environment and examine there candidate solutions for the search problem. The proposed hybrid approach adaptively splits the keywords of queries into two subsets, such that one subset is answered locally by the mobile phone, and another is offloaded to a remote server. Our experimental results verify that the hybrid approach outperforms the two other extremes.

In PIII [83], we developed a DTN solution framework, namely Ameba, for timely content delivery in DTNs by leveraging human mobility patterns in city settings. The basic idea is to leverage the mobility patterns of mobile devices in order to improve the developed forwarding utility and distributed relay algorithm. Based on the study of three DTN trace files, we found that (i) people visiting different locations exhibit strong spatial properties (that is, the participants frequently visit a small number of hot areas, and rarely visit the remaining areas), and (ii) people visiting different locations also exhibit strong temporal properties (e.g., the majority of participant visits are clustered during some specific periods).

In PIV we categorized city regions into four kinds of places, workplaces, entertainment places, residential places and other places. We show that the identification of these functional sub-areas can help us deliver data in urban DTNs more efficiently. The contribution of paper PIV is threefold. First, we find that there is a high correlation between the road networks and taxi visits. This is an important result for many applications, because the taxi visits can be seen as a proxy for the road network. In previous research, the authors have used the road network for dividing the city into subareas. With this new finding our approach can quickly divide the city into subareas without the complex road network (millions of nodes and edges) information as an input. Second, we provide a novel association rule-based method for detecting the mobility patterns (functions) of the sub-regions inside the city. Third, we can leverage the functions of sub-areas and urban mobility pattern for enhancing urban DTN routing.

1.5 Thesis Structure

This thesis consists of the original publications PI - PIV and the present introduction. Chapter 2 describes how we build a human mobility model based on two real-life GPS datasets containing approximately 20 and 10 million GPS samples with transportation mode information. Chapter 3 and Chapter 4 introduce our findings that human mobility exhibits strong spatial and temporal patterns and how can we leverage these human mobility patterns to increase the routing performance in DTNs. Chapter 5 concludes the thesis.

Chapter 2

Urban Mobility Modeling with Transportation Information

This chapter discusses the domain of human mobility modeling especially the Lévy Walks model and our urban Lévy Walks model with transportation information. This chapter answers RQ1: How can we model urban human mobility?

2.1 Human Mobility Modeling

2.1.1 Overview

Random Way Points (RWP) [10, 11, 113] or random walk models such as Brownian motion [19, 41, 55], Markovian mobility [9, 2] and Lévy Walks [88, 110, 44, 56] are the most commonly used mobility models in computer networking research.

In RWP models [10, 11, 113], the mobile nodes move randomly and freely without any restrictions. The destination, speed and direction are all chosen randomly and independently of the other nodes. In Brownian motion [19, 41, 55], the mobile nodes move with a mean flight and a mean pause time between flights. A flight is defined as the longest straight-line trip of a person from one location to another without a directional change or pause. In Brownian motion, the flights are normally distributed.

Recent research has shown that trajectories in human mobility have statistically similar features as Lévy Walks by studying the tracing of bank notes [15], cell phone users' locations [39] and GPS traces [88, 110, 44, 56]. According to the Lévy Walks model, human movement contains many short flights and some long flights, and these flights follow a power-law distribution.

Although recently human mobility has been empirically observed to exhibit Lévy flight characteristics and behaviour with power-law distributed jump size

[110, 44, 56], the fundamental mechanisms behind this behavior has not yet been fully explained. Later studies propose explanations for the emergence of the Lévy Walks pattern. In [110] Yan et al. observed that the individual human mobility patterns do not follow Lévy Walks and Lévy Walks are due to the aggregation of individual mobility patterns. The hierarchy of traffic systems [44] and road networks [56] are also possible reasons behind the Lévy Walks. Recent research results [65, 66] investigated the case of a single transportation mode (taxi) and they found that the scaling of human flights is exponential. They proposed that this is because few people tend to travel long distances by taxi due to economic considerations.

In Table 2.1 we summarize the related human mobility articles (including our paper PI) and their contributions. We note that the flight is defined as the longest straight-line trip of a person from one location to another without a directional change or pause.

2.1.2 Impact of Human Mobility Model on DTNs

In DTNs, whenever mobile devices (vehicles, mobile phones, etc.) encounter each other, they exchange content via short-range communications (e.g., Bluetooth or WiFi). Since people carry their mobile devices everywhere everyday, human mobility model plays an important role in DTNs. The impact of human mobility in DTNs has been investigated with simulations. The choice of the mobility model has a significant impact on the behaviour and performance of a DTN algorithm.

Lévy Walks provide a more accurate mobility model for DTNs compared to other existing models. Other existing models such as RWP are not based on the real human mobility studies (see Table 2.1) so that they do not reflect how people move in real life. The RWP model also does not emulate heavy-tail statistical features of human mobility. The heavy-tail tendencies of the Lévy Walks model induce heavy-tail routing delays and throughput [88], thus the routing performance in a RWP model in DTN studies tends to be overestimated compared to a Lévy Walks model.

Recent papers [106] also investigate the inherent properties of data dissemination in DTNs based the Lévy Walks Model. For example, the distribution of minimum time needed for spreading the information to a given region, or the probability bound of the earliest time at which the information arrives can also be estimated based on the Lévy Walks Model.

2.2 Lévy Walks Decomposed by Transportation Modes

In this section, we model the Lévy Walks behaviour observed in human mobility patterns by decomposing them into different classes according to the different

Publications	Dataset, Measurement Errors	Flight distribution	Transportation Mode	Explanation
[10]	None	Random	None	None
[41]	None	Normal	None	None
[15]	Bank Notes, 10-100 km	Power-Law	None	None
[39]	Celluar Tower Location, 2-3 km	Power-Law	None	Temporal and spatial regularity
[88]	GPS, 5-10 m	Power-Law	None	None
[110]	GPS, 5-10 m	Power-Law	None	Aggregated individual mobility
[44]	GPS, 5-10 m	Power-Law	None	Traffic systems
[56]	GPS, 5-10 m	Power-Law	None	Street network
[65, 66]	GPS, 5-10 m	Exponential	Yes, taxi transportation mode	Population Density
PI	GPS, 5-10 m	Power-Law and log-normal	Yes, selected transportation modes	Lévy Walks decomposed by transportation modes

Table 2.1: Comparison of recent human mobility articles and our article PI

transportation modes, namely Walk/Run, Bike, Train/Subway or Car/Taxi/Bus. According to the Lévy Walks model, human movement contains many short flights and some long flights, and these flights follow a power-law distribution. Our analysis is based on two real-life GPS datasets containing approximately 10 and 20 million GPS samples with transportation mode information. We show that human mobility can be modelled as a mixture of different transportation modes, and that these single movement patterns can be approximated by a lognormal distribution rather than a power-law distribution. Then, we demonstrate that the mixture of the decomposed lognormal flight distributions associated with each modality is a power-law distribution, providing an explanation for the emergence of Lévy Walks patterns that characterize human mobility patterns.

2.2.1 Overview

According to the Lévy Walks model, human movement contains many short flights and some long flights, and these flights follow a power-law distribution. Intuitively, these long flights and short flights reflect different transportation modalities. Figure. 2.1 shows a person's one-day trip with three transportation modalities in Beijing based on the Geolife dataset [122]. Starting from the bottom right corner of the figure, the person takes a taxi and then walks to the destination in the top left part. After two hours, the person takes the subway to another location (bottom left) and spends five hours there. Then the journey continues and the person takes a taxi back to the original location (bottom right). The short flights are associated with walking and the second short-distance taxi trip, whereas the long flights are associated with the subway and the initial taxi trip. Here a flight is the longest straight-line trip from one point to another without change of direction [88, 56]. One trail from an origin to a destination may include several different flights (Fig. 2.1). Based on this simple example, we observe that the flight distribution of each transportation mode is different.

In PI, we propose to model the Lévy Walks behavior observed in human mobility patterns by decomposing them into different classes according to the different transportation modes: Walk/Run, Bike, Train/Subway, or Car/Taxi/Bus. Our analysis is based on two large GPS datasets, the Geolife and Nokia MDC datasets (approximately 10 million and 20 million GPS samples respectively), both containing transportation mode information such as Walk/Run, Bike, Train/Subway or Car/Taxi/Bus. The four transportation modes (Walk/Run, Bike, Train/Subway and Car/Taxi/Bus) cover the most frequently used human mobility cases. We determined the flight length distributions for different transportation modes. We fitted the flight distribution of each transportation mode according to the Akaike information criteria [17] in order to find the best fit distribution.

We showed that human movement exhibiting different transportation modali-

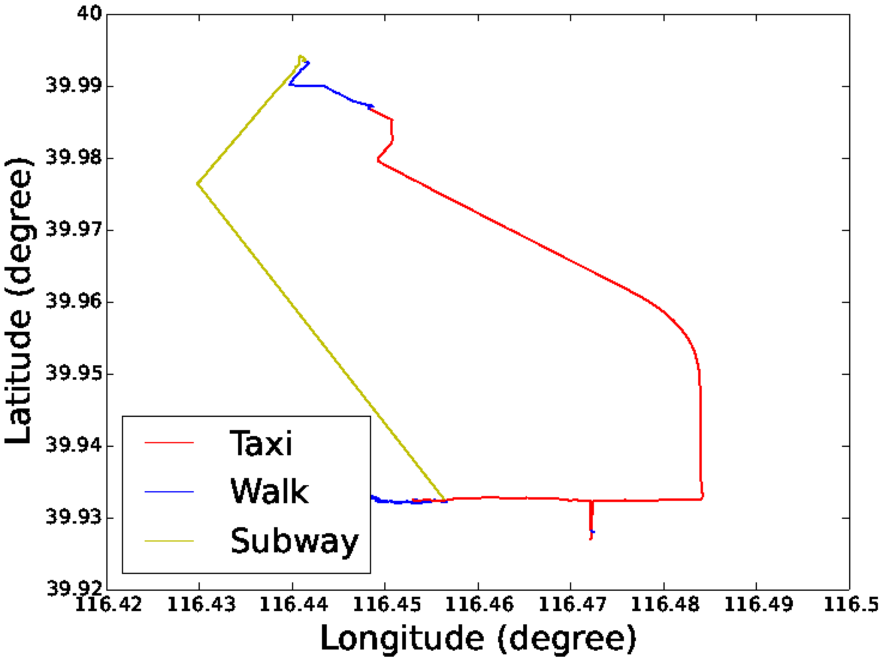


Figure 2.1: Illustration of a synthetic trail (taxi, walk, subway, walk, taxi, walk) for one day trip and their corresponding flights.

ties is better fitted with the lognormal distribution rather than the power-law distribution. We can demonstrate that the mixture of these transportation mode distributions is a power-law distribution based on two new findings: first, there is a significant positive correlation between consecutive flights in the same transportation mode, and second, the elapsed time in each transportation mode is exponentially distributed.

Datasets

Our analysis is based on two large real-life GPS trajectory datasets, the Geolife dataset [126] and the Nokia MDC dataset [60]. Both of them contain the transportation information. The key information provided by these two datasets is summarized in Table 2.2.

Geolife [126, 127, 122] is a public dataset with 182 users' GPS trajectories over five years (from April 2007 to August 2012) gathered mainly in Beijing, China. This dataset contains over 24 million GPS samples with a total distance of 1,292,951 kilometers and a total of 50,176 hours. It includes not only daily life

	Geolife	Nokia MDC
Location	Beijing	Geneva
Measurement	GPS	GPS
Number of samples	24,876,978	11,077,061
Duration	5 years	1.5 year
Accuracy	3 m	3 m
Sampling interval	1-5s	10s
Number of participants	182	200
Number of flights with transportation mode	202,702	224,723

Table 2.2: The Geolife and the Nokia MDC Human Mobility Datasets.

routines such as going to work and back home in Beijing, but also some leisure and sports activities, such as sightseeing, and walking in other cities. The transportation mode information in this dataset is manually logged by the participants.

The Nokia MDC dataset [60] is a public dataset from Nokia Research Switzerland that aims to study smartphone user behavior. The dataset contains extensive the smartphone data of two hundred volunteers in the Lake Geneva region over one and a half years (from September 2009 to April 2011). This dataset contains 11 million data points and the corresponding transportation modes.

Akaike Weights.

We use Akaike weights [101, 97, 18] to choose the best fitted distribution for each transportation mode. An Akaike weight is a normalized distribution selection criterion [17]. Its value is between 0 and 1. The larger the value is, the better the distribution is fitted.

Akaike's information criterion (AIC) is used in combination with Maximum Likelihood Estimation (MLE). MLE finds an estimator of $\hat{\theta}$ that maximizes the likelihood function $L(\hat{\theta}|data)$ of one distribution. AIC is used to describe the best fitting of all the fitted distributions,

$$AIC = -2log\left(L(\hat{\theta}|data)\right) + 2K. \tag{2.1}$$

Here K is the number of estimable parameters in the approximating model.

