ABSTRACT

HONG, SEONGIK. Human Movement Patterns, Mobility Models and Their Impacts on Wireless Applications. (Under the direction of Dr. Injong Rhee.)

Simulating human mobility is important in most areas that rely on human movements such as the performance study of mobile networks, disease outbreak control, city planning, etc. To test these applications, deploying a real testbed for a target scenario and running experiments would be the best solution but such experiments are extremely difficult. Thus to have a realistic mobility model and to run simulations with that model is invaluable. Numerous mobility models have been proposed until now but it is still questionable whether they are “realistic” enough to reflect human movement patterns. Since a mobility model is most useful when it is able to reproduce the mobility patterns realistically in any possible environments, defining “realism” and reproducing it is a key factor in mobility modeling. In this dissertation, we define the “real” human movement patterns as the heavy-tail flights and inter-contact times (ICT) which is defined to be the time duration until two mobile nodes meet again after meeting previously. These are important factors to determine the performance of human related applications since they govern the meeting probabilities among users.

We have tested 8 well-known existing models and by varying the input parameters, we tried to fit their flight and ICT distributions to the distributions observed from empirical data sets. Unfortunately, none of them can match both of them. We view this inability of reproducing the observed flight and ICT distributions is due to the insufficient description of spatial and temporal movement features. From the analysis results of 10 empirical data sets, we report the following findings, (1) human spatial gathering patterns have a fractal nature, (2) people are likely to visit destinations nearer to their current positions when visiting multiple destinations in succession and (3) each hotspot where people gather together has a lifecycle function in which the number of visitors to that hotspot varies over time and the characteristics of lifecycle functions are strongly correlated with the locations and fractal dimensions of hotspots.
In this dissertation, we show that those features are key factors to reproduce heavy-tail flight and ICT distributions that match the distributions observed from empirical data sets. Based on these observations, we propose a mobility model, called Spatio-TEmPoral model (STEP), that contains all the properties mentioned above. With this model, we could generate synthetic heavy-tail flights and ICTs that exactly match those observed from empirical data sets.

The contributions of this dissertation are as follows. First, we have collected fine-grained human movement traces using high quality GPS devices in various environments. Five sites are chosen for collecting traces. They are two university campuses (NCSU and KAIST), one metropolitan area (New York City), two theme parks (Disney World and North Carolina state fair). Second, we investigate spatial and temporal movement patterns from the empirical data sets and propose a mobility model that can represent all those patterns. The model is verified using the measured flight and ICT distributions to prove realism. Finally, we have shown the impact of each pattern on the routing performances of DTN.
© Copyright 2010 by Seongik Hong

All Rights Reserved
Human Movement Patterns, Mobility Models and Their Impacts on Wireless Applications

by
Seongik Hong

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Computer Science
Raleigh, North Carolina
2010

APPROVED BY:

Dr. George N. Rouskas
Dr. Khaled Harfoush

Dr. Steffen Heber
Dr. Injong Rhee
Chair of Advisory Committee
DEDICATION

To my wife Yeo-oak Yun

and daughter Ah-in
BIOGRAPHY

Seongik Hong received a B.S. degree in Electrical Engineering from Kyungpook National University, Korea in 1995 and a M.S. degree from Korea Advanced Institute of Science and Technology (KAIST), Korea in 1998 in electrical engineering. Since 1995, he studied in computer science and communication engineering, Kyushu university, Japan as an exchange student for a year, with the support of the Association of International Education, Japan. From 1998 to 2005, he worked as a member of technical staff with the Telecommunications and Operations Support Laboratory at KT (formerly Korea Telecom), Korea. From 2005 to 2010, he pursued his Ph.D. degree in Computer Science at North Carolina State University, Raleigh, NC. His research interests include human mobility modeling and routing algorithms in various wireless networks.
ACKNOWLEDGEMENTS

I would like to thank all people who have helped and inspired me during my doctoral study. First of all, I would like to thank my advisor, Dr. Injong Rhee, for his constructive guidance and support during my doctoral research. We had a long journey of Wednesday meetings for almost 4 years even on Christmas day and New Year’s Day. I could get many valuable comments and inspiration from the meetings. It was very fortunate for me to practice essential logical writing skills that I should have learned in my school days under his guidance. Also, I am very grateful to the other committee members, Dr. George N. Rouskas, Dr. Khaled Harfoush and Dr. Steffen Heber for their valuable comments.

This research is supported by various funding sources including the Korea Research Foundation. I would like to thank them for their support.

I would like to thank Jaewook Lee who always supports me and gives valuable and sincere advice to me. I would like to thank my friends and colleagues for their help in various ways throughout the entire duration of this thesis, especially, my companions on the Wednesday meeting journey and lab members who shared most of the time with, Minsu Shin, Ajit Warrier, Sangtae Ha, Sankararaman Janakiraman, Chisung Ahn, Seongjoon Kim, Kyunghan Lee, Hyungsuk Won, Jeongki Min, Sungro Yun and Yaogong Wang, my friends who helped me settle down at NCSU, Sooyoung Yoon, Jangeun Jun, Kyuyong Shin, Beakcheol Jang, Sangwon Hyun and Junbum Lim. Lastly, I would like to express my love and gratitude to my wife and daughter, and very special thanks to my parents and mother-in-law.
# TABLE OF CONTENTS

List of Tables ........................................................................................................ viii

List of Figures ........................................................................................................ ix

Chapter 1 Introduction ........................................................................................... 1

Chapter 2 Prior Work: Descriptions of Existing Models ................................. 10
  2.1 Random Models ............................................................................................... 11
    2.1.1 RWP ....................................................................................................... 12
    2.1.2 RD ......................................................................................................... 12
    2.1.3 BM ....................................................................................................... 12
    2.1.4 MWP ..................................................................................................... 12
    2.1.5 GM ....................................................................................................... 13
    2.1.6 SR ........................................................................................................ 13
    2.1.7 RPGM ................................................................................................... 13
    2.1.8 VT ....................................................................................................... 13
  2.2 Geographic Models ......................................................................................... 14
    2.2.1 OM ....................................................................................................... 14
    2.2.2 Manhattan/Freeway ............................................................................... 15
  2.3 Hotspot Models ............................................................................................... 16
    2.3.1 Dartmouth ............................................................................................. 16
    2.3.2 Pragma ................................................................................................ 17
    2.3.3 CMM .................................................................................................... 17
    2.3.4 ORBIT .................................................................................................. 18
    2.3.5 TCM ..................................................................................................... 18
  2.4 Social Models .................................................................................................. 19
    2.4.1 CM ....................................................................................................... 19
    2.4.2 SIMPS ................................................................................................. 19

Chapter 3 Motivations ......................................................................................... 21
  3.1 Data Sets ........................................................................................................ 22
    3.1.1 Experiments using GPS Receivers ....................................................... 22
    3.1.2 External Data Sets .............................................................................. 25
  3.2 Summary of Observations from the Empirical Data Sets ......................... 30
    3.2.1 Spatial Features .................................................................................. 30
    3.2.2 Spatio-Temporal Features ................................................................ 33
  3.3 Limitations of Existing Models .................................................................... 38
    3.3.1 Spatial Limitations ............................................................................. 38
    3.3.2 Spatio-Temporal Limitations ............................................................... 40
<table>
<thead>
<tr>
<th>Chapter 4</th>
<th>Spatial Features of Human Movements, Part I: Individual Patterns</th>
<th>46</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Overview</td>
<td>47</td>
</tr>
<tr>
<td>4.2</td>
<td>Characteristics</td>
<td>50</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Characteristic function</td>
<td>50</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Stationarity</td>
<td>52</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Diffusivity</td>
<td>53</td>
</tr>
<tr>
<td>4.3</td>
<td>Inter-Contact Times</td>
<td>57</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Infinite Space</td>
<td>57</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Finite Space</td>
<td>62</td>
</tr>
<tr>
<td>4.3.3</td>
<td>ICT Fitting</td>
<td>66</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Characteristic Time</td>
<td>68</td>
</tr>
<tr>
<td>4.4</td>
<td>Routing Performance</td>
<td>76</td>
</tr>
<tr>
<td>4.5</td>
<td>Conclusion</td>
<td>86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 5</th>
<th>Spatial Features of Human Movements, Part II: Aggregation Patterns</th>
<th>87</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Overview</td>
<td>88</td>
</tr>
<tr>
<td>5.2</td>
<td>Fractal Waypoint Maps</td>
<td>89</td>
</tr>
<tr>
<td>5.3</td>
<td>Individual Movement Patterns</td>
<td>92</td>
</tr>
<tr>
<td>5.3.1</td>
<td>LATP</td>
<td>92</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Individual Walker Model</td>
<td>96</td>
</tr>
<tr>
<td>5.4</td>
<td>Performance Evaluations</td>
<td>98</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Flights</td>
<td>98</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Inter-Contact Times and Link Durations</td>
<td>104</td>
</tr>
<tr>
<td>5.4.3</td>
<td>Routing Performance</td>
<td>107</td>
</tr>
<tr>
<td>5.5</td>
<td>Conclusion</td>
<td>111</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 6</th>
<th>Spatio-Temporal Features of Human Movements: Individual and Aggregation Patterns</th>
<th>112</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Overview</td>
<td>113</td>
</tr>
<tr>
<td>6.2</td>
<td>Spatial Features</td>
<td>115</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Fractal Dimension</td>
<td>115</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Super-Hotspot</td>
<td>117</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Sub-Hotspot</td>
<td>118</td>
</tr>
<tr>
<td>6.3</td>
<td>Temporal Features</td>
<td>121</td>
</tr>
<tr>
<td>6.4</td>
<td>Spatio-Temporal Correlations</td>
<td>127</td>
</tr>
<tr>
<td>6.5</td>
<td>STEP Mobility Model</td>
<td>130</td>
</tr>
<tr>
<td>6.5.1</td>
<td>Hotspot Model</td>
<td>130</td>
</tr>
<tr>
<td>6.5.2</td>
<td>Individual Walker Model</td>
<td>132</td>
</tr>
<tr>
<td>6.6</td>
<td>Performance Evaluation</td>
<td>136</td>
</tr>
<tr>
<td>6.6.1</td>
<td>Experimental Validation</td>
<td>136</td>
</tr>
<tr>
<td>6.6.2</td>
<td>Parameter Sensitivity</td>
<td>138</td>
</tr>
<tr>
<td>6.7</td>
<td>Conclusion</td>
<td>143</td>
</tr>
</tbody>
</table>
Chapter 7 Conclusion and Future Work ........................................ 144
References ............................................................................. 146
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Statistics of collected mobility traces from five sites.</td>
<td>23</td>
</tr>
<tr>
<td>3.2</td>
<td>AIC values: Truncated Pareto distributions give the best fit for the measured distributions.</td>
<td>24</td>
</tr>
<tr>
<td>3.3</td>
<td>Existing routing models can be categorized into four groups. Typical models in every category have been listed. None of the existing models have all the characteristics of human walks. '?' means that it is unclear from the model description.</td>
<td>36</td>
</tr>
<tr>
<td>3.4</td>
<td>Existing mobility models can be categorized by the constraints they describe. None of the existing models can faithfully describe both spatial and temporal constraints. $\triangle$ means that the constraints proposed do not produce heavy-tail flight and ICT distributions.</td>
<td>39</td>
</tr>
<tr>
<td>3.5</td>
<td>The results of Kullback-Leibler (KL) divergence tests: Each value represents the KL divergence value between the distributions from the real data sets and each model.</td>
<td>43</td>
</tr>
<tr>
<td>4.1</td>
<td>Diffusivity tendency by various parameters of TLW user mobility</td>
<td>56</td>
</tr>
<tr>
<td>4.2</td>
<td>The AIC and AW values for the ICT data from UCSD.</td>
<td>68</td>
</tr>
<tr>
<td>4.3</td>
<td>The AIC and AW values for the ICT data from Dartmouth.</td>
<td>68</td>
</tr>
<tr>
<td>4.4</td>
<td>The AIC and AW values for the ICT data from Infocom 2005.</td>
<td>71</td>
</tr>
<tr>
<td>5.1</td>
<td>The result of the Akaike test for the maximum likelihood estimation of truncated Pareto distributions (denoted Par) and exponential distribution (denoted Exp) over flights (denoted FL) and ICTs extracted from synthetically generated traces from various models whose parameters are set based on real traces obtained from four different locations (KAIST, NCSU, NYC and Disney World).</td>
<td>100</td>
</tr>
<tr>
<td>6.1</td>
<td>The results of Kullback-Leibler divergence tests: Compared with the results in Table 3.5, STEP is at least 9.6 times and 2.4 times better than other models in reproducing the measured flight and ICT distributions, respectively.</td>
<td>138</td>
</tr>
<tr>
<td>6.2</td>
<td>Parameter sensitivity of the proposed model. The PL represents power-law. The Var[DLF] means variance of an DLF.</td>
<td>138</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 1.1 Lifecycles from a real data set and existing mobility models. ............... 9
Figure 2.1 Sample traces from RWP and BM. ............................................. 11
Figure 2.2 An example of node movement in the RPGM Model, providing two snapshots at time \( T = t_0 \) (left circle) and time \( T = t_0 + \Delta t \) (right circle) [9]. .......... 14
Figure 2.3 An example of node movement in the OM Model. Shaded rectangles represent obstacles such as buildings. Dotted lines represent pathways between buildings. ....................................................... 15
Figure 2.4 The pathway graphs used in the Freeway and Manhattan models [9] ....... 16
Figure 2.5 A view of the simulated mobility using Pragma. It shows a spatial distribution of nodes in a bounded area. Each green point represents an individual mobile node, while red dots represent attractors that are located in the coordinates where individuals converge. ......................... 17
Figure 2.6 The Random ORBIT model. ......................................................... 18
Figure 2.7 An example of a social network. ..................................................... 19
Figure 3.1 Waypoint maps from two campuses. .............................................. 27
Figure 3.2 The CCDF of flights and ICTs measured from Campus I and II data sets. Various known distributions are fitted using MLE. .................. 28
Figure 3.3 Hurst values and fractal dimensions of waypoints extracted from the aggregated traces of each campus. ........................................... 29
Figure 3.4 It shows the self-similar nature of the dispersion of waypoints in the KAIST trace. At different scales, the dispersion of waypoints looks similarly bursty. ................................................................. 32
Figure 3.5 Lifecycles from a real data set and existing mobility models. .............. 37
Figure 3.6 Flight distributions of synthetic traces from various models are compared with the measured flight distribution from the Campus I data set. Plots are divided into two for clarity. .............................................. 44
Figure 3.7 ICT distributions of synthetic traces from various models are compared with the measured ICT distribution from the Campus I data set. Plots are divided into two for clarity. .............................................. 45
Figure 4.1 Sample traces of (a) TLW, (b) RWP and (c) BM ............................. 48
Figure 4.2 Our TLW model can generate synthetic traces that match the flight and pause time distributions seen from real human walk traces. .......... 50
Figure 4.3 The CDF of user displacement from their initial positions after the same amount of travel time. The RWP is most diffusive while the BM is least diffusive. The diffusion rates of Levy flights are in-between. The values in the parentheses represent Levy exponent for flight length. .......... 54
Figure 4.4 The CCDF of ICT of Levy Walk in infinite space. ............................... 61
Figure 4.5 The CCDF of ICT of Truncated Levy Walk in finite space. .................. 63
Figure 4.6  The CCDF of ICT collected from Dartmouth and UCSD. Various known distributions are fitted to the measured data distribution using the maximum likelihood estimation technique.

Figure 4.7  The CCDF of ICT collected from Infocom 2005 and simulation result by TLW. Various known distributions are fitted using MLE. TLW recreates the ICTD seen in the empirical data sets.

Figure 4.8  The ICT distributions of the TLW models. Levy walks recreate the truncated Pareto ICT distributions seen in the empirical data sets. The numbers in the figure represent the relaxation time from analyses.

Figure 4.9  A DTN routing delay according to the two hop relay algorithm. FCT and RICT can be considered as the residual times of ICT.

Figure 4.10  The ICT and DTN delay distributions with single and multiple relays. Results from both analysis and simulation are matching well. The values in parenthesis represent the number of relays.

Figure 4.11  The ICTDs of TLW ($\alpha=1.5$, $\beta=1$) according to various amounts of simulated hours. The values in parenthesis represent delivery ratio. When the truncation of the ICTD comes from the insufficient duration of the simulation, the delivery ratio is less than one.

Figure 4.12  The ICT distributions of various mobility models.

Figure 4.13  The DTN delay distributions of various mobility models and normalized 99% quantile delay with multiple relays. The numbers in the parenthesis represent the actual delays in minutes at the 99% quantile of the distributions.

Figure 5.1  Hurst parameter estimation of waypoints registered in all KAIST traces.

Figure 5.2  Hurst parameter values of waypoints in each site map. All show Hurst values higher than 0.6 except NYC.

Figure 5.3  Measuring aggregated variance of waypoints aggregated from all walk traces. We divide the area by non-overlapping $d$ by $d$ squares, and count the number of waypoints registered in each square and then normalize the sampled count by the size of the unit square. We compute the normalized variance as we increase $d$.

Figure 5.4  Delaunay triangulation of waypoints extracted from one daily trace of KAIST. In the inset, the CCDFs of the line lengths from the triangles and the flights from the same trace are plotted. The CCDF of flights and gaps match closely.

Figure 5.5  Delaunay triangulation is performed on all individual daily traces. The line segment lengths in Delaunay triangles are aggregated and their CCDFs are plotted for different walkabout sites. The CCDF of flights obtained from the corresponding traces are also plotted for comparison.

Figure 5.6  For the GPS traces of 100 participants, we measure the percentage of flights meeting the least-action criterion. We also plot the case when we exclude those flights which lengths are less than 30 meters.
Figure 5.7  Flight distributions obtained from LATP for different $a$ values performed on top of waypoints extracted from the KAIST traces and the percentage difference between the sum of flights generated from LATP and that of flights from real traces ................................. 97
Figure 5.8  Sample walk traces of various models. ......................................................... 102
Figure 5.9  Flight length distributions of synthetic traces from various models. .................. 103
Figure 5.10  The ICT distributions of synthetic traces from various models ....................... 105
Figure 5.11  The link duration distribution ................................................................. 106
Figure 5.12  Average routing delays of various protocols under CMM ............................... 109
Figure 5.13  Average routing delays of various protocols under Dartmouth, ORBIT and SLAW. ......................................................... 110

Figure 6.1  The relationship between the fractal dimensions ($D$) and the power-law slopes of corresponding gap distributions for sample fractal waypoint maps. The waypoint maps are generated by the Soneira-Peebles [80] model. ......................................................... 116
Figure 6.2  Hotspot size distributions. ............................................................................. 117
Figure 6.3  Relationship between the average number of visitors (squares) and the number of waypoints (dots) ......................................................... 118
Figure 6.4  Area of individual movement observed from the Campus I data set. ............... 119
Figure 6.5  Lifecycles from the Campus I data set and existing mobility models. ............. 124
Figure 6.6  Hotspot populations observed from the Campus I data set. Each square represents a hotspot. As the population increases, the square becomes darker. The incoming probabilities for hotspots 67 and 87 are shown. .................. 125
Figure 6.7  Hidden periodicity is discovered by eliminating the data collected during night time and weekends. ......................................................... 126
Figure 6.8  Various periodicity from hotspots in Campus I. ............................................. 126
Figure 6.9  The number of primary periods for each hotspot is represented by black squares. As the number of primary periods increases, the square becomes brighter. ......................................................... 128
Figure 6.10  The STCs observed from the four data sets. Solid lines represent linear regression results. X-axis represents the distance from the super-hotspot to each hotspot. (a)(c)(e)(g) and (b)(d)(f)(h) show the number of visitors and number of primary periods, respectively. ......................................................... 129
Figure 6.11  The primary period distributions. They follow exponential distributions. ......... 130
Figure 6.12  Relationship between the number of waypoints and the number of primary periods for hotspots ......................................................... 131
Figure 6.13  Assigned (solid line) and measured (dotted line) lifecycles for the proposed model. ......................................................... 136
Figure 6.14  A synthetic waypoint map generated by STEP hotspot model under the Campus I environment. We use $D=1.4$. ......................................................... 137
Figure 6.15  Flight and ICT distributions generated by STEP under the Campus I environment. We use $\gamma=5$. ......................................................... 140
Figure 6.16  Parameter sensitivity of STEP according to $\alpha$. We set $\gamma = 5$. .................. 141
Figure 6.17 Parameter sensitivity of STEP according to $\gamma$. We set $\alpha = 1$. . . . . . . 142
Chapter 1

Introduction
Simulating realistic mobility patterns of mobile devices is important for the performance study of mobile networks because deploying a real testbed of mobile networks is extremely difficult, and even with such a testbed, conducting repeatable performance experiments using mobile devices is not trivial. As most mobile nodes or devices (cell phones, PDAs and cars) are carried or driven by people, people are a critical factor in simulating mobile networks. Therefore, emulating the realistic mobility patterns of humans can enhance the realism in simulation-based performance evaluation of human-driven mobile networks.

“If life was so entertaining that we did not need any rest with ubiquitous supply of necessities and no responsibilities, the movement paths could be a true time-space random walk.” - T. Hagerstrand, Regional Science, vol. 24, no. 1, pp. 6-21, December 1970.

People’s movement is not random and typically governed by spatial and temporal constraints. People purposely visit a certain place at a certain time often driven by responsibilities and necessities. Capturing these constraints statistically and applying them to express realistic movements of humans in user-created virtual mobility environments for mobile network simulation is the main goal of our work.

A mobility model is most useful when it is able to govern the mobility of nodes “realistically” even in an environment that does not exist in reality. Users of mobility models may want to test their hypothesis (or verify the performance of their protocols) for many diverse environments. It is typical that users create these environments instead of modeling the testing environments from physical environments. The environments given by the users are often very sparse in the details; they may just give the size of mobility areas, a few landmarks or hotspots and the number of mobile nodes. Sometimes, these landmarks or hotspots are not even given. The job of mobility models is to create “realistic” mobility patterns for user-given environments which include these virtual environments as well as real physical environments. This requirement of mobility models rules out any models that require detailed information about the environments such as transition probability functions which govern the probability of moving from a location
to another at a given time.

Another key difficulty in mobility modeling is to define realism. When we say that a synthetically generated mobility traces are realistic, what do we mean by “realistic”? It is impossible to mimic the movement of people to every little detail. Instead, we focus on the statistical features that can significantly influence the performance of mobile networks. It is commonly accepted that the distributions of inter-contact times (ICT) \([17,22,25,38,47,54,65,75]\) and flights \([54, 75]\) are the most influential characteristics of mobility traces determining the performance of mobile networks. ICT is the time duration until a node meets another node after meeting that node previously where meeting (or contact) is defined by being in a radio range. A flight is the Euclidian distance between two waypoints visited consecutively in time by the same person. A waypoint is a geographical location where a person pauses his or her movement (a more general notion of waypoint may include the location where people change movement directions; in this dissertation, we take a more simplified notion of waypoint).

