A DATA-DRIVEN FRAMEWORK FOR MULTI-DIMENSIONAL PREDICTION PROCESSES FOR WLAN MOBILE USERS

By

JEEYOUNG KIM

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2013
To my husband Yong Nam for his love and support,
and to my mom and dad for always believing in me.
I love you.
ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor Prof. Ahmed Helmy for the continuous support and guidance throughout my Ph.D. degree. With his patience, motivation, enthusiasm, encouragement and immense knowledge he inspired me to become a better researcher. I could not have asked for a better advisor to work with and I am truly blessed to have had him as my Ph.D. advisor.

I would also like to thank my supervisory committee members Prof. Randy Chow, Prof. Shigang Chen, Prof. Sartaj Sahni and Prof. Michael Fang for their insights in asking questions that allowed me to think in different perspectives and new ideas, I thank you.

For the many colleagues I had privilege to work with; For Dr. Wei-jen Hsu, Dr. Shao-cheng Wang and Dr. Sapon Tenachaiwiwat for their advice and encouragement as seniors in the lab when I had just joined the NOMADS group. For Dr. Udayan Kumar and Dr. Gautam S. Thakur for the stimulating discussions, numerous feedbacks on my research and friendship and my other colleagues Dr. Sungwook Moon, Dr. Saeed Moghaddam, Yibin Wang and Guliz Seray Tuncay for many helpful discussions and suggestions for my research. I would also like to express my gratitude for the support of the CISE graduate advisors Mr. John Bowers, Ms. Joan Crisman and Ms. Kristina Sapp.

I do not have the words to express my love and gratitude towards my family. To my husband who loved me through thick and thin, I would have been lost without his patience and encouragement. To my dad, who has always been and will always be my role model. To my mom for being there whenever I needed her. For both of my parents’ unconditional love, support, guidance in life and for always believing in me no matter
what. To my little brother who never fails to make me smile. To my aunts and uncle for all their love. I would not be where I am today if it were not for my family.

I thank all my friends who have been there for me. There are too many to count but their friendship has always given me the boost that I needed whether they were near or far.

A special thanks goes out to all my friends and the priests at the Gainesville Korean Catholic Community, for their prayers, spiritual support and for warm words that always touched my heart. Thanks for being there when I needed a listening ear and for sharing my faith.

Last but not least I thank God for granting me this life, for all the joy and the sorrows and for guiding me through this journey. I have been blessed to have had the opportunity to pursue this Ph. D. degree.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ACKNOWLEDGMENTS</th>
<th>LIST OF TABLES</th>
<th>LIST OF FIGURES</th>
<th>ABSTRACT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-1</td>
<td>Data Sets Extracted from Dartmouth Trace</td>
<td>29</td>
</tr>
<tr>
<td>3-2</td>
<td>Number of APs and Users in each Yearlong Trace</td>
<td>29</td>
</tr>
<tr>
<td>3-3</td>
<td>Data Characteristics of University of Florida Year Long Data Sets</td>
<td>29</td>
</tr>
<tr>
<td>3-4</td>
<td>Data Characteristics of University of Florida Semester Long Data Sets</td>
<td>30</td>
</tr>
<tr>
<td>3-5</td>
<td>Data Characteristics of UF 2011 Fall Identified Smart-phones and Laptops</td>
<td>30</td>
</tr>
<tr>
<td>5-1</td>
<td>Correlation coefficients for different metrics with Markov O(2) Trace</td>
<td>49</td>
</tr>
<tr>
<td>5-2</td>
<td>Time evolution on different characteristics from yearly traces at a glance</td>
<td>49</td>
</tr>
<tr>
<td>5-3</td>
<td>Evolution and correlation between prediction accuracy of Markov O(2)</td>
<td>49</td>
</tr>
<tr>
<td>6-1</td>
<td>First order statistics on the distance error rate. Including when distance is 0 (which is a hit, Distance &gt;= 0)</td>
<td>75</td>
</tr>
<tr>
<td>6-2</td>
<td>First order statistics on the distance error rate. Not including when distance is 0 (Distance &gt; 0)</td>
<td>75</td>
</tr>
<tr>
<td>6-3</td>
<td>First order statistics on the distance error rate. Not including when it is a miss within the same building (Distance &gt; 10)</td>
<td>75</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>3-1</td>
<td>Illustration of the definition of activity range, shown as the square simulation area representing the geographic extent of the wireless mobile network. .......... 31</td>
<td></td>
</tr>
<tr>
<td>3-2</td>
<td>Illustration of a user location history of ABCBCADBCA ...................................... 31</td>
<td></td>
</tr>
<tr>
<td>3-3</td>
<td>Illustration of how the Markov O(2) will work given the user location history string of ABCBCADBCA ..................................................................................... 32</td>
<td></td>
</tr>
<tr>
<td>3-4</td>
<td>Illustration of how the LZ predictor will work given the user location history of string ABCBCADBCA ......................................................................................... 32</td>
<td></td>
</tr>
<tr>
<td>4-1</td>
<td>WLAN user prevalence for Dartmouth College Trace ........................................................ 37</td>
<td></td>
</tr>
<tr>
<td>4-2</td>
<td>VoIP device user prevalence for Dartmouth College Trace ............................................. 37</td>
<td></td>
</tr>
<tr>
<td>4-3</td>
<td>Cumulative Probability of Unique number of APs visited in Dartmouth College WLAN Trace .................................................................................................................. 38</td>
<td></td>
</tr>
<tr>
<td>4-4</td>
<td>Cumulative Probability of Unique number of APs visited in VoIP subset for Dartmouth College Trace ........................................................................................................... 38</td>
<td></td>
</tr>
<tr>
<td>4-5</td>
<td>WLAN Range Distribution for Dartmouth College Trace ................................................... 39</td>
<td></td>
</tr>
<tr>
<td>4-6</td>
<td>VoIP Range Distribution for Dartmouth College Trace .................................................. 39</td>
<td></td>
</tr>
<tr>
<td>4-7</td>
<td>Cumulative Probability of User’s Prediction Accuracy of Markov O(1) for Dartmouth College Trace ........................................................................................................... 40</td>
<td></td>
</tr>
<tr>
<td>4-8</td>
<td>Cumulative Probability of User’s Prediction Accuracy of Markov O(2) for Dartmouth College Trace ........................................................................................................... 40</td>
<td></td>
</tr>
<tr>
<td>4-9</td>
<td>Cumulative Probability of User’s Prediction Accuracy of Markov O(3) for Dartmouth College Trace ........................................................................................................... 41</td>
<td></td>
</tr>
<tr>
<td>4-10</td>
<td>Cumulative Probability of User’s Prediction Accuracy of LZ for Dartmouth College Trace .................................................................................................................. 41</td>
<td></td>
</tr>
<tr>
<td>4-11</td>
<td>Comparison of Cumulative Probability of User’s Prediction Accuracy of all Predictors for Dartmouth College WLAN Trace ........................................................................... 42</td>
<td></td>
</tr>
<tr>
<td>4-12</td>
<td>Comparison of Cumulative Probability of User’s Prediction Accuracy of all Predictors for Dartmouth College VoIP Trace ............................................................................... 42</td>
<td></td>
</tr>
<tr>
<td>5-1</td>
<td>Correlation between the AP encounter ratio and prediction using Markov O(2) on 1000 randomly sampled 02-03 users ........................................................................................................... 48</td>
<td></td>
</tr>
</tbody>
</table>
5-2 Correlation between the AP encounter ratio and prediction using Markov O(2) on 1000 randomly sampled 03-04 users ................................................................. 48

5-3 Cumulative Probability of Markov O(2) prediction accuracy for Dartmouth Quarter data sets for data collected between 2002-2006 ........................................... 50

5-4 Cumulative Probability of Markov O(2) prediction accuracy for UF semester data sets for data collected between 2007-2011 ................................................... 50

5-5 Comparison of Cumulative Probability of Markov O(2) prediction accuracy between semester long and year long traces at UF for data collected between 2007-2009 ............................................................................................ 51

5-6 Comparison of Cumulative Probability of VoIP vs WLAN Users from Dartmouth trace 2001-2004 and also UF Fall 2011 smart-phones vs. laptops this shows that the gap between the two (ultra mobile and less mobile) have gone down from 33% difference to 7% difference .............................................. 51

6-1 Framework of multi-dimensional prediction processes in a nutshell ............... 61

6-2 Comparison of the different prediction processes by controlling the input ordered by ascending order of prediction using the Oracle for UF Fall 2007 ...... 62

6-3 Improvement of Oracle over UF Fall 2007 sorted by the improvement of sectional in ascending order ........................................................................................... 62

6-4 Comparison of the different prediction processes by controlling the input ordered by ascending order of prediction using the Oracle for UF Spring 2008 ................................................................................................................ 63

6-5 Improvement of Oracle over UF Spring 2008 sorted by the improvement of sectional in ascending order ...................................................................................... 63