After determining the AIC value of each fitted distribution, we normalize these values as follows. First of all, we extract the difference between different AIC values called Δ_i,

$$\Delta_i = AIC_i - AIC_{min}. \tag{2.2}$$

Then Akaike weights W_i are calculated as follows,

$$W_i = \frac{exp(-\Delta_i/2)}{\sum_{r=1}^{R} exp(-\Delta_r/2)}. \tag{2.3}$$

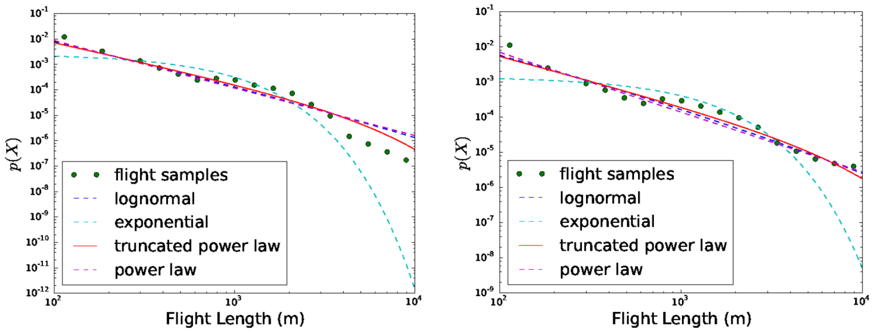
2.2.2 Power-law Fit for Overall Flight.

First, we fit the flight length distribution of the Geolife and Nokia MDC datasets regardless of transportation modes. We fit truncated power-law, lognormal, power-law and exponential distribution. We found that the overall flight length (l) distributions fit a truncated power-law $P(l) \propto l^{\alpha} e^{\gamma l}$ with exponent α as 1.57 in the Geolife dataset ($\gamma = 0.00025$) and 1.39 in the Nokia MDC dataset ($\gamma = 0.00016$) (Fig. 2.2), better than other alternatives such as power-law, lognormal or exponential. Figure. 2.2 illustrates the PDFs and their best fitted distributions according to Akaike weights. The green points refer to the flight length samples obtained from the GeoLife and the Nokia MDC dataset, while the solid red line represents the best fitted distribution according to Akaike weights. The overall flight length distribution regardless of transportation modes is well fitted with a truncated power-law distribution. The best fitted distribution (truncated power-law) is represented as a solid line and the rest are dotted lines. We use logarithm bins to remove tail noises [88, 3]. Our result is consistent with previous research ([15, 39, 88, 88, 110, 44, 56]), and the exponent α is close to their results.

We use Akaike weight for distribution fitting. The Akaike weight is a value between 0 and 1. The larger it is, the better the distribution is fitted [17, 3]. The Akaike weights of the power-law distributions regardless of transportation modes are 1.0000 in both datasets. The p-value is less than 0.01 in all our tests, which means that our results are very strong in terms of statistical significance.

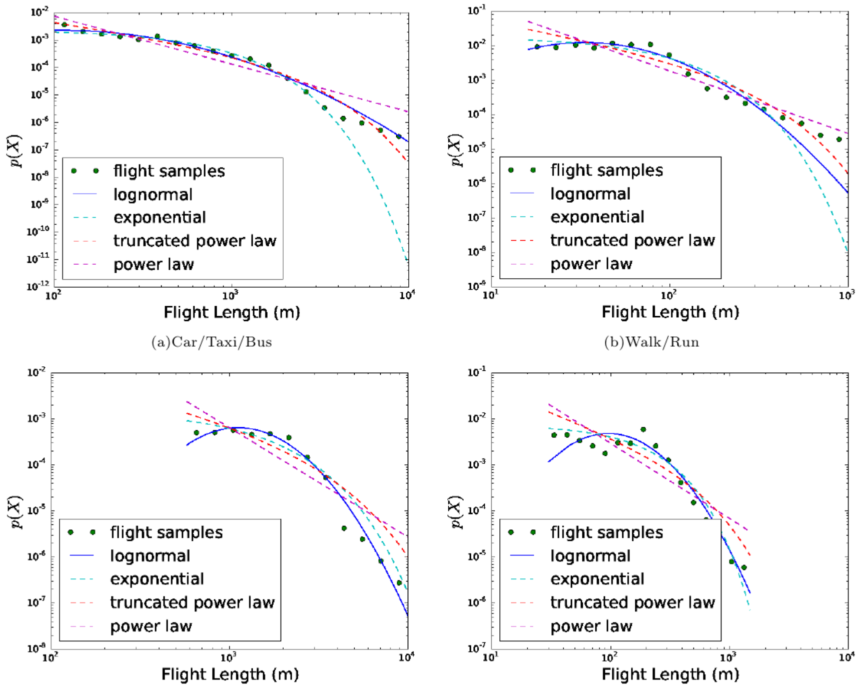
2.2.3 Lognormal Fit for Single Transportation Mode.

However, the distribution of flight lengths in each single transportation mode is not well fitted by the power-law distribution. Instead, they are better fitted with by lognormal distribution. All the segments of each transportation flight length are best approximated by the lognormal distribution with different parameters. In Fig. 2.3, we represent the flight length distributions of Walk/Run, Bike, Subway/Train and Car/Taxi/Bus in the Geolife dataset. The green points refer to the flight length samples obtained from the GeoLife, while the solid blue line represents the best fitted distribution according to Akaike weights. The flight length distribution in each transportation mode is well fitted with a lognormal distribution. The best fitted distribution (lognormal) is represented as a solid line and the rest are dotted lines.



(a)Geolife (b)Nokia MDC

Figure 2.2: Power-law fit for overall flight. (a-b) Power-law fitting of all flights regardless of transportation modes in the Geolife and the Nokia MDC dataset.



(a)Car/Taxi/Bus (b)Walk/Run

(c)Subway/Train (d)Bike

Figure 2.3: Lognormal fit for single transportation mode in the Geolife dataset. (a-d) Flight distribution of all transportation modes (Car/Taxi/Bus, Walk/Run, Subway/Train, Bike).

2.2.4 Mixture of All Transportation Modes

We characterized the mechanism of the power-law pattern with Lévy flights by mixing the lognormal distributions of the transportation modes. Previous research has shown that a mixture of lognormal distributions based on an exponential distribution is a power-law distribution [74, 49, 48, 71]. Based on their findings, we demonstrate that the reason that human movement follows the Lévy Walks pattern is due to the mixture of the transportation modes used.

We demonstrated that the mixture of the lognormal distributions of different transportation modes (Walk/Run, Bike, Train/Subway or Car/Taxi/Bus) is a power-law distribution given two new findings: first, we defined the change rate as the relative change of length between two consecutive flights with the same transport mode. The change rate in the same transportation mode is small over time. Second, the elapsed time between different transportation modes is exponentially distributed.

Lognormal in the Same Transportation Mode.

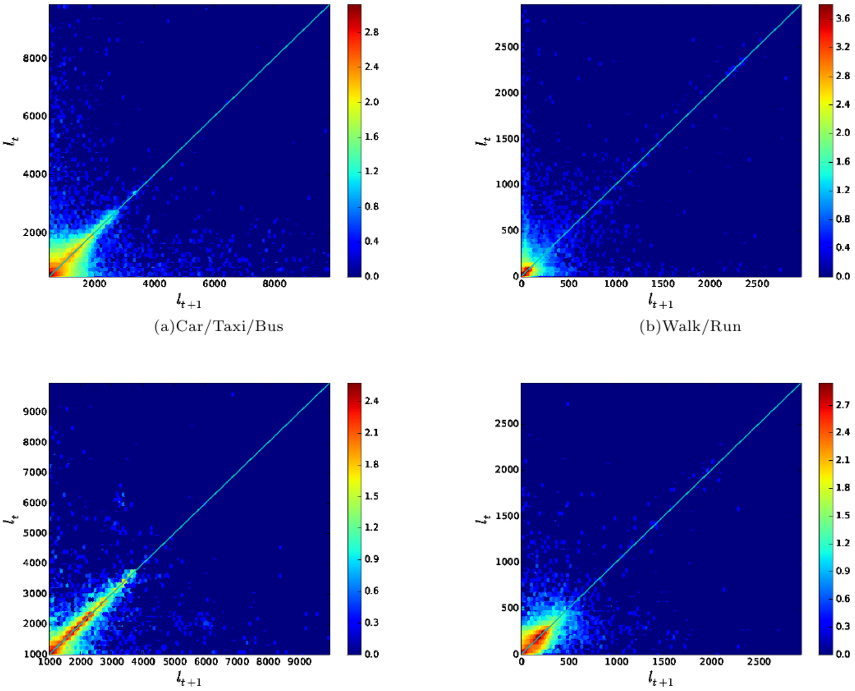
Let us consider a generic flight l_t. The flight length at the next interval of time l_{t+1}, given a change rate c_{t+1}, is

$$l_{t+1} = l_t + c_{t+1} l_t. \tag{2.4}$$

It has been found that the change rate c_t in the same transportation mode is small over time [46, 122]. The change rate c_t reflects the correlation between two consecutive displacements in one trip. To obtain the pattern of correlation between consecutive displacements in each transportation mode, we plot the flight length point (l_t, l_{t+1}) from the GeoLife dataset (Fig. 2.4). Here l_t represents the t-th flight length and l_{t+1} represents the $t+1$-th flight length in a consecutive trajectory in one transportation mode [107]. Figure. 2.4 shows the density of flight lengths correlation in Car/Taxi/Bus, Walk/Run, Subway/Train and Bike correspondingly. (l_t, l_{t+1}) are posited near the diagonal line, which identifies a clear positive correlation. Similar results are also found in the Nokia MDC dataset.

We use the Pearson correlation coefficient to quantify the strength of the correlation between two consecutive flights in one transportation mode [25]. The p value is less than 0.01 in all the cases, identifying very strong statistical significances. r is positive in each transportation mode and ranges from 0.3640 to 0.6445, which means that there is a significant positive correlation between consecutive flights in the same transportation mode, and the change rate c_t in the same transportation mode between two time steps is small.

The difference c_t in the same transportation mode between two time steps is small due to the small difference $l_{t+1} - l_t$ in consecutive flights. We sum all the contributions as follows:



(a)Car/Taxi/Bus

(b)Walk/Run

(c)Subway/Train

(d)Bike

Figure 2.4: Flight length correlation for each transportation mode. (a-d) Consecutive Flight length correlation of all transportation modes (Car/Taxi/Bus, Walk/Run, Subway/Train, Bike) in the GeoLife dataset. A high density of points are near the diagonal line $l_t = l_{t+1}$, identifying a small difference $l_{t+1} - l_t$ in the same transportation mode between two time steps.

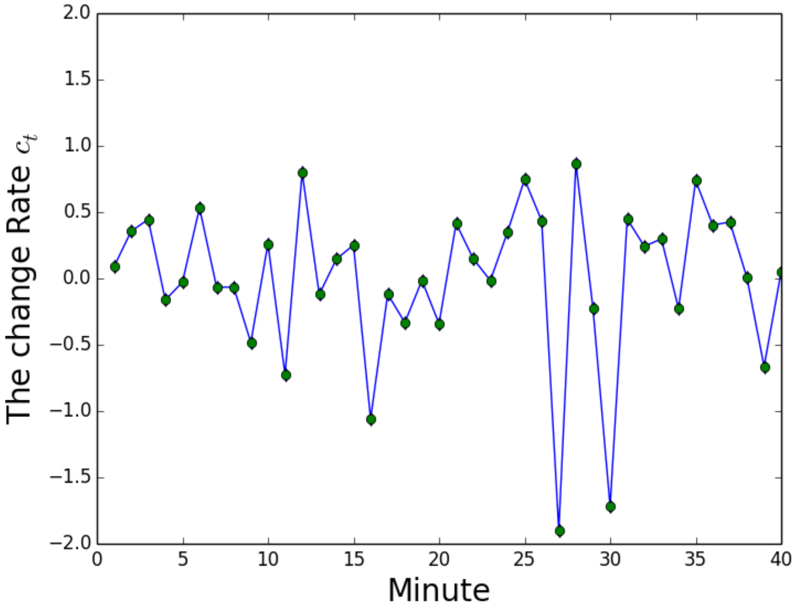


Figure 2.5: The change rate of the Car/Taxi/Bus mode in the Geolife dataset. The change rate is defined as the relative change of length between two consecutive flights with the same transport mode. From the figure we observe that the change rate are uncorrelated from one time interval to another.

$$\sum_{t=0}^{T} c_t = \sum_{t=0}^{T} \frac{l_{t+1} - l_t}{l_t} \tag{2.5}$$

$$\approx \int_0^T \frac{dl}{l} = \ln \frac{l_T}{l_0}. \tag{2.6}$$

We plot the change rate samples c_t of the Car/Taxi/Bus mode from the Geolife dataset as an example in Figure 2.5. We observe that the change rate c_t fluctuates in an uncorrelated fashion from one time interval to another in one transportation mode due to the unpredictable character of the change rate. The Pearson correlation coefficient accepts the findings at the 0.03-0.13 level with a p-value less than 0.05. By the Central Limit Theorem, the sum of the change rate c_t is normally distributed with the mean μT and the variance $\sigma^2 T$, where μ and σ^2 are the mean and variance of the change rate c_t and T is the elapsed time. Then we can assert that for every time step t, the logarithm of l is also normally distributed with a mean μt and variance $\sigma^2 t$ [89]. Note here that l_T is the length of the flight at the time T after T intervals of elapsed time. In the same transportation mode, the distribution of the flight length with the same change rate mean is lognormal, its density is given by

$$P_{singlemode}(l) = \frac{1}{l\sqrt{2\pi\sigma^2 t}} exp[-\frac{(\ln(l) - \mu t)^2}{2\sigma^2 t}], \tag{2.7}$$

which corresponds to our findings that in each single transportation mode the flight length is lognormal distributed.