The performance of mobile networks such as DTN undoubtedly depends on contact probability. Mobile nodes forward and deliver messages when two nodes are in a radio range for a sufficient amount of time. Therefore, contact probability determines the probability and latency of message deliveries. The ICT distribution governs the probability of contact as it measures the time for a node to meet another from its last meeting. Flight distributions also influence the meeting probability. It is shown in \([75, 77]\) that mobility with many long flights such as Random WayPoint (RWP) \([45]\) tends to have high meeting probability than the one with many short flights such as Brownian Motion (BM) \([29]\). Flight patterns also determine the diffusivity of node distributions which is defined to be the distance that a node travels from an origin as a function of elapsed time. Longer flights induce higher diffusivity. ICT and flight patterns together characterize the probability of the existence of routing paths \([75]\).

Recent human mobility studies have discovered that both flight and ICT patterns have a heavy-tail tendency \([38, 54, 75]\). More precisely, their distributions have power-law head and body, and then truncations (or exponential tail). The truncations are results of the bounded
areas of mobility that people tend to have [22, 47, 75]. But the heavy-tail tendencies in flight and ICT distributions have never been observed together from the same mobility traces, largely due to lack of mobility traces from which detailed statistics about both flights and ICTs can be extracted. Some traces collected using iMotes [25] are good only for ICT measurements as they measure contact history while others [75] are for flight measurements as they track the GPS location of individuals and the individuals are not tracked at the same time. Some traces are based on WiFi associations of people [37, 59] or cell-tower association of cell-phone users [34]. But these data have too low resolution in their data with a margin of errors being a few hundred or thousand meters. Furthermore, while the WiFi data do not have wide coverage areas, so mobility outside the covered area is not recorded, the cell-phone data is not available to the public any more.

It is an academic interest what type of ICT distributions a given heavy-tail flight pattern may induce. While there have been several recent theoretical studies on related topics [17, 22, 47, 54, 65, 75], it is still an unsolved problem. However, it is conjectured, and verified by simulation [38, 54, 75], that power-law flights induce power-law ICTs. What’s more, obtaining both flight and ICT distributions from the same trace has a lot of practical value since they can be used to verify a mobility model. It is easy to construct a model that matches one of the features, but not both. For instance, we can imagine a variant of the RWP model where a node randomly picks a next waypoint in a way that the distance from the current waypoint (i.e., flight) follows a given heavy-tail distribution. So as long as the input flight distribution is realistic, the generated flight distribution from the model is trivially realistic. But it cannot be verified whether the synthetically generated traces induce a realistic ICT distribution as well since there is no corresponding ICT data to the flight distribution which can be used for the verification.

In this dissertation, we have collected 7 data sets using high quality GPS receivers. For our experiments, we have chosen five different sites: two university campuses (NCSU and KAIST), one metropolitan area (New York City), one theme park (Disney World) and one state fair
(North Carolina State Fair). The participants walk most of the time in these sites and may also occasionally travel by bus, trolley, cars, or subway trains. The GPS receivers record their location information every 5 seconds within a three meter margin of error. Two of the data sets are collected at the same time over a period of one week so contact information among the volunteers as well as their location information is available in those sets. The participants were recruited from the students at NCSU and KAIST who take a computer-literacy class and they are from over 30 different departments in each campus. Our data analysis confirms that all the data traces from both experiments have heavy-tail flight and ICT distributions. This is the first empirical study confirming the existence of heavy-tail tendency in flight and ICT distributions in the same traces.

We then use the measured ICT and flight data from our experiments to verify whether existing models are capable of reproducing realistic flight and ICT distributions. We have tested 8 different mobility models and by varying various input parameters for the models, we tried to fit their ICT and flight distributions to the measured distributions. Unfortunately, none of these models can match the flight and ICT distributions observed from the real traces. Most of them do not have heavy-tail tendencies in flight or ICT distributions.

We study how we can express both heavy-tail flights and ICTs from the spatial viewpoint. From the analysis results of our GPS data sets, we found out that the waypoints that people generate have a fractal nature and people unconsciously visit relatively near waypoints first. We conjecture that those two properties are related each other and would be the causes of heavy-tail flights and ICTs. To study the mechanism more in detail, we develop a mobility model. In this model, we generate a fractal waypoint map first. Then we apply a heuristic algorithm called Least Action Trip Planning (LATP) that generates a daily trip over these waypoints minimizing the total traveling distance. In this way, we can recover heavy-tail tendencies in flight and ICT distributions. But, unfortunately, they do not match closely the observed distributions from the data sets.

We view this inability of reproducing realistic flight and ICT distributions by the existing
models is essentially due to the poor representation of spatio-temporal correlations present in real human mobility. Especially, ICTs are governed by where and when people meet. The mobility of people has regularity in their daily visiting locations and times. Most of students have daily schedules (classes and meetings) governing their visiting locations and times. For instance, classrooms have a tendency of being populated around class times with regular ups and downs in populations. Cafeterias and restaurants have typical high populations around meal times, and shopping areas or student unions tend to have more constant flows of populations with many small ups and downs.

Fig. 1.1 (a) shows our measured data of the population changes over one day time of four hotspots in a campus. We say that the function of population over time in a hotspot is the *daily lifecycle function (DLF)* of the hotspot. The daily lifecycle shown in the figure is clearly changing over time. Fig. 1.1 (b) shows the DLFs of a hotspot in other models. They are essentially flat, showing no temporal variations at all. This is because most of the models do not explicitly model the times at which a node visits a certain location, but rather choose the times randomly. We also find from our traces as well as other existing traces that waypoints are clustered together forming hotspots of various sizes. The size of a hotspot is defined by the number of waypoints within that hotspot. We find that there always exists at least one *super-hotspot* whose size dominates the mean of the hotspot sizes. This pattern is manifested by the power-law distribution of hotspot sizes. This notion is different from the clustering effect caused by the self-similarity of waypoints. What’s more, we find that the size and DLF of a hotspot is strongly correlated with the location of the hotspot. More precisely, (1) the size of a hotspot is inversely proportional to the distance of that hotspot to the super-hotspot and (2) the number of dominating frequencies in the DLF of a hotspot is proportional to that distance. We define a dominating frequency in a DLF to be a frequency component present in the function whose co-efficient is larger than a predefined threshold.

Using these findings on the spatio-temporal correlations present in the human mobility traces, we construct a new mobility model for human walks, called Spatio-TEmPoral (STEP)
Mobility. STEP first generates hotspots of waypoints in an input mobility area using a self-similar point generation technique called the Soneira-Peebles (SP) Model [80]. Then we assign a DLF to each hotspot whose dominating frequencies are set according to linear correlation functions extracted from real traces. STEP provides a set of linear correlation functions that a user of STEP can choose from. We then construct an individual walker algorithm, using that each node chooses a next waypoint to visit for a given time and location. The algorithm is based on a non-linear weight function that combines the LATP and a lifecycle-based decision making algorithm.

With simple manipulations of input parameters, we verify that STEP can easily generate a synthetic mobility trace whose statistical characteristics, in particular, flight and ICT distributions, match those of the real mobility traces. STEP takes as input a small number of parameters including the size of mobility area, the number of mobile nodes, the fractal dimension of waypoints ($D$), and two constants $\alpha$ and $\gamma$ related to its individual walker model. The fractal dimension determines the gathering patterns of waypoints and governs the power-law slope of the flight length distribution. STEP considers two parameters when choosing the next waypoint, the distance from the current waypoint and the lifecycle of the next waypoint. $\alpha$ is a parameter to determine the degree of choosing the nearest neighbor waypoint as the next destination. $\alpha$ governs the power-law tendency of the flight distribution. $\gamma$ determines the weight of lifecycles over the distance factor. By combining three inputs, $D$, $\alpha$ and $\gamma$, we could faithfully reproduce the flight and ICT distributions.

In this dissertation, we investigate “real” human movement patterns observed from empirical data sets. Based on the data analysis results, we propose a mobility model, STEP. By comparing the flight and ICT distributions generated from our model with the observed distributions from empirical data sets, we could confirm that our model faithfully reproduces reality.

The remainder of this thesis is organized as follows. Chapter 2 describes existing models in detail. Chapter 3 describes why we look into the empirical data sets, what we found from the observations of the data sets and the limitations of the existing models. Chapter 4 shows
the spatial features of the individual human movement patterns. Chapter 5 reports how people
gather together from the spatial point of view. Chapter 6 discusses the spatio-temporal features
of human movements and proposes the STEP mobility model. In chapter 7, we conclude this
dissertation and discuss our future work.
(a) Lifecycles measured from our data set during one daytime period.

(b) Lifecycles from existing models.

Figure 1.1: Lifecycles from a real data set and existing mobility models.
Chapter 2

Prior Work: Descriptions of Existing Models
Existing mobility models can be largely categorized into several groups: random and its variants, geographic, hotspot and social models. These models vary in their movement characteristics. In random and its variant mobility models, the mobile nodes move in a given area without restrictions. The destination, speed and direction are chosen randomly and independently of other nodes. In geographic models, the movements of mobile nodes are restricted by geographical obstacles such as buildings, streets or freeways. A hotspot is a region where a large number of users stay and spend a relatively larger fraction of time. Hotspot models describe the aggregation tendency of mobile users. Social mobility models are founded on social network theory. The models allow mobile nodes to be grouped together when they have social relationships among them. The group is mapped to a geographical space, with movements influenced by the strength of social ties.

2.1 Random Models

Random Way Point (RWP) [45], Random Direction (RD) [15] and Brownian Motion (BM) [29] or Random Walk (RW) are random mobility models.
2.1.1 RWP

In RWP, a mobile node chooses a random destination (waypoint) in a simulation area and moves to the waypoint with a speed chosen from a uniform distribution. In each waypoint, the user pauses for a certain period of time selected from a uniform distribution. After the pause, the mobile node chooses a new waypoint and moves there and it continues the process iteratively.

2.1.2 RD

In RD, a mobile node chooses a random direction in which to travel, then travels to the border of the simulation area in that direction. Once the simulation boundary is reached, the user pauses for a specified time, and continues the process.

2.1.3 BM

BM is a model named after the Scottish botanist Robert Brown. He noticed that pollen grains suspended in water showed a movement pattern following a zigzag path. BM characterizes the diffusion of tiny particles with a mean free path (or flight) and a mean pause time between flights. A flight is defined to be a longest straight line trip from one location to another that a particle makes without a directional change or pause. Einstein [29] first showed that the probability that such a particle is at a distance $r$ from the initial position after a time $t$ has a Gaussian distribution and thus is proportional to $\sqrt{t}$, i.e., the width or standard deviation of a Gaussian distribution. The mean squared displacement (MSD), which is defined to be the variance of the probability distribution, is proportional to $t$. It is a manifestation of the central limit theorem (CLT) as the sum of flight lengths follows a Gaussian distribution. BM is defined in a continuous space domain and RW is the another name of BM in a discrete space.

2.1.4 MWP

The Markovian Way Point (MWP) [42], Gauss-Markov (GM) [55], Smooth-Random (SR) [13], Reference Point Group Mobility (RPGM) [39] and Virtual Track (VT) [87] models can be
considered as random variant models. They have small improvements from random models in describing either spatial grouping behaviors or temporal dependencies in the velocity of users. For example, the MWP model allows adjusting pause times and momentary velocity. The distribution used to pick the next waypoint may depend on the current location, thus it implements Markovian transition probabilities among waypoints.

2.1.5 GM

The GM model is a variant of the RWP model. All the movement patterns except velocity are same as those of RWP. In the GM model, the velocity of a mobile node is assumed to be correlated over time and modeled as a Gauss-Markov stochastic process. It prohibits unrealistic abrupt velocity changes.

2.1.6 SR

The SR model is similar to the RWP model but instead of the sharp turn and sudden acceleration or deceleration, it proposes to change the speed and direction of node movement incrementally and smoothly [9, 13]. Mobility of a node may be limited by the physical laws of acceleration and velocity. The current velocity of a mobile node may depend on its previous velocity. Thus the velocities of a mobile node at different time slots are correlated. So the models such as GM and SR are called temporal dependency models [9].

2.1.7 RPGM

In RPGM, mobile nodes form several groups each of which contains one leader. All the members of a group move along their leader. Fig. 2.2 shows how the nodes move with RPGM.

2.1.8 VT

The VT model is a variant of the RPGM model. It can describe merging and splitting behaviors of groups by introducing the concept of switch stations and virtual tracks. Switch stations are
randomly placed in the simulation area and they are interconnected by virtual tracks. Groups move along the virtual tracks towards the stations. At switch stations, a group can be split into multiple groups heading toward different stations or multiple groups can be merged into a single group. In random models such as RWP, a mobile node moves independently of other nodes. Thus, these models do not capture the co-location pattern of human mobility. But using the RPGM and VT models, we can describe the property. Therefore we can call these models, group mobility models.

2.2 Geographic Models

2.2.1 OM

The Obstacle Model (OM) [43] with geographical constraints incorporates obstacles to emulate more realistic pathways of humans around obstacles using Voronoi diagrams. The obstacles are placed within a simulation area to mimic buildings. Then the authors use the Voronoi diagram [26] to construct movement paths. Mobile nodes are then randomly distributed across the paths. A destination is selected from obstacles, and the shortest path is chosen from the
starting point to the selected destination for each node. Fig. 2.3 shows a sample trace with the OM model.

**2.2.2 Manhattan/Freeway**

To emulate pathways, the Freeway and Manhattan mobility models [8] have been proposed. They restrict the movements of mobile nodes to follow the pathways. One of the simplest ways to describe geographic constraints is to restrict the mobile node movement to the pathways. Fig. 2.4 shows sample traces that describe movements in a freeway and downtown such as Manhattan. The graph based [83] model also tries to integrate geographic constraints into mobility models.
2.3 Hotspot Models

In representing social behaviors of humans, hotspot modeling is one of the most popular ways to represent collective behaviors. The Dartmouth [49], Pragma [16], Clustered Mobility Model (CMM) [56], and ORBIT [33] fall into this category.

2.3.1 Dartmouth

The Dartmouth model [49] estimates the locations and movement paths of mobile nodes from real data sets. Based on the estimated information about the users, hotspot regions and the transition probability for moving between hotspots are extracted. This model comes from real data but it requires a considerable amount of effort to generate the mobility model because hotspot locations and transition probability between hotspots must be given as input (instead of being generated). Thus, it is very hard to change the walkabout areas, the number of nodes and the locations of hotspots without any corresponding real data sets. Furthermore, since this model does not specify how nodes move within a hotspot, the generated flights are likely to be mostly long thus following an exponential distribution.
2.3.2 Pragma

Pragma [16] is a realistic mobility model that is based on the behavioral characteristics of the individuals. It relies on a simple principle of dynamic network models namely the preferential attachment [12]. They translate the concept into the spatial distribution characteristic of mobile nodes using the concept of attractors. Attractors are landmarks to which humans move; they appear for a certain period of time, do not move, then disappear. Attractors describe centers of interest for people. Humans arrive, behave in cycles until their time is up, then depart. Fig. 2.5 shows a snapshot of nodes where they move with the Pragma model.

2.3.3 CMM

Clustered Mobility Model (CMM) [56] is a model that is based on the preferential attachment theory. It first divides the simulation area into a number of subareas and uses them as attractors. Mobile nodes are assigned to a subarea using preferential attachment. The attractiveness of one area is determined by the current number of nodes assigned to that area. In CMM, a node flips a coin to determine whether to stay at the current hotspot or not. With a 50% chance, the node leaves the hot spot and randomly moves to another hot spot.
2.3.4 ORBIT

ORBIT [33] randomly creates a specified number of hotspots within a given area and a subset of hotspots is assigned to each node. A node moves only among its assigned hotspots. For movements between and within hotspots, it is completely random. Fig. 2.6 shows a diagram how the mobile nodes move with the ORBIT model.

2.3.5 TCM

In the Time-variant Community Model (TCM) [40], each node has a home location where the node periodically reappears. Each node randomly visits other locations when it does not move toward and is not at the home location. Thus TCM demonstrates a very simple temporal movement pattern since it only implements the periodic reappearance at a certain place as a temporal feature.
2.4 Social Models

Community Model (CM) [63,64] and Sociological Interaction Mobility for Population Simulation (SIMPS) [17] models fall into this category.

2.4.1 CM

CM [64] is a mobility model based on a social network theory. The CM model allows mobile nodes to be grouped together based on social relationships among individuals. The clustered group is then mapped to a topographical space. The movements of nodes are driven by the strength of the social relationships among them. In this model, the authors assumed that the strength of social ties can also be read as a measure of the likelihood of geographic co-location. They represent the social relationships in an Interaction matrix.

2.4.2 SIMPS

SIMPS [17] is also a model based on the social interaction of human motion. It uses a social graph representing social ties among individuals. Social models can be said to focus on the causes of mobility. Starting from an established research in sociology, this model defines a process called socio-station, socialize and isolate, that regulate an individual with regard to her/his own sociability level. Although the model defines only two simple behaviors, it shows
explicit collective aggregation patterns.
Chapter 3

Motivations
3.1 Data Sets

To investigate common patterns of human movements, we decided to analyze human mobility traces. Currently, numerous empirical data sets are available [1] but most of them only provide contact information among participants collected using WiFi or Bluetooth devices [25, 41]. Those data sets do not provide precise location information of device holders since Bluetooth devices do not have any functionality to record their locations. From WiFi data sets, we can estimate a user’s location from the location of the associated AP. It should incur a few hundred meters of location error [49].

To obtain human traces that contain fine-grained location information, we decided to collect new data sets using high quality GPS devices. GPS devices can record their locations with less than a few meters of error margin [2, 3]. We have collected 7 data sets from our own experiments. We complement our data sets with 3 external data sets collected from Dartmouth College [37], UCSD [59] and the INFOCOM 2005 experiment [25].

3.1.1 Experiments using GPS Receivers

The 1st Experiment

Five sites are chosen for collecting human mobility traces from September 2006 to January 2008. These are two university campuses (NCSU in the US and KAIST in south Korea), New York city (NYC), Disney World (DW) and one State Fair (SF) all in the US. The total number of traces from these sites is 226 daily traces. Garmin GPS 60CSx handheld receivers were used for data collection. The receivers are WAAS (Wide Area Augmentation System) capable with a position accuracy of better than three meters 95 percent of the time, in North America [2]. Occasionally, track information has discontinuity when bearers move indoors where GPS signals cannot be received. The GPS receivers take readings of their current positions every 10 seconds and record them in a daily tracking log. The summary of daily traces is shown in Table 3.1.

The participants at NCSU were randomly selected students who took a particular course
Table 3.1: Statistics of collected mobility traces from five sites.

<table>
<thead>
<tr>
<th>Site (# of participants)</th>
<th># of traces</th>
<th>Duration (hour)</th>
<th>Radius (km)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>min</td>
<td>avg</td>
</tr>
<tr>
<td>NCSU (20)</td>
<td>35</td>
<td>1.71</td>
<td>10.19</td>
</tr>
<tr>
<td>KAIST (32)</td>
<td>92</td>
<td>4.21</td>
<td>12.21</td>
</tr>
<tr>
<td>NYC (12)</td>
<td>39</td>
<td>1.23</td>
<td>8.44</td>
</tr>
<tr>
<td>DW (18)</td>
<td>41</td>
<td>2.17</td>
<td>8.99</td>
</tr>
<tr>
<td>SF (19)</td>
<td>19</td>
<td>1.48</td>
<td>2.56</td>
</tr>
</tbody>
</table>

in the computer science department. Every week, 2 or 3 randomly chosen students carried the GPS receivers for their daily regular activities. The KAIST traces are taken from 32 students who live in a campus dormitory. The New York city traces were obtained from 12 volunteers living in Manhattan or its vicinity. Their track logs contain relatively long distance trips. Their means of travel included cars, buses and walking. The Disney World traces were obtained from 18 volunteers who spent their Thanksgiving or Christmas holidays in Disney World, Florida, USA. The participants typically walked in the parks and occasionally rode trolleys. The state fair track logs were collected from 19 participants who visited North Carolina state fair which includes many street arcades, small street food stands and showcases. The event was very popular and attended by more than tens of thousands of people daily for two weeks. The site is completely outdoor and is the smallest among all the sites. Each participant in the state fair scenario spent less than three hours at the site.

The 2nd Experiment

Many human mobility data sets from diverse mobility environments are currently available. For example, our data from the first experiment used in TLW [75] and SLAW [54] provides GPS trace logs collected from over 100 volunteers, but each log is taken at a different time. Hui et al. [41] use contact pattern logs, collected using bluetooth devices. WiFi log data sets [37, 59] are also available. But none of the existing data sets can be used to validate both flight and ICT distributions generated by mobility models. For example, our first data set does not provide contact information since the data are not collected concurrently. Bluetooth data sets [41]
Table 3.2: AIC values: Truncated Pareto distributions give the best fit for the measured distributions.

<table>
<thead>
<tr>
<th></th>
<th>Campus I</th>
<th>Campus II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>flight</td>
<td>ICT</td>
</tr>
<tr>
<td>Exponential</td>
<td>188,878</td>
<td>219,061</td>
</tr>
<tr>
<td>Power law</td>
<td>169,012</td>
<td>170,917</td>
</tr>
<tr>
<td>Gamma</td>
<td>177,871</td>
<td>180,439</td>
</tr>
<tr>
<td>Truncated Pareto</td>
<td>167,755</td>
<td>166,892</td>
</tr>
<tr>
<td>Weibull</td>
<td>168,479</td>
<td>170,242</td>
</tr>
</tbody>
</table>

record only contact patterns hence, they do not provide spatial locations. WiFi data sets provide both spatial and contact information but their location information may contain a few hundred meters of errors.

To obtain mobility traces from which both detailed ICT and flight information can be extracted, we perform two measurement experiments. We have chosen two university campuses, KAIST and NCSU. We call both campuses as Campus I and II, respectively. Their maps and recorded waypoints are shown in Fig. 3.1. In each campus, we recruited about 100 student volunteers from over 30 different departments. All students from each campus are asked to carry the devices during the same week. Since the volunteers record their movements at the same time, we can extract very detailed flight and ICT distributions. Holux M-241 GPS loggers [3] are used for data collection with a position accuracy of better than five meters 95 percent of the time. The GPS receivers record their current positions at every five seconds.

Fig. 3.2 shows the CCDF (complementary cumulative density function) of flights and ICTs from each data set. We apply MLE to fit five known distributions, exponential, power-law, Gamma, truncated Pareto and Weibull distributions to the CCDF as done in [38]. The MLE of the truncated Pareto is performed over the x-axis range between the 1% and the 99% quantile of each distribution. We can observe visually that the truncated Pareto distribution fits best.

To quantify the degree of fitness of known mathematical distributions to observed distributions, we use the Akaike test which produces the Akaike Information Criteria (AIC) value as done in [38]. The smaller AIC value indicates a closer fit. The Akaike test results for our
measured flight and ICT distributions are shown in Table 3.2. We can see that the truncated Pareto distribution gives the best fit. This result confirms that both heavy-tail flight and ICTs may be present from the same traces.

It has been known that the waypoints of humans can be modeled by self-similar points [54]. We also checked the self-similarity of the waypoints collected from both campuses. The self-similarity is generally quantified through the Hurst parameter which can be measured by two well-known methods, the Aggregate Variance (AV) and the R/S methods [82]. The sample data are said to be self-similar if the Hurst parameter is in between 0.5 and 1. Fig. 3.3(a) shows the Hurst parameter values obtained from both tests using the aggregated traces of each campus over one-dimensional (X and Y) and two-dimensional spaces. All values are over 0.68 and can be said to show strong self-similarity. We use the same method suggested in [54,82] for the AV and R/S tests. Lee et al. [54] has shown that the Hurst parameter is a key factor to characterize the placement of waypoints.