6-6 Comparison of the different prediction processes by controlling the input ordered by ascending order of prediction using the Oracle for UF Fall 2011 laptops ........................................................................................................ 64

6-7 Improvement of Oracle over UF Fall 2011 laptop devices sorted by the improvement of sectional in ascending order ............................................................. 64

6-8 Comparison of the different prediction processes by controlling the input ordered by ascending order of prediction using the Oracle for UF Fall 2011 smart-phones .......................................................................................................... 65

6-9 Improvement of Oracle over UF Fall 2011 smart-phone devices sorted by the improvement of sectional in ascending order .................................................. 65
6-10 Comparison of Cumulative Probability of AP and building level Markov O(2) prediction for 01-02 users ........................................................................................................ 66
6-11 Comparison of Cumulative Probability of AP and building level Markov O(2) prediction for 02-03 users ........................................................................................................ 66
6-12 Comparison of Cumulative Probability of AP and building level Markov O(2) prediction for 03-04 users ........................................................................................................ 67
6-13 Comparison of Cumulative Probability of AP and building level Markov O(2) prediction for 05-06 users ........................................................................................................ 67
6-14 Cumulative Probability of Prediction Accuracy for 01 02 User Trace .................. 68
6-15 Cumulative Probability of Prediction Accuracy for 02-03 User Trace ................. 68
6-16 Cumulative Probability of Prediction Accuracy for 03-04 User Trace ................. 69
6-17 Cumulative Probability of Prediction Accuracy for 05-06 User Trace ................. 69
6-18 Cumulative Probability of Markov O(2) AP Level Prediction for each Yearly Trace at Dartmouth College ................................................................................................. 70
6-19 Cumulative Probability of Markov O(2) Building Level Prediction for each Yearly Trace .................................................................................................................. 70
6-20 Empirical CDF showing the distance error rate in meters for 50 random users from UF Fall 2007 trace (distance error rate of 0 – 500 meters shown) ............ 71
6-21 Empirical CDF showing the distance error rate in meters for 50 random users from UF Spring 2008 trace (distance error rate of 0 – 500 meters shown) ....... 71
6-22 Empirical CDF showing the distance error rate in meters (from 0-500) for 50 random users from UF Fall 2011 smart-phone traces ......................................................... 72
6-23 Empirical CDF showing the distance error rate in meters (from 0-500) for 50 random users from UF Fall 2011 laptop traces ................................................................. 72
6-24 Empirical CDF showing the distance error rate in meters for 50 random users from UF Fall 2007 trace (zoom in and show only distance error rate of 15 – 500 meters) .................................................................................................................. 73
6-25 Empirical CDF showing the distance error rate in meters for 50 random users from UF Spring 2008 trace (zoom in and show only distance error rate of 15 – 500 meters) .................................................................................................................. 73
6-26 Empirical CDF showing the distance error rate in meters for 50 random users from UF Fall 2011 smart-phone trace (zoom in and show only distance error rate of 15 – 500 meters) .................................................................................................................. 74
6-27 Empirical CDF showing the distance error rate in meters for 50 random users from UF Fall 2011 laptop trace (zoom in and show only distance error rate of 15 – 500 meters) ................................................................. 74

6-28 Correlation between the accumulated history and the AP encounter ratio of 100 random users from UF Fall 2007 ................................................................. 76

6.29 Correlation between the accumulated history and the AP encounter ratio of 100 random users from UF Spring 2008 ................................................................. 76

6-30 Pseudo code for prediction guidelines derived from the results above ............ 77
With the proliferation of numerous light weight devices along with the wide spread use of wireless local area networks (WLANs) in many public places we are connected-on-the-go nowadays more than ever. Such change, in device technology and coverage ubiquity, results in unexplored dynamics and raises several challenging questions. How are these changes affecting the behavior of mobile users? And how do these changes affect mobile user predictability and the networking protocols that utilize it?

To shed light on the changes and how protocols involving the mobility of users can change, we follow a systematic analysis methodology. First, using a three year long network trace collected from Dartmouth College, we study the user mobility and its effects on predictability of regular and ultra-mobile users, by analyzing the contrast between the mobility of the WLAN users, and four carefully selected sets of ultra-mobile users across various mobility metrics. We also investigate how these differences in mobility affect the predictability of such user’s next locations. Then, we study the evolution of user mobility using extensive network traces over a period of ten years, collected from two major Universities (University of Florida and Dartmouth College) and
also investigate a series of prediction methods in order to analyze the evolution of prediction accuracy of these WLAN users.

Based on the insights gained from these extensive analyses, we design a novel framework of a multi-dimensional prediction process that aims to improve prediction of mobile WLAN users. We also include a study on two subsets, namely a subset of the smart-phone devices and laptop devices from the UF Fall 2011 trace. This study of user mobility and predictability, followed by our multi-dimensional prediction process framework paves the way for better understanding of present day mobile users and aids in better prediction of future WLAN mobile users.
CHAPTER 1
INTRODUCTION

Wireless LAN (WLAN) traces are an important source of information that allows researchers to have a glimpse into real life human behavior. A user’s WLAN usage pattern may closely relate to real-life human behavior, and constitutes a critical research area in wireless networks. As devices become more portable (such as netbooks, smartphones, tablets, etc.) mobile users’ behavior and usage pattern is likely to change. These changes in the nature of the devices; they are portable and easier to carry around, and mobile for use on the go, subsequently allow the WLAN traces to capture more user mobility than ever before. Devices’ portability allows users to be more mobile while using the network than before. Also, access ubiquity allows users to go about their regular routines even while using their mobile devices. Previously, due to device and coverage restrictions users were forced in many cases to change their behavior and limit their mobility during network usage. How much of these changes will appear in the traces and reflect the users’ increased mobility?

We investigate a five year long WLAN trace collected over a six year period (2001-2006) at the Dartmouth College [14] and a three year long trace collected over a five year period (2007-2012) at the University of Florida (UF) [13]. For the first part of our study, we focus on a few subsets of wireless users from Dartmouth, which have been systematically selected to be more mobile than the rest of the WLAN users according to various mobility metrics (e.g., access points (APs) visited), from a three year long trace (2001-2004). We call these subsets the ultra-mobile data set and call each user in the subsets, ultra-mobile users. One of the subsets consists of VoIP device users. These users leave their devices on most of the time and the devices are light enough to walk
and talk. Hence, these users show a more mobile characteristic than laptop or other heavy device users while connected to the network. We aim to compare the behavior of highly mobile users to the general WLAN users by analyzing these traces. This sheds light on the realism of WLAN trace-based models. We also aim to examine the effect of any differences on protocol performance, e.g., prediction protocols.

When the mobility of ultra-mobile user subsets are compared to our entire WLAN trace, our results clearly indicates that there is a significant difference of APs visited and the coverage area between ultra-mobile users and general mobile users. But does such dramatic contrast in mobility affect mobile networking protocols? In order to quantify such effect we examine the accuracy of several classes of mobility prediction protocols under various conditions of realistic mobility.

We compare these different sets of traces using several different predictors including the Markov O(1), O(2), O(3) and also the LZ predictor. Our experiments indicate that the Markov O(2) is the predictor with the highest accuracy among the four predictors and the LZ has the lowest. Surprisingly, all predictors perform quite poorly with ultra-mobile users, with VoIP users having the lowest prediction rate of an average of approximately 25% correct prediction rate, compared to 60% for the general WLAN users, while the other ultra-mobile user subsets fall in between these prediction rates. These results prompt re-visiting of such algorithms for ultra-mobile users.

To further investigate the mobility and predictability trends over the more recent years, we conduct an “evolution” analysis for the second part of our study.

In Chapter 5 of our study, we increase the study period span to six years, add another set of data set collected from UF spanning four years and adopt a few more
mobility metrics. Mobility is now also measured by the number of distinct and total APs a user has visited and also the users’ encounter ratio of such APs. We focus on the Markov chain family of predictors ranging from order 1 through 3 in this study, particularly focusing on Markov O(2); the best performer in our overall study. We find that the number of distinct APs an average user visits each year increases by approximately 3 APs every year for the Dartmouth trace. We also find that the predictability decreases each year resulting in a 20% drop of prediction accuracy from the 01-02 trace to the 05-06 trace for Dartmouth and a 30% drop from fall of 2007 to fall of 2011 at UF.

This study shows that the user mobility is indeed changing over time and with the changes in the user mobility, there is also a shift in the user location predictability over time. Such trends are expected to further manifest themselves in the future with more portable devices. We have not only done the time evolution analysis over several years of WLAN traces, but we have also explored several different definitions of success for the predictors in order to better understand the results of our analysis.

