Characterization of WLAN Location Fingerprinting Systems

Jiwei Li

Master of Science by Research
Institute of Computing Systems Architecture
School of Informatics
University of Edinburgh
2012
Abstract

Wireless LAN (WLAN) location fingerprinting is an important approach for local positioning system, a complementary and alternative technology to the global positioning system (GPS). Fingerprinting is a method of localization that associates individual positions with distinct features and uses them to infer absolute locations of targets. The widespread presence of WLANs (based on IEEE 802.11 standard) makes it natural to consider WLAN signals to construct fingerprinting based localization systems, especially in indoor environments.

Although WLAN location fingerprinting is reasonably well studied and diverse approaches have been proposed, its application in practice is still quite limited owing to concerns on its robustness. This suggests the need for further research to deepen understanding of WLAN location fingerprinting and identify components that could be improved to enhance its overall robustness. In this thesis, we study the impact of different alternatives for WLAN location fingerprints, an issue that is ignored or rarely discussed in prior work. Moreover, we systematically evaluate WLAN fingerprinting systems as a whole considering different localization algorithms as well as different fingerprint types.

Our analysis highlights the impact of various features of WLAN fingerprints including visibilities of access points, virtual access points and properties of WiFi received signal strength, which is used to construct fingerprints. The results show that assumptions concerning WLAN enabled environments made by many schemes do not hold in reality. Various methods for each component of fingerprinting systems are included to find optimized combinations. And by analyzing and comparing performance of six different localization algorithms with their optimized configurations, we find that localization algorithms have limited impact. Popular algorithms have similar behaviors with appropriately chosen fingerprint types, and positioning errors are heavily dependent on the construction of fingerprints and the environment. We also find that results aggregation, a common component in positioning systems for error compensations, is not beneficial for fingerprinting.
Acknowledgements

Many thanks to my supervisor Dr. Mahesh K. Marina for his support of this research and care of my general well-beings and I genuinely appreciate all the helps from my colleagues in the WiMo group.

I would like to thank Valentin Radu, Lito Kriara, Arsham Farshad at Edinburgh and Noelia Hernández from the University of Alcalá for their supports of my works, and Saravana Manickam for his help on the writing of this thesis.

I would also like to express my deepest gratitude for my parents and girlfriend for their continuous encouraging and comforts.
Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Jiwei Li)
# Table of Contents

1 Introduction 1  
   1.1 Contributions .................................................. 2  
   1.2 Structure of the Thesis .......................................... 3  

2 Background 4  
   2.1 Classifications of LPS ........................................... 4  
   2.2 Drawbacks and Complements of GPS ............................... 6  
   2.3 WLAN Geometric Positioning ...................................... 8  
      2.3.1 Multilateration, Trilateration and Radio Propagation Modeling 9  
      2.3.2 WLAN Triangulation ....................................... 12  
   2.4 Inertial Navigation on Mobile Phones ............................ 14  
      2.4.1 Pedestrian Dead Reckoning ................................ 14  
      2.4.2 Initialization and Calibration .............................. 15  
   2.5 Signaturing ....................................................... 16  
      2.5.1 Spatial Characterization .................................. 16  
      2.5.2 Activity Recognition ...................................... 17  
      2.5.3 Augmented Reality ......................................... 18  
   2.6 Fingerprinting .................................................... 19  
      2.6.1 Off-line and On-line Phases ................................. 20  
      2.6.2 Interpolation and Crowdsourcing ............................ 23  

3 Related Work 25  
   3.1 Exploring the Space for Improvements ........................... 25  
   3.2 Analysis of WLAN Location Fingerprints ......................... 27  
   3.3 Comparative Study on WLAN Fingerprinting Systems .......... 28  

4 Key Properties of WLAN Fingerprints 30  
   4.1 Datasets design and structure .................................... 30
List of Figures

2.1 The structure of OFDM frame in IEEE802.11 . . . . . . . . . . . . . 10
2.2 Illustration from works of ancient Chinese mathematician Liu Hui . . 12
2.3 Off-line, on-line phases and fingerprinting system . . . . . . . . . . . 20
3.1 A taxonomy of radio location fingerprinting . . . . . . . . . . . . . . . 29
4.1 The user interface of site-survey application on Android . . . . . . . 33
4.2 Short NVA time series and regressions from $SLDaug13$ and $SPDaug13P$ 37
4.3 Long NVA time series and regressions from $SPDaug13$ and $SPDaug24$ 38
4.4 ECDF of NVA time series from four datasets . . . . . . . . . . . . . . 39
4.5 Summary of AP coverages on five floors . . . . . . . . . . . . . . . . 41
4.6 RSSI time series and distributions from $SLDaug13$ . . . . . . . . . 43
4.7 RSSI time series and distributions from $SPDaug13P$ . . . . . . . . . 44
4.8 RSSI time series and distributions from $SPDaug13$ . . . . . . . . . . 45
4.9 RSSI time series and distributions from $SPDaug24$ . . . . . . . . . 46
4.10 Normality and skewness test results of RSSI series on First Floor . . 47
4.11 Normality and skewness test results of RSSI series on Second Floor . 48
4.12 Normality and skewness test results of RSSI series on Third Floor . . 48
4.13 Normality and skewness test results of RSSI series on Fourth Floor . . 49
4.14 Normality and skewness test results of RSSI series on Fifth Floor . . 49
4.15 Spatial RSSI vectors of pairs of VAP on each floor . . . . . . . . . . . 55
5.1 Summary of number of fingerprints collected at cells . . . . . . . . . 59
5.2 General interface of localization engines for fingerprinting . . . . . . 61
5.3 Cell deployments on the second floor for evaluations . . . . . . . . . 65
5.4 Performances of Gaussian Modeling . . . . . . . . . . . . . . . . . . 67
5.5 Performances of Log-normal Modeling . . . . . . . . . . . . . . . . . 68
5.6 Performances of Nearest Neighbor Euclidean distance . . . . . . . . 70
5.7 Performances of Nearest Neighbor Manhattan distance ............ 71
5.8 Performances of Nearest Neighbor Mahalanobis distance ........... 73
5.9 Performances of Cosine similarity .................................. 74
5.10 Performances of varied localization engines with optimized configur-
ations ................................................................. 76
5.11 Comparison of performances of optimized systems ................. 77
## List of Tables