Transportation Mode Elapsed Time.

We define elapsed time as the time spent in a particular transportation mode; we found that it is exponentially distributed. For example, the trajectory samples shown in Fig. 2.1 contain six trajectories with three different transportation modes, (taxi, walk, subway, walk, taxi, walk). Thus the elapsed time also consists of six samples (t_{taxi1}, t_{walk1}, $t_{subway1}$, t_{walk2}, t_{taxi2}, t_{walk3}). The elapsed time t is weighted exponentially between the different transportation modes. Similar results are also reported in [65]. The exponentially weighted time interval is mainly due to a large portion of Walk/Run flight intervals. Walk/Run is usually a connecting mode between different transportation modes (e.g., the trajectory samples shown in Fig. 2.1), and Walk/Run usually takes a much shorter time than any other modes. Thus the elapsed time decays exponentially. For example, 87.93% of the walk distance connecting other transportation modes is within 500 meters and the traveling time is within 5 minutes in the Geolife dataset.

Mixture of The Transportation Modes.

Given these lognormal distributions $P_{singlemode}(l)$ in each transportation mode and the exponential elapsed time t between different modes, we make use of mixtures of distributions. We obtain the overall human mobility probability by considering that the distribution of flight length is determined by the time t, the transportation mode change rate c_t mean μ and variance σ^2. We obtain the distribution of single transportation mode distribution with the time t, the change rate mean μ and variance σ^2 fixed. We then compute the mixture over the distribution of t since t is exponentially distributed over different transportation modes with an exponential parameter λ. If the distribution of l, $p(l, t)$, depends on the parameter t. t is also distributed according to its own distribution $r(t)$. Then the distribution of l, $p(l)$ is given by $p(l) = \int_{t=0}^{\infty} p(l, t)r(t)dt$. Here the t in $p(l, t)$ is the same as the t in the $r(t)$. $r(t)$ is the exponential distribution of elapsed time t with an exponential parameter λ.

So the mixture (overall flight length $P_{overall}(l)$) of these lognormal distributions in one transportation mode given an exponential elapsed time (with an exponent λ) between each transportation mode is

$$P_{overall}(l) = \int_{t=0}^{\infty} \lambda exp(-\lambda t)\frac{1}{l\sqrt{2\pi\sigma^2 t}}exp[-\frac{(\ln(l)-\mu t)^2}{2\sigma^2 t}]dt, \tag{2.8}$$

which can be calculated to give

$$P_{overall}(l) = Cl^{-\alpha'}, \tag{2.9}$$

where the power law exponent α' is determined by $\alpha' = 1 - \frac{\mu}{\sigma^2} + \frac{\sqrt{\mu^2+2\lambda\sigma^2}}{\sigma^2}$ [49, 48, 71]. The calculation to obtain α' is given as follows:

$$\begin{aligned}
P(x) &= \int_{t=0}^{\infty} \lambda exp(-\lambda t)\frac{1}{x\sigma\sqrt{2\pi t}}exp[-\frac{(\ln(x)-\mu t)^2}{2\sigma^2 t}]dt \\
&= \frac{\lambda}{\sigma}\frac{1}{\sqrt{2\pi}}x^{-1} \\
&\int_{t=0}^{\infty} exp(-\lambda t)exp[-\frac{(\ln(x)-\mu t)^2}{2t\sigma^2}]\frac{1}{\sqrt{t}}]dt \\
&= \frac{\lambda}{\sigma}\frac{1}{\sqrt{2\pi}}x^{-1} \\
&\int_{t=0}^{\infty} exp[\frac{-(\ln(x)-\mu t)^2 - 2\lambda\sigma^2 t}{2t\sigma^2}]\frac{1}{\sqrt{t}}]dt \\
&= \frac{\lambda}{\sigma}\frac{1}{\sqrt{2\pi}}x^{-1}exp(\frac{\ln x\mu}{\sigma^2}) \\
&\int_{t=0}^{\infty} exp[-(\frac{\mu^2+2\lambda\sigma^2}{2\sigma^2})t - \frac{(\ln x)^2}{2\sigma^2}\frac{1}{t}]\frac{1}{\sqrt{t}}]dt.
\end{aligned}$$

Using the substitution $t = u^2$ gives

$$\begin{aligned}
P(x) &= \frac{\lambda}{\sigma}\frac{1}{\sqrt{2\pi}}x^{-1}exp(\frac{\ln x\mu}{\sigma^2}) \\
&\int_{u=0}^{\infty} exp[-(\frac{\mu^2+2\lambda\sigma^2}{2\sigma^2})u^2 - \frac{(\ln x)^2}{2\sigma^2}\frac{1}{u^2}]\frac{1}{\sqrt{u^2}}]2udu.
\end{aligned}$$

Let $a = \frac{\mu^2+2\lambda\sigma^2}{2\sigma^2}$ and $b = (\ln x)^2 2\sigma^2$, from the integral table we get

$$\int_{u=0}^{\infty} exp(-au^2 - \frac{b}{u^2}) = \frac{1}{2}\sqrt{\frac{\pi}{a}}exp(-2\sqrt{ab}),$$

which helps us to get the expression for $P(x)$,

$$\begin{aligned} P(x) &= \frac{\lambda}{\sigma\sqrt{\frac{\mu^2}{\sigma^2} - 2\lambda^2}} x^{-(1-\frac{\mu}{\sigma^2}+\frac{\sqrt{\mu^2+2\lambda\sigma^2}}{\sigma^2})} \\ &= \frac{\lambda}{\sigma\sqrt{\frac{\mu^2}{\sigma^2} - 2\lambda^2}} x^{-\alpha'}. \end{aligned}$$

The expression for α' is

$$\alpha' = 1 - \frac{\mu}{\sigma^2} + \frac{\sqrt{\mu^2 + 2\lambda\sigma^2}}{\sigma^2}.$$

If we substitute the parameters from the fitted distributions, we get the $\alpha' = 1.55$ in the Geolife dataset, which is close to the original parameter $\alpha = 1.57$, and $\alpha' = 1.40$ in the Nokia MDC dataset, which is close to the original parameter $\alpha = 1.39$. The result verifies that the mixture of these correlated lognormal distributed flights in one transportation mode given an exponential elapsed time between different modes is a truncated power-law distribution.

2.3 Chapter Summary

In this chapter, we have shown that human movement exhibiting different transportation modalities is better fitted with the lognormal distribution rather than the power-law distribution (Lévy Walks Model). We have demonstrated that the mixture of these transportation mode distributions is a power-law distribution based on two new findings: first, there is a significant positive correlation between consecutive flights in the same transportation mode, and second, the elapsed time in each transportation mode is exponentially distributed.

Our transportation-decomposed Lévy Walks work, combined with the travel time distribution in each transportation mode and flight correlation, can help us build a more realistic urban human mobility model. The impact of our new transportation-decomposed Lévy Walks model on DTNs is not fully examined and will be considered in future work. For example, people tend to have a larger contact duration with each other in train compared to the other transportation modes, such as a bus or walking. If we know the context of transportation information [120, 96, 86, 61, 85, 102, 92], we can set the DTN transferring time to a larger one if a person is in a in train or to a smaller one if a person is walking, for disseminating more data to the nearby nodes while reducing unnecessary data transfer energy costs.

Chapter 3

Urban Mobility Applications for DTNs

This chapter introduces routing algorithms for DTNs and how we improve the performance of DTN routing algorithms with human mobility patterns. This chapter answers RQ3: How can human mobility be used to improve network application efficiency?

3.1 Delay Tolerant Networks

In this section, we summarize the recent routing algorithms and energy modeling work on DTNs. In DTNs, whenever mobile devices (vehicles, mobile phones, etc.) encounter each other, they exchange content via short-range communications (e.g., Bluetooth or WiFi). DTNs are a promising technology for significantly reducing the mobile data traffic of cellular networks by using complementary network communication technologies (WiFi direct, Bluetooth) for delivering the data offloaded from the cellular network. Most of the time there does not exist a complete path from the source to the destination in DTNs. To deal with such opportunistic encounter based networks researchers have proposed many routing schemes. Most of them fall into three general categories, Flooding-based routing schemes [100, 4], Probability-based routing schemes [67, 95] and Social-based routing schemes [72, 50, 78, 33, 16].

A Flooding-based routing scheme simply floods the DTN network with a message. A Probability-based routing scheme determines a forwarding probability, or utility value, for deciding to which encountered node to forward a given message. A Social-based routing scheme detects communities based on node encounters and then utillizes the community structure in message delivery.

In Table 3.1 we summarize the related papers of DTNs routing algorithms (including our paper PIII) and their contributions.

Publications	Routing scheme	Algorithm	Human Mobility
[100]	Flooding-based	Whenever two nodes meet, forward all the data	None
[67]	Probability-based	Forward the data based on encounter history	Basic Encounter history
[95]	Probability-based	Spray the data and wait for its destinations	None
[50]	Social-Based	Forward the data within the community	None
[78]	Social-Based	Search the temporal community	None
[33]	Social-Based	Search the geo-community	Yes, small regions
PIII	Content-Based/Human Mobility	Forward based on utility and density	Yes, spatial/temporal pattern

Table 3.1: Comparison of DTNs routing algorithms

3.1.1 Flooding-based Routing

Epidemic

Vahdat et al. present a Flooding-based routing protocol for DTNs called Epidemic routing [100]. In this protocol each node has a buffer that is used to store messages that cannot be delivered immediately. One *Summary vector* is kept in each node that contains an index of the buffered messages. When one node meet another node, they exchange their summary vectors. After this exchange, each node determines whether or not the other node has messages that they have not seen before. In the case that new messages are detected, the node requests the new messages from the contacted node. This message exchange process is shown in Figure 3.1. Finally, the messages will be delivered to their destinations through the pairwise communications. The system uses the First-In-First-Out (FIFO) message buffering strategy.

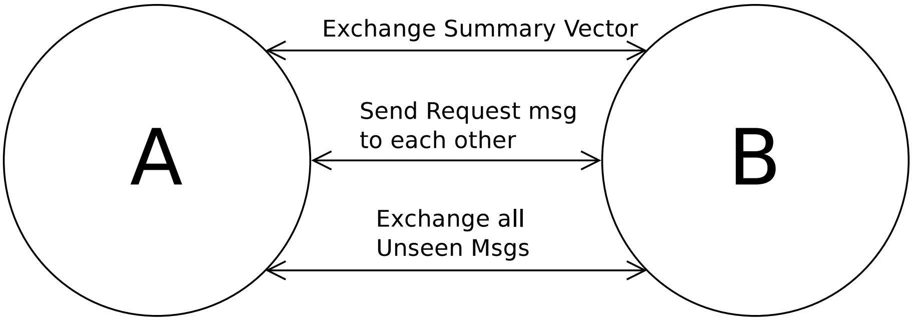


Figure 3.1: Message exchange in Epidemic Routing

The Epidemic Routing algorithm is illustrated in Algorithm 1

Algorithm 1 Epidemic Routing

```
Require: Node $n_j$, carrying $d_i$, is opportunistically encountering node $n_k$;
$n_j$ (resp. $n_k$) exchanging summary vector with $n_k$ (resp. $n_j$);
if $d_i$ is not in $n_k$'s summary vector then
    $d_i$ is forwarded to $n_k$;
end if
repeat this process until all unseen messages are exchanged between $n_k$ and $n_j$;
```

3.1.2 Probability-based Routing

ProPHET

ProPHET is a probability based routing protocol for intermittently connected networks proposed by Lindgren et al. [67]. ProPHET is short for a Probabilistic Routing Protocol using History of Encounters and Transitivity. To accomplish this, a probabilistic metric called the delivery predictability $P_{a,b} \in [0, 1]$ for the two nodes a and b, is introduced. This metric captures the probability of the two nodes meeting and sharing data. When two nodes meet, they not only exchange summary vectors, but also the delivery predictability scores. Delivery predictability is then updated in the manner given below.