Stemmed from the findings and insights gained from the extensive analysis we have designed a framework that expands the spatiotemporal space for multi-dimensional prediction processes. This is a novel framework which given multiple dimensions of prediction processes by controlling the input or output (or both), we will be able to better predict future WLAN mobile users by choosing from a set of different prediction processes rather than using a single prediction process. We expand the temporal state space by looking at different blocks of time in a day, and the spatial aspect by introducing the distance error rate as well as different granularities of success.
The rest of the document is outlined as follows. Chapter 2 discusses related work and present literature, Chapter 3 explains the data sets and metrics used throughout this study. In Chapter 4 we investigate WLAN mobile user’s mobility and predictability, and how they affect each other, in Chapter 5 we discuss our findings on the evolution study of a mobile user's mobility and predictability over the years, in Chapter 6 we discuss the insights and ways to improve the prediction process and introduce the multi-dimensional prediction process framework and the Oracle. We explore the spatiotemporal expansions using metrics such as distance error rate, different granularities of success and the time of day users are predicted. We also show a case study of smart phones vs. laptops and we conclude and discuss future work in Chapter 7.
CHAPTER 2
RELATED WORK

Related work can be found in areas of mobility modeling, location prediction, trace analysis and behavioral mining for WLAN users. Among various mobility modeling techniques, real world trace based modeling is the most realistic and is what can be called the closest to the ground truth. [2] shows a mobility model that is based on real WLAN traces, [4] also extracts a mobility model from real user traces and [30] also explore a real user trace and analyze it. Model T [5] and T++ [6] are empirical registration models derived from the WLAN registration patterns of the mobile users. They are able to formulate the inter dependence of space and time explicitly by a set of few equations.

[1] proposed a mobility model to capture time variant user mobility. In this model, they define communities that are visited often by the nodes to capture the skewed location visiting preferences, and use time periods with different mobility parameters to create the periodical re appearance of nodes at the same location. [9] also look into modeling generic WLAN users by identifying the mobility characteristics of individual users. Studies done in [29][32] shows that there exists repetitive behavioral trends in the association pattern of groups of users in large WLANs. In [7] researchers studied user mobility patterns and introduced metrics to model user mobility from a four week trace collected in a large corporate environment. They also analyzed user distribution and load distribution across access points [31]. Most of these works are directly based on WLAN traces which can be found under the MobiLib project [13]or the CRAWDAD project [14].
Researchers studied the changing usage of a mature campus wide WLAN by investigating the workload and usage of the Dartmouth WLAN trace for an 11 week period in 2001 and a 17 week period in 2003 2004 [8]. They discovered that a mature WLAN (two years old at the time of investigation) trace showed significant difference from the initial usage of the WLAN with new devices and applications such as streaming multimedia and P2P services. There is mention on mobility but the focus is on the difference in the usage of the WLAN.

The study in [3] investigated several domain independent predictors for the location prediction on the WLAN trace, but did not define mobility characteristics or propose any techniques to construct the mobility model. Based on the result, they gave some suggestions for the usage of the predictor on WLAN traces. There are a number of user mobility prediction algorithms [10][11] in the current literature that target cellular networks. These predictors are used in a different setting and for different purposes (i.e. paging scheme [10], efficient handoff [10] and resource reservation [11]. The difference including, but not limited to, the fact that a cellular device showing up in a cell that is a long distance away is very low, thus it is bounded location wise. There is also work done by [12] to improve caching paradigms that analyze the wireless information locality and association patterns on a month long campus measurement trace. The characteristics and scale of the predictions mentioned in the above literature are different from what we are working on. They aim to improve the performance of wireless infrastructures by load balancing, admission control and resource reservation whereas we are investigating the behavior of WLAN users and how this will affect the overall performance of our predictors.
In our study, we use four predictors that have already been explored in existing literature to verify how the prediction accuracy changes due to changes in user mobility. The Order k Markov predictor and Lempel-Ziv (LZ) predictor that is used in our study is well-explored and widely used in various fields of study [3][12][15][16][17][18]. We discuss our data sets and metrics in Chapter 3.
CHAPTER 3
DATA SETS AND METRICS

3.1. Data Sets

For our investigations, we use the WLAN trace collected from the Dartmouth College campus and the University of Florida campus. We make the assumption that each device belongs to one unique user, thus using the term “user” and “device” interchangeably when describing the unique MAC address we find for each device. In our more recent studies, we find that with the proliferation of smart-phones, tablets and other secondary Wi-Fi enabled devices, the number of devices nearly double the actual population (Table 3-3 and Table 3-4). However we continue to use the term “user” and “device” interchangeably throughout this study for consistency reasons. While using this trace as our standard, general WLAN user base, we also extract other ultra-mobile user data sets from these traces which include VoIP users (from Dartmouth College) and smart-phone and laptop users (from University of Florida) with more detailed descriptions of each of these data sets in the subsections 3.1.1 and 3.1.2.

3.1.1. Dartmouth College Trace

In our first study (covered in Chapter 4), we use the three yearlong Dartmouth movement trace [14] collected from 2001 to 2004. There are 13888 unique users and 623 different APs (access points) in this particular trace. The VoIP data set we use in this work is a subset of the Dartmouth WLAN trace described above and consists of 97 users. These are acquired by mapping the whole WLAN trace with a MAC to device type map, which is a list of all the MAC addresses mapped with the type of device it is by looking at the first three octets of the MAC addresses. Among these 97 users we observe two types of VoIP devices which are the Cisco7920 and Vocera devices. We
have particularly chosen VoIP devices to measure the mobility of WLAN users since VoIP devices are always on and also online, unlike other pocket PCs or PDAs that were on the market at the time (2001-2004), that may easily go into hibernate mode or may even be turned on and off frequently [8].

Along with the VoIP data set we have generated ultra-mobile test sets from the same traces in order to validate our findings. There are three test sets used in this work and they are all considered to be ultra-mobile users[18][19]. The ‘ap_200’ and ‘ap_170’ sets are both based on the number of APs a user has visited. The ‘ap_200’ set is a collection of users who have visited 200 APs or more during the length of the trace and the ‘ap_170’ set is a collection of users who have visited more than 170 APs but less than 200. The ‘range’ set is a collection of users who have covered the largest physical area during the length of the trace. This was done by studying the AP location file and calculating the area range that each user has covered. Each of these test sets has approximately 100 users each. The following Table 3-1 show the characteristics of the different data sets extracted from Dartmouth College that we used in this study at a glance.

The ultra-mobile study in Chapter 4 led us to conclude that we are indeed in need of investigating how these changes in mobile devices will affect the characteristics over time, which propelled us to further investigate the evolution of the WLAN users over time in Chapter 5.

For our next part of the study (Chapter 5), we use the three yearlong movement trace collected from the Dartmouth College from 2001 to 2004 and the syslog trace that was collected from 2005 to 2006 [14]. There are 13888 unique users and 623 different
access points (APs) in the former trace and 24399 unique users and 1270 different APs in the latter trace. The drastic change of the APs is due to a network infrastructure change that went on at Dartmouth during 2004 to 2005. They were in a transition period of going from Cisco APs to Aruba APs. While using this trace as our standard general WLAN user base, we also extract other ultra-mobile user data sets from this trace to broaden the spectrum of our study. We divide the traces into yearlong traces and the breakdown of the trace for each year is as follows; 2001-2004 is broken down from the first day of July to the last day of June of each year and 2005-2006 is from the first day of September to the last day of August. This allows us to investigate the evolution of the characteristics of the users in the WLAN trace over the time span of six years.

As shown in Table 3-2, it is easy to see the rapid growth of the number of users over the years. This shows that the number of mobile devices have almost doubled each year and again highlights the importance of such a study in order to better understand the growing and evolving mobile user community. With the rapid propagation of such affordable devices, the mobile user community is no longer a small subset of the society. This can be categorized as a characteristic of the wireless network and its users which will continue to grow in the future [8]. In order to minimize the effect that the difference in the number of users has on our studies, we have done some analysis on randomly sampled users as well as the entire year long traces.

3.1.2. University of Florida Trace

The University of Florida (UF) trace [13][26] has been collected from Fall 2007 through Spring of 2009 discontinued then resumed again starting Spring of 2011. This is a 3 year long trace collected over 5 years (and is still being collected as we speak). There was a network change that started during fall of 2008 which caused some parts
of the campus to disappear in the trace during fall of 2008 and Spring of 2009. There are some gaps present in the Spring 2009 trace and the trace collection was discontinued from summer of 2009 through fall of 2010.

In our evolution study (Chapter 5) we use the whole UF trace divided into different time lengths such as yearlong traces and semester long traces. The characteristic of the users for these traces can be found in Table 3-3 and Table 3-4.

The number of devices found in the traces has drastically changed over the years, and by looking at the numbers one can presume that the most likely reason behind the drastic change in numbers are that the number of devices each individual carries around has changed. According to the factsheet from the University of Florida Office of Institutional Planning and Research, the approximate number of student, faculty and staff combined during these years were between 50,000 to 55,000 per semester\(^1\). Thus, we can assume each individual in 2011 is on average carrying 2 devices. This is not a far-fetched idea considering that smart phones have become much more affordable and wide spread, so a user will on average have a laptop and a smart phone (or other highly mobile i.e. tablets) device.