Another common parameter to characterize the self-similarity is fractal dimension [58]. Fractal dimension is a quantity that represents how a fractal set fills a given space, as we zoom down to finer scales. A common characterization of a fractal set using this concept is provided by Minkowski dimension or box-counting dimension. We first imagine an evenly-spaced grid, and count how many boxes (grid squares) are needed to cover the set. Then as we make the grid finer, we see how this number changes. The number $N$ of boxes of size $R$ required to cover a fractal set follows a power-law, $N = N_0 \cdot R^{-D}$, where $N_0$ is a constant and $D$ represents the fractal dimension of the set. Fig. 3.3(b) shows the power-law tendencies in the number of boxes according to different box size. It shows the fractal dimensions are 1.4 and 1.5 for campus I and II, respectively.

### 3.1.2 External Data Sets

Recently, many human mobility data sets from diverse mobility environments are currently available [1]. We complemented our GPS traces with three of external data sets collected from
Dartmouth University [37], UCSD [59] and INFOCOM 2005 [25].

The Dartmouth trace data is one of the largest available sets of WiFi network traces collected since 2001. Dartmouth campus has over 190 buildings on 200 acres. Over 1,000 Cisco and Aruba APs are installed to completely cover the campus area. The traces collected contain data from nearly 10,000 users. Whenever clients authenticate, associate, roam, disassociate or de-authenticate with an AP, a syslog message is recorded, containing a timestamp in seconds, the client MAC address, the AP name and the event type. Since the Dartmouth trace data provides the locations of some APs, we can estimate the location of a client using the location of the associated AP. But it might incur a few hundred meters of location error [49].

The UCSD data set is collected from 275 freshmen PDA users for an 11 week period in 2002. The freshmen were the initial students in a new college on the university campus. Each PDA was equipped with a WiFi 802.11b network card. We can identify users according to their registered wireless card MAC addresses. But the mapping is anonymous. Since the UCSD trace data also provides the locations of APs, we can estimate the location of a user using the location of the associated AP. The UCSD data set is more exhaustive than the Dartmouth set, since it logs all reachable APs for each client at each time slot, while the Dartmouth data only logs the associated AP.

The INFOCOM 2005 data set was collected during the IEEE INFOCOM 2005 conference in Miami where 41 iMotes were carried by attendees for 3 to 4 days. In this experiment, only the contacts among devices are recorded since the device does not have any localization equipment such as a GPS. iMotes contacts were classified into two groups: iMotes recording the meetings with other iMotes are classified as internal contacts, while meetings with other Bluetooth devices (not the iMotes) are called external contacts. The internal contacts represent the data transfer opportunities among participants.
Figure 3.1: Waypoint maps from two campuses.
Figure 3.2: The CCDF of flights and ICTs measured from Campus I and II data sets. Various known distributions are fitted using MLE.
(a) The Hurst parameter values (with 95% and 99% confidence interval) by the AV and R/S methods. Their values indicate the self-similarity of waypoints. The X-axis represents the site and method used, e.g., ‘I (AV)’ represents the Campus I and aggregated variance method.

(b) The fractal dimensions by the box-counting method.

Figure 3.3: Hurst values and fractal dimensions of waypoints extracted from the aggregated traces of each campus.
3.2 Summary of Observations from the Empirical Data Sets

In this section, we show the characteristics of human movements extracted from empirical human traces. We have analyzed multiple data sets collected from various environments such as university campuses, a metropolitan area and theme parks. From all the results, we could confirm the same tendencies. Therefore, we can say that these properties do not rely on a specific condition and the following are general properties that are universally valid.

3.2.1 Spatial Features

Truncated power-law flights

Biologists [7, 71, 86] have found that the mobility patterns of foraging animals such as spider monkeys, albatrosses (seabirds) and jackals can be commonly described in what physicists have long called *Levy Walks*. The term Levy walks was first coined by Schlesinger et al. [78] to explain atypical particle diffusion not governed by Brownian motion (BM). BM characterizes the diffusion of tiny particles with a mean free path (or flight) and a mean pause time between flights. A *flight* is defined to be the longest straight line trip from one location to another that a particle makes without a directional change or pause. In Levy walks, flights follow a power law distribution.

In [74,75], we have shown that statistical patterns of human walks observed within a radius of tens of kilometers. We use mobility track logs obtained from the first experiment 3.1.1. From the data analysis of our traces, we find the followings:

- The mobility patterns of the participants in these outdoor settings have features congruent with those of Levy walks; their flight distributions and pause time distributions closely match truncated power-law distributions.

- There exist some deviations from pure Levy walks occurring due to various factors specific to human mobility including geographical constraints such as roads, buildings, obstacles and traffic. These deviations are manifested in our traces in the form of flight truncations.
which may make the flight distribution appear like heavy-tailed or even short-tailed at times.

Our results are supported by [20, 34] which are conducted in different scales using tracking of bank notes and cellphone locations.

Several studies [49, 74, 75] also show that the pause time distributions of human walks show heavy-tail tendencies.

**Heterogeneously bounded mobility areas**

Gonzalez et al. [34] report that people mostly move only within their own confined areas of mobility and different people may have widely different mobility areas. They used two data sets to explore the mobility pattern of individuals. The first set ($D_1$) consisted of the movement trajectories for 100,000 mobile phone users during a six-month period. Each time a user makes or receives a call or a message, the location of the routing cell tower was recorded. They reconstruct the users’ trajectories based on those data. To confirm that the obtained results were not affected by the irregular call pattern, they also studied a data set ($D_2$) that captured the location of 206 mobile phone users. Their calls are recorded every two hours for an entire week. In both $D_1$ and $D_2$ data sets, the spatial resolution was determined by the density of over 100 mobile towers. The average coverage of the cell towers were approximately $3km^2$, and over 30% of the towers covered an area of $1km^2$ or less.

They found that the distribution of displacement of all users is well approximated by a truncated power law with cut-off values of $400km$ and $80km$ for each data sets $D_1$ and $D_2$, respectively.

Compared with bank note results shown in [20], while the bank note results capture a convoluted results by heterogeneous users, the results of [34] mean that individual human beings show significant regularity. This tendency comes from the fact that humans return to a few locations, such as home or work. This tendency is also supported by [47].
Figure 3.4: It shows the self-similar nature of the dispersion of waypoints in the KAIST trace. At different scales, the dispersion of waypoints looks similarly bursty.

**Fractal waypoints**

The bursty dispersion of waypoints implies that people tend to swarm near a few popular locations and their popularity measured by the number of waypoints shows high burstiness: the popular locations tend to be very popular while most other areas are not.

Fig. 3.4 plots the aggregated waypoints of KAIST while we zoom in to smaller areas (denoted by small boxes) in the map. The patterns of swarming distinctively appear similar independent of their zoom resolution (or scale). Thus we can say that the waypoint sets humans make are *fractal sets*. A fractal is defined to be “a rough or fragmented geometric shape that
can be split into parts, each of which is (at least approximately) a reduced-size copy of the whole” [57]. It is also called self-similarity.

Fractals are easily found in nature. Examples include coastlines, trees, ferns, mountain ranges, stellar matters, meteorites, moon craters and systems of blood vessels. It is very interesting the waypoints that humans make also have the fractal characteristics. The fractal nature of waypoints plays an important role in producing power-law flight length distributions. This fact will be explained in detail in chapters 5 and 6.

3.2.2 Spatio-Temporal Features

Truncated power-law inter-contact times

DTN provides a challenging environment in which communications between nodes are intermittent [31]. DTN does not assume that there exists connectivity between nodes at a certain point in time. When the nodes are disconnected, the packets are stored and forwarded through intermittent contacts established by the mobility of the nodes. In this type of networks, ICT is a key determinant of routing performance.

Many simulation and theoretical studies of DTN routing (e.g. [35, 77]) have long assumed that the ICT distribution (ICTD) of human walks follows an exponential distribution. Exponential distributions make mobility analysis tractable and the simulation results with popular mobility models [23] such as the RWP and RD models can easily produce exponentially decaying ICT distributions. But recently, empirical studies [25] show that this assumption is wrong especially in the context of human mobility: the ICTD of human walks contains a power law tendency. Under the assumption that the ICTD has a power law tail, the DTN routing delays approach infinite because of the presence of infinitely long inter-contact times.

Recently, [47] shows that for random walks with home coming tendency, there exists a characteristic time until which the ICTD has a power law tendency and after which it decays exponentially. Concurrently, [22] also shows that mobility within a finite area is also a cause of the exponential decay of the ICTD for random walks. These finding on the exponential ICTD
tails imply that the delay of opportunistic routing algorithms should be finite in contrast to the infinite delay under the power law ICTD assumption.

We do not yet have mathematical models to describe the dichotomy of ICTD that are easy enough like exponential distributions for the performance analysis of DTN routing. To solve this problem, we analyze three empirical data sets of human ICT [38]. We observe from Maximum Likelihood Estimation (MLE) and Akaike test [50] results that the ICTD closely follows a truncated Pareto distribution. Truncated Pareto distributions have a truncation point that corresponds to the characteristic time in the ICTD. We show that the closed form expression of DTN routing delay can be induced from the ICTD model.

**Spatio-Temporal Correlations**

The mobility of people has regularity in their daily visiting locations and times. For example, most of students have daily schedules (classes and meetings) governing their visiting locations and times. Classrooms have a tendency of being populated around class times with regular ups and downs in populations, cafeterias and restaurants have typical high populations around meal times, and shopping areas or student unions tend to have more constant flows of populations with many small ups and downs.

Fig. 3.5 (a) shows our measure data of the population changes over one day time visiting four popular hotspots in a campus. The DLF shown in the figure is clearly changing over time. Fig. 3.5 (b) shows the DLF of a hotspot in other models. They are essentially flat, showing no temporal variations at all.

We find that the size and location a hotspot and the characteristics of its DLF are strongly correlated as follows.

- The size of a hotspot is inversely proportional to the distance of that hotspot to the super-hotspot.
- The number of dominating frequencies in the DLF of a hotspot is proportional to that distance.
Using these findings on the spatio-temporal correlations present in the human mobility traces, we construct the STEP mobility model. STEP and the spatio-temporal correlations will be explained in more detail in Chapter 6.

In summary, Table 3.3 shows how the existing models satisfy the characteristics of human movements mentioned above. F1 to F5 represent the followings.

- (F1) Truncated power-law fight lengths and pause-times
- (F2) Heterogeneously bounded mobility areas
- (F3) Truncated power-law inter-contact times
- (F4) Fractal waypoints
- (F5) Spatio-Temporal Correlations
Table 3.3: Existing routing models can be categorized into four groups. Typical models in every category have been listed. None of the existing models have all the characteristics of human walks. '?' means that it is unclear from the model description.

<table>
<thead>
<tr>
<th>Category</th>
<th>Models</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>RWP [45]</td>
<td>F1 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F2 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F3 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F4 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F5 N</td>
</tr>
<tr>
<td>Random</td>
<td>RD [15]</td>
<td>F1 N</td>
</tr>
<tr>
<td>Variants</td>
<td></td>
<td>F2 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F3 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F4 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F5 N</td>
</tr>
<tr>
<td>Random</td>
<td>BM(RW) [29]</td>
<td>F1 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F2 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F3 Y</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F4 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F5 N</td>
</tr>
<tr>
<td>Random</td>
<td>MWP [42]</td>
<td>F1 N</td>
</tr>
<tr>
<td>Variants</td>
<td></td>
<td>F2 N</td>
</tr>
<tr>
<td></td>
<td>GM [55]</td>
<td>F3 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F4 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F5 N</td>
</tr>
<tr>
<td>Random</td>
<td>SR [13]</td>
<td>F1 N</td>
</tr>
<tr>
<td>Variants</td>
<td></td>
<td>F2 N</td>
</tr>
<tr>
<td></td>
<td>RPGM [39]</td>
<td>F3 Y</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F4 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F5 N</td>
</tr>
<tr>
<td>Random</td>
<td>VT [87]</td>
<td>F1 N</td>
</tr>
<tr>
<td>Variants</td>
<td></td>
<td>F2 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F3 Y</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F4 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F5 N</td>
</tr>
<tr>
<td>Geographic</td>
<td>Freeway [8]</td>
<td>F1 N</td>
</tr>
<tr>
<td></td>
<td>Manhattan [8]</td>
<td>F2 N</td>
</tr>
<tr>
<td></td>
<td>OM [43]</td>
<td>F3 ?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F4 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F5 N</td>
</tr>
<tr>
<td>Hotspot</td>
<td>Dartmouth [49]</td>
<td>F1 N</td>
</tr>
<tr>
<td></td>
<td>Pragma [16]</td>
<td>F2 N</td>
</tr>
<tr>
<td></td>
<td>CMM [56]</td>
<td>F3 N</td>
</tr>
<tr>
<td></td>
<td>ORBIT [33]</td>
<td>F4 Y</td>
</tr>
<tr>
<td></td>
<td>TCM [40]</td>
<td>F5 N</td>
</tr>
<tr>
<td>Social</td>
<td>CM [64]</td>
<td>F1 N</td>
</tr>
<tr>
<td></td>
<td>SIMPS [17]</td>
<td>F2 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F3 Y</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F4 N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F5 N</td>
</tr>
</tbody>
</table>
Figure 3.5: Lifecycles from a real data set and existing mobility models.
3.3 Limitations of Existing Models

3.3.1 Spatial Limitations

Numerous mobility models have been proposed [9]. They can be categorized in several groups according to spatial and temporal constraints they represent as shown in Table 3.4. Spatial constraints are the constraints that a model explicitly imposes to control flight patterns. Temporal constraints are pertaining to ICT patterns. A mobility model provides an algorithm to choose the next waypoint from the current waypoint. This algorithm governs the spatial and temporal features of a mobility model. When the next waypoint are chosen without any constraint, e.g., randomly chosen from a uniform distribution, we say that the model does not put any constraints on node movement.

In random and its variants models (e.g., [29, 45]), the nodes move in a simulation area without any restrictions so no spatial and temporal constraints are represented. Geographic and hotspot models (e.g., [9,43]) describe a type of spatial constraints that humans likely show. For example, the Manhattan model [9] uses a grid map and nodes move in horizontal or vertical direction on the map. Clustered Mobility Model (CMM) [56] uses the preferential attachment theory [11] to describe human gathering patterns.

It has been shown that if the next waypoint for a node is selected uniformly from a simulation area, the flight distribution follows an exponential distribution [54,75,77]. For example, in RWP [45], a node chooses a random destination in a simulation area. The resulting flight distribution of RWP follows an exponential distribution. Most of the existing models except SLAW [54] adopt similar random selection mechanisms to choose the next waypoint. For example, ORBIT [33] creates a number of hotspots at randomly selected locations. Movements between and within hotspots are randomly driven. In the Dartmouth model [49], a node generates a set of waypoints, uniformly distributed in a bounded box, when it moves between hotspots. In the CMM model [56], nodes move toward a uniformly selected location within the chosen hotspot. In the Community Model (CM) [65] and Time-variant Community Model
Table 3.4: Existing mobility models can be categorized by the constraints they describe. None of the existing models can faithfully describe both spatial and temporal constraints. $\Delta$ means that the constraints proposed do not produce heavy-tail flight and ICT distributions.

<table>
<thead>
<tr>
<th>Category</th>
<th>Models</th>
<th>Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Spatial</td>
</tr>
<tr>
<td>Random</td>
<td>RWP [45]</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>BM(RW) [29]</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>TLW [75]</td>
<td>$\times$</td>
</tr>
<tr>
<td>Random</td>
<td>RPGM [39]</td>
<td>$\times$</td>
</tr>
<tr>
<td>Variants</td>
<td>GM [55]</td>
<td>$\times$</td>
</tr>
<tr>
<td>Geographic</td>
<td>Manhattan [9]</td>
<td>$\Delta$</td>
</tr>
<tr>
<td></td>
<td>OM [43]</td>
<td>$\Delta$</td>
</tr>
<tr>
<td>Hotspot</td>
<td>Dartmouth [49]</td>
<td>$\Delta$</td>
</tr>
<tr>
<td></td>
<td>CMM [56]</td>
<td>$\Delta$</td>
</tr>
<tr>
<td></td>
<td>ORBIT [33]</td>
<td>$\Delta$</td>
</tr>
<tr>
<td></td>
<td>TCM [40]</td>
<td>$\Delta$</td>
</tr>
<tr>
<td></td>
<td>SLAW [54]</td>
<td>$\circ$</td>
</tr>
<tr>
<td></td>
<td>SWIM [60]</td>
<td>$\circ$</td>
</tr>
<tr>
<td>Social</td>
<td>CM [65]</td>
<td>$\times$</td>
</tr>
<tr>
<td></td>
<td>SIMPS [17]</td>
<td>$\times$</td>
</tr>
</tbody>
</table>

(TCM) [40] models, after communities are determined, they are associated to a location in the simulation area with a uniform distribution. Other models that are not mentioned here such as [16, 17] also use similar random selection mechanisms. So the resulting flights are not heavy-tail. SLAW [54] is the only model that does not choose the next waypoint randomly using a uniform distribution. It shows that a self-similar placement of waypoints and the LATP algorithm which assigns more weight on the waypoints near to the current waypoint in choosing the next waypoint induce a power-law flight distribution.

None of the existing models except TCM add any temporal constraints. TCM demonstrates a very simple temporal constraint. In this model, each node has a home location where the node periodically reappears. Each node randomly visits other locations when it does not move toward and is not at the home location. But the model does not satisfy the spatial constraint since it does not put any restriction on visiting the other locations and the next waypoint
is chosen randomly from a uniform distribution. Also it is not proven whether the temporal constraint of TCM enforces a heavy-tail ICT distribution. As shown above, numerous mobility models have been proposed but as far as we know none of the existing models can describe both spatial and temporal features of human movements.

### 3.3.2 Spatio-Temporal Limitations

In this section, we run mobility simulations using various mobility models to examine whether the exiting models can produce the flight and ICT distributions similar to the measured distributions from our traces. We examine RWP [45], TLW [75], Dartmouth [49], CMM [56], ORBIT [33], CM [65], TCM [40] and SLAW [54]. We tested using both Campus I and II data sets. Both tests produce similar results. To save the space, we present only Campus I test.

**Setup**

For simulation of various models, we use the following setup. These parameters are common to all models. We fix the simulation areas to be approximately the same as the measurement site of Campus I which is 2km by 2km. We assume that the entire area is divided into a number of equal-sized square cells of 200m by 200m. The transmission range of each node is set to 50 meters as in [54,75]. 70 nodes are simulated for 56 hours and the first 24 hours of simulation results are discarded to avoid transient effects. The speed of every user is set to 1 m/s for simplicity as in [54,75]. We use the pause time distribution obtained from Campus I traces.

For hotspot models such as the Dartmouth, CMM, SLAW, ORBIT, CM and TCM models, we use the same square cell to define each hotspot region. For SLAW, hotspots are defined to be the waypoint set in each square cell. When a model requires the use of real trace information, we use the ones extracted Campus I traces. For example, in the Dartmouth model, we use the transition probabilities measured from the Campus I traces as input. For SLAW, we use the Hurst parameter of 0.75 to generate a fractal waypoint map which we obtained from the Campus I traces.
For other input parameters to each model, we run many experiments and find the values that match the measured flight and ICT distributions best. In CMM, clustering exponent, $c_e$, determines the level of preferential attachment. When $c_e$ is equal to 0, nodes randomly choose their next waypoints from the whole area with the uniform distribution. As $c_e$ increases, the preferential attachment patterns become stronger so nodes tend to gather around in the biggest hotspot. We found that $c_e = 2$ produces the best result. In CM, rewiring probability, $p_r$, determines the probability of interaction with the nodes in other communities. Initially, each community has a group of nodes and each node in a community is connected to each other within the community by an edge. With $p_r$, each edge is broken and reconnected to a node in the other communities. As we increase $p_r$, the tails of the flight distribution become longer. $p_r = 0.1$ produce the best matching to the measured distributions. In SLAW, when the distance exponent $\alpha$ becomes larger, then a node is more likely to choose the nearer unvisited waypoint. $\alpha = 4$ produces the best matching heavy-tail tendencies to the given distributions. In TCM, $p_l$ is the probability to choose the local epoch. A node has two different modes of movement: local and roaming epochs. In a local epoch, the mobility of the node is confined within its community. In a roaming epoch, the node is free to move in the whole simulation area. At the end of each epoch, the node chooses the next epoch to be local with probability $p_l$. We set $p_l = 0.75$ to generate the best matching flight and ICT distributions.

**Flight and ICT distributions**

Fig. 3.6 shows that the flight distributions from the existing models do not match the measured distribution except TLW and SLAW (Note that we show the plots in semi-log scale since it can show the difference more clearly.) We perform the Kullback-Leibler (KL) divergence test [51] to determine the point where the flight and ICT distributions from each model match the measured distributions best. The KL-test provides a measure of the distance between two probability distributions. As we vary the input parameters of each model, we check the value of the average KL distance of flight and ICT distributions. We consider that the best matching value of each
parameter can be obtained when the average KL distance becomes the minimum.

We conjecture that the discrepancies between the measured flight and ICT distributions and those from each model come from the random selection mechanism mentioned in Section 3.3.1. When the nodes choose their next waypoints from a uniform distribution, the probability to have relatively short flights (e.g., less than 100 m) becomes small. We can also confirm this from Fig. 3.6. For example, in the real traces the proportion of flights less than 100 m is about 0.8. But with the RWP model, the generated flights are larger than 100 m.

TLW and SLAW seem to match the measured flight distribution better than other models. In TLW, the next waypoint is selected in a way that the resulting flight distribution follows a given heavy-tail distribution. But TLW is a random model so it cannot describe interactions among mobile nodes, which means it does not have hotspots where people gather together. SLAW first generates a self-similar waypoint map and mobile nodes move along a subset of the waypoints using the LATP algorithm. In this way, SLAW could reproduce the heavy-tail tendency present in the measured flight distribution. But we have found that the flight distribution by SLAW shows only heavy-tail tendencies and it is not a good match to the measured distribution. In addition, we set the distance exponent \( \alpha = 4 \) to obtain the best fit, which implies that nodes move mostly to the nearest waypoint. It is our common sense that people do not always move to the nearest location since most people have daily schedules governing their visiting locations.

Fig. 3.7 shows the ICT distributions obtained from various models. We can see that none of the existing models produce good match to the measured distribution. One of the main causes of this discrepancy is the inaccurate representation of the flight distribution. As we can see from Fig. 3.7, the proportion of small ICTs between 10 and 50 minutes generated by the existing models is much less than that from the real traces. It has been known that short ICTs are made by short flights [38,75]. The ORBIT model seems to match the measured ICT distribution better than others since with this model mobile nodes perform many short movements within hotspots.
Table 3.5: The results of Kullback-Leibler (KL) divergence tests: Each value represents the KL divergence value between the distributions from the real data sets and each model.