The calculation of delivery predictabilities has three parts. Firstly, the metric is updated whenever a node is encountered, so that the often encountered node will have a higher delivery predictability than a less frequently encountered node. The following equation defines the delivery predictability,

$$P_{(a,b)} = P_{(a,b)_{old}} + (1 - P_{(a,b)_{old}}) \times P_{init},$$

where the P_{init} parameter is set to 0.75 by default.

If a pair of nodes does not meet each other for a while, they are less likely to be good forwarders to each other, thus their delivery predictability should be decreased. This behaviour is given by the following equation:

$$P_{(a,b)} = P_{(a,b)_{old}} \times \gamma^k,$$

where $\gamma \in [0, 1)$ is a constant and k is the number of time units that have elapsed since last time that the metric was aged. The time unit can differ and be defined based on the application and the expected delays.

The transitive property complements the basic delivery predictability metric by capturing the transitivity of the encounters. If a frequently meets b and b frequently encounters c, then b is a suitable node for forwarding messages from a to c. The following equation defines this property:

$$P_{(a,c)} = P_{(a,c)_{old}} + (1 - P_{(a,c)_{old}}) \times P_{(a,b)} \times P(b, c) \times \beta,$$

where $\beta \in [0, 1]$ is a scaling constant that affecting how large an impact the transitivity should have on the delivery predictability, the default value is 0.25.

When two nodes encounter each other, they will first update the delivery predictabilities. A message is forwarded to the other node if the delivery predictability of the destination of the message is higher in the other node. A forwarded message is not deleted at the source node if there is enough buffer space for the message. This message buffering results in the probability-flooding behaviour of the protocol. The queue management used in ProPHET is FIFO.

The ProPHET Routing algorithm is illustrated in Algorithm 2.

Algorithm 2 ProPHET Routing

```
Require: Node $n_j$, carrying $d_i$, is opportunistically encountering node $n_k$;
$n_j$ (resp. $n_k$) exchanging summary vector and delivery probability with $n_k$ (resp. $n_j$);
$n_j$ (resp. $n_k$) update its delivery probability;
if the delivery probability of $d_i$ in $n_k$ is higher than in $n_j$ then
    $d_i$ is forwarded to $n_k$;
end if
repeat this process until all unseen messages are checked and/or forwarded between $n_k$ and $n_j$;
```

The PRoPHETv2 [40] protocol was developed based on the lessons learned with ProPHET in various simulation scenarios. This new version improves the design of the original protocol. The new version retains the original idea and presents minor modifications to the evolution calculations. The following equation gives the new delivery predictability:

$$P_{(a,b)} = P_{(a,b)_{old}} + (1 - P_{(a,b)_{old}}) \times P_{enc},$$

where P_{enc} takes into account the time since an information exchange was performed with the node. This change reduces possible distortion from intermittent wireless connections that are frequently established and then disconnected. The protocol improvements increase the protocol performance especially for heterogeneous network mobility scenarios.

3.1.3 Social-based Routing

BUBBLE Rap

Pan et al. proposed a Social-based forwarding algorithm for DTNs called BUBBLE Rap in 2008 [50]. They observed that human interaction is heterogeneous both in terms of popularity and communities, and developed a BUBBLE Rap protocol based on these observations for improving the forwarding efficiency.

There are two intuitions behind this algorithm. First, people have different popularities in society, so the first part of this algorithm is to bubble the messages to the more popular nodes. Secondly, people form communities in their social lives, and people in the same community have more chances to meet others in the same community. Thus the second part aims for the efficient spreading of messages inside the community.

Community detection is a key problem in a Social-based routing protocol. Many community detection methods have been proposed and examined in the

Bubble Rap algorithm, e.g., K-CLIQUE by Palla et al. [76], weighted network analysis by Newman et al. [73] and distributed community detection by Pan et al. [53].

The BUBBLE Rap algorithm first sends a message to a more popular node globally that represents a community. When the message reaches its destination community, it will be forwarded within the community. The detailed 'bubble' and forwarding processes are carried out as follows. If a node has a message destined for another node, this node first bubbles this message through the hierarchical ranking tree using the global ranking until it reaches a node inside the same community as the initiator node. Then the nodes inside the community will use the local ranking tree to continuing bubbling the message until it reaches its destination or the message expires.

The BUBBLE Rap Routing algorithm is illustrated in Algorithm 3.

Algorithm 3 BUBBLE Rap Routing

```
Require: Node $n_j$, carrying $d_i$, is opportunistically encountering node $n_k$;
if $n_j$ is $d_i$'s destination community then
    if $n_k$ is also in $d_i$'s destination community
    and
    Local Rank of $n_k$ is higher than that of $n_j$
     then
        $d_i$ is forwarded to $n_k$;
    end if
else
    if $n_k$ is in $d_i$'s destination community
    or
    Global Rank of $n_k$ is higher than that of $n_j$
     then
        $d_i$ is forwarded to $n_k$;
    end if
end if
repeat this process until all unseen messages are checked and/or forwarded between $n_k$ and $n_j$;
```

3.1.4 Energy Modeling

In this section we discuss the energy costs in DTNs that relates to RQ2: How to model and optimize mobile phone energy usage with applications supporting both local and remote processing?

Pan et al. [42] built a prototype called *Opp-Off* and verified the availability of mobile phone communication during a short contact. Bluetooth and WiFi are two common local wireless communication technologies found on most smartphones, and thus are the two possible techniques to be used for building DTNs. WiFi scanning will significantly reduce the battery life of a fully charged phone, on the other hand, Bluetooth scanning will not drain battery life very quickly. Compared to WiFi, Bluetooth may be a better candidate for DTN communication. During a 30 seconds contact of two Bluetooth devices, the maximum number of transferred bytes is about 1,517.58KB, and the average number of transferred bytes is 563.25KB [42]. If the mobile phones have a longer communication range, they may have a higher probability to transfer more data during their opportunistic communication. The above result suggests that it is feasible to implement a DTN using a short-period communication technology, such as Bluetooth.

In PII, we studied energy-aware keyword searches [80, 12, 7] on mobile phones and proposed three approaches. The proposed hybrid approach adaptively splits the keywords of queries into two subsets, such that one subset is answered locally by the mobile phone, and another is offloaded to a remote server. Our experimental results indicate that the hybrid approach outperforms the two other extremes.

This result is useful for creating hybrid DTN applications that can leverage remote servers for reducing energy consumption in low power situations. For example, if we consider two mobile phones encountering each other, a hybrid approach for distributing and processing part of the data locally and the rest of the data on a remote support server will reduce the energy consumption significantly. It is also possible for other mobile phones to act as servers and utilize only local communications when offloading tasks in the network; however, the servers need to have access to the necessary application specific data.

The optimisation framework in PII indicates that server assisted keyword search and indexing can result in significant energy savings for mobile devices. Our framework focuses on the mobile cloud environment; however, we anticipate that a similar design is also useful in optimising keyword and content based DTNs. The main assumption is that the mobile or fixed network server assisting the mobile device has sufficient data for performing the offloaded operation. This assumption is reasonable for modern mobile devices that have extensive cloud synchronisation features.

With the proposed framework, a DTN node receiving a full-text document query can tune the ratio between local and remote processing for the matching operation in order to meet the energy budget. We anticipate that this offloading of queries and certain routing table functions can be beneficial for very constrained low-power DTN nodes; however, further experimentation is needed for the multi-device routing environment. The main contribution of the article is a generic framework for offloading that allows the tuning of the local versus remote processing with applications in mobile search that include DTNs.

3.2 Ameba Routing Algorithm

3.2.1 Overview

We designed and developed a solution framework, Ameba, for timely content delivery in DTNs by leveraging human mobility patterns in a city setting [83]. The basic idea of the system is to leverage the mobility information [21, 31, 54] of mobile devices for determining forwarding utilities for encounters and a content-aware relay algorithm that builds on the utilities [14, 20, 23, 24, 28, 29, 68, 82].

Ameba is able to (i) leverage the optimal routing hop count for each content, (ii) capture human mobility patterns, to deliver content towards the needed nodes. In this section, we mainly introduce the (ii), how can we use human mobility patterns in the Ameba algorithm. Here briefly introduce the overview of the Ameba algorithm in this sub-section.

In DTNs, the roles of mobile devices can be *publishers* (sources), *subscribers* (destinations) or *intermediate carriers*. Publishers publish content of specific topics. Subscribers register subscription filters (containing defined topics) to receive the needed content. In addition to publishers and subscribers, mobile devices can act as carriers to relay content. Mobile devices are typically equipped with short range interfaces (e.g., Bluetooh or Wi-Fi) to detect and communicate with each other. When mobile devices encounter each other, the data are exchanged opportunistically, and relayed from the publishers to the subscribers with the help of the intermediate carriers.

First, Ameba leverages the distribution of content and assigns a larger hop counter for the highly popular content demanded by more subscribers [83]. In this way, more nodes act as intermediate carriers of popular content, and subscribers have more chance to receive the content in a timely manner.

Second, Ameba develops a metric, namely the forwarding utility, to identify (i) which nodes are interested in the content and (ii) how fast the encountered node can forward the content towards subscribers [83]. Based on the developed utility, Ameba selects the best carriers to forward the content, and adaptively creates the copies of an content for timely delivery.

In Figure 3.2 we give an overview of the Ameba routing algorithm. We have publishers (S1, S2, S3, ...) publishing data items for each topic, such as weather forecast or news. The subscribers (N1, N2, N3, ...) are nodes that are interested in these data items. The data items are exchanged opportunistically, and relayed from the publishers to the subscribers with the help of the intermediate carriers (r1, r2, r3, ...). First, Ameba leverages the popularity distribution of the data and assigns a larger hop counter for popular data items that are demanded by more subscribers. This process of determining the topic-based hop counters is performed offline before the dissemination of the data items in the DTN network at runtime.

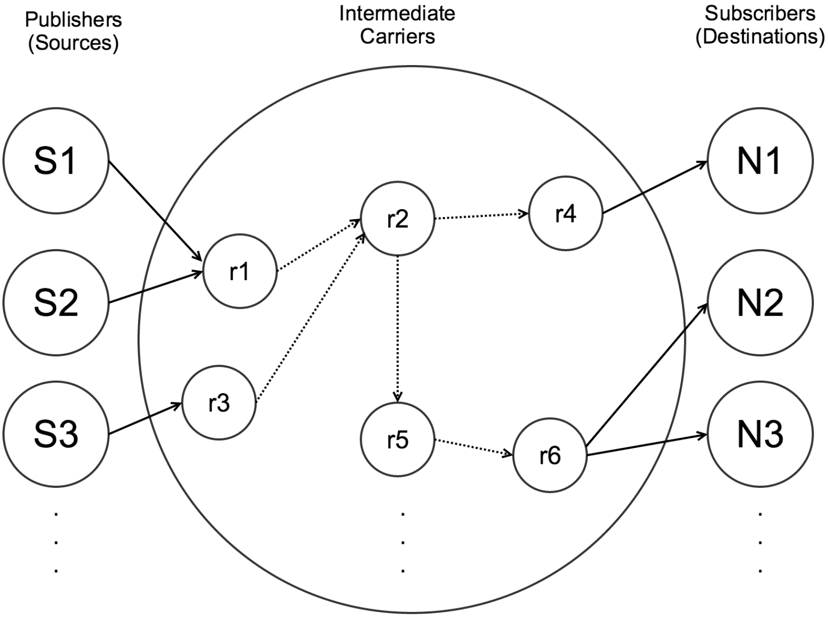


Figure 3.2: Ameba Routing algorithm

The runtime dissemination is started with the selection of the best carriers for efficient data forwarding. Human mobility patterns can be utilized for selecting carriers with high mobility in order to improve forwarding efficiency in the DTN network at runtime. In a city there are hot areas such as shopping centers or workplaces. The 'hot areas' gather most of the people and exhibit strong temporal regulations, the persons who visit these hot areas during peak times are optimal potential carriers for DTN data. The nodes with higher probability visiting the hot areas and meeting the subscribers are the optimal carriers for disseminating the data. We show how we utilize the knowledge of the spatial and temporal patterns of a city to select carriers and improve routing efficiency in the following sections.

3.2.2 Datasets

We use three real-world datasets (Table 3.2) of human mobility traces to motivate and demonstrate the efficiency of our solution. (i) The Infocom06 dataset [21] contains opportunistic Bluetooth contacts between 98 iMotes, 78 of which were distributed to Infocom06 participants and 20 of which were static. (ii) The MIT Reality trace [31] comprises 95 participants carrying GSM enabled cell-phones over a period of 9 months. (iii) In the UCSD dataset [70], 274 WiFi-enabled

Experimental data set	INFOCOM06	Reality	UCSD
Context settings	Conference	City	City
Device	iMote	Nokia 6600	PDA
Network type	Bluetooth	Bluetooth	WiFi
Number of devices	78	97	274
Duration of trace	4 days	9 months	2 months
Granularity	120s	300s	120s
Number of Areas	20	213	287
Number of internal contacts	191,336	54,667	195,364

Table 3.2: Summary of three real-world traces

PDAs were respectively used by 274 freshmen to log nearby Access Points (APs) for an 11-week period between Sep 22, 2002 and Dec 8, 2002 and a contact was recorded when two devices are associated to the same AP. Besides, we also used the location information of phones (e.g., the GSM cellular tower in the MIT reality trace).