In our framework (Chapter 6) we introduce two subsets of the UF trace, which is a data set comprised of smart phones and laptops. We have conducted a survey to collect the unique manufacturers first 3 octets of MAC addresses among smart phone users. Through the survey we were able to identify 30 unique octets (Organizationally Unique Identifier – OUI) from the total of 95 responses received. These surveys were

\(^1\) http://www.ir.ufl.edu/
conducted from December 2011 to January 2012. Table 3-5 shows the data characteristics of these two subsets of users from the UF Fall 2011 trace.

In Chapter 4, we focus on the mobility and predictability of WLAN users. How does one measure mobility? How does one define success in a prediction model? In order to answer these questions, we will next discuss the metrics and approaches that are taken in order to be able to measure the mobility and predictability of WLAN users.

3.2. Metrics

3.2.1. Mobility Metrics

How do we measure mobility? Several metrics can be used but it is unclear which of the metrics is best suited for our study. We discuss the perceived metrics we used to measure mobility in this sub section.

3.2.1.1. Distinct number of APs

This is a metric used commonly in previous work such as [2][7]. This may seem intuitive however we have experimented in a systematic manner in order to verify that the number of unique access points a user has visited can indeed be used as a method to measure mobility. We found that indeed a user who visits only a handful of APs shows less mobility while online than those who visit a large number of APs.

3.2.1.2. Prevalence

Prevalence is a mobility metric proposed in [2][6][7], which indicates the time that a user spends at a given AP, as a fraction of the total amount of time they spend on the network. Higher prevalence indicates that a user has spent more time on a certain AP, and thus is deemed less mobile and lower prevalence means that a user has spent less time on a given AP and is described as more mobile.
3.2.1.3. Activity range

The activity range is another mobility metric we investigate and is defined as the smallest square area which covers all the access points the user has visited in a single activity, where a single activity is denoted as the time when a user logs on to the network until the user logs out of the network. Figure 3-1 shows an illustration of how the activity range is defined, where the circles implicate locations and the arrows indicate the user’s visitation path in a single activity. The area of the outer square which is the smallest square covering all the access points will be the activity range.

3.2.1.4. AP encounter ratio

As we extended our study to different data sets in both time (expanded over 10 years) and space (Dartmouth College and UF) we realized that due to the dynamically changing network (network upgrades every 3 years on average with an astonishing addition of APs with every upgrade) we needed a mobility metric that does not only rely on the absolute number of APs. The AP encounter ratio is a calculated metric of the ratio of distinct number of APs visited over the total number of encounters with APs. This metric shows a clear and consistent negative correlation (Figure 5-1 and Figure 5-2) between predictability and mobility and is also used as a means to define mobility in our study.

3.2.2. Predictability Metrics

In our study, we used domain independent predictors to predict the location of a WLAN user. In subsections 3.2.2.1 and 3.2.2.2 we briefly discuss the prediction algorithms we used.
3.2.2.1. Markov family of predictors

We focus on the well-known Markov chain algorithm for our predictors in this study. Markov chain assumes that a location can be predicted from the current context which is the sequence of the k most recent symbols in the location history. Depending on what k is, i.e. if we use 3 most recent symbols (locations) for the Markov chain predictor we call this predictor Markov Order 3 and denote it as Markov O(3).

The Markov chain model represents each state as a context, and transitions represent the possible locations that follow that context. Thus, if we were to use the Markov O(3) predictor, we would look at the 3 most recent locations of the user and try to predict the next location that the user will visit. In our study, we use the Markov O(1), Markov O(2) and Markov O(3) predictors [3][12][17][18][19]. Figure 3-3 shows an example of what the Markov O(2) predictor does when given a user location history of ABCBCADBCA as illustrated in Figure 3-2. Based on the past 2 locations Markov O(2) builds its own graph and tries to predict where the user will move to next depending on the 2 most recent locations.

3.2.2.2. Lempel-Ziv algorithm

The Lempel-Ziv algorithm is a famous data compression algorithm which we use in our prediction method so that it predicts in the case when the next symbol in the produced sequence is dependent on only its current state (but does not have to correspond to a string of fixed length). The length of the string may vary and is allowed to grow up to infinity. This is similar to Markov O(k) however, k is not fixed and may grow to infinity [3][15][17]Error! Reference source not found.[18][19][20][23]. Figure 3-4 shows an example of what the LZ predictor looks like as history builds.
Table 3-1. Data Sets Extracted from Dartmouth Trace

<table>
<thead>
<tr>
<th>Year</th>
<th>Labels (Characters)</th>
<th>Number of Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001-2004</td>
<td>WLAN (all users in trace)</td>
<td>13439</td>
</tr>
<tr>
<td></td>
<td>VoIP (Voice over IP users in trace)</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>AP_200 (users visiting more than 200 Aps)</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>AP_170 (users visiting more than 170 Aps and less than 200 Aps)</td>
<td>127</td>
</tr>
<tr>
<td></td>
<td>range (users covering the largest physical area range)</td>
<td>113</td>
</tr>
</tbody>
</table>

Table 3-2. Number of APs and Users in each Yearlong Trace

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of APs</th>
<th>Number of Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001-2002</td>
<td>516</td>
<td>2898</td>
</tr>
<tr>
<td>2002-2003</td>
<td>554</td>
<td>6370</td>
</tr>
<tr>
<td>2003-2004</td>
<td>572</td>
<td>11369</td>
</tr>
<tr>
<td>2005-2006</td>
<td>1270</td>
<td>24399</td>
</tr>
</tbody>
</table>

Table 3-3. Data Characteristics of University of Florida Year Long Data Sets

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of APs</th>
<th>Number of Devices</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007-2008</td>
<td>812</td>
<td>49898</td>
</tr>
<tr>
<td>2008-2009</td>
<td>665</td>
<td>81678</td>
</tr>
<tr>
<td>2011</td>
<td>1868</td>
<td>158703</td>
</tr>
</tbody>
</table>
### Table 3-4. Data Characteristics of University of Florida Semester Long Data Sets

<table>
<thead>
<tr>
<th>Semester</th>
<th>Number of APs</th>
<th>Number of Devices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fa_07</td>
<td>643</td>
<td>49898</td>
</tr>
<tr>
<td>Sp_08</td>
<td>760</td>
<td>35600</td>
</tr>
<tr>
<td>Su_08</td>
<td>662</td>
<td>38711</td>
</tr>
<tr>
<td>Fa_08</td>
<td>585</td>
<td>54332</td>
</tr>
<tr>
<td>Sp_09</td>
<td>580</td>
<td>43410</td>
</tr>
<tr>
<td>Sp_11</td>
<td>1475</td>
<td>84995</td>
</tr>
<tr>
<td>Fa_11</td>
<td>1673</td>
<td>119272</td>
</tr>
</tbody>
</table>

### Table 3-5. Data Characteristics of UF 2011 Fall Identified Smart-phones and Laptops

<table>
<thead>
<tr>
<th>Semester</th>
<th>Number of APs</th>
<th>Number of Devices</th>
<th>Device Type</th>
<th>Number of Devices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall 2011</td>
<td>1673</td>
<td>119272</td>
<td>Smart-phones</td>
<td>19300</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Laptops</td>
<td>23929</td>
</tr>
</tbody>
</table>
Figure 3-1. Illustration of the definition of activity range, shown as the square simulation area representing the geographic extent of the wireless mobile network.

Figure 3-2. Illustration of a user location history of ABCBCADBCA
Figure 3-3. Illustration of how the Markov $O(2)$ will work given the user location history string of ABCBCADBCA

Figure 3-4. Illustration of how the LZ predictor will work given the user location history of string ABCBCADBCA
CHAPTER 4
WLAN USER MOBILITY AND PREDICTABILITY

In this study, we compare the mobility characteristics of WLAN traces and ultra-mobile user traces from several different aspects.\* We also look into several prediction techniques in order to study how the sharp contrast in mobility between ultra-mobile users and WLAN users affects different protocols that use these data sets such as location (i.e., next access point the user will visit) prediction.

The evaluation metrics for mobility in this study include prevalence, the number of distinct access points visited by a given user and the activity area range. We also investigate Markov $O(1)$, Markov $O(2)$, Markov $O(3)$ and the Lempel-Ziv(LZ) algorithms as our prediction methods. The results of the mobility comparison and the prediction algorithm on each of our set of traces are shown in the subsequent sections 4.1., and 4.2.