<table>
<thead>
<tr>
<th></th>
<th>Campus I</th>
<th>Campus II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Flight</td>
<td>ICT</td>
</tr>
<tr>
<td>RWP</td>
<td>0.6300</td>
<td>0.2048</td>
</tr>
<tr>
<td>TLW</td>
<td>0.0125</td>
<td>0.0430</td>
</tr>
<tr>
<td>Dartmouth</td>
<td>0.0485</td>
<td>0.0161</td>
</tr>
<tr>
<td>CMM</td>
<td>0.3686</td>
<td>0.0505</td>
</tr>
<tr>
<td>ORBIT</td>
<td>0.1042</td>
<td>0.0148</td>
</tr>
<tr>
<td>CM</td>
<td>0.2639</td>
<td>0.0175</td>
</tr>
<tr>
<td>TCM</td>
<td>0.2124</td>
<td>0.0491</td>
</tr>
<tr>
<td>SLAW</td>
<td>0.0224</td>
<td>0.0171</td>
</tr>
</tbody>
</table>

We conjecture that the second cause of the discrepancy is the lack of descriptions on temporal human movement patterns. The flight distribution is governed by only the spatial movement patterns but the ICT distribution is also determined by temporal features of movements since contacts occur when people go to the same place at the same time. As we have shown in Table 3.4, none of the existing models faithfully describe the temporal features of movement patterns.

As a summary, we perform the Kullback-Leibler (KL) divergence test [51] to quantify the closeness of the flight and ICT distributions generated by various mobility models to the measured distributions. The model that gives the minimum KL value means the best fit for the measured distribution. The table also contains the simulation results obtained from Campus II environments. The parameters used are measured from the Campus II data set. Table 3.5 shows that TLW and SLAW show relatively better matching results than other models in reproducing the measured flight and ICT distributions.

In this section, we have seen that none of the existing models match their flight and ICT distributions to the measured distributions. We view this inability is due to the fact that the existing models do not faithfully describe both spatial and temporal movement patterns of humans. In the next section, we investigate those patterns from empirical data sets.
Figure 3.6: Flight distributions of synthetic traces from various models are compared with the measured flight distribution from the Campus I data set. Plots are divided into two for clarity.
Figure 3.7: ICT distributions of synthetic traces from various models are compared with the measured ICT distribution from the Campus I data set. Plots are divided into two for clarity.
Chapter 4

Spatial Features of Human Movements, Part I: Individual Patterns
4.1 Overview

In this chapter, we focus on the individual movement patterns, i.e., heavy-tail or truncated power-law flight and pause time distributions. To evaluate the impacts of those patterns, we propose a simple model, Truncated Levy Walk (TLW), that contains both properties. We evaluate its characteristics such as ICTs and Delay Tolerant Network (DTN) routing performances.

\[ p(l) \text{ and } \psi(t) \text{ represent flight and pause time probability density distributions in TLW, respectively. Then their asymptotic behaviors can be expressed as follows} [46]. \]

\[ p(l) \sim |l|^{-(1+\alpha)} \quad (4.1) \]
\[ \psi(t) \sim t^{-(1+\beta)}, \text{where } t > 0 \quad (4.2) \]

\( \alpha \) and \( \beta \) have a value between 0 and 2. When \( \alpha \) (or \( \beta \)) is 2, \( p(l) \) (or \( \psi(t) \)) becomes a Gaussian distribution. When \( \alpha \geq 2 \), the model becomes BM due to the central limit theorem [79]. Flights and pause times cannot exceed certain values, \( f_{max} \) and \( p_{max} \), respectively. Fig. 4.1 illustrates sample traces of TLW, RWP and BM.

In TLW, a step is represented by four variables, flight length \( (l) \), direction \( (\theta) \), flight time \( (\Delta t_f) \), and pause time \( (\Delta t_p) \). Our model selects flight lengths and pause times randomly from their PDFs \( p(l) \) and \( \psi(\Delta t_p) \) which are Levy distributions with coefficients \( \alpha \) and \( \beta \), respectively. The following defines a Levy distribution with a scale factor \( c \) and exponent \( \alpha \) in terms of a fourier transformation,

\[ f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx-|ct|^\alpha} \, dt \quad (4.3) \]

For \( \alpha = 1 \), it reduces to a Cauchy distribution and for \( \alpha = 2 \), a Gaussian with \( \sigma = \sqrt{2}c. \) Asymptotically, for \( \alpha < 2 \), \( f_X(x) \) can be approximately by \( \frac{1}{|x|^\alpha} \). We allow \( c, \alpha \) and \( \beta \) to be simulation parameters. We use a uniform angle distribution as shown in most of our data and we do not add any geographical artifacts (other than boundaries) in our synthetic model. The flight speed of our model is set by the following relation between flight times and flight lengths:
Figure 4.1: Sample traces of (a) TLW, (b) RWP and (c) BM
\[ \Delta t_f = kl^{1-\rho}, \quad 0 \leq \rho \leq 1 \] where \( k \) and \( \rho \) are constants. In one extreme, when \( \rho \) is 0, flight times are proportional to flight lengths and it models the constant velocity movement. In another extreme, when \( \rho \) is 1, flight times are constant and flight velocity is linearly proportional to flight lengths. In our measurement data, the relation is best fitted with \( k = 18.72 \) and \( \rho = 0.79 \) when \( l < 500m \), and with \( k = 1.37 \) and \( \rho = 0.36 \) when \( l \geq 500m \).

Based on the above model, we generate synthetic Levy-walk mobility tracks with truncation factors \( \tau_l \) and \( \tau_p \) for flights and pause times respectively in a confined area as follows. First, the initial location of a walker is picked randomly from a uniform distribution in the area. At every step, an instance of tuple \((l, \theta, \Delta t_f, \Delta t_p)\) is generated randomly from their corresponding distributions. If \( l \) and \( \Delta t_p \) are negative or \( l > \tau_l \) or \( \Delta t_p > \tau_p \), then we discard the step and regenerate another step. We repeat this process after the step time \( \Delta t_f + \Delta t_p \). Until the end of the simulation, we generate the tuples repeatedly.

Now we verify whether TLW can synthetically generate the statistical features we have observed in our traces. Figs. 4.2 (a) and (b) show statistical distributions of flights and pause-time matching each scenario (we do not show the matching of NCSU data as it is similar to that of KAIST). To produce these traces, we set the simulation area by the same size of each corresponding scenario. We then vary the values of \( \alpha \) and \( \beta \) to find synthetic traces that have similar flight length and pause-time distributions of each scenario. We do not add any geographical constraints other than the simulation area (i.e., we set \( \tau_l \) to infinity) and any flight that goes outside the area is abandoned and a new flight is generated. We set the truncation of pause time \( (\tau_p) \) using the same values we obtained from the traces. Our synthetic traces show strikingly similar flight and pause time distributions seen from our real traces. This show the versatility of our model.
Figure 4.2: Our TLW model can generate synthetic traces that match the flight and pause time distributions seen from real human walk traces.

4.2 Characteristics

4.2.1 Characteristic function

Over the last few decades, physicists have been investigating Levy distributions. A characteristic function for a stable Levy distribution in Eq.(4.1) is defined as follows [46].

\[
L(z) = \exp\{-c|z|^{\alpha}[1 + i\text{sign}(t)\tan(\frac{\pi \alpha}{2})] + iz\}
\]  

(4.4)
The variable \( c \) is a scale factor characterizing the width of the distribution; \( m \) gives the peak position; \( \delta \) is a skewness parameter which characterizes the asymmetry of the distribution (\( \delta = 0 \) gives a symmetric distribution); \( \text{sign}(t) \) is just the sign of \( t \).

For our purpose, symmetric distributions are sufficient. Also we assume a maximum at \( t = 0 \), leading to \( m = 0 \), then the simplified version of Eq.(4.4), the characteristic function \( L_s(z) \), becomes

\[
L_s(z) = \exp\{-c|z|^\alpha\}. \tag{4.5}
\]

The idea of truncating a Levy distribution at some typical scale \( 1/\eta \) was mainly born in the analysis of financial data [76]. Its characteristic function can be defined as follows.

\[
T(z) = \exp\left\{-c\frac{(\eta^2 + z^2)^{\alpha/2}\cos(\alpha\text{arctan}|z|/\eta) - \eta^\alpha}{\cos(\frac{\alpha\pi}{2})}\right\} \tag{4.6}
\]

The distribution reduces to a Levy distribution for \( \eta \to 0 \) and to a Gaussian for \( \alpha = 2 \),

\[
T_s(z) \to \exp\{-c|z|^\alpha\} \tag{4.7}
\]

which is the same form as Eq.(4.5).

A truncated Levy distribution has finite variance so the sum of truncated Levy flights should converge to a normal process. But it is shown that until some amount of summation of independent stochastic variables the result does not converge to the expected Gaussian behavior [76]. It is known that there exists a clear crossover point between Levy and Gaussian regimes and it is a function of the Levy exponent and cutoff point. The crossover point makes it possible to define Levy and Gaussian regimes and with a relatively big cutoff, we can consider that the behavior of a TLW would be same as that of the Levy Walk (LW) with same parameters (Note that we denote the original Levy walk as LW). In the following sections, we consider the TLW with a big enough truncation point, so we can use most LW results for analyzing a TLW.
4.2.2 Stationarity

We are interested in showing the existence of the stationary distribution of a TLW and its limiting probability in 2-D area. We consider two boundary conditions, wrap around and reflection boundary. For the node distribution, the RWP model is said to have a nonuniform \([14]\) distribution but RD model has a uniform stationary distribution \([66]\). Let’s consider a user moving on a square 2-D area. We follow the Markov Chain (MC) modeling process for an RD \([66]\) to form a MC for the TLW. For the simplicity, let’s first prove for a wrap around model. We assume that at time \(T_j\) we select a flight length and a direction. The variable \(j = T_{j+1} - T_j, j \geq 0\) is the duration of the \(j\)-th movement; \(\theta_j\) denotes a new direction selected at time \(T_j; j \geq 1\). \(l_j\) means flight length which follows a distribution \(G_l(.), \) here, \(G_l(.)\) is a TLW. Let’s define a movement vector, \(\{y_j := (\tau_j, l_j, \theta_j)\}\). Then the state of the system can be represented by the vector

\[ Z(t) := (X(t), Y(t)) \]

where \(Y(t) := (R(t), L(t), \theta_t), R(t)\) is the remaining travel time at time \(t; L(t)\) is remaining flight length at time \(t; \theta_t\) is the direction at time \(t\). Then we can say that there exists a stationary distribution for a TLW.

**Theorem 1.** Uniform stationary distribution exists for the location of a TLW in a finite domain.

**Proof.** Firstly, let’s show \(\{y_j\}\) is an MC. Without loss of generality, we can assume that \(\tau_j, l_j, \theta_j\) have discrete values and they are finite. Then, \(y_j\) is finite since \(\tau_j, l_j, \theta_j\) are finite. It is apparent that the movement vector \(\{y_j\}\) is an MC since \(y_j\) does not depend on \(y_k\) where \(k < j - 1\). \(y_j\) is aperiodic since there can exist odd cycle for \(y_j\). And without loss of generality we can assume \(y_j\) is irreducible. Then by the convergence theorem in \([30]\), \(y_j\) has a unique stationary distribution. Secondly, let’s show that \(\{z_j := Z(T_j)\}\) is also an MC. Let’s say \(z_j = (x_j, y_j)\). It is shown above that \(\{y_j\}\) is an MC. \(x_{j+1} = x_j + l_j(\cos \theta_j, \sin \theta_j) - [x_j + l_j(\cos \theta_j, \sin \theta_j)]\), it means \(x_j\) depends only on the previous values of \(l_j\) and \(\theta_j\). So \(x_j\) is an MC. Then \(z_j\) is
an MC since both $x_j$ and $y_j$ are MCs. Then, $\{z_j\}$ is finite, aperiodic and irreducible with the same reason as in the reasoning of $\{y_j\}$, hence, $\{z_j\}$ is an MC. Finally, let’s show that $X(t)$ is a semi-Markov process and of which uniform stationary distribution exists. $X(t) = X(T_j) + \{(t - T_j)l_j/\tau_j\}(\cos \theta_j, \sin \theta_j) - \lfloor X(T_j) + \{(t - T_j)l_j/\tau_j\}(\cos \theta_j, \sin \theta_j) \rfloor$, which means it is not a Markovian process; i.e., the future trace of a user is dependent on the past history and possesses a memory since the user continues in a fixed direction for a specified distance between turning points. But this walk possesses some similarities to Markovian processes. Since, at each turning points, the position of the user’s next turning point is chosen independently of its past history. Thus it can be said to form a class of semi-Markovian processes [28]. Since the MC $\{z_j\}$ is aperiodic and irreducible and the expected travel time for a flight is finite, the stationary distribution for $Z(t)$ exists, which means the stationary distribution for $X(t)$ also exist by [66]. And by the same assumptions it is uniform.

**Corollary 1.** With a reflection boundary condition, Theorem 1 still holds.

**Proof.** Following the process of [66], we can also easily show the same properties of Theorem 1 hold for reflection model.

### 4.2.3 Diffusivity

One of the important characteristics of Levy walks is its high diffusivity [74,75]. Diffusivity can be defined to be the variance of the displacement between the current position at time $t$ and a previous position at time $t_0$. Fig.4.3 shows the amount of displacement for various mobility models plotted in CDF. In this simulation, all mobility models use the same velocity and pause time distributions, and traces are taken from the same area. It can be shown that the difference in displacement patterns comes from the difference in their flight distributions.

The Mean Squared Displacement (MSD), which is defined to be the variance of the displacement between the current position at time $t$ and a previous position at time $t_0$, is a measure of diffusion. Fig.4.3 shows the amount of displacement for various mobility models plotted in
Figure 4.3: The CDF of user displacement from their initial positions after the same amount of travel time. The RWP is most diffusive while the BM is least diffusive. The diffusion rates of Levy flights are in-between. The values in the parentheses represent Levy exponent for flight length.

Cumulative Distribution Function (CDF) forms. The MSD is highly related to the amount of displacement.

Assume $X(t)$ is the location of a user with $X(0) = 0$. For an ordinary diffusive process, the MSD scales with $t$ as $\langle X^2(t) \rangle \sim t$ in the large $t$ limit. If $\langle X^2(t) \rangle \sim t^\gamma$, we have the anomalous diffusion with $0 < \gamma < 1$ and $1 < \gamma < 2$ corresponding to sub-diffusive and super-diffusive cases, respectively. When $\gamma = 1$, it is called normal diffusion. We provide two basic propositions for the MSD. Let $A(t)$ and $B(t)$ be the positions of user $A$ and $B$ in 2-dimensional (2-D) space at time $t$, respectively. We assume that $\langle A(t) \rangle = \langle B(t) \rangle = 0$, and $A(t)$ and $B(t)$ are independent with each other. Then the following propositions hold.

**Proposition 1.** Let $C(t) = A(t) - B(t)$ where $A(t)$ and $B(t)$ have same value of $\gamma$, i.e. $\langle A(t)^2 \rangle \sim t^\gamma$ and $\langle B(t)^2 \rangle \sim t^\gamma$ in 2-D space. Then $C(t)$ satisfies

$$\langle C(t)^2 \rangle \sim t^\gamma.$$
Proposition 2. Let $A_X(t) = A(t) \cos \theta_t$ which means $A_X(t)$ in 1-D is the projection of $A(t)$ to $x$ axis. Let’s say $\langle A(t)^2 \rangle \sim t^\gamma$ and $A(t)$ and $\cos \theta_t$ are independent each other. Then $A_X(t)$ satisfies

$$\langle A_X(t)^2 \rangle \sim t^\gamma.$$ 

Proof. $\langle C(t)^2 \rangle = \langle (A(t) - B(t))^2 \rangle = \langle A(t)^2 \rangle + \langle B(t)^2 \rangle \sim t^\gamma$ \qed

We will also consider the LW mobility model since a number of useful results are shown for LW models and under some conditions those results can be applicable to the TLW. In the previous section, we show the relationship between LW and TLW, and the condition to use the results of LW for the TLW model. Consider an LW model with constant velocity, power law flight length distribution $p(t)$, and power law waiting time distribution $\psi(t)$. Their asymptotic behavior can be represented as follows [46].

$$p(t) \sim |t|^{-(1+\alpha)} \quad (4.8)$$

$$\psi(t) \sim t^{-(1+\beta)}, \text{where } t > 0 \quad (4.9)$$

The variable $\alpha$ and $\beta$ are the indices of the distributions which give the exponent of the asymptotic power-law behaviors; $\alpha$ and $\beta$ have a value between 0 and 2. The special case of $\alpha = 2$ or $\beta = 2$ gives the Gaussian distribution, respectively. In Sec. 4.3.1, we consider a continuous time random walk (CTRW) which has a power law flight length and a power law pause time, the relationship between $\gamma$ and power law coefficients of $\alpha$ and $\beta$ is known [85,88]. We will explain the CTRW in detail in Sec. 4.3.1.

Theorem 2. Assume a CTRW of which flight length and pause time distribution can be represented by Eq.(4.8) and Eq.(4.9) respectively. Then it is known that the diffusivity $\gamma$ of that
Table 4.1: Diffusivity tendency by various parameters of TLW user mobility

<table>
<thead>
<tr>
<th>Parameter</th>
<th>more diffusive</th>
<th>less diffusive</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\rightarrow 0$</td>
<td>$\rightarrow 2$</td>
</tr>
<tr>
<td>max flight length</td>
<td>$\rightarrow \infty$</td>
<td>$\rightarrow 0$</td>
</tr>
<tr>
<td>flight length scale</td>
<td>bigger</td>
<td>smaller</td>
</tr>
<tr>
<td>$\beta$</td>
<td>$\rightarrow 2$</td>
<td>$\rightarrow 0$</td>
</tr>
<tr>
<td>max pause time</td>
<td>$\rightarrow 0$</td>
<td>$\rightarrow \infty$</td>
</tr>
<tr>
<td>pause time scale</td>
<td>smaller</td>
<td>bigger</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>$\rightarrow 2$</td>
<td>$\rightarrow 0$</td>
</tr>
</tbody>
</table>

$CTRW$ can be computed as follows.

$$
\gamma = \begin{cases} 
\min(2, 2 - \alpha + \beta) & \alpha < 2, \quad \beta < 1 \\
\min(2, 3 - \alpha) & \alpha < 2, \quad \beta \geq 1 \\
\beta & \alpha \geq 2, \quad \beta < 1 \\
1 & \alpha \geq 2, \quad \beta \geq 1 
\end{cases}
$$

Proof. Refer to [85,88].

Diffusivity tendencies of various TLW parameters can be summarized as in Tab. 4.1. It is apparent to have more bigger flights make user behavior more diffusive. In contrast, longer pause time tend to make less diffusive behaviors.
4.3 Inter-Contact Times

4.3.1 Infinite Space

It is shown that any type of discrete time random walks in an infinite space without any boundary make the ICT power law decay [22]. Recently it is shown that human mobility and pause time distribution tend to follow power law [74,75], which means continuous time random walk (CTRW) model would be better to explain human mobility. We will extend existing result of human ICTD with the CTRW models to show the relationship between the bounds for power law slope and the diffusivity of underlying mobility.

Diffusion and FPT

For a stochastic process, the first passage time (FPT) is defined as the time $T_F$ when the process, starting from a given point, reaches a fixed point for the first time, which is a measure of diffusion [72]. Consider a stochastic process $X(t)$ with $X(0) = 0$. The FPT $T_F$ to the point $X = a$ is defined as

$$ T_F = \inf \{ t : X(t) = a \}. $$

In this section, we will consider a general CTRW which can produce both subdiffusive ($0 < \gamma < 1$) and superdiffusive processes ($1 < \gamma < 2$) as well as ordinary diffusion ($\gamma = 1$) depending on the form of Levy type flight length and Pareto waiting time distribution as used in [72,75]. The CTRW model is based on the idea that the length of a given jump and the waiting time elapsing between two successive jumps are drawn from a pdf $\psi(x,t)$ where $x$ and $t$ denote location and time [61,62]. The advantage of the CTRW model lies in the fact that we can efficiently manipulate the heavy tail characteristics of human flight length and pause time shown in [75].

The following results shown in [72] for the asymptotic behavior of the FPT probability
density functions (PDFs) for the above CTRW will be used in our proof of the main result.

\[
f(t) \sim \begin{cases} 
  t^{1-\gamma/2} & \text{for large } t \\
  \frac{r}{t^{(4\gamma)/(4-2\gamma)}} \exp\left(-\frac{d}{t^{\gamma/(2-\gamma)}}\right) & \text{for small } t
\end{cases}
\]  

(4.10)

where,

\[
r = \frac{a\gamma \sqrt{4K}}{2\sqrt{\pi} (2-\gamma)} \left( \frac{a\gamma}{2\sqrt{K}} \right)^{(\gamma-1)/(2-\gamma)}
\]

\[
d = \frac{2 - \gamma}{2} \left( \frac{\gamma}{2} \right)^{(\gamma-1)/\gamma} \left( \frac{a}{\sqrt{K}} \right)^{2/(2-\gamma)}
\]

K is a generalized diffusion constant. We make two comments. First, note that for \( \gamma = 1 \) Eq.(4.10) reduces to Eq.(4.11) which is the well known law for the ordinary Brownian motion.

\[
f(t) \sim \begin{cases} 
  t^{-3/2} & \text{for large } t \\
  \frac{a}{\sqrt{4\pi K t^3}} \exp\left(-a^2/4Kt\right) & \text{for small } t
\end{cases}
\]  

(4.11)

We will use an approach introduced by [22] to get a bound for ICT distribution (ICTD) using FPT distribution (FPTD). The FPTD \( f(t) \) is characterized by the power law relation in Complementary Cumulative Distribution Function (CCDF) form for large \( t \) in the following Lemma.

**Lemma 1.** *FPTD in CCDF form of a CTRW of which the diffusivity is \( \gamma \) can be represented as follows.*

\[
F(t) \sim t^{-\gamma/2}
\]  

(4.12)

*Proof.* Shown above. \( \square \)

Lemma 1 becomes the well known 1/2 scaling law for the ordinary Brownian motion for which \( \gamma = 1 \). For small \( t \) behavior in Eq.(4.10), simulation results show the exponential term is dominant for \( t \lesssim 10 \), which is not shown due to the lack of space. As we see in Eq.(4.12),
FPTD is a direct function of diffusion, \( \gamma \).

**Power Law ICT**

To the best of our knowledge, the general solution for the ICTD with anomalous diffusion has not been known. But it is known that the asymptotic bound for the ICTD can be computed by using the FPTD [22]. In this section, we get the lower bound of 2-D ICTD problem using the FPTD problem in 1-D semi-infinite interval. We use the same notation as in Sec.4.2.3. We denote \( A(t) \) and \( B(t) \) be the independent positions of users \( A \) and \( B \) at time \( t \). We define \( C(t) = A(t) - B(t) \), the vector difference between the positions of two users in 2-D area.

**Lemma 2.** The ICTD problem of \( A(t) \) and \( B(t) \) in 2-D area is analogous to the FPTD problem of corresponding \( C(t) \) to the origin from near origin in 2-D area.