<table>
<thead>
<tr>
<th>Table Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Features of measuring devices</td>
<td>31</td>
</tr>
<tr>
<td>4.2</td>
<td>802.11 AP modules deployed in informatics forum</td>
<td>31</td>
</tr>
<tr>
<td>4.3</td>
<td>Data structure of tables for stationary measurements</td>
<td>34</td>
</tr>
<tr>
<td>4.4</td>
<td>Basic information of datasets of stationary measurements</td>
<td>34</td>
</tr>
<tr>
<td>4.5</td>
<td>Basic information of sit survey datasets</td>
<td>35</td>
</tr>
<tr>
<td>4.6</td>
<td>Data structure of aggregated site-survey datasets</td>
<td>35</td>
</tr>
<tr>
<td>4.7</td>
<td>Statistical features of NVA time series from four datasets</td>
<td>39</td>
</tr>
<tr>
<td>4.8</td>
<td>Summary of AP with continuous coverages</td>
<td>40</td>
</tr>
<tr>
<td>4.9</td>
<td>Number of appearances of representative AP in four datasets</td>
<td>42</td>
</tr>
<tr>
<td>4.10</td>
<td>Statistical features of RSSI series of LOSap (00:80:48:57:C7:A3)</td>
<td>43</td>
</tr>
<tr>
<td>4.11</td>
<td>Statistical features of RSSI series of NLOSap (00:24:97:83:3B:C0)</td>
<td>44</td>
</tr>
<tr>
<td>4.12</td>
<td>Statistics of normality and skewness test for five floors</td>
<td>50</td>
</tr>
<tr>
<td>4.13</td>
<td>Pair correlations between strength, constancy and stability</td>
<td>51</td>
</tr>
<tr>
<td>4.14</td>
<td>Correlations between Shapiro-Wilk test values and other features</td>
<td>51</td>
</tr>
<tr>
<td>4.15</td>
<td>Basic features of readings from VAP</td>
<td>53</td>
</tr>
<tr>
<td>4.16</td>
<td>Pearson correlation coefficients between each pair of VAP</td>
<td>53</td>
</tr>
<tr>
<td>4.17</td>
<td>Cosine similarities between each pair of VAP</td>
<td>53</td>
</tr>
<tr>
<td>4.18</td>
<td>Basic information of selected pairs of VAP</td>
<td>54</td>
</tr>
<tr>
<td>4.19</td>
<td>Basic information of selected pairs of VAP</td>
<td>54</td>
</tr>
<tr>
<td>5.1</td>
<td>Statistics of performances of Gaussian Modeling</td>
<td>66</td>
</tr>
<tr>
<td>5.2</td>
<td>Statistics of performances of Log-normal Modeling</td>
<td>69</td>
</tr>
<tr>
<td>5.3</td>
<td>Statistics of performances of Nearest Neighbor Euclidean distance</td>
<td>72</td>
</tr>
<tr>
<td>5.4</td>
<td>Statistics of performances of Nearest Neighbor Manhattan distance</td>
<td>72</td>
</tr>
<tr>
<td>5.5</td>
<td>Statistics of performances of Nearest Neighbor Mahalanobis distance</td>
<td>75</td>
</tr>
<tr>
<td>5.6</td>
<td>Statistics of performances of cosine similarity</td>
<td>75</td>
</tr>
<tr>
<td>5.7</td>
<td>Pearson correlation coefficients of errors for varied systems</td>
<td>78</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Localization, or positioning, is an old domain of human knowledge. And approaches for localization have kept evolving. In recent decades, however, we witnessed a proliferation of localization technologies like the global positioning system (GPS), while location based services (LBS) are receiving increasing attention and are applied in various contexts, like health, social networks, work, etc. At the same time, local positioning systems (LPS) are important complement and alternative to GPS. Unlike GPS, they do not provide global coverage but can be more accurate, efficient and adaptive to local environments.

Wireless LAN (WLAN) location fingerprinting is an important approach for LPS, especially in indoor environments where WLAN is commonly deployed. The method uses WLAN signals covering the space of interest to create fingerprints at different known locations during an initial site survey process; the fingerprint database so created is looked up during the operational phase using a newly collected fingerprint to find the closest match as the location estimate. WLAN location fingerprinting has received much attention from both industry and academia in recent years as it leverages existing WiFi infrastructure and is easy to implement. Although reasonably well studied, this method is not widely put in practice given concerns on its limited accuracy and reliability. The unsatisfactory performance of WLAN location fingerprinting not only limits its own applicability but also that of hybrid systems using it in conjunction with other methods. This makes us believe that further research to deepen understanding on WLAN location fingerprinting is needed and motivates the work presented in this thesis.

The goal of this thesis is to conduct a detailed characterization of WLAN location fingerprinting systems. This goal can be viewed from two aspects. First, influenc-
ing features regarding the engineering of WLAN location fingerprint itself need to be better studied. Majority of previous works in this areas mainly focus on the design of localization algorithms while assuming ideal properties of fingerprints. However, the construction of WLAN fingerprint and its realistic features could have significant impact and be the source for unsatisfactory performance. Second, given various techniques proposed in this area, systematic and comparative evaluation of WLAN fingerprinting systems is needed to identify optimized configurations and study relations between location estimations from different systems.

This thesis focuses on the above mentioned aspects. This work is based on post-processing of carefully collected data for WLAN fingerprints and emulation of different fingerprinting systems. The work presented in this thesis spans three phases which are:

- **Measurements**: Both stationary and mobile WLAN sensing measurements using smart-phone and laptop are collected which serve as data for analysis and system evaluations.

- **Analysis**: Analyzing collected measurements to identify key influencing features for the engineering of WLAN fingerprint in realistic situations.

- **Emulations**: Emulating fingerprinting systems with various configurations to evaluate and analyze their relative performance.