3.2.3 Human Mobility Patterns

Based on the study of three DTN trace files (Table 3.2), we find that (i) people visiting different locations exhibit strong spatial properties (that is, the participants frequently visit a small number of hot areas, and rarely visit the remaining areas), and (ii) people visiting different locations also exhibit strong temporal properties (e.g., the majority of participant visits are clustered during some specific periods). To the best of our knowledge, this is the first work that examines the population density of different areas in DTNs and exploits it for content dissemination. The location information could be passively gathered by GPS, WiFi or other positioning solutions [81, 27, 99].

We find that *The 'area density' distribution is very heterogeneous.* The area density is defined as follows. When a node n_i (e.g., a mobile phone carried by a user) visits an area j, we say *one visit* v_{ij} occurs. The visit was logged by iMote in the Infocom06 dataset or by cellular tower in the MIT Reality dataset, or logged by the APs in the UCSD dataset. Based on the logged visits, we calculate the area density p_j of a area j by $p_j = \frac{\sum_{i=1}^{N} v_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{R} v_{ij}}$, where N is the total number of nodes and R is the total number of areas. By the definition p_j, an area j becomes more dense, (i.e., a higher p_j), when more users visit the area j. The intuition behind it is that people tend to visit 'hot areas' such as the university campus during weekdays or a football stadium during match days, rather than moving randomly.

Taking the MIT Reality trace [31] as an example, it comprises of 95 partici-

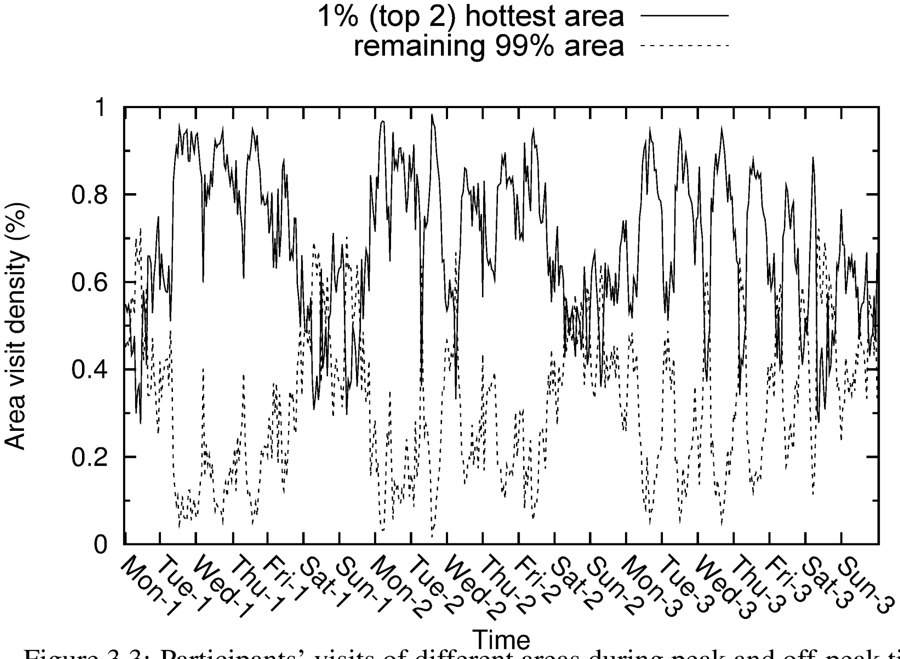


Figure 3.3: Participants' visits of different areas during peak and off-peak times

pants carrying GSM-enabled cell-phones over a period of 9 months in the Boston area while recording their locations. First, we extract the cumulative distribution (CDF) of all areas visits in the MIT Reality dataset. Note that we assume these locations in the Reality dataset are equal in area size. We set the observation period to 3 weeks in the Reality dataset (see Figure 3.3), 'Mon-1' represents Monday in week 1, etc. Here 'hot areas' are the areas where their cumulative visits account for at least 65% of total visits.

We find that the visits of the participants are highly clustered in the hot areas. In the MIT Reality dataset the top two hot areas (1%) occupy 70% of participants visits. Actually these top two hot areas in the Reality dataset are the 'Work' places such as the MIT Media lab and Sloan Business School. We plot the hourly visits in 'hot areas' in Figure 3.3 and we find that in the urban settings those students tend to visit their work places (hot areas) during daytime on weekdays (peak time) and their visits will drop significantly during night on weekdays or weekends (off-peak time).

3.2.4 Ameba Routing Algorithm

Since the 'hot areas' gather most of the people and exhibit strong temporal regulations, the persons who visit these hot areas during peak times are potential carriers for DTN data. We design and implement the Ameba algorithm by leveraging such human mobility patterns.

We use the percentage similarity $f_{j,k}$ to predict the chance that two mobile devices n_j and n_k will encounter each other. We define an encounter event as follows: two nodes encounter each other in the same area if their arrival time at such an area are within a specific period. For illustration we use one minute in the three datasets. The interval is chosen to be shorter than the device discovery interval in order to avoid the synchronous device discovery periods.

We compute the percentage similarity $f_{j,k}$ as follows: depending upon the similarity of the locations that n_j and n_k have respectively visited, we compute $f_{j,k} = 1 - \frac{\sum_{i=1}^{C} |V_{ji} - V_{ki}|}{2}$, where C is the total number of areas, V_{ji} indicates the percentage of n_j's visits at the area i and V_{ki} indicates the percentage of n_k's visits at the area i.

To clearly illustrate the intuition of $f_{j,k}$, we give an example. We consider $C = 5$ areas. A user n_j visits the areas with $10, 10, 10, 20$ and 50 times respectively, and another user n_k visits the areas with $25, 0, 0, 20$ and 5 times respectively. We then have $f_{j,k} = 1 - \frac{|0.1-0.5|+|0.1-0|+|0.1-0|+|0.2-0.4|+|0.5-0.1|}{2} = 0.4$. Based on this example, if n_j and n_k visit the same areas more frequently, we have higher percentage similarity $f_{j,k}$.

The percentage similarity $f_{j,k}$ captures the possibility that two nodes will visit the same place in the future. In Ameba, the percentage similarity $f_{j,k}$ incorporates the mobile devices n_j and n_k interests of the associated topic t_i to calculate the element utility u_i^k and u_i^j (the details of calculation is described in PIII [83], Section 4). Each element utility u_i^j, is computed as a number inside the range $[0.0, 1.0]$. The utility u_i^j measures the goodness of n_j to successfully relay a message d_i (with a topic t_i) towards the nodes that are interested in d_i. A larger u_i^j indicates that n_j has more chance to relay d_i successfully to the nodes that are interested in d_i. The Ameba algorithm is illustrated in Algorithm 4.

Here V_k is the sum of the visit percentage V_{jk} of user n_k's visits at hot areas during peak times. When V_k is higher, then n_k visits the hot areas more frequently. The intuition of $V_k > \lambda$ is as follows. When n_k more frequently visits the hot areas (due to a higher V_k), the node n_k, though not interested in d_i, could help to forward d_i to the nodes that are interested in d_i. Such forwarding is useful because many users frequently encounter each other at the hot areas during peak times.

Algorithm 4 Ameba Routing

```
Require: Node $n_j$, carrying $d_i$, is opportunistically encountering node $n_k$;
if $n_k$ is interested in $d_i$, or $V_k > \lambda$ during peak times; then
    $d_i$ is forwarded to $n_k$;
end if
if the element utility $u_i^k$ is larger than the element utility $u_i^j$; then
    $n_k$ keeps a copy of $d_i$;
end if
if $u_i^k > \mu_i^j$, $\mu_i^j$ is the largest element utility of the topic $t_i$ among all
nodes that $n_j$ ever encountered.; then
    $n_j$ removes $d_i$;
else
    $n_k$ does not keep a copy of $d_i$;
end if
```

3.2.5 Evaluation

We compare Ameba (Algorithm 4) with the Epidemic routing scheme [100] (chapter 3.1.1, Flooding-based algorithm), ProPHET [67] (Chapter 3.1.2, Probability-based algorithm), and Bubble Rap [50] (Chapter 3.1.3, Social-based algorithm). We use the MIT reality dataset to simulate the mobility pattern of DTN nodes, similar results has been found in the other two datasets [83].

During the experiment, we measure the average of the following metrics.

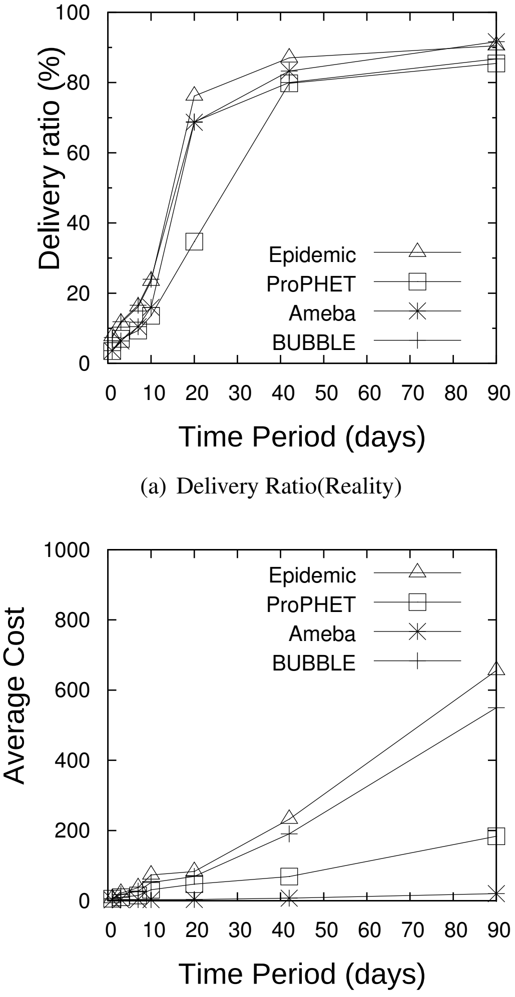
- Delivery ratio: the average ratio of the number of successfully delivered destinations to the total number of destinations.

- Average cost: the average number of content transmissions (including transmissions for duplicated copies) used to deliver a data item. Thus, the average cost measures the average *overhead* to deliver the data item.

Our simulation shows that Ameba is able to achieve a comparable delivery ratio to a Flooding algorithm (Epidemic [100]) but with only 3.5% overhead (see Figure 3.4). This result is mainly due to the new location-based algorithm that captures the movement of the users and the forwarding of content to users who often visit 'hot areas' during peak times. The mobility-pattern-based optimizations improve the information dissemination performance by leveraging information pertaining to spatial and temporal encounters.

3.3 Chapter Summary

In this chapter, we summarized recent DTN routing algorithms and presented an overview of our Ameba routing algorithm. In DTNs, most of the time there



(a) Delivery Ratio(Reality)

(b) Average Cost(Reality)

Figure 3.4: Delivery ratio and overhead MIT Reality data traces

does not exist a complete path from the origin to the destinations. To deal with such networks researchers have proposed many routing schemes. Most of them fall in to the three general categories, Flooding-based routing schemes [100, 4], Probability-based routing schemes [67, 95] and Social-based routing schemes [72, 50, 78, 33].

We proposed a solution framework, called Ameba, for timely data delivery in DTNs based on human mobility patterns. We find that urban human mobility exhibits strong spatial and temporal patterns. We leverage such human mobility patterns to derive an optimal routing algorithm that minimizes the hop count while maximizing the number of needed nodes in DTNs. Simulation results with experimental traces indicate that Ameba achieves a comparable delivery ratio to a Flooding algorithm, but with only a 3% lower overhead.

Chapter 4

Automatic City Region Analysis for Urban Routing

This chapter discusses the domain of automatic city region analysis with human mobility patterns and how we use the functional sub-areas of a city to improve DTN algorithm. This chapter answers RQ4: To what extent does an urban human mobility model improve network application efficiency?

4.1 Automatic City Region Analysis

4.1.1 Overview

There are different functional regions in cities such as tourist attractions, shopping centers, workplaces and residential places. The human mobility patterns for different functional regions are different, e.g., people usually go to work during daytime on weekdays, and visit shopping centers after work. In this chapter, we analyse urban human mobility patterns and infer the functions of the regions in three cities. The analysis is based on three large taxi GPS datasets [116, 115, 79, 4] in Rome, San Francisco and Beijing containing 21 million, 11 million and 17 million GPS points respectively.