**Proof.** Let’s assume both \( A(t) \) and \( B(t) \) move in a Cartesian coordinate system which defines a point uniquely in a plane through two numbers, the \( x \) and \( y \) coordinate. Let’s assume further that the system is divided into sub-squares, called cells. When only the two users are in a same cell, they can communicate each other. Assume \( C(0) = origincell \) and at \( t = 0^+ \) \( C(t) \) is not at the origin cell. Suppose at \( t = T_F \), \( C(t) \) arrives at the origin cell for the first time after \( t = 0^+ \), then we can consider \( T_F \) is ICT between \( A(t) \) and \( B(t) \), which is the FPT of \( C(t) \). So the ICTD between \( A(t) \) and \( B(t) \) in 2-D area has same distribution as the FPTD of \( C(t) \) to the origin from near origin in 2-D area.

Since the FPTD for 2-D anomalous diffusion is not known yet, we make a relationship between 2-D and 1-D FPT problem to use Lemma 2.

**Lemma 3.** The FPT of \( C(t) \) in 2-D area and the corresponding \( C_X(t) \), which means the projection of \( C(t) \) to \( x \)-axis, in 1-D area have the following relationship

\[
FPT_{C(t)} \geq FPT_{C_X(t)}.
\]
Proof. The $FPT_{C(t)}$ and $FPT_{C_X(t)}$ define the FPT of $C(t)$ in 2-D and the FPT of $C_X(t)$, respectively. Without the loss of generality, we can assume that $C(0) = \text{origin}$ and at $t = 0^+$, $C(t)$ is in the positive side of x-axis, not at the origin. For the $C(t)$ in 2-D to return to the origin, $C_X(t)$ should cross the origin first, which means $FPT_{C(t)} \geq FPT_{C_X(t)}$.

Now we have a relationship between ICT of $A(t)$ and $B(t)$ and FPT of $C_X(t)$, which means an ICT problem in 2-D can be described with the corresponding FPT problem in 1-D. Specifically, the latter is quite well known problem, the FPT problem in a semi-infinite interval $[0, \infty)$ such that there exists an absorbing boundary at $x=0$ and $C_X(t)$ starts from $x=1$.

**Theorem 3.** Let’s say $T_I$ denote ICT between $A(t)$ and $B(t)$. Then, ICTD of $A(t)$ and $B(t)$ satisfies

$$P\{T_I > t\} \geq Ct^{-\gamma/2} \text{for large } t.$$ (4.13)

**Proof.** Lemma 1 shows the FPT problem depends on $\gamma$ of the underlying mobility model. By Proposition 1, $\gamma$ of $C(t)$ is same to the $\gamma$ of $A(t)$ and $B(t)$. By Proposition 2, $\gamma$ of $C_X(t)$ is same to the $\gamma$ of $C(t)$. The two propositions state that the diffusivity is preserved through the process. Then, by Lemma 2 and Lemma 3, $P\{T_I > t\} = FPT_{C(t)} \geq FPT_{C_X(t)} = Ct^{-\gamma/2}$. This completes the proof.

**Remark 1.** Theorem 3 shows the tail distribution at least has power law distribution with slope between 0 and 1 when the $\gamma$ is in between 0 and 2, which is different from the result of [22]. They used the well-known Sparre-Andersen theorem. Sparre-Andersen theorem states that for any discrete-time random walk process with independent steps chosen from a continuous, symmetric arbitrary distribution, the FPT PDF decays asymptotically as $n^{-3/2}$, where $n$ is the number of steps [84]. This asymptotics was obtained under the assumption of exponential waiting time distribution of the single steps. But in reality human related movement cannot be modeled with discrete time model. So we get a better bound of ICT using CTRW. For example, the authors in [25] get the power law slope of 0.9 with experiments in Toronto, iMote Intel and iMote Cambridge which is not possible with the lower bound using Sparre-Andersen theorem. But with
Figure 4.4: The CCDF of ICT of Levy Walk in infinite space.

Theorem 3 it can be shown to be possible to get that slope with highly super-diffusive mobility models.

Remark 2. Theorem 3 says the diffusivity of underlying mobility model can be a good measure for ICT. Existing mobility models emphasize the difference in movement patterns but the meaning of Theorem 3 states that knowing of their diffusivity will help assume the distribution of ICT rather than the movement patterns.

Fig.4.4 shows numerical results on the ICTD with various exponents of $\alpha$ and $\beta = 0.5$. In this case, $\gamma = 2.5 - \alpha$ by Eq.(2), so we can see various ICT distributions according to the value of $\gamma$. The simulation is run for 300 hours and contact information is checked every 1 minute. In the simulation, 200 users start moving with the TLW at the origin initially. For an infinite space simulation, stationary distribution does not exist. After every contact the space look like same infinite space for every user, so the same initial point makes sense. The speed of every user is set to constant 1m/s for simplicity. Transmission range of mobile devices are set to 250m. The scale factors of flight length and pause time are set to 10, 1, respectively. Most of the parameters used here are from [74, 75]. The simulation results are consistent with the
statement of Theorem 3.

4.3.2 Finite Space

It is shown that in a finite space upper bound of ICT tail should follow exponential distribution and simply removing the boundary in a simple two-dimensional (2-D) isotropic random walk model, we can see power law ICTD [22]. And in [47], the authors find that there is a characteristic time, beyond which the distribution decays exponentially. Up to this time, the ICTD follows a power law. This power law finding was previously used to support the hypothesis that ICT has a power law tail. What is not obvious with existing work is not the tail part of exponential decay but the power law part to the characteristic time. In this section, we show that ICTD would follow power law until relaxation time $T_R$ (i.e. before exponential decay begins) and the slope of it is distorted and shows difference from that in infinite space.

Power Law

Fig. 4.5 shows numerical results on the ICT distributions with various exponents of $\alpha$ with $\beta$. We can see exponential decay at the tails and the power law slopes are distorted in contrast to the Fig. 4.4. The simulation area is fixed to 3.5x3.5 km$^2$, truncation points are set to 3.5 km and 8 hour for flight length and pause time, respectively. We use the same size of area as used in [75] and assume for human daily life the maximum pause time of 8 hours is quite reasonable. Initially every user is distributed uniformly in the area since we already know the stationary distribution is uniform from the Theorem 1. This setting of initial positions will make a sound simulation [18]. Other settings for the simulations are same for Fig. 4.4.

To explain the dichotomy in Fig. 4.5, let’s first think of an FPT problem in one dimensional string with a random walk mobility model [73]. The string has a mixed boundary conditions of reflection at $x = L$ and absorption at $x = 0$. If a user starts close to $x = 0$, the effect of the distant reflection boundary will not be apparent until the relaxation time in that system. The FPTD should decay as in the infinite system, which means the corresponding ICTD should
Figure 4.5: The CCDF of ICT of Truncated Levy Walk in finite space.
have at least heavy tail. In this case the power law slope should be the same as in infinite area case. But Unlike the distribution in infinite space, the power law slope of ICTD was found to be highly distorted in finite spaces as shown in Fig. 4.5. For example, $\gamma = 2(\alpha = 0.5, \beta = 1)$ in Fig.4.5 show a slope of 0.1 but the same $\gamma$ in Fig.4.4 shows a slope of almost 0.9. That is to say, higher diffusivity makes steeper ICTD in infinite space, but the tendency in finite space is totally reversed.

To the best of our knowledge, exact form of ICTD or even FPTD for superdiffusion in a finite space is not known yet. Instead, from the result of extensive simulation, we could find an empirical formula between the power law slope and $\alpha$.

$$\text{slope} \sim \begin{cases} 0.38\alpha - 0.3 & 1.5 < \alpha \leq 2 \\ 0.24\alpha - 0.09 & 1 < \alpha \leq 1.5 \\ 0.15\alpha & 0 < \alpha \leq 1 \end{cases}$$

(4.14)

The power law slope in finite space would be a function of every parameter listed in Tab. 4.1. But for a space between around 1 km$^2$ and 100 km$^2$, which cover most of the existing empirical ICT data sets [25,37,59,74,75], $\alpha$ is dominant, and other parameters such as $\beta$ show little impact to the power law slope.

**Exponential Decay**

In [20], the authors estimate the time to reach equilibrium for a 2-D Levy flight in a finite space. After this time the users should have reached an equilibrium distribution. Let’s call this point a relaxation time, $T_R$. We will show the characteristic time and the relaxation time have qualitatively same meaning. Let’s consider a renewal process with renewal interval $T_R$ then theorem 4 holds.

**Theorem 4.** Assume a mobility model of which stationary distribution exists. ICTD of the mobility model follows exponential decay after the relaxation time have past.
Proof. Let’s consider a renewal process with renewal period of $T_R$ and an initial condition at $t_0$. By the definition of relaxation time, at $T_R + t_0$ the locations of users are totally independent of the locations at $t_0$. Then the tail should follow geometric distribution which is analogous to exponential distribution. For a detailed mathematical representation of ICTD with renewal process, refer to [22, 44].

Relaxation Time

Let’s assume a square with a size of $L^2$. Single step radial distribution of TLW can be approximated by the truncated version of Eq.(4.1). By Bachelier’s equation [46](chap 3), the location after $N^{th}$ flight is the convolution of each transition probability. Convolution in time domain means multiplication in frequency domain, so from Eq.(4.7) we can get a characteristic function Eq.(4.15).

$$T_s(z)^N = \exp\{-cN|z|^\alpha\}. \quad (4.15)$$

$N$ is $t/\delta t_f$ where $\delta t_f$ denotes the typical time duration of a single flight. The relaxation time in a confined region is provided by the $N_m$-th mode.

$$k_{N_m} = \frac{2\pi}{L} N_m \quad (4.16)$$

Insert Eq.(4.16) into Eq.(4.15) with $N = t/\delta t_f$ to obtain

$$T_s(z)^N = \exp\{-\frac{t}{\delta t_f} - \frac{c}{2\pi\alpha}\}. \quad (4.17)$$

By the definition of relaxation time, after $T_{R'}$ in Eq.(4.17) the distribution is said to reach stationary state where there is no pause time.

$$T_{R'} = \frac{\delta t_f}{c(\frac{2\pi}{L})^\alpha} \quad (4.18)$$
Considering pause time will make Eq.(4.18) to

\[ T_R = T_R' \left[ \frac{T_R'}{\delta t_f} \right] \delta t_p. \]  

(4.19)

\( \delta t_p \) denotes the typical time duration of a single pause. \( \delta t_f \) and \( \delta t_p \) can be computed from the PDF of Truncated Pareto distribution as follows [4]. \( f_{min}, f_{max}, p_{min} \) and \( p_{max} \) represent minimum flight length, maximum flight length, minimum pause time and maximum pause time, respectively.

\[
\delta t_f = \begin{cases} 
\frac{f_{min}f_{max}(\ln f_{max} - \ln f_{min})}{f_{max} - f_{min}} & \text{if } \alpha = 1 \\
\alpha f_{max}^\alpha f_{min}^\alpha - f_{max}^\alpha - f_{min}^\alpha + f_{max}^\alpha & \text{else}
\end{cases}
\]

\[
\delta t_p = \begin{cases} 
\frac{p_{min}p_{max}(\ln p_{max} - \ln p_{min})}{p_{max} - p_{min}} & \text{if } \beta = 1 \\
\frac{\beta p_{min}p_{max}^\beta - p_{max}p_{min}^\beta}{\beta p_{max}^\beta - \beta p_{min}^\beta + p_{max}^\beta} & \text{else}
\end{cases}
\]

Finally, we need to get the best \( N_m \) for \( T_R \) in Eq.(4.16), since Eq.(4.19) gives too rough bound of relaxation. Using more exact \( T_R \) for normal diffusion with reflecting boundary [5], we can get the value of \( N_m = 13 \), which is consistent to abnormal diffusion cases too.

4.3.3 ICT Fitting

In order to evaluate routing algorithms for human driven DTN, it is necessary to investigate the ICTD of real human walk traces. We analyze three empirical data sets of human ICT. Those data sets are traces taken in UCSD [59], Dartmouth University [37] and Infocom 2005 [41]. The UCSD data records mobility patterns of 275 wireless PDA users within a campus WiFi network for the duration of 11 weeks. The Dartmouth data contains thousands of laptop/PDA users using campus WiFi networks over years. We pre-process one month worth of data from the
UCSD and Dartmouth traces to construct ICTD. We choose one-month traces because routing delay longer than one month would be useless for any possible applications of human oriented networks. In addition, we show next section that, one month is long enough periods for packets to be delivered considering the size of the area where the traces are taken and the diffusivity of users in both campuses. The experiment in Infocom 2005 has inter-contact information of 41 iMotes (Bluetooth devices) carried by attendees of the conference for 3 to 4 days.

We apply Maximum Likelihood Estimation (MLE) [50] to fit to the complementary cumulative density function (CCDF) of the produced ICTDs from the traces five well-known distributions, exponential, Lognormal, power-law, gamma, Weibull and truncated Pareto distributions. The gamma and Weibull distributions are considered good candidates to describe the dichotomy patterns. The gamma distribution has both a power law and exponential terms in its PDF and the CCDF of Weibull is known to be a stretched exponential form which also has a dichotomy. A log-normal distribution is a probability distribution of a random variable whose logarithm is normally distributed. The truncated Pareto distribution also has a power law tendency at the head part and decays exponentially at the tail. MLE is performed over the x-axis range between one minute and the 99% quantile of each data set. Figs. 4.6 and 4.7 show the CCDF of the ICTDs from each data set and the best MLE fittings for every distribution. Although the MLE itself does not give quantitative information on the closest matching distribution, we can observe visually the truncated Pareto distribution fits the best.

To get more concrete result on the best fitting distribution, we perform Akaike test. The Akaike test gives an estimate of the expected, relative distance between the fitted distribution and the unknown true distribution that generated the observed data [50]. In this test, the best fitting distribution is chosen to be the one with the minimum value of Akaike’s Information Criteria (AIC). Or if the value of Akaike Weight (AW) which is the relative likelihoods of each distribution has the closest value to 1, it is the best fitting distribution.

To quantify the degree of the best fitting, we perform the Akaike test [50]. The Akaike test gives an estimate of the expected, relative distance between the fitted distribution and
Table 4.2: The AIC and AW values for the ICT data from UCSD.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>AIC</th>
<th>AW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>315,970</td>
<td>0</td>
</tr>
<tr>
<td>Lognormal</td>
<td>265,500</td>
<td>0</td>
</tr>
<tr>
<td>Power law</td>
<td>255,290</td>
<td>0</td>
</tr>
<tr>
<td>Gamma</td>
<td>276,640</td>
<td>0</td>
</tr>
<tr>
<td>Truncated Pareto</td>
<td>253,550</td>
<td>1</td>
</tr>
<tr>
<td>Weibull</td>
<td>270,940</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.3: The AIC and AW values for the ICT data from Dartmouth.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>AIC</th>
<th>AW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>27,349,688</td>
<td>0</td>
</tr>
<tr>
<td>Lognormal</td>
<td>26,608,392</td>
<td>0</td>
</tr>
<tr>
<td>Power law</td>
<td>26,878,457</td>
<td>0</td>
</tr>
<tr>
<td>Gamma</td>
<td>26,893,923</td>
<td>0</td>
</tr>
<tr>
<td>Truncated Pareto</td>
<td>26,288,007</td>
<td>1</td>
</tr>
<tr>
<td>Weibull</td>
<td>26,786,551</td>
<td>0</td>
</tr>
</tbody>
</table>

the unknown true distribution obtained from the observed data. In this test, the best fitting distribution is the one with the minimum Akaike’s Information Criteria (AIC) value and with Akaike Weight (AW) the closest value to 1.

The AIC and AW values for the three data sets are shown in Tables 4.2, 4.3 and 4.4. They find the Truncated Pareto distribution to be the best fitting distribution. From these results, we conclude that the ICTD of human walks can be represented by a truncated Pareto distribution.

### 4.3.4 Characteristic Time

In this section, we quantify the characteristic time discussed in [47]. It has been known that the ICTD for the RWP and RD model has exponential tail [22, 44]. Their proofs are based on that the node movements can form a renewal process. For example, in the RWP model, the longest time duration for two nodes to finish two jumps can be interpreted as a renewal interval, say \( \zeta \). Since in the RWP model a node chooses a random destination in a simulation area so the
Figure 4.6: The CCDF of ICT collected from Dartmouth and UCSD. Various known distributions are fitted to the measured data distribution using the maximum likelihood estimation technique.
Figure 4.7: The CCDF of ICT collected from Infocom 2005 and simulation result by TLW. Various known distributions are fitted using MLE. TLW recreates the ICTD seen in the empirical data sets.
locations of two nodes at $t_0$ are independent of the locations at $t_0 - \zeta$.

Now let’s form a similar renewal process for the TLW model. In the stationary regime, the locations of each node are independent of the locations at initial time. Therefore, the renewal interval $\zeta$ is the duration for the locations of nodes to reach the stationary state from an initial state. In the viewpoint of diffusion process, stationary (equilibrium) state is defined in a space between reflecting boundaries, in which the probability density for node locations does not disappear, it just relaxes to equilibrium. Since the equilibrium state is reached by the effect of boundaries, we can say that before $\zeta$ have past, the effect of boundary will not be apparent [73]. Since it is shown that in infinite space without a boundary, the ICTD for any random walk model follows a power law [22], we can see that the power law ICTD appears at first and it became an exponential decay after $\zeta$ have past for the TLW model. This formulation can be applied to any mobility model that has the stationary regime.

Relaxation time between reflecting boundaries is the time elapsed until that the initial distribution of node locations reach the equilibrium state. From the above discussion, we can see that the relaxation time corresponds to the characteristic time. From now on, we use both the relaxation time and the characteristic time. It is known that how the relaxation time can be computed for normal diffusion [5], we include the result for completeness.

$$\tau_r(x|x_0) \equiv \rho^x(x)^{-1} \int_0^\infty [p(x,t|x_0) - \rho^x(x)]dt \quad (4.20)$$
\( p^\varepsilon(x) \equiv \lim_{t \to \infty} p(x,t|x_0) \) is the equilibrium distribution. The relaxation time is defined as \( \tau(x|x_0) \) where \( x_0 \) defines the initial location. \( p(x,t) \) denotes a distribution function at the location \( x \) at time \( t \). Then we can say the function \( p(x,t) \) is governed by the diffusion equation

\[
\frac{\partial p}{\partial t} = -\mathcal{L}_x(p). \tag{4.21}
\]

The diffusion operator in a \( d \) dimensional space can be written as

\[
\mathcal{L}_x = -x^{1-d}\frac{\partial}{\partial x}x^{d-1}D(x)e^{-\beta V(x)}\frac{\partial}{\partial x}e^{\beta V(x)}. \tag{4.22}
\]

\( D(x) \) is the diffusion coefficient; \( V(x) \) is a potential function and \( \beta = 1/k_BT \) where \( k_B \) is Boltzmann’s constant and \( T \) is the absolute temperature. Eq.(4.21) can be written as a continuity equation

\[
x^{d-1}\frac{\partial p}{\partial t} = -\partial(x^{d-1}j)/\partial x. \tag{4.23}
\]

The definition of the flux \( j(x,t) \) is evident from Eqs.(4.21) and (4.22). At the boundaries we impose reflecting boundary conditions

\[
j(a,t) = 0, \tag{4.24}
\]

where \( a \) denotes a boundary. Our simulation results for normal diffusion show that the solution from Eqs.(4.21) and (4.24) is quite accurate. To extend it to the anomalous diffusion, we need to change Eq.(4.21) to fractional Fokker-Planck equation (for sub-diffusion) or Levy fractional Fokker-Planck equation (for super-diffusion) as shown in [62]. But instead of solving the notoriously complex differential equations, we decide to use a simpler and intuitive approach. We make some modifications to the approach shown in [20], and combine it with the approach mentioned above. We show that it presents a quite good solution for both normal and anomalous diffusion.
In [20], the authors estimate the time to reach equilibrium for a 2-D Levy flight in a finite space. We also assume a square area with the size of $L^2$. Single step radial distribution of the TLW model can be approximated by the truncated version of Eq.(4.1). By Bachelier’s equation [46](chap 3), the location after $N^{th}$ flight is the convolution of each transition probability. Convolution in time domain corresponds to the multiplication in frequency domain, so from Eq.(4.7) we can get a new characteristic function, Eq.(4.15).

$$T_s(z)^N = \exp\{-cN|z|^\alpha\}$$ \hspace{1cm} (4.25)

The number of flights, $N$, can be represented by $t/t_f$ assuming zero pause time where $\delta t_f$ denotes the typical time duration of a single flight. Then the relaxation time in a confined region is provided by the $N_m$-th mode.

$$k_{N_m} = \frac{2\pi}{L}N_m$$ \hspace{1cm} (4.26)

Insert Eq.(4.26) into Eq.(4.25) with $N = t/\delta t_f$ to obtain

$$T_s(z)^N = \exp\{-\frac{t}{\delta t_f}\frac{1}{c(\frac{2\pi N_m}{L})^\alpha}\}.$$ \hspace{1cm} (4.27)

By the definition of relaxation time, after $\zeta'$ in Eq.(4.27), the distribution is said to reach stationary state. $\zeta'$ can be represented by the following equation.

$$\zeta' = \frac{\delta t_f}{c(\frac{2\pi N_m}{L})^\alpha}$$ \hspace{1cm} (4.28)

If we consider non-zero pause time, Eq.(4.28) becomes

$$\zeta = \zeta' + \left[\frac{\zeta'}{\delta t_f}\right]\delta t_p.$$ \hspace{1cm} (4.29)

$\delta t_p$ denotes the typical time duration of a single pause. $\delta t_f$ and $\delta t_p$ can be computed from the PDF of truncated Pareto distribution as shown in Eq.(4.30). $x_{\min}$ and $x_{\max}$ represent the
minimum and maximum flight length (or pause time).

\[
E(X) = \begin{cases} 
\frac{\ln x_{\text{max}} - \ln x_{\text{min}}}{x_{\text{min}}^{-1} - x_{\text{max}}^{-1}} & \text{if } \alpha = 1 \\
\frac{\alpha - 1}{\alpha} \frac{x_{\text{min}}^{-\alpha} - x_{\text{max}}^{-\alpha}}{x_{\text{min}}^{-\alpha} - x_{\text{max}}^{-\alpha}} & \text{else}
\end{cases}
\] (4.30)

Finally, we need to get the best mode number \((N_m)\) in Eq.(4.26) since Eq.(4.29) with the first mode which is generally used to get the relaxation time gives too rough value. For normal diffusion, the solution of Eqs.(4.21) and (4.24) is \(L^2/(3D)\), where \(D\) is a diffusion constant \(m^2/sec\). It corresponds to the solution of Eq.(4.29) with \(N_m = 6\). Fig.4.8 shows the analyses results using \(N_m = 6\) are consistent with the simulation results. The simulation is run for 600 hours in a 3.5 by 3.5 \(km^2\) square area. The contact information is checked every 1 minute. In the simulation, 50 users start moving with the TLW model, and the initial position of every node has a uniform distribution which is the stationary distribution of the model. We discard the first 100 hours of simulation results to avoid transient effects as shown in [19]. The speed of every user is set to 1 m/s for simplicity. Transmission range of mobile devices is set to 250m, which is the typical value of WiFi. The scale factors of flight length and pause time are set to 10 and 1, respectively. Most of the parameters are from [75].
Figure 4.8: The ICT distributions of the TLW models. Levy walks recreate the truncated Pareto ICT distributions seen in the empirical data sets. The numbers in the figure represent the relaxation time from analyses.
4.4 Routing Performance

In delay tolerant networks (DTN), mobile nodes may establish on and off connectivity with their neighbors and the rest of the network. Therefore, store-and-forward is the main paradigm of routing in such networks where communication transpires only when two devices are in a radio range. We call the time period that two nodes are in a radio range the contact time of the two nodes. One of the most widely studied routing algorithms in DTN is two-hop relay routing [35] where a source node sends a message (or a sequence of data packets) to the first node it contacts and then that first node acts as a relay and delivers the message when it contacts the destination node of the message. Here the period between the time that the message has originated and the time that the message is delivered to the relay node is called first contact time (FCT) to a relay and the period after that to the time the message is delivered to the destination is called remaining inter-contact time (RICT) between the relay and destination.