4.1. Mobility Comparison

Prevalence is one of the mobility metrics proposed in [6], and defined in Chapter 3. Figure 4-1 and 4-2 show that VoIP users are more mobile than WLAN users, since the bars are lower, indicating that the users are spending less time at a given AP compared to the overall time that they spend online. Especially for the rightmost bar which indicates prevalence higher than 0.95, the WLAN trace shows a much higher trend than the VoIP trace set. This can be interpreted as a larger portion of users in the WLAN trace has spent most of their time on only one AP compared to the users in the VoIP trace.

\* This chapter adapted from [19][20][21][23]. For more detail on this work refer to [19][20][21][23].
The number of APs a user visits is a second mobility metric we look into. A user who visits only a handful of APs show less mobility while on-line than those who visit a large number of APs. Figure 4-3 and 4-4 shows the distribution (CDF) of WLAN and VoIP users’ number of access points visited for each user. With the average number of APs a VoIP user visits being 146 compared to that of the WLAN user which was 36, we can see there is a huge difference in mobility. The median number of APs visited for WLAN users was 17 whereas for VoIP users it was 131. You can clearly see in Figure 4-3 that more than 70% of WLAN users access less than 50 APs, whereas the VoIP users are more evenly distributed and 60% of the population accesses more than 100 APs as indicated in Figure 4-4.

Figure 4-5 and 4-6 show the activity range distribution for WLAN and VoIP users. The percentage of VoIP users having a larger area of activity range is higher than that of the WLAN users. As indicated in Figure 4-5, you can see that 90% of the user population in the WLAN trace stays inside a 1 square kilo meter area range whereas only a little more than 50% stays inside a 1 square kilometer area for VoIP users as shown in Figure 4-6.

4.2. Predictability Comparison

To study the effect of the sharp contrast in mobility and behavioral characteristics between VoIP and other WLAN users on networking protocols, we analyze a set of well-known prediction algorithms explained in Chapter 3 with the various sets of traces we have in our study, namely the subsets which constitute of users who have visited more than 200 APs (ap_200), more than 170 but less than 200 APs (ap_170) and visited an area that is larger than 1km\(^2\)(range).
We have run the Markov O(1), O(2) and O(3) predictors along with the LZ predictor [3][12][17][18][19][23] for each of the test sets we have, and also for the VoIP trace set and the whole body of the WLAN trace. We also compared the accuracy of all four predictors with the VoIP trace data to see which one has the best performance. Accuracy is measured as percentage of correct predictions of the next AP to visit.

As shown in Figures 4-7 through 4-10, we can see that the WLAN trace always has the best prediction accuracy overall, for each of the different predictors, with an average of about 60% accuracy. The VoIP trace, by contrast had the worst prediction accuracy for all the predictors with an average of approximately 25% accuracy. The ultra-mobile sets ap_170, range and ap_200 each had an average of 50%, 47%, and 40% prediction accuracy respectively. From Figures 4-7 through 4-10, we can see that the best accuracy can be no more than 80% for VoIP users, while there can be more than 95% accuracy for WLAN users.

When we were first conducting our experiment, we expected that the range of the physical area that each user covered would be a better criterion to measure mobility than the number of APs visited, since we consider a user to be more mobile when that user covers more ground. Hence, we expected that the ‘range’ set would return bad prediction accuracy. Surprisingly, the ‘range’ set always exhibits performance between the other two test sets (ap_200 and ap_170), which indicates that the users that covered larger areas physically, most likely have visited an average of 200 APs during their lifetime.

To explain this result, intuitively the users that had visited less APs also had a better prediction rate than that of the users who had visited more APs. The difference of
the prediction accuracy between the two data sets (ap_170 and ap_200) is always around 10% near the median.

As for the comparison of the predictors on the VoIP data set and WLAN trace, as shown in Figures 4-11 and 4-12, the LZ predictor showed the worst prediction rate and the Markov O(2) showed the best prediction accuracy by a very minimal difference from the Markov O(1). Markov O(3) did not show a good prediction and these results indicate that a larger data structure and higher complexity does not help in making better predictions. However, the four predictors that are used in this study do not provide good prediction for the VoIP data set, although they are showing a very similar trend regardless of the mobility of the user.

In Chapter 5 we discuss the evolution of WLAN mobile user characteristics over time, and how these changes affect user location predictability.
Figure 4-1. WLAN user prevalence for Dartmouth College Trace

Figure 4-2. VoIP device user prevalence for Dartmouth College Trace
Figure 4-3. Cumulative Probability of Unique number of APs visited in Dartmouth College WLAN Trace

Figure 4-4. Cumulative Probability of Unique number of APs visited in VoIP subset for Dartmouth College Trace
Figure 4-5. WLAN Range Distribution for Dartmouth College Trace

Figure 4-6. VoIP Range Distribution for Dartmouth College Trace
Figure 4-7. Cumulative Probability of User's Prediction Accuracy of Markov O(1) for Dartmouth College Trace

Figure 4-8. Cumulative Probability of User's Prediction Accuracy of Markov O(2) for Dartmouth College Trace
Figure 4-9. Cumulative Probability of User’s Prediction Accuracy of Markov O(3) for Dartmouth College Trace

Figure 4-10. Cumulative Probability of User’s Prediction Accuracy of LZ for Dartmouth College Trace
Figure 4-11. Comparison of Cumulative Probability of User’s Prediction Accuracy of all Predictors for Dartmouth College WLAN Trace

Figure 4-12. Comparison of Cumulative Probability of User’s Prediction Accuracy of all Predictors for Dartmouth College VoIP Trace
CHAPTER 5
EVOLUTION OF WLAN USER MOBILITY AND PREDICTABILITY

Now that we have analyzed the different characteristics of ultra-mobile users and also studied the effect it has on prediction, the next question is, in which direction will user mobility evolve in the future? What is the trend in user mobility and how in turn, will this affect the location prediction of these users? In Chapter 5, we investigate the above matters by first dividing our data set into each year long data set as shown in Table 3-2. Then we apply our mobility metrics in order to extract the characteristics and trend over time. We continue to use the four predictors mentioned in Chapter 3 and 4 but with a focus on Markov O(2) since it has continuously proved to have the best prediction accuracy among the four.

5.1. Evolution of User Mobility

In this part of the study, we use the distinct number of APs a user has visited and also the AP encounter ratio as mentioned in Chapter 3, as the mobility metrics. User mobility is difficult to capture with numbers, but in this section we investigate the correlation between our mobility metric and prediction accuracy to see how closely related they are. First, in Figure 5-1 and 5-2 we show the correlation between the AP encounter ratio and prediction accuracy of Markov O(2) which has been found to best predict WLAN users among all four predictors in the study we have done in Chapter 4. These two graphs were produced using a thousand randomly sampled users from the 02-03 and 03-04 trace [14].

We have quantified the correlation coefficient between the AP encounter ratio and the Markov O(2) prediction accuracy in Table 5-1. As shown, there is a clear and

* This chapter is adapted from [16][17][18][16]. For more detail on this work refer to [16][17][18].
consistent negative correlation between predictability and mobility. The correlation graph for the other years show similar patterns.

In Table 5-2 we can see at a glance, how the mobility metrics evolve from year to year. The calculated number of users is the actual number of users that are considered in the calculation of the average and median access points and these are the users that have enough history accumulated (appeared at least 500 times in the trace) to run the predictors. We can easily see that the average number of distinct APs as well as the median of distinct APs is growing, which means the users are showing higher mobility. However we must keep in mind that the total number of APs can change with network upgrades. This is also the case for the 05-06 Dartmouth trace where there was a drastic increase in the number of APs where it went from 572 to 1270 (Table 3-2). With a network change almost every 3 years, there was a need to introduce a new mobility metric i.e. the AP encounter ratio.

In Figure 5-3, we can see that this coincides with our results for the lower prediction accuracy, as described in Chapter 4, that the more mobile a user is the less predictable they are. In section 5.2., in order to compare the performance of different predictors we use the prediction accuracy metrics which define the percentage of correct prediction for each user and study the change in predictability over the years by looking at the evolutionary trend of the WLAN mobile users.

5.2. Evolution of User Predictability

In Chapter 4 we have shown that predictability and mobility in WLAN users have a significant relationship. We showed that the more mobile a user is the less predictable and vice versa. Among the predictors we used, the predictor with the best performance
was the Markov O(2) predictor. However, even Markov O(2) only had an approximate of 60% prediction accuracy which we did not find satisfactory.

Table 5-3 shows a binning of users for each year long trace according to their prediction accuracy using Markov O(2) AP level prediction. It shows the average AP encounter ratio for each user category of prediction accuracy, where the 0~10% category is the group of users who have 10% or less prediction accuracy, 10~30% category users have more than 10% but less than 30% accuracy and so on. There were no users with prediction accuracy 10% or lower for the 01-02 and 02-03 trace. We find that the trend is the more predictable the user, the lower the average AP encounter ratio. Note that the average distinct number of APs visited is growing each year at a pace of approximately 3 APs every year, indicating that the users are indeed becoming more mobile over time.