### 1.1 Contributions

This thesis presents interesting and novel findings on both targeted aspects.

Firstly, it provides analysis of the engineering of WLAN location fingerprints with focuses on influencing features which have been ignored in previous studies including visibilities of access points (AP) and virtual access points (VAP). Our results show that temporal and spatial properties of AP visibilities are unstable and incoherent, contrary to common assumptions made by existing schemes. And AP visibilities are also affected by hardware heterogeneity.

VAP is a recent development of WLAN technologies increasingly used in WLAN enabled environments. For the first time, we discuss impacts of VAP on performances of fingerprinting, especially location discriminability. The analysis demonstrates significant temporal and spatial similarities of VAP hosted on physical AP platforms. This
thesis also studies various properties of received signal strength indication (RSSI) time series used in constructing fingerprints. We also validate the negative effects of line of sight (LOS) paths [1] with further evidence.

Secondly, this thesis presents systematic evaluation of fingerprinting systems as a whole considering the key underlying components. Different from previous studies, it first decomposes a typical WLAN fingerprinting scheme into individual functionalities and focus on most critical components: fingerprints selection, localization engines and results aggregation. Seven methods for fingerprints selection and six localization engines are considered, including both existing and novel approaches. Different combinations of methods are tested to identify optimized configurations. The performance analysis not only focuses on comparisons in terms of accuracy but also relates the actual location estimates from different systems, leading to a better understanding of errors in fingerprinting systems and their determinants. We find that common localization algorithms have similar behaviors with suitably chosen fingerprints selection methods and results aggregation is not beneficial for fingerprinting. The results prove that localization performance is more dependent on the environment and the engineering of fingerprints.

In summary, a thorough and comprehensive study on WLAN location fingerprinting is conducted afresh with analysis from novel perspectives, making it a unique contribution to this area.

1.2 Structure of the Thesis

The next chapter presents a comprehensive survey on modern positioning technologies with relevant background material not only on fingerprinting but also other approaches for LPS. It first introduces different classifications of diverse LPS schemes based on varied criteria to better differentiate fingerprinting from other methods. Based on this we elaborate major methods of LPS with a focus on fingerprinting, introducing its relative advantages, basic concepts and state of the art. Chapter 3 follows with discussions of previous works specifically related to this research, analysis of fundamental aspects of WLAN fingerprinting and evaluations.

Chapter 4 explains the measurements phase and its methodology as well as presents analysis of collected data regarding the engineering of WLAN fingerprints. Chapter 5 describes our emulations of fingerprinting systems, and presents our results. The conclusions of this research and future work are discussed in Chapter 6.
Chapter 2

Background

Diverse techniques and ideas can be utilized to construct LPS. To further explore in this area while highlighting potential breakthroughs and avoid crossing efforts, it is necessary to examine and summarize previous works and results in a systematic way, and also to find connections to GPS. This chapter presents an overall documentary study on most remarkable existing approaches for localization, introducing classifications, comparisons and relationships among them. Here we provide a well rounded context of the development of LPS, while focusing on fingerprinting, which is still one of the most practical solutions for indoor localization and contains large space to improve.

Section 2.1 presents classifications based on different dimensions for current LPS schemes. Then we examine drawbacks of GPS and approaches that assist it with LPS components in Section 2.2. In Section 2.3, 2.4 and 2.5, we discuss geometric, inertial and signaturing schemes for LPS respectively, in order to provide a comprehensive survey on this area. In section 2.6, we focus on fingerprinting, introducing various aspects and approaches to it.

2.1 Classifications of LPS

Due to their diversity, LPS schemes can be categorized from different perspectives. The classification could be based on various criteria like methodology, dependency on equipments, operating scenarios, targeted users and outputs. Here we categorize indoor localization systems along three dimensions. First, we consider the output of a localization system which is usually generalized as locations or positions, but actually could be divided into different groups. Most common outputs includes:
• **Absolute positions**: Positions with coordinates information which could be directly mapped to physical space. The absolute location could either be represented in discrete space as cells or grids, or in continuous space with coordinates.

• **Relative positions**: The geometrical relative location by angle or distance to certain reference point (RP), in other word, the absolute position of the user could be obtained only if absolute positions of RP are known.

• **Logical positions**: Identifying user’s position without information directly corresponding to physical space but associated with certain patterns including motions, activities and landmarks. Physical location could then be given by retrieving associate spatial features of those patterns.

The output of a localization system is closely related to the methodology or algorithms the system uses to localize. The calculation of positions is the core engine of localization. We examine various existing methods and sort them into following categories:

• **Geometric**: System based on geometric calculation including triangulation, trilateration and multilateration, to give user’s relative position to certain RP. When the absolute position of RP are known, the absolute position of user could be given.

• **Inertial navigation system (INS)**: A common navigation method widely used on vehicles like ships and aircrafts which continuously calculate the position and orientation of objects by dead reckoning (DR) without the need of external RP.

• **Signaturing**: System utilizing user identifications or spatial signatures corresponding to environmental characteristics or artificial landmarks to localize, which could be based on various technologies including RFID, motion detections, optoacoustic and humiture sensors.

• **Fingerprinting**: Constructing a fingerprint map which would be queried explicitly for localization. Often referred as but different from signaturing, fingerprinting provides more fine grained mapping of fingerprints to absolute positions instead of logical positions.

The vast development of mobile computing, wireless and sensing technologies provides various platforms and tools for localization systems for general or specific scenarios, and fertilize the diversity of LPS. It is necessary to examine technologies and
platforms that localization systems are based on. Commonly used platforms for LPS includes:

- **Sensors**: The most diverse and abundant group of technologies found in localization schemes. The proliferation of sensing technologies give opportunities to develop localization schemes through various methods on scenarios from consumer devices to sensor networks.

- **Cellular network**: Using cellular networking signal such like GSM as location indicators for both indoors and outdoors.

- **FM**: FM broadcasting signals are less susceptible to interferences like human presence, multi-path and fading because of its low frequency. Thus it is suitable to be utilized in localization.