In a dense network, FCT is typically negligible and RICT dominates the message delay. One way to characterize RICT is to measure the inter-contact time (ICT), the time period between two successive contact times of the same two nodes. Since it is difficult to measure RICT from real mobility traces, ICT have been used to characterize RICT [25].

It is known that the ICT of human mobility exhibits strong tendency for power-law distributions [25]. The result is interesting because [22,77] showed that RWP produces exponentially decaying ICT in 2 dimensional (2-D) area, implying human mobility cannot be modeled by RWP, which seems obvious given the mobility patterns in RWP. Recently, it is shown that there is a characteristic time until which the ICT distribution (ICTD) follows power law and after that time ICTD decays exponentially [47]. They present evidence that the return time of a mobile node explain the observed dichotomy in simple 1-D random walk models. Concurrently, [22] shows the finite space plays an important role in creating the exponential tail of the ICTD regardless of the mobility models. In this section, we claim that the distribution of ICT with dichotomy is well fitted by a truncated form of power law distribution such as the positive side of Truncated Levy Walk (TLW) or Truncated Pareto model suggested in [4,75,76].
We show the ICTD from the data collected in UCSD, Dartmouth and Infocom 2005 [25,37,59] and fitted distribution using Truncated Pareto model in Figs. 4.6 and 4.7. We use Maximum Likelihood Estimation (MLE) as used in [38,75].

The relationship between ICT and RICT in two-hop relay algorithm can be represented by the relationship between the recurrence time and residual time as shown in Fig.4.9. The recurrence time is the time duration between two successively recurrent events, and the residual time is the time until the next event happens from an arbitrary point in time. Let $T$ and $R$ denote the recurrence and residual time process, respectively. Then it is known that the following properties hold [53].

\[
P(R > y) = \frac{E(T) - \int_0^y P(T > x)dx}{E(T)} \quad (4.31)
\]

\[
f_R(y) = \frac{P(T > y)}{E(T)} \quad (4.32)
\]

In the stationary regime, we can assume enough time has past so that all the nodes have met each other at least once before a message transfer starts. Since FCT is negligible in a dense network, the routing delay of a message is approximately equal to RICT between the relay and
destination nodes. Let $X$, $Y$ and $Z$ denote ICT, RICT and routing delay, respectively. In case that $X$ follows a truncated Pareto distribution with coefficient $\gamma$ between 0 and 2, the CCDF and expectation of $X$ can be represented as follows [4].

\[
P(X > x) = \frac{x_{\min} (x^{-\gamma} - x_{\max}^{-\gamma})}{1 - (x_{\min}/x_{\max})^\gamma}
\]

\[= \frac{(x^{-\gamma} - x_{\max}^{-\gamma})}{1 - x_{\max}^{-\gamma}}
\]

(4.33)

Here, $x_{\min}$ and $x_{\max}$ denote the minimum and maximum ICTs, respectively. For simplicity, we will assume $x_{\min} = 1$ in this section simplifying the above equations in Eqs.(4.33) and (4.34). Eq.(4.34) is a simplified form of Eq.(4.30).

\[
E(X) = \begin{cases}
\ln x_{\max} - 1, & \text{if } \gamma = 1 \\
\frac{\gamma - 1}{\gamma - 1 - x_{\max}^{-\gamma}}, & \text{else}
\end{cases}
\]

(4.34)

From Eqs.(4.31), (4.33) and (4.34), we get a closed form of routing delay distributions for the two hop relay protocol with a single relay as follows.

\[
P(Z > t) \simeq P(Y > t)
\]

\[= \frac{E(X) - \int_1^t P(X > x)dx}{E(X)}
\]

\[= \frac{E(X) - \frac{(t-1)(1-\gamma)-x_{\max}^{\gamma}(t^{1-\gamma}-1)}{(\gamma-1)(x_{\max}^{-\gamma}-1)}}{E(X)}
\]

\[= \frac{g(\gamma) + x_{\max}^{\gamma}(t^{1-\gamma}-1) - (t-1)(1-\gamma)}{g(\gamma)}
\]

\[= h(\gamma, x_{\max})
\]

(4.35)
\( g(\gamma) \) denotes \( \gamma(x_{\max}^\gamma - x_{\max}) \). Eq.(4.35) shows that the routing delay is a function of \( \gamma \) and \( x_{\max} \). From Eq.(4.32), we know the PDF of the RICT, so we can get a closed form expression of the expectation for \( Z \) using Eqs.(4.33) and (4.34).

\[
E(Z) \simeq E(Y) = \int_0^{x_{\max}} y \frac{P(X > y)}{E(X)} dy = \frac{1}{E(X)(1 - x_{\max}^{-\gamma})} \left( \frac{x_{\max}^{2-\gamma} - 1}{2 - \gamma} - \frac{x_{\max}^{-\gamma}(x_{\max}^2 - 1)}{2} \right) \quad (4.36)
\]

Numerical evaluation from Eqs.(4.34) and (4.36) shows that the expectation of \( Z \) is almost ten times bigger than that of \( X \). It is consistent with inspection paradox [70] [25] because \( Y \) is dominated by the large samples of \( X \). Note that Eq.(4.36) implies that the expectation is always finite in contrast to the results of infinite delays under the power law ICTD [25].

In a generalized relaying algorithm [25], the source distributes the message to the first \( m \) relays that it contacts. The routing delay is the time till any copy of the message is delivered to the destination. In this case \( Z \) can be represented by \( \min(Y_1, \ldots, Y_m) \) where \( Y_i \) represents the RICT by the \( i^{th} \) relay. Using simple probabilistic manipulation, we can see that the CCDF of \( Z \) is a multiplication of CCDFs of \( Y_1 \) through \( Y_m \). Then the closed form of the CCDF of \( Z \) can be shown as below. \( h_i(\gamma, x_{\max}) \) represents the result of Eq.(4.35) by the \( i^{th} \) relay.

\[
P(Z > t) = P(Y_1 > t) \cdots P(Y_m > t) = h_1(\gamma, x_{\max}) \cdots h_m(\gamma, x_{\max}) = h(\gamma, x_{\max})^m \quad (4.37)
\]

Fig. 4.10 shows the results of above analysis and corresponding simulation. We can see the analysis results are consistent with the simulation results by TLW. The simulation setups are the same as used in Fig.4.8.
So far, we have assumed that FCTs are negligible. We now consider a case that this assumption is not valid. We need to compute FCT and add it to RICT to get the exact delay. FCT can be considered as the minimum RICT from the source node to the other nodes. Let $W$ represent FCT. Then $W = \min(Y_1, \cdots, Y_{N-1})$ where $N$ is the total number of nodes.

$$P(W > t) = P(Y_1 > t) \cdots P(Y_{N-1} > t)$$

$$= h_1(\gamma, x_{max}) \cdots h_{N-1}(\gamma, x_{max})$$

$$= h(\gamma, x_{max})^{N-1} \quad (4.38)$$

We covered the cases that $X$ follows a truncated Pareto distribution, but this formulation can also be used for any distribution of $X$ such as exponential distributions with RWP. To get the closed form of routing delays for RWP, we only need to replace Eqs. (4.33) and (4.34) with the corresponding values from exponential distributions.
In addition, we study the impact of insufficient duration of measurement (or simulation) to the delivery ratio of messages. We already know that the power law ICTD experiences an exponential decay due to the effect of the boundaries, home coming tendency or the renewal interval formed by the stationarity of node locations. But another reason of the exponential decay is insufficient measurement (or simulation) duration. Since we cannot observe ICT bigger than the measurement (or simulation) duration, the ICTD look like decaying exponentially at the tail. In this case, the ICTD still fits the truncated Pareto distribution since the original power law part of the ICTD is truncated by insufficient measurement duration as shown in Fig.4.11. Even though the ICTD is distorted by insufficient time, the routing delay distribution can be calculated by the same process introduced in the previous section. But there should be a difference in the delivery ratio since the measurement (or simulation) is ended before some messages reach their destinations.

Fig.4.11 shows an example of insufficient simulation time. The simulation has the same setup as in Figs.4.8 and 4.10. Both 600 hours and 300 hours of TLW make the same ICTDs,
but for other cases the ICTD is truncated earlier and the delivery ratio is below one. From this figure, we can conclude that if the observed characteristic time is different from the analysis result by Eq.(4.19), the truncation is incurred by insufficient simulation duration. In this case, we can predict the delivery ratio is less than one.

The earlier measurement studies on ICT report power-law distributions of ICT with human mobility with slopes in the range between 0.3 and 0.9 [25]. By varying the parameters of $\alpha$ and $\beta$ of TLW mobility model, we are able to generate ICT distributions with the similar characteristics as in [25] by simulation. [25] reports power-law slopes of 0.3 from the INFOCOM trace [41] and 0.4 from the UCSD trace [59].

To see the relation between TLW and ICT, we measure the ICT and routing delays from simulation setups that mimic the environments of UCSD. Fig. 4.12 shows the result. In the UCSD simulation, we fix the simulation area to 3.5 km by 3.5 km, maximum flight length to 3 km and maximum pause time to 28 hours. These values are chosen based on the data from [59]. 40 nodes are simulated in both scenarios for 300 hours. We set the scale factors for flight lengths and pause times to 10 and 1, respectively. For all the simulation, we assume infinite buffer and that the message transmission rate is instant. These assumptions are used to isolate the effect of mobility patterns on the performance of DTN routing.

We also simulate RWP and BM in the same setup as the UCSD environment to compare the results. BM’s ICT distribution shows 0.5 power-law slope while RWP’s shows an exponential decay. The ICT distribution patterns of various mobility models are closely related to their diffusion rates examined in Fig. 4.5. The more diffusive the mobility is, the shorter tail its ICT distribution becomes. To confirm this pattern, we run Levy walks with different values of $\alpha$ while fixing $\beta$ to one. Fig. 4.12 shows that as $\alpha$ gets smaller, the tail distribution of ICT becomes shorter.

The routing delays of the TLW models tend to have high delays but the routing delays of RWP still show a short tail distribution. To see the effect of flight length distributions on routing delays more clearly, we measure routing delays in the simulation runs used for Fig. 4.12.
Fig. 4.12: The ICT distributions of various mobility models.

Fig. 4.13 (a) shows the result from which the following can be observed. BM tends to have very larger delays than any other models while RWP, as expected, shows the smallest delays because its probability of long flights is highest. The Levy walk models show their patterns in-between the two extremes: as we increase $\alpha$, their delays get closer to BM’s and as we reduce $\alpha$, they get closer to RWP.

The heavy tail distribution of routing delays may imply that many nodes experience similar long routing delays and that use of more relays (or copies of messages) may not necessarily improve the performance drastically. In a generalized relaying algorithm, the source distributes the message to the first $m$ relays that it contacts. The routing delay is the time till any copy of the message is delivered to the destination. Fig. 4.13 (a) shows the DTN routing delays of various models when one relay is used, and Fig. 4.13 (b) shows the 99% quantile delays of the same models normalized by their corresponding one-relay delays as we add more relays. As expected, BM hardly achieves this goal; the delay does not improve so much as the number
of relays increases, since every relay takes long time to meet the destination. However, we are
surprised to find that all our Levy walk models including the one with $\alpha = 1.5$ which shows
fairly similar delay patterns as BM for one relay case, show almost the same improvement
ratio as RWP as we add more relays. This implies that while in RWP, most nodes travel long
distances frequently, in Levy walks, although not all nodes make such long trips, there exist,
with high probability, some nodes within the mobility range of the source nodes that make such
long trips. This contributes to the great reduction of the delays even with a small number of
relays.
Figure 4.13: The DTN delay distributions of various mobility models and normalized 99% quantile delay with multiple relays. The numbers in the parenthesis represent the actual delays in minutes at the 99% quantile of the distributions.
4.5 Conclusion

This chapter finds that human walks contain statistically similar features to Levy walks including heavy-tail flight and pause time distributions. Combined with the results from [20] and [34], our result shows a scale-free nature of human mobility even beyond the scale of a few thousand kilometers. Based on these findings, we proposed a simple model, TLW, to evaluate the impacts of heavy-tail flight and pause time distributions on the inter-contact times and DTN routing performances.

From the ICT and routing performance analysis results, we find the following:

- We could confirm that the heavy-tail flights induce the heavy-tail ICT distributions.
- The heavy-tail ICT distributions of human walks can be modeled by a truncated Pareto distribution. From this model, we can induce the closed form expression of the DTN routing delay distribution. While the existence of dichotomy in the ICT distribution qualitatively suggests finite routing delays, we have shown it quantitatively.
- The TLW mobility model enables us to quantify the characteristic time of human walks using the relaxation time theory. It can be used to predict the message delivery ratio.

We view that our work is an important step for performance evaluation in human-driven DTN environments. Though there exist many studies on the patterns of the ICT distributions of human walks, they do not suggest quantitative approaches. Our work also points out that the ICT distribution patterns observed from empirical data sets can be generated by use of heavy-tail flights.
Chapter 5

Spatial Features of Human Movements, Part II: Aggregation Patterns
5.1 Overview

As shown in chapter 3, human aggregation patterns have a fractal nature and in chapter 4 we show that the flights of individuals have a heavy-tail tendency. In this chapter, we detail how to describe both power-law flights and the fractal aggregation patterns in synthetic traces. Due to mutual dependency among various parameters such as the degree of self-similarity and the characteristics of flight distributions (e.g., the power-law slope of the distribution), controlling them concurrently is hard. To show both properties are related to each other, we adopt the following approach to the problem. We first generate fractal waypoints using a technique similar to a fractional Gaussian noise (fGn) or Brownian Motion generation (fBm) techniques (e.g., [67, 68]) over a 2 dimensional space. Then we leverage fundamental properties of fractal points to generate power-law flights on top of them. We prove analytically that fractal points induce power-law gaps where gaps are defined to be inter-spacing among neighboring fractal points. Our analysis recovers algebraic relations between the Hurst parameter of fractal points and the power-law slope of the corresponding gap distributions. For details, please refer to [54].

It has been empirically known that the gaps among fractal points follow a power-law distribution [32,57]. Typically gaps over multiple dimensions are measured by Delaunay triangulation [32]. Delaunay triangulation for a set \( P \) of points in the plane is a triangulation \( DT(P) \) such that no point in \( P \) is inside the circumcircle of any triangle in \( DT(P) \). We experimentally verify that the line lengths of Delaunay triangles drawn on top of the waypoints extracted from real human walk traces have a power-law distribution and furthermore, their power-law slopes are almost identical to those of flight distributions from real traces. This strongly suggests that people plan their trips over known destinations (if we view waypoints as destinations) in a gap-preserving manner where they visit the nearby destinations first before visiting farther destinations. This trip planning is in fact a heuristic to the well-known traveling salesman problem whose objective is to minimize the total distance of travel and aligns well with the least-action principle of Maupertuis [27] where all objects are moving to the direction of minimizing their discomfort. In a human walk case, the discomfort is distance. It is intuitive that when visiting
a place, people strive to reduce the distance of travel by visiting all the nearby destinations before visiting farther destinations unless some high priority events such as appointments force them to make a long distance trip even in the presence of unvisited nearby destinations. The least-action principle has led us to developing a trip planning algorithm called *Least Action Trip Planning* (LATP). We show that LATP can recover the same flight distribution observed in real traces within around 10% error margin.

To show that the LATP algorithm on top of a fractal waypoint map induces power-law flights, we propose a mobility model, called Self-similar Least Action Walk (SLAW). SLAW satisfies the spatial features of human walks (from $F_1$ to $F_4$) shown in section 3.2. LATP and fractal waypoints satisfy $F_1$ and $F_4$. To complete the trace generation process while satisfying $F_2$ and $F_3$, we combine LATP with an individual walker model. To enforce heterogeneously bounded walkabout areas among walkers, SLAW develops an individual walker model to restrict the mobility of each walker to a predefined sub-section of the total area. It is done by selecting a subset of hotspots and restrict the movement of each walker to its own designated set of hotspots. To add randomness, it also allows walkers to move out of their predefined walkabout areas occasionally with some controlled probability. We verify that this walker model combined with LATP and fractal waypoints generate power-law ICTs (satisfying $F_3$).

More important, SLAW can realistically represent the regular patterns of human daily mobility. Since it is typical that people maintain a fixed routine of daily mobility such as going to an office, expressing these patterns are important. We verify later that the combination of fractal waypoints, LATP and the individual walk model effectively create this regularity as well as randomness and spontaneity in human mobility.

### 5.2 Fractal Waypoint Maps

In this section, from the analysis of the GPS traces of human walks, we report that the waypoints of humans can be modeled by fractal points. To show this, we define a *waypoint* to be the GPS location where a participant stays more than 30 seconds within a circle of 5 meter radius of
Figure 5.1: Hurst parameter estimation of waypoints registered in all KAIST traces.
Figure 5.2: Hurst parameter values of waypoints in each site map. All show Hurst values higher than 0.6 except NYC.

Figure 5.3: Measuring aggregated variance of waypoints aggregated from all walk traces. We divide the area by non-overlapping \(d\) by \(d\) squares, and count the number of waypoints registered in each square and then normalize the sampled count by the size of the unit square. We compute the normalized variance as we increase \(d\).
that location. For each site, we plot the waypoints registered by every walk trace of that site. We call these points *aggregated waypoints*.

To measure the self-similarity in the dispersion of waypoints, we divide the site map into a grid of unit squares (initially of 5 by 5 meters). We count all the waypoints within each square and then normalize the count by the area of the square. We measure the variance in these normalized count samples and call it *aggregated variance*. Fig. 5.3 illustrates the method. If there exists a long-range dependency in the samples, the aggregated variance should not decay faster than -1 in a log-log scale as we increase the size of the square. To see this, we plot aggregated variance in a log-log scale as we increase the square size and measure its absolute slope $\beta$. The *Hurst parameter* of the samples is $1 - \beta/2$. The sample data are said to be *bursty* or *long-range dependent* (and therefore, self-similar) if the Hurst parameter is in between 0.5 and 1. Aggregated variance can also be computed over one dimension by mapping waypoints to $X$ or $Y$ axis of the map. In this case, we use a line instead of a square.

Fig. 5.1 shows the Hurst parameter measured from the aggregated waypoints of KAIST. These values show a very strong long-range dependency with a Hurst value larger than 0.8. Fig. 5.2 plots the Hurst values measured from all the sites with their 95% confidence intervals. All the sites except NYC show a high degree of burstiness while the NYC traces show only slight burstiness. This outlier is, we conjecture, due to the very small number of participants relative to the size of the area and the number of registered waypoints are relatively small. Except NYC, the burstiness of the waypoints is evident independent of their site locations although the degree of burstiness may vary from one site to another.

### 5.3 Individual Movement Patterns

#### 5.3.1 LATP

We showed that fractal points induce a power-law gap distribution [54]. To find out how the power-law gaps are related to the actual human flights, we compare the flight distributions
obtained from real traces and the gap distributions induced by the waypoints from the same traces. We find a strong similarity between these two distributions, especially in terms of their slopes and shapes. Fig. 5.4 show that the Delaunay triangles on top of a daily trace of one participant, its line length CCDF forming the triangles and the CCDF of flights extracted from the same trace. We also perform the Delaunay triangulation on each individual trace and aggregate all the resulting triangle lines from all the traces. Fig. 5.5 plots the resulting CCDFs. Their similarity with the flight distributions from the corresponding traces is strikingly impressive.

The similarity in the power-law slopes of gaps and flight distributions suggests interesting aspects about the order which a walker visits waypoints for a given set of waypoints. Obviously people are not conscious about gaps when they travel. However, as we can see from Fig. 5.4, Delaunay triangles are formed among neighboring waypoints as Delaunay triangulation produces a planar graph where edges intersect only at their endpoints. From this, we conjecture that people might be minimizing the traveling distance. This makes sense intuitively when we
Figure 5.5: Delaunay triangulation is performed on all individual daily traces. The line segment lengths in Delaunay triangles are aggregated and their CCDFs are plotted for different walkabout sites. The CCDF of flights obtained from the corresponding traces are also plotted for comparison.

are given a priori multiple destinations at different distances, we often strive to minimize the total distance of travel by first visiting nearby locations before visiting farther locations. In fact, this “greedy” way of trip planning is similar to a heuristic to the traveling salesman problem whose objective is to minimize the total distance of travel and aligns well with the least action principle of Maupertuis [27].

To measure how much real human traces follow the least action principle, we measure the degree that people chooses their next destination based on the distance. This can be estimated as follows. We first measure the flight-to-nearest-waypoint ratio. For a given flight from $x$ to $y$, suppose $k$ is the nearest unvisited waypoint from $x$. The flight-to-nearest-waypoint ratio is the ratio of $||x - y||$ over $||x - k||$. We then define the least-action criterion: for a given flight, it tests if its flight-to-nearest-waypoint ratio is less than some threshold. Fig. 5.6 plots the percentage of flights meeting the least-action criterion in real traces for all participants when the threshold ratio $(r)$ is less than 2. On average, 53% of flights meet the criterion. However, if we consider
Algorithm 1  

Least action trip planning (LATP) algorithm with a distance weight function $d^{-\alpha}$

- $V$: set of all vertices (waypoints)
- $V'$: set of all visited vertices
- $s \in V$: starting vertex
- $c \in V$: current vertex

$c \leftarrow s$

$V' \leftarrow \{c\}$

**while** $V' \neq V$ **do**

- Calculate distances to all unvisited vertices, $d(c, v) = \| c - v \|_2$ for all $v \in V - V'$
- Calculate probability to move to all unvisited vertices, $P(c, v) = \left\{ \frac{1}{d(c, v)} \right\}^\alpha$ for all $v \in V - V'$
- Choose a next vertex $v' \in V - V'$ according to the probabilities $P(c, v)$

$c \leftarrow v'$

$V' \leftarrow V' \cup \{c\}$

**end while**

that humans are less sensitive to the distance when next destinations are all nearby. So if we exclude those flights whose length is less than a short distance (say 30 meters), we get more than 87% of flights meeting the criterion on average. This indicates that most people in our traces use distance as an important metric for deciding the next waypoint, substantiating our conjecture.