Figure 5-3 shows the Markov O(2) prediction accuracy for the Dartmouth trace divided in quarters. This shows a unique trend where some quarters in the same year show close coupling that is almost overlapping whereas they show 3%-5% drop in prediction accuracy for different years. The prediction accuracy deteriorates as time passes. Each trace is approximately 10 weeks long and is probably the smallest time period of data we have investigated. We have filtered out data that had less than 500 prediction-worthy users, which included all the summer quarters along with 2001 quarters. In Chapter 6, Figure 6-18 shows that the Markov O(2) prediction accuracy for each yearlong Dartmouth trace also shows a similar trend in the prediction accuracy deteriorating over the years. There is a need to investigate why such a unique trend has appeared in this particular group of data and we will explore that in the future.
We have studied and explored the evolution trend of the Dartmouth College trace. Next we will investigate the UF trace. Being a larger campus than Dartmouth, the trace collected at UF has many more APs and much more devices than that collected from Dartmouth. The difference in the traces introduces a richness and diversity to the data while being collected from a University campus gives these two traces some similarities in characteristic. Figure 5-4 shows the evolution of UF traces split into semester long data sets. Again the consistent degradation of the prediction accuracy over time is present with the one exception of Spring 2009 almost overlapping with Fall 2008. Further investigation showed us this anomaly was due to a 6 week gap in the Spring 2009 trace.

How will these traces that are split into semester long and quarter long time periods compare to the yearlong WLAN trace? In Figure 5-5 we show the comparison of the UF semester long and yearlong trace. You can easily see that the yearly trace has worse prediction accuracy in both 2007-2008 and 2008-2009 by approximately 8% to 20% depending on which semester you are comparing it with. This is a quantitative result of the intuitive assumption that users on campuses will be more predictable from semester to semester than year to year due to the aspects of students having different schedules for each semester.

Some evolutionary studies can be seen in section 6.2.1 when we discuss the different granularity of location prediction accuracy. The results shown in that study holds the trend shown in this section, further strengthening our argument that WLAN mobile users are evolving towards a less predictable state. Thus motivates the need to revisit WLAN user location prediction. We also find an interesting evolutionary trend that
involves the more mobile and less mobile users (i.e. VoIP and smart phones vs. laptops and regular WLAN traces) in Figure 5-6. Although we have validated in length that the users are becoming less predictable as time goes by, exactly one subset of users do not fit in this category and that is the smart-phone trace. Compared to the VoIP trace (which is also ultra mobile and light weight similar to smart-phones) the smart-phones show a better prediction accuracy of approximately 18% with the evolutionary trend going the other way around. The predictability gap between the ultra mobile and less mobile users may be shrinking and we can hypothesize that this is due to smart-phones having more capabilities compared to the VoIP devices and that they may be in some ways replacing the role of what only laptops used to do i.e. checking email, surfing the web, etc.. This would be an interesting topic to look into in the future to see whether this will become a new trend that overwrites the current ones.

We have done an extensive study using two large data sets collected over a ten year time period and have done evolution analysis on the mobility and predictability of mobile users. We will discuss the insights we have gained in the various and systematic analyses we have completed and introduce a novel framework of multi-dimensional prediction process to better predict future WLAN mobile users in Chapter 6.
Figure 5-1. Correlation between the AP encounter ratio and prediction using Markov O(2) on 1000 randomly sampled 02-03 users.

Figure 5-2. Correlation between the AP encounter ratio and prediction using Markov O(2) on 1000 randomly sampled 03-04 users.
### Table 5-1. Correlation coefficients for different metrics with Markov O(2) Trace

<table>
<thead>
<tr>
<th>Correlation with Markov O(2) Prediction Accuracy</th>
<th>01-02</th>
<th>02-03</th>
<th>03-04</th>
<th>05-06</th>
</tr>
</thead>
<tbody>
<tr>
<td>District Number of AP</td>
<td>-0.534</td>
<td>-0.504</td>
<td>-0.516</td>
<td>-0.559</td>
</tr>
<tr>
<td>AP Encounter Ratio</td>
<td>-0.589</td>
<td>-0.594</td>
<td>-0.502</td>
<td>-0.570</td>
</tr>
</tbody>
</table>

### Table 5-2. Time evolution on different characteristics from yearly traces at a glance

<table>
<thead>
<tr>
<th></th>
<th>01-02</th>
<th>02-03</th>
<th>03-04</th>
<th>05-06</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of Users</td>
<td>2898</td>
<td>6370</td>
<td>11369</td>
<td>24479</td>
</tr>
<tr>
<td>Total Number of Users</td>
<td>834</td>
<td>2317</td>
<td>3765</td>
<td>2148</td>
</tr>
<tr>
<td>Average of distinct APs</td>
<td>50.17</td>
<td>53.67</td>
<td>57.18</td>
<td>64.47</td>
</tr>
<tr>
<td>Median of distinct APs</td>
<td>41</td>
<td>44</td>
<td>48</td>
<td>59</td>
</tr>
</tbody>
</table>

### Table 5-3. Evolution and correlation between prediction accuracy of Markov O(2)

<table>
<thead>
<tr>
<th>Users binned according to prediction accuracy</th>
<th>01-02</th>
<th>02-03</th>
<th>03-04</th>
<th>05-06</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 ~ 10%</td>
<td>0</td>
<td>0</td>
<td>0.1325</td>
<td>0.0851</td>
</tr>
<tr>
<td>10% ~ 30%</td>
<td>0.1072</td>
<td>0.0998</td>
<td>0.1050</td>
<td>0.1028</td>
</tr>
<tr>
<td>30% ~ 50%</td>
<td>0.0609</td>
<td>0.0653</td>
<td>0.0552</td>
<td>0.0829</td>
</tr>
<tr>
<td>50% ~ 70%</td>
<td>0.0392</td>
<td>0.0337</td>
<td>0.0335</td>
<td>0.0505</td>
</tr>
<tr>
<td>70% ~ 90%</td>
<td>0.0249</td>
<td>0.0204</td>
<td>0.0221</td>
<td>0.0247</td>
</tr>
<tr>
<td>90% ~ 100%</td>
<td>0.0098</td>
<td>0.0084</td>
<td>0.0083</td>
<td>0.0079</td>
</tr>
</tbody>
</table>
Figure 5-3. Cumulative Probability of Markov O(2) prediction accuracy for Dartmouth Quarter data sets for data collected between 2002-2006

Figure 5-4. Cumulative Probability of Markov O(2) prediction accuracy for UF semester data sets for data collected between 2007-2011
Figure 5-5. Comparison of Cumulative Probability of Markov O(2) prediction accuracy between semester long and year long traces at UF for data collected between 2007-2009.

Figure 5-6. Comparison of Cumulative Probability of VoIP vs WLAN Users from Dartmouth trace 2001-2004 and also UF Fall 2011 smart-phones vs. laptops this shows that the gap between the two (ultra mobile and less mobile) have gone down from 33% difference to 7% difference.
CHAPTER 6
FRAMEWORK DESIGN ON MULTI-DIMENSIONAL PREDICTION PROCESSES

In Chapter 6 we shall introduce a new process to improve prediction for future mobile users, including smart-phone and highly mobile users. Looking at the analysis from Chapter 4 and 5 future WLAN traces are likely to have more highly mobile population introduced into them, thus, the need to improve WLAN mobile user prediction will grow in the future.

6.1. Insights and Improvements

With the ubiquity of wireless networks and the introduction of more capable mobile devices (PDAs, smart-phones, tablets, etc.), users are perceived to be distinctly more mobile. These differences in mobility characteristics (different degrees of mobility) significantly affect user location predictability and will likely affect all protocols that utilize WLAN user traces [26]. As seen in Chapter 5, prediction algorithms showed poorer performance over time and as seen in Chapter 4 more mobility perceived in users directly resulted in less predictability. With this extensive study of investigating two data sets spanning 10 years, we can say with confidence that more mobility is expected in the future. This means that we need to continuously monitor and analyze the WLAN trace and mobile users to understand such an intriguing evolving dynamic. The evolving dynamic include but are not limited to the changes in the underlying network, devices and user online behavior. Such insights urge us to open the door for revisiting and improving prediction for modern day network users.

We have already shown the need to seek for better prediction methods for the WLAN user trace. There are numerous avenues we can explore to try to improve user location prediction and are planning to explore a select few of them. There are largely
two ways to improve predictors without actually changing the algorithm; first, we can control the trace that is being fed into the predictor, and second, we can control the prediction process itself. Using the first method, we can expand and explore different temporal dimensions by cutting the input trace so that we predict weekdays only, certain days of the week or certain time of day, etc. or expand the different spatial dimensions such as calculating distance error between the predicted location and actual location, investigate changes in predictability when using different granularity of locations to predict (i.e. AP level vs. Building level) [16][17].