- **WLAN**: The proliferation of WLAN infrastructure and devices especially 802.11, give birth to many WLAN based localization schemes with different methodologies which mostly are signaturing and fingerprinting.

The classification could be based on another criteria. The requirement and expectation of localization accuracy and reliability is different in scenarios which acquire two types of equipments:

- **Specific equipments**: Devices designed specifically for localization which usually provide higher accuracy and are adaptive to difficult conditions.

- **Off the shelf**: Utilizing existing infrastructure and devices such like wireless networks, consumer devices with ease to deploy and promote. The localization would be comparatively software defined and low cost.

With those categorizing dimensions, we could analyze current LPS schemes more systematically and explore the space of improvements from a more comprehensive perspective.

### 2.2 Drawbacks and Complements of GPS

Although already widely accepted and deployed, GPS has its inherent drawbacks and limits which make GPS unsuitable in many scenarios. Those limitations include:
• **Economics**: The price and energy cost of GPS components is usually much higher than its alternatives. Though some GPS systems are inexpensive, cheap solutions may have significantly reduced quality. And it is usually very energy consuming. For instance, GPS on smart phones consumes much large portion of energy than other integrated sensing and communicating components [2]. Even with the high cost, GPS on consumer devices cannot provide desired performance.

• **Accuracy**: The accuracy of GPS is usually not enough when small scale of space and fine grained location discrimination are considered, especially on nonprofessional devices with poor quality. And due to its inherent large scale, GPS is difficult to maintain and update, while in many scenarios, the space of interest for localizing is highly dynamic.

• **Speed and reliability**: Mobile phone GPS users usually experience slow updating of their locations. And since GPS requires LOS paths to satellites and stable signal coverage, GPS is not practical or available in many cases.

Thus, techniques assisting GPS with LPS components have received much attention from both industry and academia in recent years.

An approach widely used on mobile phone navigations is cellular networking based localization. The global deployment and general availability of cellular networks make it an intuitive option for localization and assisting GPS. As other RF technologies, various methodologies could work on cellular signals including fingerprinting [3]. But due to the broad range of coverage in large scale of space, to collect and maintain sufficient fingerprinting data for a cellular network based localization is extremely difficult.

Thus most practical solutions for cellular networks are geometry based, same as GPS, which do not require prior knowledge of the space and explicit data collections. Multilatamation, also known as hyperbolic navigation, is a very common and traditional method first introduced by the Gee system [4] in the second world war. And is widely used for cellular network based localization [5, 6] for its ease to construct without the need of clock synchronization.

Another common approach is WiFi signaturing which is included in commercial applications like Google location based services (LBS) and facebook check in service. The system would acquire databases of WiFi service set identifications with their deployed locations which are used as signatures to position user at a building level accuracy.
Apart from commercial schemes, other GPS assisting systems have been proposed leveraging more sophisticated techniques. The AAMPL system [7] argues that the accuracy of GPS component on consumer devices is usually not sufficient enough for contextual information retrieving from physical locations. For instance, several meters of errors may lead to completely different results in contextual space such like neighboring shops and buildings. Since modern LBS such like location based advertising is highly dependent on spatial contexts, those errors could severely affect the performance and reliability of LBS. Instead of increasing the positioning accuracy in meters, AAMPL leverage integrated sensors on mobile phones and activity recognition [8] to create contextual signatures, through which it could complement GPS in more complicated spaces like urban areas.

GAC [9] also proves that energy consumption of GPS module on mobile phones is much larger than other internal components through their measurements on HTC Dream, and propose a hybrid localization system with a focus on energy efficiency. The GAC system mainly uses low energy consuming sensors like accelerometer and compass to perform inertial navigation and periodically querying the GPS for more accurate positioning to reduce the cumulative errors of INS, which is a common method for calibrating INS and is elaborated further in Section 2.4.

### 2.3 WLAN Geometric Positioning

Geometry based positioning techniques is perhaps the most traditional and widespread methodology for positioning, used from ancient navigation to GPS. Although the methodology is old, due to progress in technologies, opportunities have emerged to apply geometric methods on new platforms like WLAN. Thus geometric methods are still a popular approach in the development and research of LPS, especially WLAN based systems.

Since it is very different from to fingerprinting, geometric positioning is often considered as a major alternative to fingerprinting techniques. A main difference between geometric positioning and fingerprinting is that it generates relative positions while fingerprinting reports absolute positions.

The rest of this section introduces most typical geometric positioning methods and their latest implementations on IEEE802.11 platform.
2.3.1 Multilateration, Trilateration and Radio Propagation Modeling

As mentioned before, multilateration, or hyperbolic navigation, is a popular approach for radio navigation systems and widely seen in cellular network based LPS. The calculation of multilateration is based on the difference in distance of the user to two or more RP. When the difference in distance and positions of RP is known, we can form a hyperbolic curve in 2D space or a hyperboloid in 3D. When additional pairs of RP and the difference in distance between them are provided, more hyperbolic curves or hyperboloids could be generated, the position could be detected by finding the intersection between them.

The process of determining locations by distances based measurements and geometry could be referred by a more general term of trilateration. Distance relation of user to RP could form various geometric objects including circles, spheres and triangles. A position could be narrowed down to by finding intersections of multiple objects. Trilateration is widely used in practical applications including GPS.

2.3.1.1 Timing based Approaches

Although the methodology is straightforward, implementing it on radio platforms is non-trivial. The question could be narrowed down to distances measurements through radio signaling. In this context, we consider multilateration and trilateration separately since in radio navigation systems, they depend on different metrics respectively:

- **Time of Arrival (TOA)**: Computing distance through radio transit time and wave speed.
- **Time Difference of Arrival (TDOA)**: Foundations of multilateration systems, computing difference in distances to multiple RP by measuring difference in transit time.

TOA/TDOA metrics are both based on time reference of received signal to be transformed to distances, and must reply on clock synchronization to get accurate timing. Traditionally, timing is performed by specific clock synchronization component integrated in radio systems like cellular network. However, modern wireless technology enable TOA/TDOA based localization based on off the shelf WLAN infrastructure. In [10], the authors propose a timing synchronization method based on OFDM framework of IEEE802.11, shewed in Figure 2.1. By computing auto-correlations of
received OFDM packets and finding the peak of correlations by both short training symbols (STS) and long training symbols (LTS), packets could be synchronized.