We construct a new trip planning algorithm called *Least Action Trip Planning (LATP)* that given a set of waypoints to visit, decides the order in which a person visits. Algorithm 1 gives a pseudo-code of LATP. The algorithm selects a next waypoint to visit based on a weight function of $1/d^a$ where $d$ is the distance from the current waypoint to an unvisited waypoint and $a$ is a constant deciding the distance weight. If $a$ is infinite, then the algorithm always chooses the nearest unvisited waypoint and if it is zero, then it randomly chooses the next waypoint. Fig. 5.7 shows the flight distributions obtained from LATP for different $a$ values performed on top of waypoints extracted from the KAIST traces and the percentage difference between the sum of flights generated from LATP and that of real flights from traces. In all cases, their difference is within a margin of around 10% when $a$ is between 1 and 3. The figure shows that LATP
generates traces very similar to the original traces.

5.3.2 Individual Walker Model

For a given input area $S$, our fractal waypoint generation generates a set $W$ of the waypoints. We model an individual walker model that selects a subset of $W$ and specifies the order in which those selected waypoints are visited. When selecting these waypoints, we need to be careful. Fractal waypoints have a tendency of creating bursty hotspots of various sizes dispersed over $S$. If waypoints are uniformly selected from $W$, then it is mostly like that all walkers will traverse through most hotspots and will not satisfy $F_2$. To satisfy $F_2$, we need to assign different walkabout areas to different walkers and restrict each walker to move only around their designated areas. We develop a heuristic individual walker model to model these behaviors.

The algorithm works as follows. We first build hotspots of waypoints by transitively connecting waypoints within a radius of 100 meters (typical WiFi outdoor transmission range). Let $C = \{c_i, i = 1, n\}$ be the hotspot set and $|c_i|$ be the number of waypoints and $T$ be the
Figure 5.7: Flight distributions obtained from LATP for different $\alpha$ values performed on top of waypoints extracted from the KAIST traces and the percentage difference between the sum of flights generated from LATP and that of flights from real traces.

the total number of waypoints in $S$. We assign a weight $|c_i|/T$ to each hotspot $i$. Then each walker $k$ chooses 3 to 5 hotspots (the exact number is controlled by the input) randomly from $C$ with probability proportional to these weights – the hotspot with a higher weight gets the higher chance to be picked. Let $C_k$ be the set of the selected hotspots. For each hotspot, it uniformly chooses 5 to 10% of waypoints in each hotspot randomly (the exact proportion is an input value); two different walkers are allowed to have the same waypoints. Let $W_k$ be the set of waypoints walker $k$ has selected from $S$. It also picks a starting waypoint from $W_k$ from which it always starts its daily trip.

To add some randomness in his travel, each day, a walker $k$ picks some additional waypoints $W'_k$ as follows; it first chooses one new hotspot $c'$ randomly (ignoring weights) not in $C_k$ and $W'_k$ is the waypoints randomly picked from $c'$ (about 5 to 10% of waypoints in $c'$). At the beginning of each day, walker $k$ starts from its starting point and throughout the day, makes a one-round trip visiting all waypoints in $W_k \cup W'_k$ using LATP. ($W_k$ does not change each day, but $W'_k$ does). It uses a truncated power-law pause-time distribution to decide the amount time to stay.
at each waypoint. In the end of the day, it comes back to its starting point. The average pause time is adjusted so that the whole trip will end within a period of 12 hours.

Since each walker $k$ always makes daily trips over a fixed set of $W_k$, its mobility area is bounded and also since each walker picks its set randomly, they tend to have different mobility area. In addition, we allow walkers to deviate from these waypoints by making them to pick some new waypoints additionally from the other hotspots not in $C_k$. This allows walkers without any overlapping hotspots occasionally meet, thus having some long ICTs. Those with overlapping hotspots may have regular periodic contacts, depending on the transmission ranges or the time they arrive to the hotspots.

We apply the hotspot weights when selecting $C_k$ to build some sense of community among all walkers. Because of fractal waypoints, some hotspots are very large – so many walkers are likely to visit them. These hotspots are emulating the common popular gathering places for all participants such as student union, dormitory, shopping, street malls or classrooms. Next section, we verify that SLAW with this individual walker model produces power-law ICTs ($F3$).

### 5.4 Performance Evaluations

#### 5.4.1 Flights

For validation, we run simulations using various mobility models. We fix the simulation areas to be approximately the same as the measurement sites in [75]. The transmission range of each node is varied from 25 meters to 150 meters. If not explicitly said, it is set to 50 meters. 50 nodes are simulated for 200 hours and the first 50 hours of simulation results are discarded to avoid transient effects. The speed of every user is set to 1 m/s for simplicity. We use a truncated Pareto distribution as pause time distribution of which the minimum and maximum values are 30 seconds and 700 minutes, respectively.

For comparison with SLAW, we choose four other models. Dartmouth, CMM and ORBIT models from social models and TLW from random models as shown in Chapter 2. Those models
have their own advantages. Dartmouth is based on extracted information from real data sets and it can be assumed to reproduce human walk patterns better than other models. CMM implements more realistic gathering property using the preferential attachment theory. ORBIT model is the only one where the movements of a user is bounded. In later sections, we will explain how SLAW recreates the patterns of human walks compared with the results from these other models.

For simulation of various models, we use the following setup. In the Dartmouth model, the hotspots are formed by the same method as in SLAW. It uses the same waypoints extracted from real traces [75] to build hot spots and also uses the transition probability obtained from the same traces. Note that since this model requires using these information, the simulation involving the Dartmouth model uses the same waypoint map that we obtained from the real traces in [75].

In the CMM model, the level of preferential attachment depends on the parameters such as the number of nodes and clustering exponent. We set the clustering exponent of the biggest hotspot to 0.5.

In ORBIT, we vary the size of one side of hubs from 200 to 500 meters corresponding to the size of each simulation site while fixing the number of hubs. Each user selects the same number of hubs for daily travels as the number of hot spots chosen by individual walkers in SLAW.

Fig. 5.8 shows the sample traces of various mobility models. It is clearly visible that SLAW generates traces similar to real GPS traces. In the Dartmouth model [49], walkers visit every hotspot with non-zero transition probability. In ORBIT [33], each user travels a fixed set of hotspots (i.e., hubs) daily in a random order with a uniform probability. TLW [75] is a random model so it does not have common hotspots for users. CMM [56] uses one popular hotspot using preferential attachment but it also makes users visit every place in a given area. The traces of TLW and CMM are not shown to save space.

We now verify how well SLAW models the statistical features of human mobility. Since SLAW explicitly models $F_2$ and $F_4$, we focus only on truncated power-law flights and ICTs
Table 5.1: The result of the Akaike test for the maximum likelihood estimation of truncated Pareto distributions (denoted Par) and exponential distribution (denoted Exp) over flights (denoted FL) and ICTs extracted from synthetically generated traces from various models whose parameters are set based on real traces obtained from four different locations (KAIST, NCSU, NYC and Disney World).

<table>
<thead>
<tr>
<th>Model</th>
<th>ICT</th>
<th>FL</th>
<th>ICT</th>
<th>FL</th>
<th>ICT</th>
<th>FL</th>
<th>ICT</th>
<th>FL</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLAW_{real}</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
</tr>
<tr>
<td>SLAW_{syn}</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
</tr>
<tr>
<td>CMM</td>
<td>Exp</td>
<td>Exp</td>
<td>Exp</td>
<td>Exp</td>
<td>Exp</td>
<td>Exp</td>
<td>Exp</td>
<td>Exp</td>
</tr>
<tr>
<td>ORBIT</td>
<td>Par</td>
<td>Exp</td>
<td>Par</td>
<td>Exp</td>
<td>Par</td>
<td>Exp</td>
<td>Par</td>
<td>Exp</td>
</tr>
<tr>
<td>TLW</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
<td>Par</td>
</tr>
</tbody>
</table>

(F1 and F3) - ICT results will be shown in Section 5.4.2. Figure 5.9 shows the flight distributions from various models and also from the measured GPS traces. SLAW is performed both on the waypoint maps generated synthetically by our fractal point generation technique and those extracted from real traces. SLAW produces very closely matching flight distributions (both synthetic and read maps) to that from the GPS traces. The Akaike test [21] tells whether the generated flight distributions fit power-law distributions (e.g., Pareto) or exponential distributions. Table 5.1 shows the result of the Akaike test [21] between Pareto and exponential distributions. In all cases, the flight distributions generated by SLAW are closer to truncated Pareto than exponential. All the other models except TLW do not produce power-law flights.

SLAW can be an important tool for emulating human walk behaviors. The applications of our work does not stop at mobile network; it can be applied to accurate urban planning, traffic forecasting and biological and mobile virus spread analysis where emulating human mobility is essential. SLAW is also as simple to work with as random mobility models such as RWP and BM because it needs only a small number of input parameters such as the size of the walkabout area, the number of walkers, and the Hurst value used for generating fractal waypoints. Note that many sophisticated models (e.g., [16, 33, 49, 56]) require detailed information such
as mobility transition probability tables and location and sizes of hot spots which are only available from real traces. SLAW does not require any real walk traces for generating synthetic traces.
Figure 5.8: Sample walk traces of various models.
Figure 5.9: Flight length distributions of synthetic traces from various models.
5.4.2 Inter-Contact Times and Link Durations

Fig. 5.10 shows the distributions of ICTs among 50 nodes. Since we do not have any ICT traces corresponding to the GPS traces in [75], we cannot verify the realism of these ICT distributions. However, we can verify whether the ICT distributions follow a truncated power-law pattern. Table 5.1 shows the result of the Akaike test on the ICT distributions. It shows that the ICTs of SLAW, TLW and ORBIT fit better to power-law distributions while the ICTs of the other models fit better to exponential distributions. For the NYC traces, Dartmouth also shows Pareto fitting. SLAW tends to show many occurrences of short ICTs because of hot spots and regularity of visits. The ICTs of ORBIT shows significantly higher occurrences of very long ICTs. This is because in ORBIT, those with non-overlapping orbits do not meet at all while the others may also meet rarely because of randomness in picking the waypoints within their own orbits. The ICTs of CMM and Dartmouth have exponential distributions in most cases and tend to have much more occurrences of long ICTs than SLAW (note that the scales are log). This is because of their randomness in choosing the next clusters (or hot spots) to visit.

Link duration (LD) distribution is another property that human walks have. In [24, 25], it is shown that the LD distributions of four data sets which include UCSD and Dartmouth data sets show power law decay. Fig. 5.11 shows LD distributions generated by various models. All the models make power law decaying LD distributions. We see this is due to the fact that we use power law pause time distributions for all the models since it is known that short LDs are created by two nodes passing by each other. Long LDs are created by two nodes that pause within each others’ transmission range [52]. In this case, LD between two nodes that have power law pause time distributions should have longer tail than those with exponential or Gaussian pause time distributions. Note that the LD distribution by SLAW has the longest tail due to the bursty waypoints even though other models have the same pause time distributions. Bursty waypoints increases the probability of long LDs since more nodes tend to stay within a restricted area. It is a very important feature since longer LD leads to higher throughput in wireless networks.
We have seen flight length, ICT and LD distributions that SLAW generates are similar to those from empirical results. We consider these results as evidences of our model’s similarity to real human movement patterns. ICT and LD distributions are also very important in viewpoint of estimating routing performances since they greatly impact the performance of wireless networking applications and can be used as general performance metrics independent of routing protocols [8, 25, 38, 75].
Figure 5.11: The link duration distribution
5.4.3 Routing Performance

To measure how these mobility patterns influence the performance of mobile network protocols, we study the performance of DTN routing under SLAW and various other mobility models. Our study indicates that SLAW realistically brings out the unique performance features of various DTN routing protocols. Especially, compared to random mobility models, it provides a clear performance differentiation between memoryless and memory-full protocols where memory-full protocols utilize past contact history information among nodes to predict the future contact probability. Examples of these protocols are abundant (e.g., [6,10,81]).

We test the following five DTN routing protocols. We generate one message bundle between each of randomly selected 100 source and destination pairs. All transmissions are assumed to be reliable and instantaneous when the communicating nodes are within a transmission range. To maximize the effect that mobility models have on routing performance, we assume that all nodes keep the entire history of past contacts with other nodes. All results are averaged over 40 runs.

Random forwarding [81], Direct transmission [81], PRoPHET [6], LET [36] and ECT. In LET (last encounter time), a forwarding node of a message picks, as a next relay, the node with the most recent history of meeting the destination of the message among its current neighboring nodes. Each node updates its neighbor set at every minute. ECT (expected contact time) is a new metric we developed. It computes the expected time that a node meets the destination by subtracting the last encounter time from the expected inter-contact time which is computed by averaging the past inter-contact times with the destination.

We can categorize these protocols as memoryless and memory-full protocols. Random forwarding and direct transmission are memoryless as they do not use any past meeting history information. The other protocols are memory-full as they all use past contact information to predict the future probability of meeting the destination.

We find that the routing performance of TLW, RWP and CMM has almost the same patterns although their average delays. For brevity, we show the result of CMM only in Figure 5.12.
In these models, both memoryless and memory-full protocols perform almost the same. This pattern happens because the mobility of nodes in these models is highly random so the prediction of memory-full protocols is not effective. The lack of performance differentiation among various types of protocols limits the usefulness of these models for mobile network simulation.

Dartmouth (Figure 5.13 (a)) shows results similar to CMM as the performance of various protocols except LET is not very distinguishable. The pattern can be explained as follows. The probability that Dartmouth nodes jump to any other hot spots is determined by the transition probability; so any node can jump to any other hot spots as long as the transition probability is non-zero. This causes some randomness in visiting hot spots and likewise, ICT patterns similar to those in random mobility models such as CMM which are manifested in its exponentially distributed ICTs. However, because of higher transition probability to visit bigger hot spots, Dartmouth exhibits much shorter delays than CMM and TLW. The low performance of LET in Dartmouth is because Dartmouth represents a hot spot as a single waypoint (i.e., a cluster reduces to one point) which has a side-effect of increasing pause times at a hot spot (since all the pause times for a hot spot are aggregated for all the points inside a hot spot). Thus, when a forwarding node meets a new node with a shorter LET than its LET, it is likely that the new node has just arrived to that hot spot. Thus the new node is more likely to stay in that hot spot much longer than the forwarding node, thus causing a longer delay to meet the destination next time. These features are an artifact of rather unusual and unrealistic setups of hot spots.

ORBIT (Figure 5.13 (b)) show a clear performance differentiation among different protocols. In ORBIT, nodes in non-overlapping “orbits” (i.e., they do not share a common hot spot) do not meet at all. The only way to deliver messages between two non-overlapping orbits is through the other nodes with overlapping orbits with the destination. This means that direct and random forwarding can perform really poorly. On the other hand, in ORBIT, memory-full protocols can perform much better. The performance difference between ECT and LET are relatively small compared to the difference between random forwarding and LET. This is because in ORBIT, the nodes with long LETs are likely to meet the destination fairly rarely.
due to only a small overlap in their orbits. Since each node in ORBIT moves like RWP among hot spots in its orbit, a small overlap results in a very long ICT and thus likely to have long LETs. Thus, the nodes with shorter LETs are likely to have more overlapping orbits with the orbit of the destination. Thus, choosing these nodes as relays leads to short routing delays. A similar argument is applicable for ECT because those with long LETs are likely to have long ECTs in ORBIT.

The mobility patterns of SLAW are quite different from those of ORBIT. The most salient feature is that SLAW has much shorter routing delays. SLAW has much more regularity in their trip patterns than any other models because it uses LATP for the selected set of waypoints and each node visits almost the same set of waypoints every day. In this type of scenarios, relaying to a node with short LETs can be detrimental because those nodes who just met the destination are likely to have long ECTs. That means that choosing, as relays, those nodes with short ECTs would always result in shorter routing delays because expected ICTs are very accurate because of the regularity in trip. The performance of PRoPHET is not as good as ECT because PRoPHET updates its probability only after meeting a destination. Thus, its behavior is slightly similar to that of LET. In SLAW, the power-law ICT distributions play a significant
role. Due to the inspection paradox property of the renewal process and the power-law ICTs, when a node meets another node, it is more likely to meet a node with a long ICT. If it has a short ICTs, then ECT would perform as well as LET. But with long (predictable) ICTs, those nodes with short LETs are not likely to meet the destination for a long time. Thus ECT can perform much better than LET. The fact that ECT performs best indicates that the regularity of trip patterns is well represented in SLAW without loss of inherent statistical features such as power-law flight and ICT distributions.

Figure 5.13: Average routing delays of various protocols under Dartmouth, ORBIT and SLAW.
5.5 Conclusion

In this chapter, we have presented a new mobility model, called SLAW, that captures the spatial patterns of human mobility found in real human mobility traces, especially the self-similar human gathering property. We report many pieces of empirical evidences that the movement of people can be expressed very well using fractal waypoints. We present confirming data for the use of the least action principle in human trip planning. Based on this, we develop a simple heuristic algorithm called LATP that generates heavy-tail flights on top of fractal waypoints. Combining with heterogeneously bounded walkabout areas, we can successfully reproduce many statistical features important to the study of mobile network performance. Our routing performance study indicates that SLAW effectively expresses mobility patterns arising from people with some common interests or within a single community such as students on the same university campus or people in theme parks where people tend to share common gathering places. We find that LATP over heterogeneously bounded areas realistically expresses some periodicity in the daily mobility of humans.
Chapter 6

Spatio-Temporal Features of Human Movements: Individual and Aggregation Patterns
6.1 Overview

It is impossible to mimic the movement of people to every little detail. Instead, we focus on the statistical features that can significantly influence the performance of mobile networks. ICT is the time duration until a node meets another node after meeting that node previously where meeting (or contact) is defined by being in a radio range. It is commonly accepted that the distributions of ICTs [17, 22, 25, 38, 47, 65, 75] and flights [54, 75] are the most influential characteristics of mobility traces determining the performance of mobile networks.

Several empirical studies have shown that both ICT and flight distributions have heavy-tail tendencies [25, 54, 75]. However, both features have never been observed together from the same mobility traces. This is because of lack of mobility traces that record both spatial locations and contact information. Some traces collected using iMotes [25] contain only contact information, and other traces [54, 75] contain only GPS locations of individual persons, but each trace is taken at a completely different time so no contact information can be extracted. There are also traces based on WiFi associations of people [37, 59] or cell-tower association of cell-phone users [34]. But these data have too low resolution in their data with a margin of errors being a few hundred or thousand meters since we have to estimate the location of a node using the associated AP or cell tower’s location. The lack of traces containing both contact and location information makes it very difficult to verify a mobility model because while a model can be shown to match one feature from mobility traces it cannot be proven to match both features.

To get mobility traces that contain both locations and contact information, we conducted two measurement experiments (the 2nd experiment in section 3.1), each involving simultaneous GPS tracking of about 100 students from a different university campus. Since all the participants in each experiment are tracked at the same time, we can measure both flights and ICTs among these traces. Our data analysis confirms that heavy-tail flight and ICT distributions are observed from both experiments. This is the first traces confirming the existence of heavy-tail tendency in flight and ICT distributions in the same traces.

We then use the measured flight and ICT data from our experiments to verify whether the
existing models are capable of reproducing those distributions. We have tested eight mobility models and by varying various input parameters of the models, we fit their flight and ICT distributions to the measured distributions using MLE (maximum likelihood estimation), a common technique used for distribution fitting. Unfortunately, none of these models can match the flight and ICT distributions observed from the real traces. It is not trivial to match both distributions. For instance, we can fix the flight distribution by selecting the next waypoint so that the resulting flight from the current waypoint to the next one is from a given flight distribution. TLW [75] uses this method. However, this method does not control the contact times between two nodes, which determine the ICT distribution. Conversely, a model may fit the ICT distribution by controlling the contact times of nodes, but it is unclear how the model can control waypoint selections to produce a specific flight distribution.

We view this inability of reproducing realistic flight and ICT distributions is essentially due to the poor representation of spatio-temporal correlations (STCs) present in real human mobility as shown in 3.2. While flights are a spatial feature, ICTs are a spatio-temporal feature governed by where and when people meet. The STC implies that the DLFs are highly correlated to the locations of the hotspots and this relationship is consistent irrespective of the mobility environments.

Based on these observations, we propose a spatio-temporal mobility model, STEP. STEP allows flight and ICT distributions to be flexibly modeled by adjusting a few input parameters. This property of STEP can be used to produce synthetic traces whose flight and ICT distributions match closely those from the real traces. The strength of STEP lies in its generality where while maintaining realistic flight and ICT distributions, it can vary the other properties of the output traces such as the size of mobility area, the number of nodes, the locations and number of hotspots, to generate a diverse set of mobility traces.
6.2 Spatial Features

As shown in Table 3.5, TLW and SLAW provide the best matching results to the measured flight and ICT distributions. Both models have shown that they can reproduce heavy-tail flights and ICTs in their original work [38, 54, 75]. But as we can see in Figs. 3.6 and 3.7, their results still show some discrepancies. In this section, we investigate the causes of those discrepancies from the spatial viewpoint.

While TLW is a random model that chooses its flight from a given heavy-tail distribution, SLAW reveals the underlying mechanism of heavy-tail flights. It shows that distance minimizing trip planning over a fractal waypoint map induces heavy-tail flights. But we conjecture that SLAW in its original form misses some important features since SLAW cannot match the measured flight and ICT distributions. In the rest of this section, we explain the spatial features that the SLAW model missed.

6.2.1 Fractal Dimension

Waypoint sets of people have been known to be fractal [54]. To generate a fractal waypoint map, SLAW uses the Hurst parameter to control the characteristics of waypoint placement. But we have found that the Hurst parameter is not commonly used for multi-dimensional fractal sets and is not easily defined or characterized for spatial point processes. We use the fractal dimension (D) [58] as a parameter to control its fractal characteristics.

Mandelbrot [58] has shown that for a Levy flight in a two dimensional space of which the flight follows the distribution \( Pr(X > x) = x^{-D} \), the fractal dimension of the resulting waypoints is given by the fractal dimension \( D \). This result strongly suggests that there is a close relationship between the exponent of a power-law flight distribution and the fractal dimension of the corresponding waypoint map.

Flights are formed by connecting waypoints. The order in which a mobile node visits these locations determines his flight patterns. Then what aspects of fractal waypoints induce a heavy-tail flight distribution? To find this relation, we study the characteristics of the gaps formed
among these waypoints. Gaps are defined to be inter-spacing among neighboring waypoints. But it is not easy to define a gap in a two dimensional space. Delaunay triangulation is often used to measure two dimensional gaps [54]. Delaunay triangulation for a set $P$ of points in the plane is a triangulation $DT(P)$ such that no point in $P$ is inside the circumcircle of any triangle in $DT(P)$.

Fig. 6.1 shows the relationship between the fractal dimensions of waypoint maps and the corresponding gap distributions. We use Soneira-Peebles (SP) model [80] to generate a fractal waypoint map. The SP model is one of the well-known fractal point process generation methods. To measure gaps, we apply Delaunay triangulation on the waypoints and aggregate all the resulting triangle lines. We can see that as we increase the fractal dimension of waypoint maps from 0.5 to 2, the power-law exponent of gaps also increases from 0.5 to 2. It is shown that the gap distribution of the waypoints is closely related to the flight distribution of human movements [54]. Thus we can say that the flight distribution can be controlled by the fractal dimension of the waypoint map.
6.2.2 Super-Hotspot

In this section, we investigate the characteristics of hotspot size which is defined to be the number of waypoints in each hotspot. We use the same grid definition for a hotspot used in Section 3.3.2. Fig. 6.2 shows the distribution of hotspot size has a power law tendency. It implies there is a super-hotspot that has the largest number of waypoints. The super-hotspot also means that it has the largest number of visitors since Fig. 6.3 shows there is a linear correlation between the average number of visitors and the number of waypoints. The existence of single super-hotspot is important in determining the heavy-tail flight and ICT patterns since a large number of waypoints are aggregated in a hotspot, many short flights occur in that hotspot (Note that power-law patterns come from many short samples and a relatively small number of large samples.) If there are many super-hotspots, the number of flights among them (which are mostly long) increases. Then the proportion of long flights increases and the flight distribution may show an exponential tail.