Using the second method we could put different weights on the decision tree branches using data such as duration that the user accessed the location or implement a timer to put more weight on locations that were recently visited, and locations visited longer ago will decay and lose weight. Instead of a hit or a miss as a prediction success or fail, we could try to predict the gray area as well. We could weigh in how much confidence we have in predicting a certain location for a user i.e., user will go to location X with a 78% chance.

Based on these insights we have expanded our study to attempt to improve prediction of WLAN mobile users by expanding the spatiotemporal dimensions such as time of day (temporal) and calculating distance error and different granularities of prediction (spatial) by introducing the novel framework of multi-dimensional prediction processes discussed in the following section.

6.2. Multi-Dimensional Prediction Process Framework

In section 6.2. we introduce a way to better predict future WLAN mobile users, namely the multi-dimensional prediction process framework. This framework was
motivated strongly from the findings and insights we have gained from the extensive analysis done on the 10 year long WLAN traces from collected from Dartmouth and UF.

First we introduce the concept of “the Oracle”. The Oracle is the all-seeing-eye that is one step ahead and already know what the multi-dimensional prediction processes will predict, and thus will always choose the best prediction accuracy given all the choices of the multi-dimensional prediction processes. The Oracle essentially chooses the predictor that will predict correctly given an instance of time, and if none of the prediction processes is a hit, the Oracle will choose the most favorable (i.e. closer distance, same building, etc.). Thus even for the same user the Oracle may switch between different prediction processes to try to provide the best possible prediction accuracy given all the different prediction processes to choose from. The multi-dimensional prediction process is shown in Figure 6-1.

The input is controlled such that the data it is fed is the entire trace, the sectional trace (looking at time blocks of the day) or in different granularity such as the AP level or building level. There may be numerous ways to control the input data as discussed in the section above but we investigate these three in our work. These inputs will go through different prediction algorithms and by looking into the history of each of the user and trying to classify these users in meaningful ways guidelines are provided that will act as a control element which will enable this multi-dimensional prediction process to successfully choose the best predictor for any given user at any given time. As for the output we look at different granularities (AP level and building level) as well as the distance error in the case the predictor we have chosen makes the wrong prediction.
6.2.1. Temporal Expansions

Figures 6-2, 6-4, 6-6 and 6-8 show the prediction accuracy comparison for the Oracle, whole trace and sectional trace for 100 randomly selected users from 4 different traces collected at UF (Fall 2007, Spring 2008, Fall 2011 smart-phones, Fall 2011 laptops)[24]. The WHOLE trace is our base data which is the WLAN trace in its whole form and not expanded in anyway. The sectional trace is a trace that has been expanded using a temporal expansion by looking at different time blocks of the day for each user. In this case we have looked at four 6 hour time blocks from midnight-6am, 6am-12pm, 12pm-6pm and 6pm-midnight. This sectional trace prediction is made by only looking into the same time block for users. The Oracle always picks the better prediction process of these different prediction processes and thus results in an envelope like fashion that show the best prediction among all the different prediction processes. Figures 6-3, 6-5, 6-7 and 6-9 show how much improvement the Oracle made over the whole or sectional traces. There is a definite trend that we can see from these figures. The improvement rate over time (from 2007 to 2011) has risen drastically, where the maximum improvement for 2007-2008 data sets were 12-13% but that of the 2011 data sets increased nearly threefold resulting in a 30-40% improvement for the sectional traces. Following this trend, we can conclude that the Oracle will become increasingly beneficial in the future as it will help improve the dropping prediction accuracy over time.

6.2.2. Spatial Expansions

Here we discuss the spatial aspects of state expansion by showing the results for different granularities of prediction and also of distance error rate.
6.2.2.1. Different granularities of prediction

We explore different granularities of “success” by looking at the AP level prediction as well as the building level prediction. In Figure 6-10 through 6-13 we compare the AP level prediction and the building level prediction accuracy. One can clearly see that by using prediction on the building level it shows significant improvement over using the AP level prediction. Thus, expanding the spatial dimension of our prediction process will gain up to a 20% increase (for 05-06) and up to 10% increase (for 01-02) in location prediction accuracy over the AP level prediction. We can also see that the improvement from the AP level prediction to the building level prediction is significantly bigger with time evolution.

After providing the comparison of different levels of prediction granularity, we shall study how these different definitions affect the time evolution of the user predictability. First, we note that among the predictors we use for each yearly trace, the Markov O(2) is clearly still the best prediction algorithm for the traces in our study. One can easily see in Figures 6-14 through 6-17 that across each yearly trace the Markov O(2) prediction proves to have the most accurate prediction result for the WLAN users. We noted in our experiments that with Markov O(1) prediction, less people may have lower prediction accuracies than those of Markov O(2) but this trend is quickly overturned once the predictor gains enough history. However, in order to have 3 previous steps followed by a certain step is harder to achieve thus, the Markov O(2) has the best tradeoff of the two and performs the best among the three. This validates our previous findings of Markov O(2) predictor being the best among the ones explored in this study time and time again.
In Figures 6-18 and 6-19 we show the time evolution of the yearly traces from 2001 to 2006 for the AP level prediction and the Building level prediction respectively. We can easily spot that the building level prediction has smaller difference of prediction accuracy between each yearly trace compared to the AP level prediction. Another interesting trait is that whereas the AP level prediction shows less predictability as time evolves. This, however, is not necessarily the case for the building level prediction. Although the building level prediction rate for 02-03 and 03-04 are almost identical, we can still see that at one point 03-04 actually has better predictability than 02-03. Also we can observe that 01-02 has less predictability than both 02-03 and 03-04. Interestingly the 05-06 trace consistently shows the least predictability over both levels of prediction.

6.2.2.2. Distance error rate

In 6.2.2.2, instead of predicting a hit or a miss we also look at how far off we missed when we did miss, by looking at the distance error rate. The distance was calculated by mapping the building coordinates to the APs belonging to each building and then calculating the distance between two buildings. For the purpose of differentiating an actual hit (distance 0) and a miss which occurred in the same building (distance 0) we give the miss that occurred in the same building a uniform distance of 10 meters since this is shorter than the shortest distance between two buildings and not too short that it is an insignificant distance. We have looked at 4 different set of traces (UF Fall 2007, Spring 2008, Fall 2011 smart-phones and laptops) and 50 randomly selected users for each of the subset of the UF trace. First we show the empirical CDF graph in Figure 6-20, 6-21, 6-22 and 6-23 with user error distance rate from 0 (a hit) to 500 meters. Figures 6-22 and 6-23 that looks at the laptop and smart-phone subsets for
the UF fall 2011 semester shows a highly dense CDF graph compared to Figures 6-20 and 6-21 which shows that of the fall 2007 and Spring 2008 UF trace.

We zoom in on the Figures 6-21 through 6-23 by looking at those misses which occur outside of the same building. (distance of 15 to 500 meters) As shown in Figures 6-24 and 6-25 for the fall 2007 and Spring 2008 trace show clusters which means for the misses that happen most likely occurred for the same distances and shows these higher and sharper knees, whereas in Figures 6-26 and 6-27 the knees become lower and smaller creating almost a curve-like graph instead of a step-like one. This means whatever the misses may be, they are very evenly distributed and when a miss occurs it is not likely for the same distance. We can assume this comes from the changing dynamics of the WLAN network. The network itself by introducing a larger number of APs may be the cause or the devices being light-weight as they are, are tightly coupled with users [27] while online and allows the trace to capture better mobility which leads to lower predictability and the evenly distributed distance error CDF graph. In Tables 6-1 through 6-3, we have calculated the mean, standard deviation, max and min of a subset of Fall 2007, Spring 2008, the laptop trace and smart-phone trace. The mean and the standard deviation show a steady trend of the distance increasing in all the different conditions (i.e. regardless of including 0 distances or not, etc.). This is an interesting finding since the campus is going through network changes and by adding more APs in a limited area, one would think that the distance of error for individual instances would go down. However as you can see in Tables 6-1 through 6-3, the distance error is continuously growing larger. Even within the same year of users, the Fall 2007 and the laptop trace shows more clusters than that of the Spring 2008 or smart-phone trace
which are less predictable, there is a strong correlation between the distribution of distances and prediction accuracy along with user mobility.