In [11, 12], WLAN positioning systems based on OFDM packets synchronization are proposed. With timing reference, multilateration with TDOA, or in some contexts, differential time difference of arrival (DTDOA) is used to determine user’s location.

Although it is possible to implement timing based positioning on WLAN, the practicality of the system is very limited because of its inherent drawbacks. Just as GPS, timing based distance measurements require LOS paths from user to transmitter which are usually not obtained in environments where WLAN is deployed like indoor spaces. And the need for low layer packets information usually cannot be met by WLAN devices like smart phones.

### 2.3.1.2 Radio Propagation Modeling

An alternative methodology for distance measuring with radio signaling could be used to construct WLAN positioning systems which is based on a RSSI. By modeling the radio propagation, the relation between transmitting distance and RSSI could be established. Due to the complexity of WLAN enabled environments like indoor spaces and urban areas which usually contain large amount of objects and users that cause interferences and multi-paths, Accurate modeling is very challenging. However, by designing models adaptive to environmental variations, WLAN positioning system based on propagation modeling is made possible. The basic model of radio propagation is the log distance path loss (LDPL) model described in Eqn 2.1, where \( P_d \) is the RSSI seen at the distance \( d \). \( P_0 \) is the RSSI seen at a reference distance \( d_0 \) which is known, usually set to be 1 meter. \( r \) is the pass loss exponent indicating how steep the RSSI would fall with distance. And \( X_g \) is a variable representing environmental variations.

\[
P_d = P_0 - 10r \log \frac{d}{d_0} + X_g
\]  

(2.1)
One remarkable early effort of WLAN positioning is RADAR [13]. Although it is mostly known for its fingerprinting component which is mentioned in later Sections, it also propose a comparative propagation modeling approach utilizing a modified model derived from LDPL which focusing on handling indoor wall attenuation described in Eqn 2.2.

\[ P_d = P_0 - 10r \log \frac{d}{d_0} + C \cdot WAF \]  

(2.2)

The \( X_g \) in LDPL is modified to \( C \cdot WAF \) where \( WAF \) is the \textit{wall attenuation factor} and \( C \) is the estimated number of obstructions (wall) between user and transmitter. Both two parameters are learned empirically with sufficient training samples.

In EZ [14], a similar approach is proposed. The significance of EZ system is the locations of transmitters (AP), the path loss exponent and attenuation parameter as well as user’s location are supposed to be unknown and would be retrieved from RSSI data. Without heavily modifying the LDPL model itself, EZ focus on the computational and practical challenges to implement LDPL on WLAN platform and systematically deploys auxiliary components to assist the positioning system, which mainly include following parts:

- \textit{Genetic algorithm (GA) and constrained random solution generation:} Since the large number of unknown quantities, it is difficult to find solutions of set of LDPL equations with traditional methods. In EZ, GA is used to solve equations and the initial solutions space of GA is generated randomly but constrained by trilateration.

- \textit{Relative gain estimation:} Heterogeneous devices may report different RSSI for identical signals. To overcome device heterogeneity, EZ add a device variation parameter \( G^k \) identifying the receiver gain for each user.

- \textit{AP selecting:} To reduce the computational burden, an AP selecting algorithm clusters AP with similar signaling and a representative AP is selected from each cluster. AP selection is important for improving efficiency of WLAN positioning system and is explored in more details in later Sections.

With the help of assisting components, adaptivity to environmental and infrastructure heterogeneity and overall performance of the system could be largely improved. Yet there are difficulties for practical radio propagation modeling approach that still need to be addressed including:
Chapter 2. Background

12

- **Modeling completeness**: WLAN signal propagation is affected by many factors including crowds, temporal variations, obstructions, and WLAN coverage is usually dynamic. Thus a complete and well rounded model for WLAN propagation is still an open question.

- **Computational complexity**: In WLAN positioning systems, usually large amount of AP are considered each of which is a signal transmitter. And to overcome affected factors like multi-path and channel interferences, auxiliary computations should be included. Thus the overall computational burden could severely affect the efficiency of the system.

2.3.2 WLAN Triangulation

The other metric for geometric positioning systems is angle, instead of distances. The use of angles to RP to determine locations is referred as triangulation, a traditional technique still widely used today. An ancient example of triangulation is shewed in Figure 2.2.

![Figure 2.2: Illustration from works of ancient Chinese mathematician Liu Hui](image)

Triangulation has many practical applications on radio platforms. However, most of radio triangulation systems require specific equipments which is able to detect the angle of arriving radio signal. The emerging topic of applying triangulation on WLAN infrastructures has received attentions recently.
2.3.2.1 Directional Analysis

The directional analysis technique of radio signal is to determine radio angle of arrival (AOA) by analyzing directional impacts on it from objects, especially human bodies. Thus the relative orientation of the user to the transmitter (AP) would have considerable impact on received signal. The most convincing evidence is that human body contains a high proportion of water which makes it a significant absorber of WLAN signals.

In [15, 16], the user would rotate himself and record the received RSSI from a LOS AP with different orientations. And then by modeling the correlation between the relative angle of the user to the AP and the RSSI, further RSSI samples could be associated with directions and an intuitive policy of direction detecting is picking up the opposite direction of the lowest RSSI, which is experimental validated and utilized in [16].

However, this policy is contradicted by experiment results shewed in [15], which indicates the rotating RSSI samples from different users and devices show an identical pattern. And a more adaptive model can be described in 2.3, where \( p \) is the distance between human body and the phone and \( b \) is the width of the body.

\[
\beta \approx 180^\circ - 2 \arctan \frac{2p}{b} \quad (2.3)
\]

The directional analysis and modeling techniques of WLAN signal AOA make it possible to deploy triangulation system on basic WLAN infrastructures and devices. However, the directional modeling itself is problematic and contradicting experiments results have been presented. And the correlation between orientations and RSSI samples is largely location related and LOS path to the AP is often essential to accurate modeling, which limits the practicality of such approaches.