We categorized the city regions into four kinds of places, workplaces, entertainment places, residential places and other places. First, we provide a new quad-tree region division method based on the taxi visits. Second, we use the association rule to infer the functional regions in these three cities according to temporal human mobility patterns. Third, we show that these identified functional regions can help us delivering data in network applications, such as urban Delay Tolerant Networks (DTNs), more efficiently. The new functional-regions-based DTNs algorithm achieves up to 183% improvement in terms of delivery ratio.

We infer the temporal human mobility of these three cities based on the taxi trips. Since taxis carry people to different places at different times, taxi trips reflect how people move in a city. We define a taxi's trip as two GPS points from picking up the passenger until dropping off the passenger. The taxi trips reflect the actual human movement inside the city.

First, we use the taxi visits as a baseline for dividing the sub-areas. We provide a new quad-tree sub-area division algorithm based on the taxi visits. Then we use association rules to infer the functions of different sub-areas in these three cities according to traveling patterns [98, 35]. We show that the identification of these functional sub-areas is useful in improving the efficiency of DTN routing in urban environments [32, 42, 53].

4.1.2 Datasets

A variety of urban human mobility data have been gathered and published due to the significant growth of sensing technologies and large-scale computing infrastructures. An urban human mobility dataset can be utilized for many urban applications. For example, Zheng et al. use the urban taxi dataset for monitoring the air pollution in a large city [123], Hemminki et al. use the mobile sensor data to detect the transportation modes [46]. In Table 4.1, we summarize the recent human-mobility-based urban applications (including our paper PIV) and their contributions.

We use three large taxi GPS trajectory datasets in our work, the Rome dataset, the San Francisco dataset and the Beijing Dataset. We summarize the key information in these three datasets in Table 4.2. All of the three datasets contain the following information: taxi id, timestamp and position (longitude, latitude). In the taxi mobility patterns, the drivers typically either move to pick up or drop off customers, or stay in parking areas while waiting for new customers.

The San Francisco dataset [79] is a public dataset from the Exploratorium that aims to study the invisible economic, social, and cultural trends of the city. The dataset contains extensive GPS data of five hundred Yellow Cab vehicles in the San Francisco region over one month (from 17th May 2008 to 10th June 2008). This dataset contains 11 million data points and the corresponding timestamps.

The Rome dataset [4] is a public dataset containing mobility traces of 316 taxi cabs in Rome over 30 days. Each taxi driver had a tablet that was set to retrieve the GPS position every 7 seconds after which the position was sent to a central server.

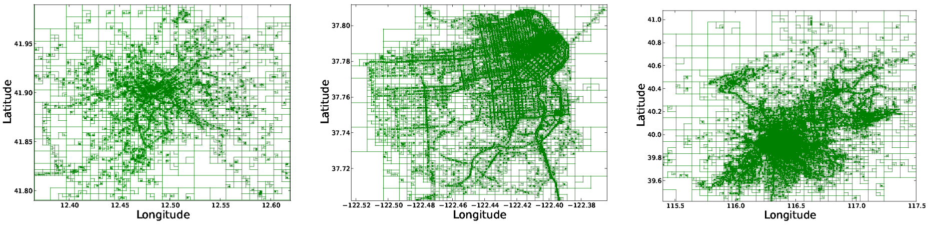
The Beijing dataset [116, 115] is a public dataset gathered by Microsoft Research Asia. It records the GPS trajectories of 10,357 taxis in Beijing from Feb.2 to Feb.8, 2008. There are about 15 million GPS points in this data set, and the total distance for each trajectory reaches up to 9 million kilometers.

Publications	Dataset	Applications	Methodology
[123]	Beijing taxi	Air pollution monitoring	Semi-supervised learning
[124]	Geolife	Transportation mode detection	Inference Model
[126]	Geolife	Social networks	Data Mining
[127]	Geolife	Tourist recommendation	Inference Model
[46]	Transportation data	Transportation mode detection	Machine Learning
[30]	Nokia MDC	Next place prediction	Nonlinear time series analysis
PIV	Geolife/Nokia MDC	Region function analysis/DTNs	Data Mining

Table 4.1: Summary of recent human-mobility-based urban applications.

	Rome	San Francisco	Beijing
Measurement	GPS	GPS	GPS
Number of samples	11,219,955	21,817,851	17,586,065
Duration	1 month	1 month	1 week
Sampling Interval	64 s	9 s	177 s
Number of taxis	536	316	10357

Table 4.2: Taxi Mobility Datasets



(a) Rome (b) San Francisco (c) Beijing

Figure 4.1: Original GPS samples in the three cities

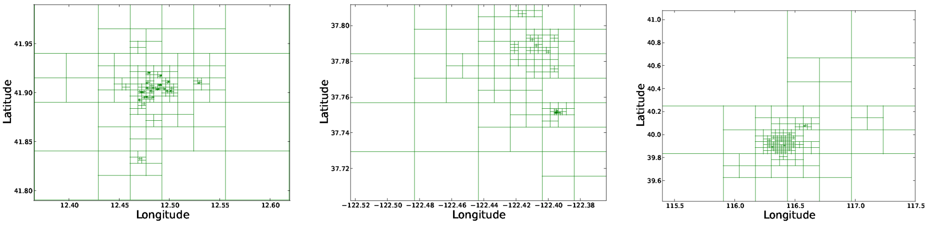
4.1.3 Quad-tree Division Based on Taxi Visits

We use taxi visits as a baseline for dividing the sub-regions. Fig. 4.1 shows the GPS samples of the three cities. We use the quad-tree [34, 109, 104] for dividing the city into different regions. We set the subdivision threshold as 1% of total visits inside the cities. If the number of visits in a sub-area is larger than 1% of the total visits, we further divide the sub-area into four equal-sized smaller sub-areas. This process continues until all the sub-areas have equal to or smaller than 1% of total visits. Fig. 4.2 shows the sub-areas after the division, we obtain 367 sub-areas in Rome, 211 sub-areas in San Francisco and 259 sub-areas in Beijing.

4.1.4 Inferring Hot Areas

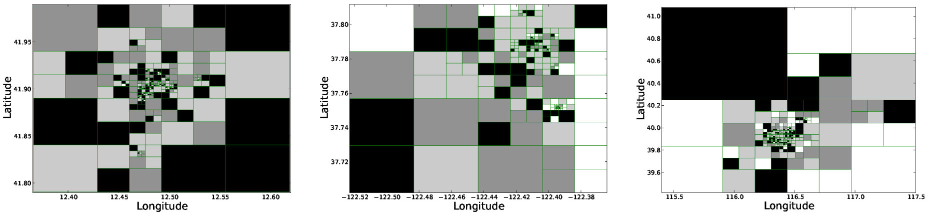
In this section, we provide an Apriori-based [1] function detection method for the sub-areas inside the city.

We utilize the knowledge of urban human mobility patterns to identify the functional regions of three cities, Rome, San Francisco and Beijing. A functional region [114, 117, 94] is a region (we use grid here) that has a specific characteristic such as tourist attractions, shopping centers, educational areas, workplaces or residential places. The human mobility patterns for different function regions are different. People usually go to work during daytime weekdays, visit the en-



(a) Rome (b) San Francisco (c) Beijing

Figure 4.2: Quad-Tree based region division in the three cities. The threshold is 1% of total taxi visits



(a) Rome (b) San Francisco (c) Beijing

Figure 4.3: Function identification in the three cities. The black regions represent workplaces, the light grey regions represent residential places, the dark grey regions represent entertainment places while the white represent all the other places.

tertainment places such as shopping centers after work, and stays at home during the night. Such temporal human mobility patterns can help us to identify the functional regions of a city.

First, we group the taxi visits for the sub-areas every hour and convert the taxi visits into a boolean table for the Apiori algorithm [1]. For example, suppose there are five sub-areas and three taxis. Each taxi's visits to these five areas are 2,3,0,2,1, 1,0,0,1,2 and 0,0,0,2,2. We can build a boolean table consisting of (1,1,0,1,1),(1,0,0,1,1), (0,0,0,1,1). Here each transaction (row) is the list of the locations that the taxis visits.

After building the boolean table, we use an Apriori algorithm [1] to generate frequent item sets for each hour. Apriori is an algorithm for finding the frequent item sets and learning association rules for transactional datasets. Here in our taxi visits boolean table, each transaction (row) is the list of the locations that the taxi visits within one hour, the frequent item sets are the sets that people usually visit within one hour. We set the threshold as 0.2 to find the frequent item sets in the first place.

After generating the frequent item sets (popular places people visit) for each hour, we divide the city into four kinds of places, the workplaces which people usually visit during work time, the entertainment places where people usually stay during entertainment time, the residential places where people usually stay during home time and the other places with no identical mobility patterns. We define work time as the daytime on weekdays (08:00-17:00, Monday to Friday), the entertainment time as the evening on weekdays and daytime and evening on weekends (17:00-23:00, Monday to Friday and 08:00-22:00, Saturday to Sunday), and home time as the night time all week (23:00-08:00 Monday to Friday).

We plot the cities according to the different functions detected above in Fig. 4.3. We use four colors to identify the functions inside the city. The black regions represent the workplaces, the light grey regions represent the residential places, the dark grey regions represent the entertainment places while the white represent all other places.

4.2 Data Dissemination (DTNs) in the Regions

We can utilize urban mobility patterns in different functional regions and across regions for enhancing DTNs routing algorithms. The key idea is to map content to functional sub-areas and then select carriers that optimize the information delivery to the destinations. For example, a person that is going to visit functional-areas during a peak time (for example, workplaces during daytime on weekdays) has a high chance of meeting a person that may need the carried content targeted for those areas. The functional regions can also be viewed as hubs for forwarding messages to more people. We provide two simple functional-region-based urban routing algorithms and we show that they achieve up to 183% improvement in terms of delivery ratio compared with a random-routing algorithm.

4.2.1 Target-set Problem

DTN routing protocols aim to route data (e.g., a weather forecast notification) from the data sources to the destinations through opportunistic and delay tolerant data exchanges that typically are based on short-range wireless communications, such as Bluetooth.

In DTNs, the data is not typically distributed to all devices, but only to a subset of the devices. The devices in the subset then further distribute the data to other devices through an opportunistic DTN communication. How can we choose the initial target-set for maximizing the number of devices that further receive the desired data is called the target-set problem in DTNs [42].

4.2.2 Oracle-based, History-based, and Random Algorithms

We present our two simple functional region-based algorithms called the Oracle-based (Greedy) algorithm and the History-based (Heuristic) algorithm for solving the target-set problem in DTNs. To evaluate the performance of the Oracle-based and History-based algorithms, we compare them with the Random algorithm. In the Random algorithm, the initial group of carriers is chosen randomly.

In the Oracle-based algorithm, the initial subset of data carriers consists of people who have the highest probability of visiting the hot-areas in the city. Note that here the hot-areas represent workplaces during work time, entertainment places during entertainment time and residential places during night. The Oracle-based algorithm provides an upper bound for the target-set problem.

The History-based algorithm is similar to the Oracle-based algorithm, with the exception that a taxis probability of visiting hot areas is obtained from the historical data traces. We use the taxi visit data of the same hour in the previous day as the historical data. This strategy is motivated by the observation that human mobility has regularity. For example, a person usually goes to work-places at 8 am in the morning, and returns for dinner time. If the person has visited the work-places at 8 am in the previous weekdays, he has a high probability of visiting the work-places at 8 am on the current weekday. The person would be a good candidate carrier for data to the work-places.

4.2.3 Evaluation of Functional Region-based Algorithms

We use the Beijing dataset for evaluating the three algorithms, similar results has been found in the other two datasets. When two taxis visit (enter) the same region in Beijing at the same time, we consider it as an encounter event and these two taxis can exchange the data they carried. We randomly choose 100 taxis (100 subscribers) out of the 10,357 taxis as the subscribers who wish to receive messages (e.g., weather notifications) from the DTN. Then 100 copies of the message to be delivered are distributed to 100 taxis (100 publishers). If these 100 messages are all delivered to the subscribers, we count the delivery ratio as 100%. The initial target-set of the 100 publishers are chosen according to the three algorithms, the Oracle-based algorithm, the History-based algorithm and the Random algorithm.

After obtaining the message, each publisher will keep the message until it meets a subscriber and forwards the message. We extract the taxi visits of two hours to evaluate our algorithm under different scenarios. One is during ”2008-02-03 (Sunday) 15:00:00 - 16:00:00” and the other one is ”2008-02-05 (Tuesday) 15:00:00-16:00:00”.

During ”2008-02-03 (Sunday) 15:00:00-16:00:00” these entertainment places are usually the hot-areas in the terms of people visits, while during ”2008-02-05 (Tuesday) 15:00:00-16:00:00” these workplaces are usually the hot-areas. To obtain the historical information for our History-based algorithm, we take the taxi visits from ”2008-02-02 (Saturday) 15:00:00 - 16:00:00” as a baseline for entertainment places and ”2008-02-04 (Monday) 15:00:00 - 16:00:00” as a baseline for work places.