Figure 6.2: Hotspot size distributions.
6.2.3 Sub-Hotspot

In this section, we investigate the geographical area around which an individual moves within a hotspot. We divide a hotspot into a smaller grid (e.g. 20m by 20m) and count the number of waypoints of each user in each cell. We call the cell that has the largest number of waypoints, a center cell. In Fig. 6.4(a), we plot the number of waypoints normalized by the number of waypoints of the center cell according to the distance from the center cell. From the figure, we can see that waypoints of each individual are highly aggregated in a region of radius less than...
about 50 meters from the center cell.

We also found that those waypoints from individuals are collocated in a small region and there are multiple such small regions in each hotspot. We call those regions *sub-hotspots*. Fig. 6.4(b) shows the locations of center cells for each user in a hotspot observed from the Campus I data set. The figure shows that center cells from users are highly co-located at small regions, i.e., sub-hotspots. Possible sub-hotspots are indicated by circles in the figure. The concept of sub-hotspots suggests that the movements of humans are restricted within a small area even though the area of a hotspot is large. We might be able to describe this tendency by reducing
the area of hotspots. But if neighboring hotspots are selected for the same user, the total area from the neighboring hotspots becomes large, which means the probability of longer flights increases.

The contributions of sub-hotspots in describing human movements are twofold: (1) the head of the flight distribution shows stronger power-law tendency since the available area for a node can visit becomes smaller and it makes the proportion of small flights increase, and (2) the tail of the ICT distribution becomes longer since even two nodes are in the same hotspot they might not be within the radio range if they are in different sub-hotspots.
6.3 Temporal Features

In the previous section, we investigated the spatial features observed from our data sets. Until now, most mobility models have focused on the spatial features. But would it be enough to describe human movements? Could we reproduce the ICT distribution that is also governed by temporal characteristics of human movements? We present one of the evidences that show the existing models do not describe meeting patterns among nodes. Fig. 6.5 (a) shows our measured data of the population change, i.e., DLF, over one day time visiting four hotspots in the Campus I data set. The DLF shown in the figure is clearly changing over time. The result suggests that there are temporal variations in human gathering patterns, more specifically, at certain times hotspots are not “hot” any more. Fig. 6.5 (b) shows the DLFs of sample hotspots in the existing models. They are essentially flat, showing no temporal variations at all. This is because most of the models do not explicitly model the times at which a node visits a certain location, but rather choose the times randomly.

Fig. 6.6 shows another evidence of time-varying property of human movements. The figures represent the populations of the hotspots observed from the Campus I data set at different time slots. The color of each hotspot represents the level of population. As the population increases, the square becomes darker. We also have shown the incoming probabilities for hotspots 67 and 87. The incoming probability is defined to be the proportion of transitions to each hotspot over all transitions. We can clearly observe that both hotspot populations and transition probabilities have time-varying characteristics. For example, at 3 pm we can see that the hotspots near the hotspot 87 become less popular but the hotspots near the hotspot 67 become more popular. The incoming transition probability is also increased by almost four times for the hotspot 67 at 3 pm than the one at 1 pm.

In this section, we identify temporal characteristics of human movement patterns. We complement our data sets with two existing WiFi data sets collected from Dartmouth [37] and UCSD [59] campuses. Note that we have used four data sets collected from different sites since we want to obtain fundamental properties that do not rely on a specific condition.
To describe the temporal variations in hotspot population, we need a method to represent when mobile nodes arrive at and depart from a hotspot and how the population of the hotspot varies over time. To show this variation, we defined the DLF which is a time series function of a hotspot population. The goal of this section is to characterize the DLFs of hotspots. It has been well known that any time series function can be represented by sums of trigonometric functions with different periodicity. We can extract periodicity information from the measured DLFs using Fourier transform. Eqs. 6.1 and 6.2 represent the discrete Fourier transform (DFT) and inverse DFT functions, respectively. The sequence of $N$ complex numbers $x_0, ..., x_{N-1}$ is transformed into an $N$ complex number sequence $X_0, ..., X_{N-1}$ by DFT. $i$ represents the imaginary unit.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn}, k = 0, ..., N - 1.$$  \hspace{1cm} (6.1)$$

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{\frac{2\pi i}{N} kn}, n = 0, ..., N - 1.$$ \hspace{1cm} (6.2)$$

We also use the periodicity to characterize the DLFs for hotspots. Kim et al. [48] analyzed a WiFi data set and show a result on Access Point (AP) popularity in terms of periodicity. It is shown that APs have two dominant periods, 24 hours and 1 week. But AP popularity should show shorter periodicities. Because, for example, in classrooms students should gather around and depart more frequently than one day period. The results of [48] that every place has the same one day and one week periodicity do not seem to give satisfactory information. Fig. 6.7 shows how we can obtain more precise periodicity information.

We first define a primary period as the period that have the biggest amplitude ($A_b$) or bigger amplitudes than a certain threshold. We set the threshold to $T_p \cdot A_b$ where $T_p$ is a constant less than 1. We have tested various values of $T_p$ between 0.4 and 0.8 but they show similar tendencies. The results below are from $T_p = 0.5$.

Fig. 6.8 shows the Fourier transform results observed from four hotspots in the Campus I data set. We looked at the data collected during daytime since we asked volunteers to record only the daytime movements. Figs. 6.8 (a) and (b) represent hotspots that have 1 and 2 primary
periods. Figs. 6.8 (c) and (d) represent hotspots that have more than 3 primary periods. The figure shows that hotspots have different primary periods, which means the primary periods can be considered as the temporal characteristics of hotspots. It is an important fact since it means that if we can assign synthetic primary periods to hotspots then we can recover a corresponding time series function of DLF using the inverse DFT in Eq. 6.2.
Figure 6.5: Lifecycles from the Campus I data set and existing mobility models.
Figure 6.6: Hotspot populations observed from the Campus I data set. Each square represents a hotspot. As the population increases, the square becomes darker. The incoming probabilities for hotspots 67 and 87 are shown.
Figure 6.7: Hidden periodicity is discovered by eliminating the data collected during night time and weekends.

Figure 6.8: Various periodicity from hotspots in Campus I.
6.4 Spatio-Temporal Correlations

We have seen how we could characterize the temporal characteristics of hotspots using their primary periods. In this section, we consider relationships between the primary periods of hotspots and the spatial locations of them. To see whether there is any relationship, we visually plotted the number of primary periods in each hotspot according to their locations in Fig. 6.9. As squares become darker, the number of primary periods decreases. From the figure, we can say that a few darkest squares are placed around the center of the area. The colors of the squares are darker at the places nearer to the darkest hotspots. Thus we can conjecture that there might be a relationship between the number of primary periods of hotspots and the distance from the super-hotspot.

Fig. 6.10 (a) and (b) show the average number of visitors and the number of primary periods of each hotspot according to the distance from the super-hotspot. Fig. 6.10 (a) shows that the hotspot population is inversely proportional to the distance from the super-hotspot. The solid lines represent the linear regression results. This tendency implies that people usually stay at places near the super-hotspot and visit farther places less. Fig. 6.10 (b) shows that the number of primary periods is proportional to the distance from the super-hotspot. This means that people tend to visit hotspots near to the super-hotspot with higher regularity since as the number of primary periods decreases, the DLFs show more regular patterns.

Figs. 6.10 (c) and (d) show the results from the Campus II data set. Figs. 6.10 (e)(f) and (g)(h) show the results from the Dartmouth and UCSD data sets, respectively. All the results confirm the same tendencies.

Note that this property does not depend on a specific displacement of hotspots. For example, in the Campus I environment, the super-hotspot falls on the dormitory. Near the super-hotspot, there are lecture halls and restaurants for students. But in the Campus II environment, the main library and nearby lecture halls form the super-hotspot.
Figure 6.9: The number of primary periods for each hotspot is represented by black squares. As the number of primary periods increases, the square becomes brighter.
Figure 6.10: The STCs observed from the four data sets. Solid lines represent linear regression results. X-axis represents the distance from the super-hotspot to each hotspot. (a)(c)(e)(g) and (b)(d)(f)(h) show the number of visitors and number of primary periods, respectively.
6.5 STEP Mobility Model

In this section, we describe our model in detail. We first describe how to generate a synthetic waypoint map. We show how to form hotspots and assign DLFs to each hotspot. Then we propose the STC-aware LATP algorithm that describes how the node moves along the selected waypoints.

6.5.1 Hotspot Model

We use Soneira-Peebles (SP) model [80] to generate a fractal waypoint map. We modify the SP model to implement the super-hotspot by controlling hotspot size to follow a power-law distribution as shown in Fig. 6.2. After generating a waypoint map, we build hotspots using the grid method.

To generate the DLFs of hotspots, we need to assign primary periods for each hotspot. If we could find out a relationship between hotspots and their primary periods from a waypoint map, we can generate synthetic DLFs from a synthetic waypoint map. We first look at the relationship between the number of waypoints and the number of primary periods for each
Figure 6.12: Relationship between the number of waypoints and the number of primary periods for hotspots.

hotspot. As shown in Fig. 6.12, the number of primary periods of each hotspot is inversely proportional to the number of waypoints. We have applied several functions to fit the data and could see that the following rational function matches best.

\[
f(x) = \frac{p}{x + q}.
\]  

The value of \( p \) and \( q \) are 10,300 and 2,493 for the Campus I data set and 4,906 and 890 for the Campus II data set. Using this relationship, we assign one or multiple primary periods to
each hotspot. At least one of the primary periods should be 1 day to represent daily life-cycle. Fig 6.11 shows the distributions of primary periods observed in our GPS traces and the two WiFi data sets. It shows that they follow exponential distributions and their mean are half day. Thus other primary periods can be picked up from an exponential distribution.

Now we describe the above procedure more formally. The number of primary periods for a hotspot $k$ is set to be inversely proportional to
$$W_t W_h k + (W_t - W_h s),$$
where $W_t$, $W_h k$ and $W_h s$ represent the number of total waypoints, the number of waypoints in a hotspot $k$ and the super-hotspot, respectively. Then, we generate a DLF for each hotspot by using the inverse Fourier transform.

Now we have time series functions of DLFs for each hotspot. Then we need to set their amplitudes. As we already know that the number of visitors for each hotspots is proportional to the number of waypoints in that hotspot from Fig. 6.3, we use the relationship to set the amplitudes of the DLFs. For the super-hotspot, we set the average number of visitors, $U_h s$, is equal to the total number of users. For other hotspot $k$, the average number of visitors, $U_h k$, is set to $U_h s \cdot (W_h k / W_h s)$. The generated DLFs are normalized to have an average $U_h k$ for each hotspot $k$.

### 6.5.2 Individual Walker Model

For a given area $S$, we generate a set $W$ of waypoints as in Section 6.5.1. Then we model an individual walker model that selects a subset of $W$ and specifies the order in which those selected waypoints are visited. Firstly, we assign different walkabout areas to different mobile nodes and restrict each node to move only around its designated area. Each node $k$ chooses 5 to 7 hotspots randomly with probability proportional to the hotspot size. Let $C_k$ be the set of the selected hotspots.

When selecting hotspots, each node needs to choose a sub-hotspot if the selected hotspot consists of multiple small hotspots. To divide a hotspot into multiple sub-hotspots, we divide the area $S$ into finer grid (e.g. 20m by 20m) and count the number of waypoints in each cell. By plotting the number of waypoints in each cell according to their locations, we get a
3-dimensional surface plot of the number of waypoints. We define the local peaks from the surface as sub-hotspot centers. Then we divide the hotspots using a Voronoi diagram as in [43]. A Voronoi diagram is a decomposition of a space determined by distances to a specified point. Each sub-hotspot contains all the waypoints closer to that sub-hotspot center than to any other sub-hotspot centers.

Then we uniformly chooses 5 to 10% of waypoints in each hotspot (or sub-hotspot if exists) randomly. Let $W_k$ be the set of waypoints node $k$ has selected from $W$.

Now we need an algorithm that decides the order in which a node visits a set of selected waypoints. Each individual $k$ move along its selected waypoint set $W_k$ with the Algorithm 2. The algorithm is based on the LATP that SLAW [54] suggests. The LATP algorithm can be formally described by the following equation:

$$
P(w_c, w) = \frac{1}{d(w_c, w)^\alpha} \frac{1}{\sum w d(w_c, w)^\alpha} \text{ for all } w \in W - W',
$$

where $d(w_c, w)$ is the distance from $w_c$ to an unvisited waypoint $w$ and $\alpha$ is a constant. The equation describes the probability with which a node moves from the current waypoint, $w_c$, to a next waypoint, $w$.

Our goal for the individual node model is to add the impact of the DLF to Eq. 6.4. It can be done by two alternative ways: (1) addition and (2) multiplication. The first one can be represented by $A(w_c, w) = \gamma A_I(w_c, w) + A_d(w_c, w)$, where $A_I(w_c, w)$ and $A_d(w_c, w)$ represent attractiveness to $w$ from $w_c$ by the DLF and LATP, respectively. The $A(w_c, w)$ represents the resultant attractiveness. $\gamma$ is a constant to describe a relative weight on $A_I(w_c, w)$. But in this way the resulting attractiveness can be distorted. For example, when $A_I(w_c, w) = 0$, the node should not be able to move to $w$. But by the addition method, $A(w_c, w)$ still have a non-zero value if $A_d(w_c, w)$ is not zero. The following multiplication method does not incur
such a problem:

$$P(w_c, w) = \frac{l_w(t)^\gamma 1}{\sum_w l_w(t)^\gamma 1} d(w_c, w)^\alpha$$

for all $w \in W - W'$, \hspace{1cm} (6.5)

where $l_w(t)$ represents the DLF value of a waypoint at time $t$. All the waypoints in the same hotspot share the same DLF value. $l_w(t)$ defines the number of nodes that has to be in the hotspot.

**Algorithm 2** *STC-aware LATP algorithm*

\begin{itemize}
\item $\alpha$: distance weight
\item $\gamma$: lifecycle weight
\item $l_w(t)$: lifecycle function of the waypoint $w$
\item $W$: set of all waypoints
\item $W'$: set of all visited waypoints
\item $w_s \in W$: starting waypoint
\item $w_c \in W$: current waypoint
\item $w_c \leftarrow w_s$
\item $W' \leftarrow \{w_c\}$
\end{itemize}

**while** $W' \neq W$ **do**

Calculate distances to all unvisited waypoints,

$$d(w_c, w) = \|w_c - w\|_2$$

for all $w \in W - W'$

Calculate probability to move to all unvisited waypoints,

$$P(w_c, w) = \frac{l_w(t)^\gamma 1}{\sum_w l_w(t)^\gamma 1} d(w_c, w)^\alpha$$

for all $w \in W - W'$

Choose a next waypoint $w' \in W - W'$ according to $P(w_c, w)$

$w_c \leftarrow w'$

$W' \leftarrow W' \cup \{w_c\}$

**end while**

Algorithm 2 selects a next unvisited waypoint to visit based on a probability function $P(w_c, w)$ that uses weighted functions $d(w_c, w)^\alpha$ and $l_w(t)^\gamma$. In the preferential attachment theory [11], the attractiveness of a location relies on the current number of nodes who are in the place. But with this theory, we cannot describe ups and downs of the hotspot population.
since when the population begins to grow it will never fall down. Instead, we use predefined $l_w(t)$ as the degree of attraction so we can control the population of hotspots better. By adding $l_w(t)$ to LATP, we can describe the tendency of a node to follow the schedule. For example, during the mealtime, the DLF value of a restaurant becomes large then the people around the area will gather around.

To add some randomness in his daily travel, a node $k$ replaces one of the hotspots in $C_k$ at the start of each trip. This allows nodes without any overlapping hotspots occasionally meet, thus having very long ICTs.
6.6 Performance Evaluation

6.6.1 Experimental Validation

In this section, we evaluate our model. Firstly, we see whether our model can faithfully reproduce DLFs as measured in Campus I. We run simulations using the real waypoint map and real DLF functions. Other setups are same as described in Section 3.3.2. The result is the average of 10 simulation runs. Fig. 6.13 shows the DLFs of a sample hotspot measured from STEP and real data, respectively. From this figure, we can confirm that the DLF of our synthetic model (green dotted line) closely matches the assigned DLF (blue solid line) measured from Campus I.

We evaluate our model by reproducing the same flight and ICT distributions as the measured ones. All the setups are same as the Campus I environments. Firstly, we generate a synthetic waypoint map. Fig. 6.14 (a) shows the waypoint map generated by the STEP hotspot model using the fractal dimension of 1.4 which is measured from Campus I. In the figure, we can see the the placement of waypoints is similar to that in Fig. 3.1. In addition, the hotspot size

---

Figure 6.13: Assigned (solid line) and measured (dotted line) lifecycles for the proposed model.
distribution follows a power-law distribution as shown in Fig. 6.14 (b). Then we assign DLFs to each hotspot.

We run mobility simulations and check resulting flight and ICT distributions. Figs. 6.15 (a) and (b) show how flight and ICT distributions change as we vary $\alpha$ from 0 to 4. When $\alpha$ is larger than 4, the flight and ICT distributions do not show much difference. We can see that STEP faithfully reproduces the measured flight and ICT distributions.

In Table 6.1, the KL test [51] results for the flight and ICT distributions generated by STEP
Table 6.1: The results of Kullback-Leibler divergence tests: Compared with the results in Table 3.5, STEP is at least 9.6 times and 2.4 times better than other models in reproducing the measured flight and ICT distributions, respectively.

<table>
<thead>
<tr>
<th></th>
<th>Campus I</th>
<th>Campus II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flight</td>
<td>0.0012</td>
<td>0.0015</td>
</tr>
<tr>
<td>ICT</td>
<td>0.0074</td>
<td>0.0086</td>
</tr>
</tbody>
</table>

Table 6.2: Parameter sensitivity of the proposed model. The PL represents power-law. The Var[DLF] means variance of an DLF.

<table>
<thead>
<tr>
<th>Flight PL tendency</th>
<th>ICT</th>
<th>Var[DLF]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head</td>
<td>Head PL tendency</td>
<td>Tail length</td>
</tr>
<tr>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>Tail (between hotspots)</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>Head</td>
<td>↑</td>
<td>↑</td>
</tr>
<tr>
<td>Tail length</td>
<td>↑</td>
<td>↑</td>
</tr>
</tbody>
</table>

are shown. The results suggest that STEP generates much closer flight and ICT distributions to the measured flight and ICT ones than other models.

6.6.2 Parameter Sensitivity

Table 6.2 shows the parameter sensitivity of STEP. It shows the impact of α and γ on the flight and ICT distributions and the variance of DLFs.

Figs. 6.16 (a) and (b) show how flight and ICT distributions change as we vary the value of α. As Lee et al. [54] have shown, α determines the power-law tendency of the flight distribution. When α is close to zero, the flight distribution follows an exponential distribution. But as we increase α, it comes close to a power-law distribution. α has the same impact on the ICT distribution. As we increase α from zero to infinity, the ICT distribution changes from an exponential to a power-law distribution. α does not have any effect on lifecycles.

γ governs the flights between hotspots (i.e. long flights) since waypoints in the same hotspot have the same DLF value. So the impact of γ is mostly negligible for the head part of the flight and ICT distributions since they are determined by short flights occurred within hotspots. γ has a very interesting impact on the tail distribution of flights. As we increase γ, DLFs become
dominant in determining the next waypoint to visit in Eq. 6.5. Since $\gamma$ governs the occurrences of long flights, the tail of the flight distribution becomes to fit the real distribution as shown in Fig. 6.17 (a).

Fig. 6.17 (b) shows the impact of $\gamma$ on the ICT distribution. As we increase $\gamma$, we can see that the tail becomes longer. When $\alpha$ and $\gamma$ are small, nodes meet each other frequently in the super-hotspot, which makes the short tail ICT distribution. But as we increase $\gamma$, nodes do not meet so frequently since nodes spread to different hotspots that have large DLF values.

For DLFs, $\gamma$ is the only factor of impact. As we increase $\gamma$, we can see the variance of a DLF becomes larger, which means DLFs of hotspots become more dominant in determining the next waypoint.
Figure 6.15: Flight and ICT distributions generated by STEP under the Campus I environment. We use $\gamma=5$. 
Figure 6.16: Parameter sensitivity of STEP according to $\alpha$. We set $\gamma = 5$. 

(a) Flight distributions

(b) ICT distributions
Figure 6.17: Parameter sensitivity of STEP according to $\gamma$. We set $\alpha = 1$. 
6.7 Conclusion

In this chapter, we have presented the empirical evidence of the spatio-temporal correlation of human movement patterns and propose a mobility model called STEP that captures those correlations as well as the spatial features shown in chapters 4 and 5. We analyze multiple sets of human movement data to ensure that the results do not rely on any specific environment. By faithfully describing those features, we succeed in reproducing heavy-tail flight and inter-contact time distributions observed from empirical data sets. To generate synthetic traces, we first construct a fractal waypoint map. We divide the map into hotspots and generate life-cycles for each hotspot. Then we run the STC-aware LATP algorithm to determine each user’s movements.

STEP is the first model that extracts spatio-temporal correlations of human movements from empirical data sets and explicitly describes the movements in the dimensions of space and time.
Chapter 7

Conclusion and Future Work
In this dissertation, we have analyzed 10 empirical data sets collected from various environments. From the data analysis results, we have found that (1) human walking patterns contain statistically similar features to Levy walks including heavy-tail flight and pause time distributions, (2) people generate fractal waypoints and move in a greedy fashion to minimize the total traveling distance over the waypoints and (3) the fractal waypoints can be divided into multiple hotspots, the number of visitors to that hotspot (i.e., lifecycle) varies over time with cyclic ups and downs, and the characteristics of lifecycle functions are strongly correlated with the spatial locations of hotspots. Based on these findings, we have proposed a new mobility model, STEP.

STEP contains all the properties that we have observed from empirical data sets, i.e., heavy-tail flight lengths and pause-times, heterogeneously bounded mobility areas, heavy-tail inter-contact times, fractal waypoints and spatio-temporal correlations. This is the first such model that incorporates all of these properties.

In future work, we intend to apply our model to various applications that rely on spatio-temporal movement patterns. Familiar strangers [69] is a good example. They represent individuals who are recognized from regular activities, but with whom one does not interact. Familiar strangers are observed often in public urban areas. For instance, we see a person at a subway station every morning. We do not know him but if that person fails to appear, we notice. To describe this type of meeting, we have to faithfully describe the spatial and temporal visiting patterns of humans. Familiar strangers play an important role in forwarding messages or spreading viruses since they belong to different communities. By using spatial and temporal information of hotspots and individuals provided by our model, we can evaluate the impact of familiar strangers on mobile networking applications or even on viral infection studies.
REFERENCES


146