2.3.2.2 Smart Antennas

The development of WLAN signaling technologies makes it possible to construct WLAN triangulation in a hardware defined way. Smart antennas, also known as multiple antennas or multiple-input multiple-output (MIMO), is already widely enabled on WLAN devices. And smart antennas based indoor localization systems have been proposed. In recent years, WLAN AP manufacturers use smart antennas to bolster the capacity of performance of WLAN. At the same time, smart antennas technology provide platform to deploy AOA positioning on WLAN.
In [17], beam-forming on smart antennas is leveraged to detecting AOA. The system is implemented on PC client equipped with smart antennas and able for beam-forming. The beam-forming weight $w$ is pre assigned for each angle and restored on the client. By 3 clients receiving signals at different locations, the targeted transmitter is localized.

The ArrayTrack [18] is implemented on Rice WARP FPGA based wireless platforms. And differing from [17], the positioning is conducted on transmitters instead of clients. Training packets is first send to build a model estimating signal strength with a function of angles which then used to detecting AOA at running time.

### 2.4 Inertial Navigation on Mobile Phones

In recent years, the world witnessed the rapid development and proliferation of mobile computing and smart phones which largely influence the social culture and technology orientations. The significance of modern mobile phones is its integrated sensing technologies, making it a capable mobile sensing platform. Most commonly integrated sensors include accelerometer and compass. While enabling various interesting applications, those sensors can also be used to develop inertial navigation on mobile platform. Inertial navigation is an important approach for mobile phone based positioning system and is often combined with fingerprinting to construct hybrid systems.

#### 2.4.1 Pedestrian Dead Reckoning

The foundation of inertial navigation is DR, which is the recursive process of determining an user’s current location by measuring his movement and referencing his previously determined location. DR is widely used in air planes, ships and other automotive systems.

Pedestrian dead reckoning (PDR) is a conception derived from DR with a focus on tracking pedestrians, instead of vehicles in traditional scenarios, which raises particular challenges distinct from other DR systems. Differing from automotive systems, where object’s movement can easily be measured through inherent components or additional devices like autometers, the movement tracking on pedestrians itself, same as on animals, is less straightforward and the design of efficient movement tracking techniques is non-trivial. For pedestrian movement tracking, major methods used in applications include following options.
• **Speed estimation**: Estimating the speed of the user directly in a time interval

• **Step counting**: Designed specifically for PDR, detecting user’s speed indirectly by counting walking steps and obtaining a typical step length of the user.

PDR is usually implemented by attaching a measuring device on user’s body. Most existing schemes using specific equipments attached on different parts of the user’s body including shoes [19], belt [20] and helmet [21], which are all using step counting technique. The accuracy of step counting usually outperform that of speed estimation and with specific devices attached to user’s body, step counting is straightforward to implement. However, step counting mechanism is largely dependent on the position of the attachment of the device, making it less flexible and unsuitable for general consumer devices, which user may hold in many different ways. For mobile phone PDR application, both methods are used in different schemes. Yet the accuracy of PDR on mobile phones is not comparable to those based on specific devices.

### 2.4.2 Initialization and Calibration

As a recursive process, inertial navigation need an initial position to start tracking. And since the errors would cumulate through iterations, inertial navigation in practice would fix its determined position, referred to as calibration, periodically or opportunistically. And both initialization and calibration can be narrowed down to a single task of getting position fed from outside the inertial navigation system.

For mobile phone based PDR system, since ideally the localization should be done automatically without the user indicating his own position to assist it, many technologies have been combined with PDR to construct the whole system. In CompAcc [22], GPS is used to initialize the system and a digital map making available paths is used to adjusting positions determined by PDR. The process of recording geographical constraints of the place of interest and use it to fixing positioning is often referred as map matching (MM) and is widely used in hybrid positioning systems as a assisting component.

In Escort [23] and FTrack [24], user encounters detected by blue-tooth or audio signals. The PDR is used to detect user’s movement in the segment between encounters which would reposition the location. Activity recognition and augmented reality techniques are also used respectively to assist those systems.

Mobile phone based PDR is usually designed for indoor localizations which require higher accuracy and more frequent positioning queries. The most popular and intuitive
approach for this purpose is complementing PDR with WLAN positioning, usually fingerprinting. Various schemes have been proposed based on this intuition [25, 26, 27, 28, 29].

Although phone based PDR has limited accuracy and reliability, since integrated sensors on mobile phones are usually low quality and step counting or speed estimation is difficult to implement on consumer devices, it is widely used and proposed as a complementary component in hybrid systems because of the availability and relatively low energy consumption of accelerometer and compass sensors on mobile phones.

2.5 Signaturing

The concept of spatial signaturing is simple and often confused with fingerprinting. Actually signaturing method share many similarities with and closely related to fingerprinting which commonly causes confusions. Thus it is helpful to examine the nature of signaturing method and distinguish it from fingerprinting. We could find following distinct features of signaturing compared to other methods, especially fingerprinting:

- **Logical position**: Signaturing scheme is not able to give spatial information directly. Instead, a signaturing system could only capture the logical relation between user and signatures which is usually boolean. By enquiring location data of those signatures, the system could estimate user’s location indirectly.

- **Context retrieving**: Signaturing is inherently more suitable for contextual positioning since it capture additional information other than locations.

- **Low accuracy**: Since a signature main cover a range of positions, signaturing system usually contain low level of accuracy compared to other methods reporting absolute or relative positions.

Signaturing is widely used in LPS since it leverages spatial characterization and contexts which are aspects where LPS holds advantages. And location signatures can be constructed by various techniques.