Fig. 4.4 shows the delivery ratio of the three algorithms. The delivery ratio is the average ratio of the number of data messages that were successfully delivered divided by the total number of messages. We observe that the Oracle-based algorithm outperforms the other two as it provides the upper bound. The History-based algorithm increases the deliver ratio with up to 183% improvement compared with a Random algorithm. This is mainly because that these taxis carrying data (e.g., today’s weather forecast) visiting the hot-areas (e.g., workplaces during daytime on weekdays) have a higher chance of meeting a person that needs the required data (e.g., today’s weather forecast). These functional regions can be viewed as hubs for forwarding messages to more people.

4.3 Chapter Summary

In this chapter, we analyzed urban human mobility patterns and presented a technique for inferring the functions of the sub-areas in Rome, San Francisco and Beijing. We categorize the city regions into four kinds of functional places, workplaces, entertainment places, residential places and the other places. We use association rules to infer the functions of different sub-areas in these three cities according to the temporal travel patterns. We show that the identification of these functional sub-areas can help us to deliver the data in urban DTNs with a 183%

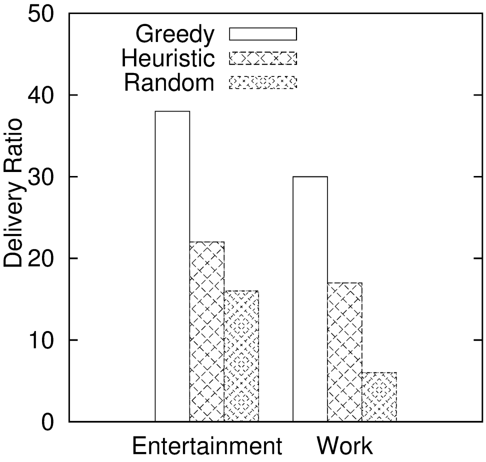


Figure 4.4: Delivery ratio of Oracle-based (Greedy), History-based (Heuristic), and Random Algorithms.

increase of the delivery ratio.

Since this is still an ongoing work, some important parts are still missing. Adding Points of Interest (POIs) will further increase the accuracy of the region function identification. A comparison with other traditional transportation engineering works is needed for the evaluation of our method. The computational improvement of the quad-tree method has not yet been evaluated. We will consider these in future work.

Chapter 5

Conclusions

In this thesis we address two key challenges presented by mobile applications, namely urban mobility modeling and its network applications (DTNs). These two topics pertaining to urban mobility examined in the thesis support the design and implementation of efficient information dissemination solutions for various applications, such as content offloading to the DTN and environmental monitoring.

First, we model urban human mobility as a Lévy Walks and decomposed it with transportation mode information such as Walk/Run, Bike, Train/Subway or Car/Taxi/Bus. We show that in each transportation mode, the flight length follows log-normal distribution and the mixture of these single-transportation modes is a power-law distribution. The travel time in each transportation mode is exponentially distributed and the flight fluctuations are identically and independently distributed. Our transportation mode decomposed Lévy Walks model deepens the understanding of human mobility in the city environment with different transportation modes. This contribution answers RQ1, how can we model urban human mobility?

The transportation mode decomposed Lévy Walks model is important for applications relying on human mobility models, for example network simulations, network planning applications, and applications running on top of urban DTNs. For example, according to our model, human movement exhibits different spatial and temporal patterns in each transportation mode. For DTNs this can be used to estimate the contact time of two or more devices sharing the same transportation mode.

The impact of our transportation mode decomposed Lévy Walks model on DTNs has not been investigated yet. Transportation modes have a profound impact on the performance of a DTN network, because their spatial and temporal properties differ. Our results motivate the design of an adaptive DTN routing algorithm that would change the routing parameters and dynamics based on the transportation mode. Such an algorithm could conserve energy by optimizing network scans and utilize the transportation patterns in maximizing data dissemination speed and accuracy. The transportation mode can be detected at runtime with efficient algorithms as discussed in Chapter 2. We plan to investigate the adaptive DTN algorithm in future work.

Second, we find that urban human mobility exhibits strong spatial and temporal patterns in frequently visited hot areas. We leverage such human mobility patterns to derive an optimal routing algorithm (Ameba) in DTNs to maximize the delivery ratio and minimize the overhead. Hot areas, such as workplaces or big shopping malls, are ideal places for transferring data in DTNs. People gathered in the same place have a higher chance of meeting the person with the desired message (data) in DTNs. Simulation results with experimental traces indicate that Ameba achieves a delivery ratio comparable to a Flooding-based algorithm, but with only 3% overhead. By proposing the Ameba routing algorithm, we answer RQ3, how can human mobility be used to improve network application efficiency?

One item that demands further attention is the social aspect of the hot-area-based Ameba. The prediction of social events, such as a workshop meeting or watching a football game, would enable the algorithm to leverage anticipating the formation of hot areas. We plan to conduct a social-spatial analysis of the human mobility datasets in future work.

Third, we analyze the temporal urban mobility patterns and infer the functions of the sub-areas in three cities. The identification of these functional sub-areas can be utilized to increase the efficiency of urban DTN applications. People have strong temporal regularity, e.g., people usually go to work during the daytime on weekdays, and visit shopping centers after work. Based on this temporal information we can infer the functional regions of the cities. These identified functional regions can improve the performance of urban network applications, such as urban DTNs. The proposed functional-regions-based DTN algorithm achieves up to 183% improvement compared to a random scheme in terms of the delivery ratio. This answers RQ4: To what extent does an urban human mobility model improve network application efficiency?

More and more urban human mobility datasets are published in the era of the Internet of Things. These urban mobility datasets can help us to better understand how people move in the urban environment. As the number and density of urban areas increases, there are more opportunities for urban network applications, such as DTNs. In this thesis, we show that with the help of urban human mobility patterns, we are able to significantly increase the efficiency and coverage of network applications (DTNs).

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pages 487–499, 1994.

[2] I. F. Akyildiz, Y. Lin, W. Lai, and R. Chen. A new random walk model for PCS networks. *IEEE Journal on Selected Areas in Communications*, 18(7):1254–1260, 2000.

[3] J. Alstott, E. Bullmore, and D. Plenz. powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, 9(1), 01 2014.

[4] R. Amici, M. Bonola, L. Bracciale, A. Rabuffi, P. Loreti, and G. Bianchi. Performance assessment of an epidemic protocol in vanet using real traces. volume 40, pages 92 – 99, 2014. Fourth International Conference on Selected Topics in Mobile Wireless Networking (MoWNet2014).

[5] D. Balcan and A. Vespignani. Phase transitions in contagion processes mediated by recurrent mobility patterns. *Nat Phys*, 7(7):7, 2011.

[6] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky. Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1):74–82, 2013.

[7] S. M. Beitzel, E. C. Jensen, A. Chowdhury, O. Frieder, and D. A. Grossman. Temporal analysis of a very large topically categorized web query log. *JASIST*, 58(2):166–178, 2007.

[8] V. Belik, T. Geisel, and D. Brockmann. Natural human mobility patterns and spatial spread of infectious diseases. *Phys. Rev. X*, 1:011001, Aug 2011.

[9] C. Bettstetter. Mobility modeling in wireless networks: categorization, smooth movement, and border effects. *Mobile Computing and Communications Review*, 5(3):55–66, 2001.

[10] C. Bettstetter, G. Resta, and P. Santi. The node distribution of the random waypoint mobility model for wireless ad hoc networks. *IEEE Trans. Mob. Comput.*, 2(3):257–269, 2003.

[11] C. Bettstetter and C. Wagner. The spatial node distribution of the random waypoint mobility model. In *Mobile Ad-Hoc Netzwerke, 1. Deutscher Workshop UBer Mobile Ad-Hoc Netzwerke WMAN 2002*, pages 41–58. GI, 2002.

[12] A. R. Bharambe, M. Agrawal, and S. Seshan. Mercury: supporting scalable multi-attribute range queries. In *SIGCOMM*, pages 353–366, 2004.

[13] C. Boldrini, M. Conti, and A. Passarella. Modelling data dissemination in opportunistic networks. In *Proceedings of the third ACM workshop on Challenged networks*, CHANTS '08, New York, NY, USA, 2008. ACM.

[14] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and zipf-like distributions: Evidence and implications. In *INFOCOM*, pages 126–134, 1999.

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

[16] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine. Maxprop: Routing for vehicle-based disruption-tolerant networks. In *INFOCOM 2006. 25th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 23-29 April 2006, Barcelona, Catalunya, Spain*, 2006.

[17] K. Burnham and D. Anderson. Model selection and multi-model inference: A practical information-theoretic approach. Springer, 2010.

[18] K. P. Burnham and D. R. Anderson. Multimodel inference understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[19] T. Camp, J. Boleng, and V. Davies. A survey of mobility models for ad hoc network research. *Wireless Communications and Mobile Computing*, 2(5):483–502, 2002.

[20] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Trans. Mob. Comput.*, 6(6):606–620, 2007.

[21] A. Chaintreau, A. Mtibaa, L. Massoulié, and C. Diot. The diameter of opportunistic mobile networks. In *CoNEXT*, page 12, 2007.

[22] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring millions of footprints in location sharing services. *ICWSM*, 2011:81–88, 2011.

[23] G. Chockler, R. Melamed, Y. Tock, and R. Vitenberg. Constructing scalable overlays for pub-sub with many topics. In *PODC*, 2007.

[24] G. Chockler, R. Melamed, Y. Tock, and R. Vitenberg. Spidercast: a scalable interest-aware overlay for topic-based pub/sub communication. In *DEBS*, 2007.

[25] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Routledge, 2 edition, July 1988.

[26] V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, and A. Vespignani. Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *PLoS Med*, 4(1):e13, 01 2007.

[27] I. Constandache, X. Bao, M. Azizyan, and R. R. Choudhury. Did you see bob?: human localization using mobile phones. In *MOBICOM*, pages 149–160, 2010.

[28] P. Costa, C. Mascolo, M. Musolesi, and G. P. Picco. Socially-aware routing for publish-subscribe in delay-tolerant mobile ad hoc networks. *IEEE Journal on Selected Areas in Communications*, 26(5):748–760, 2008.

[29] F. M. Cuenca-Acuna and T. D. Nguyen. Text-based content search and retrieval in ad-hoc p2p communities. In *NETWORKING Workshops*, pages 220–234, 2002.

[30] M. D. Domenico, A. Lima, and M. Musolesi. Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing*, 9(6):798–807, 2013.

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

[32] K. Fall. A delay-tolerant network architecture for challenged internets. In *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM '03, pages 27–34, New York, NY, USA, 2003. ACM.

[33] J. Fan, J. Chen, Y. Du, W. Gao, J. Wu, and Y. Sun. Geo-community-based broadcasting for data dissemination in mobile social networks. *IEEE Transactions on Parallel and Distributed Systems*, pages 1–1, 2012.

[34] R. A. Finkel and J. L. Bentley. Quad trees a data structure for retrieval on composite keys. *Acta informatica*, 4(1):1–9, 1974.

[35] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

[36] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB JournalThe International Journal on Very Large Data Bases*, 20(5):695–719, 2011.

[37] J. B. Goddard. Functional regions within the city centre: A study by factor analysis of taxi flows in central london. *Transactions of the Institute of British Geographers*, pages 161–182, 1970.

[38] S. Goh, K. Lee, J. S. Park, and M. Y. Choi. Modification of the gravity model and application to the metropolitan seoul subway system. *Phys. Rev. E*, 86:026102, Aug 2012.

[39] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[40] S. Grasic, E. Davies, A. Lindgren, and A. Doria. The evolution of a dtn routing protocol - prophetv2. In *Proceedings of the 6th ACM Workshop on Challenged Networks*, CHANTS '11, pages 27–30, New York, NY, USA, 2011. ACM.

[41] R. Groenevelt, E. Altman, and P. Nain. Relaying in mobile ad hoc networks: The brownian motion mobility model. *Wireless Networks*, 12(5):561–571, 2006.

[42] B. Han, P. Hui, V. A. Kumar, M. V. Marathe, J. Shao, and A. Srinivasan. Mobile data offloading through opportunistic communications and social participation. *IEEE Transactions on Mobile Computing*, 99(PrePrints), 2011.

[43] H. Han, X. Yu, and Y. Long. Discovering functional zones using bus smart card data and points of interest in beijing. *CoRR*, abs/1503.03131, 2015.

[44] X.-P. Han, Q. Hao, B.-H. Wang, and T. Zhou. Origin of the scaling law in human mobility: Hierarchy of traffic systems. *Phys. Rev. E*, 83:036117, Mar 2011.

[45] J. Härri, F. Filali, and C. Bonnet. Mobility models for vehicular ad hoc networks: a survey and taxonomy. *Communications Surveys & Tutorials, IEEE*, 11(4):19–41, 2009.

[46] S. Hemminki, P. Nurmi, and S. Tarkoma. Accelerometer-based transportation mode detection on smartphones. In *SenSys*, page 13, 2013.