6.2.3. Guidelines

Here we will discuss an initial set of guidelines we have discovered that will aid in improving the prediction accuracy for WLAN mobile users. Note that these guidelines may not directly apply to traces not used in this study. However, the framework itself and method is generic, and is applicable to generate new sets of guidelines as needed for other traces and data sets. Figure 6-28 and 6-29 each shows all the occurrences of the two processes that we use (sectional and whole) when for 100 random users selected from UF Fall 2007 and Spring 2008 respectively, the predicted location is different from each other but one of them is a hit. In other words WHOLE predicts user A at instance t goes to location X next, and SECTIONAL predicts user A at the instance t will go to location Y next. Only in the case the correct next location is among the prediction processes’ predicted location (i.e. X or Y is the correct next location) available will it appear in Figures 6-28 and 6-29. The x axis shows the amount of history accumulated at the time this instance happens and y axis is the AP encounter ratio (i.e. user mobility) at the time. Figures 6-28 and 6-29 definitely show a trend that the more history is accumulated the smaller the AP encounter ratio is and with less history the AP encounter ratio can go up to nearly 1. It also shows that if the history is below 500 and AP encounter ratio is above 0.1 it is better to go with the SECTIONAL prediction process and if AP encounter ratio is below 0.02 it is better to pick WHOLE. With this guideline in place as the control for the framework illustrated in Figure 6-1, feeding into the prediction algorithms we expect to achieve a better prediction accuracy compared to choosing one or the other prediction method. We provide pseudo code for these
prediction guidelines in Figure 6-30. Poracle is our proposed prediction oracle implementation (to be validated and improved in the future work). In the code it is the predicted next AP (or location) by our proposed oracle.
Figure 6-1. Framework of multi-dimensional prediction processes in a nutshell
Figure 6-2. Comparison of the different prediction processes by controlling the input ordered by ascending order of prediction using the Oracle for UF Fall 2007

Figure 6-3. Improvement of Oracle over UF Fall 2007 sorted by the improvement of sectional in ascending order
Figure 6-4. Comparison of the different prediction processes by controlling the input ordered by ascending order of prediction using the Oracle for UF Spring 2008

Figure 6-5. Improvement of Oracle over UF Spring 2008 sorted by the improvement of sectional in ascending order
Figure 6-6. Comparison of the different prediction processes by controlling the input ordered by ascending order of prediction using the Oracle for UF Fall 2011 laptops

Figure 6-7. Improvement of Oracle over UF Fall 2011 laptop devices sorted by the improvement of sectional in ascending order
Figure 6-8. Comparison of the different prediction processes by controlling the input ordered by ascending order of prediction using the Oracle for UF Fall 2011 smart-phones.

Figure 6-9. Improvement of Oracle over UF Fall 2011 smart-phone devices sorted by the improvement of sectional in ascending order.
Figure 6-10 Comparison of Cumulative Probability of AP and building level Markov O(2) prediction for 01-02 users

Figure 6-11. Comparison of Cumulative Probability of AP and building level Markov O(2) prediction for 02-03 users
Figure 6-12. Comparison of Cumulative Probability of AP and building level Markov O(2) prediction for 03-04 users

Figure 6-13. Comparison of Cumulative Probability of AP and building level Markov O(2) prediction for 05-06 users
Figure 6-14. Cumulative Probability of Prediction Accuracy for 01 02 User Trace

Figure 6-15. Cumulative Probability of Prediction Accuracy for 02-03 User Trace
Figure 6-16. Cumulative Probability of Prediction Accuracy for 03-04 User Trace

Figure 6-17. Cumulative Probability of Prediction Accuracy for 05-06 User Trace
Figure 6-18. Cumulative Probability of Markov O(2) AP Level Prediction for each Yearly Trace at Dartmouth College

Figure 6-19. Cumulative Probability of Markov O(2) Building Level Prediction for each Yearly Trace
Figure 6-20. Empirical CDF showing the distance error rate in meters for 50 random users from UF Fall 2007 trace (distance error rate of 0 – 500 meters shown)

Figure 6-21. Empirical CDF showing the distance error rate in meters for 50 random users from UF Spring 2008 trace (distance error rate of 0 – 500 meters shown)
Figure 6-22. Empirical CDF showing the distance error rate in meters (from 0-500) for 50 random users from UF Fall 2011 smart-phone traces

Figure 6-23. Empirical CDF showing the distance error rate in meters (from 0-500) for 50 random users from UF Fall 2011 laptop traces
Figure 6-24. Empirical CDF showing the distance error rate in meters for 50 random users from UF Fall 2007 trace (zoom in and show only distance error rate of 15 – 500 meters)

Figure 6-25. Empirical CDF showing the distance error rate in meters for 50 random users from UF Spring 2008 trace (zoom in and show only distance error rate of 15 – 500 meters)
Figure 6-26. Empirical CDF showing the distance error rate in meters for 50 random users from UF Fall 2011 smart-phone trace (zoom in and show only distance error rate of 15 – 500 meters)

Figure 6-27. Empirical CDF showing the distance error rate in meters for 50 random users from UF Fall 2011 laptop trace (zoom in and show only distance error rate of 15 – 500 meters)
Table 6-1. First order statistics on the distance error rate. Including when distance is 0 (which is a hit, Distance >= 0)

<table>
<thead>
<tr>
<th>All Distance (including hit)</th>
<th>Fall 2007</th>
<th>Spring 2008</th>
<th>Laptop</th>
<th>Smartphone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>7.09</td>
<td>17.15</td>
<td>23.98</td>
<td>40.69</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>47.87</td>
<td>95.78</td>
<td>121.35</td>
<td>177.12</td>
</tr>
<tr>
<td>Max</td>
<td>2694.09</td>
<td>3241.48</td>
<td>3221.35</td>
<td>6346.67</td>
</tr>
<tr>
<td>Min</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 6-2. First order statistics on the distance error rate. Not including when distance is 0 (Distance > 0)

<table>
<thead>
<tr>
<th>All Misses</th>
<th>Fall 2007</th>
<th>Spring 2008</th>
<th>Laptop</th>
<th>Smartphone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>53.84</td>
<td>116.3</td>
<td>108.94</td>
<td>146.37</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>122.02</td>
<td>225.12</td>
<td>240.08</td>
<td>312.07</td>
</tr>
<tr>
<td>Max</td>
<td>2694.09</td>
<td>3241.48</td>
<td>3221.35</td>
<td>6346.67</td>
</tr>
<tr>
<td>Min</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 6-3. First order statistics on the distance error rate. Not including when it is a miss within the same building (Distance > 10)

<table>
<thead>
<tr>
<th>All Misses outside of buildings</th>
<th>Fall 2007</th>
<th>Spring 2008</th>
<th>Laptop</th>
<th>Smartphone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>177.35</td>
<td>203.52</td>
<td>288.63</td>
<td>290.13</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>190.19</td>
<td>274.56</td>
<td>335.06</td>
<td>399.73</td>
</tr>
<tr>
<td>Max</td>
<td>2694.09</td>
<td>3241.48</td>
<td>3221.35</td>
<td>6346.67</td>
</tr>
<tr>
<td>Min</td>
<td>20.68</td>
<td>20.06</td>
<td>18.34</td>
<td>18.34</td>
</tr>
</tbody>
</table>
Figure 6-28. Correlation between the accumulated history and the AP encounter ratio of 100 random users from UF Fall 2007

Figure 6-29. Correlation between the accumulated history and the AP encounter ratio of 100 random users from UF Spring 2008

76
Figure 6-30. Pseudo code for prediction guidelines derived from the results above
CHAPTER 7
CONCLUSION AND FUTURE WORK

The dynamically changing WLAN trace characteristics, with the introduction of lighter more mobile devices with increased capabilities in communication, computation, storage and sensing, changes in the network as well as the user behavior is affecting the protocols using these traces. We show that the change in the traces and mobility indeed affect protocols such as user location prediction by doing an extensive study and analysis on user mobility, predictability and evolution of 10 years worth of real life trace collected from two major campuses. We learned that users are becoming more mobile, and less predictable as time goes by.

Based on these findings we built a multi-dimensional prediction process framework that will allow us to better predict the future WLAN mobile users who are deemed to becoming less and less predictable. By introducing multiple dimensions we provide different prediction processes to this framework that will allow anyone using this framework to customize the prediction process for users in order to achieve the best prediction accuracy possible based on all the information it has.

For our future work we plan to extend our framework and also collaborate with another colleague to predict user’s web visitation using NetFlow traces [33][33]. This will be an exciting area to explore especially not only because there are many challenges in dealing with such a large data (each day’s worth of data can be as large as few hundred gigabytes) but also that this can be used to create better predicted caching such as some applications are already putting to use such as [25][28].
As mentioned numerous times throughout this study, there is a most definite need to revisit predictors and any other protocols that use these user mobility traces, which are using changing dynamically due to the ever evolving nature of these traces.
LIST OF REFERENCES


BIOGRAPHICAL SKETCH

Jeeyoung Kim received her Ph. D. from the Computer and Information Science and Engineering department at the University of Florida in Gainesville, Florida in the summer of 2013. She was a research assistant of the Mobile Wireless Networks Design and Testing Group (NOMADS) laboratory under the advisement of Prof. Ahmed Helmy. She received her Master of Science in Engineering from the Computer and Information Science department at the University of Pennsylvania in Philadelphia, Pennsylvania and her Bachelor of Science and Engineering degree from the Computer Science department at Kyungpook National University in Daegu, Republic of Korea. Her research interests lie in wireless network mobile user online behavior of real-life WLAN traces, analyzing the mobility and predictability of WLAN users.