2.5.1 Spatial Characterization

As mentioned previously, by capturing environment characteristics or creating artificial checking points, spatial signatures could be made. A common technology for constructing spatial signatures is wireless sensor network (WSN).
A remarkable early effort is the active badge system [30], which designs a tag for every individual users in the form of "active badge" that actively emits unique codes identifying the user at a fixed frequency. The signal will be picked up by a WSN deployed around the place of interests. In this typical signaturing system, the nodes of the WSN of which the positions and coverages are known act as spatial signatures and Pulse-width modulated infrared (IR) is used to signaling between badges and sensor. The master station connected to the WSN checks whether a badges is within the "sight" of a sensing node, thus detecting the logical position of the badge. Apart from IR, many other sensing technologies could be used to construct a signaturing system including RFID, humiture sensors, optoacoustic sensors etc.

However, the cost of spatial characterization including collecting environmental data and deploying sensor networks is usually high. And since its dependency on the environment, it is heavily affected by environment heterogeneity. In many cases, users would prefer systems that are more flexible and adaptive on different spaces and could be implemented on existing infrastructures.

2.5.2 Activity Recognition

The vast development of mobile sensing platforms like latest smart phones makes it possible to use motion detections on off the shelf consumer devices. When carrying a mobile device, the user could create distinct patterns of sensor readings according to different activities which could be mapped based on its contextual information. As it is mentioned previously, activity recognition techniques could be used to create contextual signatures. Accelerometer is mostly used to sense the patterns.

In AAAML [7], the activity recognition is conducted by two steps. At first, the raw accelerometer samples would be classified using Bayesian classifier to extract basic motion information such like standing and sitting. Based on those basic motions, the first stage classification summarize the user’s activity with 3 features:

- **Percentage of standing samples**
- **Average variations of coordinates while standing**
- **Total number of samples**

The output is represented as 3 dimension vectors which would be further classified using nearest neighbor (NN) method. In localization, samples are usually represented
as vectors and many common localization algorithms are vector based like NN. More
details are provided in later sections.

In [31], more detailed and well rounded study on activity recognition with cell
phones is presented. The raw accelerometer readings are also processed to a more
abundant vector containing 6 features. And in the further classification, those features
are classified into 6 particular activities are considered. In this work, individual dif-
ferences are also considered, 29 different users have been invited to the experiments
which show that identical patterns of different individuals could be recognized.

It should be noted that all the above classification methods are based on a super-
vised way which means that the targeted activities to be classified are already listed
before the classification.

In [32], a hybrid signaturing system on mobile phones is presented in which ac-
celerometer is also included and act as a core feature for constructing signatures.
However, the classification is conducted in an unsupervised way where no training
set samples are provided. The system use k-means method to cluster samples carrying
similar patterns.

The use of activity recognition for signaturing have considerable benefits. How-
ever, there are still many challenges remaining and drawbacks that are difficult to
overcome. For activity recognition, there are certain difficulties to put it into prac-
tice including:

- **Collecting training samples**: Activity signatures are much more complicated
  than other forms and it require large amount of sensor readings. Thus training
  samples for activity recognition are difficult to collect and maintain.

- **Mapping signatures**: Activity signatures are difficult to map into physical spaces
  since an activity like walking or ascending involves a sequence of motions and
  locations. For unsupervised recognition, since the ground-truth is not obtained
  through collecting training samples, to map signatures become even harder.

### 2.5.3 Augmented Reality

Perhaps the most intuitive signature of a location is its appearance and the most popular
way to detect a location is to visually recognize it. Although years ago, to detect
the location just by taking a picture is like an intriguing vision of the future, modern
imaging sensors and computer vision technologies have revealed the possibility. And
visual recognition of locations have already explored by many. The visual information
retrieved from physical environment through computing devices is often referred as augmented reality. And a core feature of augmented reality is associate the location information with image objects.

In Object Positioning System (OPS) [33], smart phone cameras, GPS and inertial sensors together with modern computer vision technology are used to localize objects shewed in pictures as well as the user outdoors. The user is required to take several pictures of the object from different angles in order to provide enough information to be processed by the structure from motion (SFM) component, which extracts the 3D structure of the object, mostly buildings in this context, through diverse views of it. It could also indicate the relative positions of the object to the user who’s position is reported by GPS and inertial sensors including compass and accelerometer. Thus through triangulation and trilateration the position of the object can be obtained. Contrarily, if the object’s location is known to the system and used as RP, the user’s location could be detected without using GPS.

Although augmented reality based approaches could be used to develop interesting applications, they have limited practicality. The granularity of such a localization system is very large which make it unsuitable for small scale and fine grained situations like indoor localization. The accuracy of recognition is not sufficient enough and could cause severe errors. The time-energy consumption of computer vision process running on the phone is non-trivial. And the need for diverse images of the object taken from reliable locations cannot always be met.

2.6 Fingerprinting

For radio based small scale LPS, such like indoor localization systems, fingerprinting is one of the most suitable approach, since the wide coverage of radio signal including FM, WLAN or GSM in indoor spaces make it an ideal platform for location fingerprinting. Here we define fingerprinting as the process of associate absolute locations in the place of interests with distinct marks. In a fingerprinting system, the locations would be represented discretely as cells, or grids. By estimating multiple likely cells and finding centroid of them, an estimation in continuous space could also be given.
2.6.1 Off-line and On-line Phases

The basic structure of fingerprinting system is simple. A fingerprint map should first be constructed where each cell is associated with distinct fingerprints, which will be queried later with new sensing samples related to fingerprints. The procedure could be summarized as two phases.

- **Off-line training phase**: The division and planning of cells on the place of interests, and collection of raw sensing for each cells used to form spatial fingerprints and construct a fingerprinted map, or fingerprints datasets.

- **On-line testing phase**: The runtime localization phase, where new sensing samples are collected and used to query the fingerprint map to get location estimation.

Thus the working procedure of WLAN fingerprinting can be summarized as three steps.

- **Construction of fingerprint map/training dataset**: The storage and management of spatial fingerprints data.

- **Run time sensing**: The collection of run time samples indicating user’s location.

- **Fingerprinting system**: The system handling both off-line and on-line data and estimate the user’s location.