[47] 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.

[48] B. A. Huberman and L. A. Adamic. Evolutionary dynamics of the World Wide Web. *Condensed Matter*, Jan. 1999.

[49] B. A. Huberman and L. A. Adamic. The nature of markets in the world wide web. *Quarterly Journal of Economic Commerce*, Jan. 2000.

[50] P. Hui, J. Crowcroft, and E. Yoneki. Bubble rap: social-based forwarding in delay tolerant networks. In *MobiHoc*, pages 241–250, 2008.

[51] P. Hui, A. Lindgren, and J. Crowcroft. Empirical evaluation of hybrid opportunistic networks. In *COMSNETS*, pages 1–10. IEEE, 2009.

[52] P. Hui, R. Mortier, M. Piórkowski, T. Henderson, and J. Crowcroft. Planet-scale human mobility measurement. In *Proceedings of the 2nd ACM international workshop on hot topics in planet-scale measurement*, page 1. ACM, 2010.

[53] P. Hui, E. Yoneki, S. Y. Chan, and J. Crowcroft. Distributed community detection in delay tolerant networks. In *Proceedings of 2nd ACM/IEEE international workshop on Mobility in the evolving internet architecture*, MobiArch '07, pages 7:1–7:8, New York, NY, USA, 2007. ACM.

[54] S. Ioannidis, A. Chaintreau, and L. Massoulié. Optimal and scalable distribution of content updates over a mobile social network. In *INFOCOM*, 2009.

[55] S. Ioannidis and P. Marbach. A brownian motion model for last encounter routing. In *INFOCOM 2006. 25th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 23-29 April 2006, Barcelona, Catalunya, Spain*, 2006.

[56] B. Jiang, J. Yin, and S. Zhao. Characterizing the human mobility pattern in a large street network. *Phys. Rev. E*, 80:021136, Aug 2009.

[57] E. P. Jones, L. Li, J. K. Schmidtke, and P. A. Ward. Practical routing in delay-tolerant networks. *Mobile Computing, IEEE Transactions on*, 6(8):943–959, 2007.

[58] W.-S. Jung, F. Wang, and H. E. Stanley. Gravity model in the korean highway. *EPL (Europhysics Letters)*, 81(4):48005, 2008.

[59] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth. Understanding human mobility from twitter. *CoRR*, abs/1412.2154, 2014.

[60] N. Kiukkonen, B. J., O. Dousse, D. Gatica-Perez, and L. J. Towards rich mobile phone datasets: Lausanne data collection campaign. In *Proc. ACM Int. Conf. on Pervasive Services (ICPS), Berlin.*, 7 2010.

[61] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. A survey of mobile phone sensing. *Communications Magazine, IEEE*, 48(9):140–150, 2010.

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

[63] Q. Li, S. Zhu, and G. Cao. Routing in socially selfish delay tolerant networks. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE, 2010.

[64] X. Li, G. Pan, Z. Wu, G. Qi, S. Li, D. Zhang, W. Zhang, and Z. Wang. Prediction of urban human mobility using large-scale taxi traces and its applications. *Frontiers of Computer Science*, 6(1):111–121, 2012.

[65] X. Liang, J. Zhao, L. Dong, and K. Xu. Unraveling the origin of exponential law in intra-urban human mobility. *Sci. Rep.*, 3, Oct 2013.

[66] X. Liang, X. Zheng, W. Lv, T. Zhu, and K. Xu. The scaling of human mobility by taxis is exponential. *Physica A: Statistical Mechanics and its Applications*, 391(5):2135–2144, 2012.

[67] A. Lindgren, A. Doria, and O. Schelén. Probabilistic routing in intermittently connected networks. *Mobile Computing and Communications Review*, 7(3):19–20, 2003.

[68] H. Liu, V. Ramasubramanian, and E. G. Sirer. Client behavior and feed characteristics of rss, a publish-subscribe system for web micronews. In *Internet Measurment Conference*, pages 29–34, 2005.

[69] E. Manley. Identifying functional urban regions within traffic flow. *Regional Studies, Regional Science*, 1(1):40–42, 2014.

[70] M. McNett and G. M. Voelker. Access and mobility of wireless PDA users. *Mobile Computing and Communications Review*, 9(2):40–55, 2005.

[71] M. Mitzenmacher. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[72] M. Motani, V. Srinivasan, and P. Nuggehalli. Peoplenet: engineering a wireless virtual social network. In *MOBICOM*, 2005.

[73] M. E. J. Newman. Analysis of weighted networks. *PHYS.REV.E*, 70:056131, 2004.

[74] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323, 2005.

[75] S. Ni and W. Weng. Impact of travel patterns on epidemic dynamics in heterogeneous spatial metapopulation networks. *Physical Review E*, 79(1):016111, Jan. 2009.

[76] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *NATURE*, 435:814, 2005.

[77] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 344–353. ACM, 2013.

[78] A. K. Pietiläinen and C. Diot. Dissemination in opportunistic social networks: the role of temporal communities. In *MobiHoc*, pages 165–174, 2012.

[79] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser. A parsimonious model of mobile partitioned networks with clustering. In *Communication Systems and Networks and Workshops, 2009. COMSNETS 2009. First International*, pages 1–10, Jan 2009.

[80] M. Pitkanen, T. Karkkainen, J. Greifenberg, and J. Ott. Searching for content in mobile dtns. In *Seventh Annual IEEE International Conference on Pervasive Computing and Communications, PerCom 2009, 9-13 March 2009, Galveston, TX, USA*, pages 1–10, 2009.

[81] T. Pulkkinen and P. Nurmi. Awesom: Automatic discrete partitioning of indoor spaces for wifi fingerprinting. In *Pervasive*, pages 271–288, 2012.

[82] W. Rao, L. Chen, and A. W. Fu. On efficient content matching in distributed pub/sub systems. In *INFOCOM*, 2009.

[83] W. Rao, K. Zhao, P. Hui, Y. Zhang, and S. Tarkoma. Maximizing timely content delivery in delay tolerant networks. *IEEE Transactions on Mobile Computing*, 14(4):755–769, 2015.

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

[85] S. Reddy, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Determining transportation mode on mobile phones. In *Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on*, pages 25–28. IEEE, 2008.

[86] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)*, 6(2):13, 2010.

[87] I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong. On the levy-walk nature of human mobility. In *INFOCOM 2008. 27th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 13-18 April 2008, Phoenix, AZ, USA*, pages 924–932, 2008.

[88] 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.

[89] S. M. Ross. *Stochastic Processes (Wiley Series in Probability and Statistics)*. Wiley, 2 edition, Feb. 1995.

[90] N. Scafetta. Understanding the complexity of the lévy-walk nature of human mobility with a multi-scale cost/benefit model. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(4):043106, 2011.

[91] M. Shin, S. Hong, and I. Rhee. Dtn routing strategies using optimal search patterns. In *Proceedings of the third ACM workshop on Challenged networks*, pages 27–32. ACM, 2008.

[92] T. Sohn, A. Varshavsky, A. LaMarca, M. Y. Chen, T. Choudhury, I. Smith, S. Consolvo, J. Hightower, W. G. Griswold, and E. De Lara. Mobility detection using everyday gsm traces. In *UbiComp 2006: Ubiquitous Computing*, pages 212–224. Springer, 2006.

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

[94] R. Song, W. Sun, B. Zheng, and Y. Zheng. Press: A novel framework of trajectory compression in road networks. In *VLDB*, September 2014.

[95] T. Spyropoulos, K. Psounis, and C. S. Raghavendra. Spray and wait: an efficient routing scheme for intermittently connected mobile networks. In *WDTN*, 2005.

[96] L. Stenneth, O. Wolfson, P. S. Yu, and B. Xu. Transportation mode detection using mobile phones and gis information. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 54–63. ACM, 2011.

[97] M. R. Symonds and A. Moussalli. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using akaikes information criterion. *Behavioral Ecology and Sociobiology*, 65(1):13–21, 2011.

[98] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[99] P. Tournoux, J. Leguay, F. Benbadis, V. Conan, M. D. de Amorim, and J. Whitbeck. The accordion phenomenon: Analysis, characterization, and impact on DTN routing. In *INFOCOM 2009. 28th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 19-25 April 2009, Rio de Janeiro, Brazil*, pages 1116–1124, 2009.

[100] A. Vahdat and D. Becker. Epidemic routing for partially-connected ad hoc networks. In *Duke Technical Report CS-2000-06*, July 2000.

[101] E.-J. Wagenmakers and S. Farrell. Aic model selection using akaike weights. *Psychonomic bulletin & review*, 11(1):192–196, 2004.

[102] H. Wang, F. Calabrese, G. Di Lorenzo, and C. Ratti. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 318–323. IEEE, 2010.

[103] P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González. Understanding road usage patterns in urban areas. *Scientific reports*, 2, 2012.

[104] S. Wang and M. P. Armstrong. A quadtree approach to domain decomposition for spatial interpolation in grid computing environments. *Parallel Computing*, 29(10):1481–1504, 2003.

[105] S. Wang, M. Liu, X. Cheng, and M. Song. Routing in pocket switched networks. *Wireless Communications, IEEE*, 19(1):67–73, 2012.

[106] S. Wang, X. Wang, X. Cheng, J. Huang, and R. Bie. The tempo-spatial information dissemination properties of mobile opportunistic networks with levy mobility. In *IEEE 34th International Conference on Distributed Computing Systems, ICDCS 2014, Madrid, Spain, June 30 - July 3, 2014*, pages 124–133, 2014.

[107] X.-W. Wang, X.-P. Han, and B.-H. Wang. Correlations and scaling laws in human mobility. *PLoS ONE*, 9(1):e84954, 01 2014.

[108] L. Willemen, P. H. Verburg, L. Hein, and M. E. van Mensvoort. Spatial characterization of landscape functions. *Landscape and Urban Planning*, 88(1):34–43, 2008.

[109] Y. Xia, Y. Liu, Z. Ye, W. Wu, and M. Zhu. Quadtree-based domain decomposition for parallel map-matching on gps data. In *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*, pages 808–813. IEEE, 2012.

[110] X.-Y. Yan, X.-P. Han, B.-H. Wang, and T. Zhou. Diversity of individual mobility patterns and emergence of aggregated scaling laws. *Sci. Rep.*, 3, Sep 2013.

[111] S. Yang, X. Yang, C. Zhang, and E. Spyrou. Using social network theory for modeling human mobility. *Network, IEEE*, 24(5):6–13, 2010.

[112] E. Yoneki, P. Hui, S. Chan, and J. Crowcroft. A socio-aware overlay for publish/subscribe communication in delay tolerant networks. In *Proceedings of the 10th ACM Symposium on Modeling, analysis, and simulation of wireless and mobile systems*, pages 225–234. ACM, 2007.

[113] J. Yoon, M. Liu, and B. Noble. Random waypoint considered harmful. In *Proceedings IEEE INFOCOM 2003, The 22nd Annual Joint Conference of the IEEE Computer and Communications Societies, San Franciso, CA, USA, March 30 - April 3, 2003*, 2003.

[114] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, pages 186–194, 2012.

[115] J. Yuan, Y. Zheng, X. Xie, and G. Sun. Driving with knowledge from the physical world. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pages 316–324, 2011.

[116] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang. T-drive: driving directions based on taxi trajectories. In *18th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2010, November 3-5, 2010, San Jose, CA, USA, Proceedings*, pages 99–108, 2010.

[117] N. J. Yuan and Y. Zheng. Segmentation of urban areas using road networks. Technical Report MSR-TR-2012-65, July 2012.

[118] K. Zhao, M. Musolesi, P. Hui, W. Rao, and S. Tarkoma. Explaining the power-law distribution of human mobility through transportation modality decomposition. *Nature Scientific Reports*, 2015.

[119] M. Zhao, L. Mason, and W. Wang. Empirical study on human mobility for mobile wireless networks. In *Military Communications Conference, 2008. MILCOM 2008. IEEE*, pages 1–7. IEEE, 2008.

[120] Y. Zhao. Mobile phone location determination and its impact on intelligent transportation systems. *Intelligent Transportation Systems, IEEE Transactions on*, 1(1):55–64, 2000.

[121] Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma. Understanding transportation modes based on gps data for web applications. *TWEB*, 4(1), 2010.

[122] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. Understanding mobility based on gps data. In *UbiComp*, pages 312–321, 2008.

[123] Y. Zheng, F. Liu, and H. Hsieh. U-air: when urban air quality inference meets big data. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 1436–1444, 2013.

[124] Y. Zheng, L. Liu, L. Wang, and X. Xie. Learning transportation mode from raw gps data for geographic applications on the web. In *WWW*, pages 247–256, 2008.

[125] Y. Zheng, Y. Liu, J. Yuan, and X. Xie. Urban computing with taxicabs. In *Ubicomp*, pages 89–98, 2011.

[126] Y. Zheng, X. Xie, and W.-Y. Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

[127] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *WWW*, pages 791–800, 2009.