![Figure 2.3: Off-line, on-line phases and fingerprinting system](image)
Each step can be further decomposed to individual components. The fingerprinting system is the main process of positioning and could be decomposed to procedures and modules including fingerprint selection, localization algorithm, results aggregation etc. Fig 2.3 presents a diagram of the procedure. The detailed systematic evaluation of those various functionalities is presented in Chapter 5.

As mentioned before, for indoor localization, fingerprinting systems are commonly based on radio platforms including WLAN, FM or GSM. Most of existing approaches in this area only focus on the positioning algorithm, while in this thesis, we study the totality of WLAN location fingerprinting schemes, evaluating the engineering of WLAN fingerprints itself, the construction and maintenance of fingerprinted maps as well as the influential features for designing a fingerprinting system. Thus a more well rounded and comprehensive research on fingerprinting is made.

A remarkable early effort on fingerprinting is RADAR [13], as introduced before, which also includes trilateration and radio propagation modeling aspect. However, the focus of RADAR is fingerprinting and it suggests WLAN fingerprinting outperform radio trilateration for indoor localization. And RADAR is one of the earliest schemes that propose the conception of off-line phase and on-line phase. The datasets are constructed based on broadcasting UDP packets, recording timestamps, RSSI as well as sampling locations. The system use NN to determine positions, which is one of most widely used positioning algorithm in fingerprinting.

Another noticeable NN based fingerprinting scheme is presented in [34], which puts efforts on combining FM and WLAN signals to construct an uniform fingerprints. The contribution mainly lays on the engineering of hybrid radio location fingerprints capturing both FM and WLAN signals, and including signal noise ratio (SNR), multi-path indicator and frequency offset instead of only RSSI. Similar to it, our study also discusses specifically on the properties of fingerprints but with a focus on WLAN.

PinLoc [35] also explore the possibility to leverage alternative radio features other than RSSI to construct fingerprints and proposes the use of channel frequency response (CFR) in WLAN OFDM communication. The CFR in OFDM communication is a function of frequency for modeling transmitted symbols described as $H(f)$ in Eqn 2.4, where $X(f)$ and $Y(f)$ represent transmitted symbol and modulated symbol respectively. On each location, every transmitter (AP) create a vector of CFR corresponding to different carrier frequencies which forms a fingerprint. NN algorithm is also used to determine locations.

\[
Y(f) = H(f)X(f)
\]  

(2.4)
Apart from WLAN and FM, GSM based fingerprinting systems have also been proposed in [36, 37] with similar methodologies. On the other dimension, the design of localization algorithms has kept moving forward. The Horus system [38] is widely considered as one of the most practical WLAN fingerprinting system and often used as a benchmark for comparative studies. Instead of NN, Horus develops a system based on probabilistic modeling with an assumption that a temporal distribution of RSSI at a fixed location from a certain transmitter should be Gaussian, or normal distribution notated as $\mathcal{N}(\mu, \sigma^2)$, where $\mu$ stands for the mean and $\sigma$ stands for the standard deviation. Now considering a run time WLAN sensing sample described in Eqn 2.5. The vector contains RSSI readings from $k$ AP.

$$S = (s_1, \ldots, s_k)$$ (2.5)

The localization problem is to find a location $x$ in the fingerprinted map maximizing the probability $P(x/S)$. According to Bayes theorem, we have an equivalence as Eqn 2.6. And the probability $P(S/x)$ could be computed by aggregating the probability based on each of the $k$ AP RSSI distribution, as described in Eqn 2.7.

$$\arg\max_x [P(x/S)] = \arg\max_x [P(S/x)]$$ (2.6)

$$P(S/x) = \prod_{i=1}^{k} P(s_i/x)$$ (2.7)

Thus, the most likely location for that sample could be determined. Although the normality of RSSI distribution itself is questionable and would be discussed later, Horus proposes an important methodology of using probabilistic analysis to determine locations which is leveraged in many other approaches. Apart from the core positioning methods, Horus also proposes assisting components that could improve the overall performance of localization including locations clustering and multiple estimations averaging which are explored in more details in Chapter 5.

Apart from Horus, other probabilistic based schemes have also been developed. In [39], a Bayesian graphical modeling approach is proposed. And in [40], Monte Carlo method is leveraged for determining the most likely position. All those probabilistic approaches share many similarities with Horus including the dependency on the assumption of normal distribution of RSSI.

Another widely used technique to determine location empirically is machine learning. Without the prior assumption of normal distribution of RSSI samples, machine learning methods estimate location in a heuristic way. The widely used machine learning technique, support vector machine (SVM) is leveraged in [41]. There are many
other machine learning algorithms used for fingerprinting including FURIA [42] and hidden Markov model (HMM) [43].

The recently emerging compressed sensing theory [44] in statistics has received attentions from various areas given its potential in signal processing and data retrieval. And localization approaches leveraging compressed sensing have been proposed in [45, 46, 47]. The algorithm is relatively complicated but here we give a brief introduction of it. To utilize compressed sensing, the data should satisfy two requirements:

- **Sparsity:** The signal of the object should be sparse. In fingerprinting, a signal of an user is his location derived a set of possible locations \( X = (x_1, \ldots, x_k) \). Since normally only one location should be given as an estimation, the sparsity is satisfied.

- **Incoherence:** The two basis representing the object should be as incoherent as possible. In fingerprinting, the grids \( X \) and vector of RSSI readings \( S \) are the two basis representing the user which are usually incoherent.

Since in the scenario of fingerprinting localization system, those two requirements are satisfied, it is intuitive to apply compressed sensing on fingerprinting, which could reduce efforts of fingerprints collection and computation costs. Given the constructed matrix and a new sample, the signal (location) of the user could be reconstructed through the simple \( L_1 \) minimization.

However, there is another assumption which compressed sensing based fingerprinting schemes are based on that all grids of interests should share a same set of visible AP. In other words, a constant set of AP should cover the whole place of interests. As shewed in Chapter 4, this assumption does not commonly hold in WLAN environments. In [45, 47], this drawback is overcame by first clustering the space of interest into subspaces which share same AP coverages. However, according our experiments, the WLAN coverage could be highly dynamic and clustering based on AP visibilities is often not reliable. For its limited practicality, compressed sensing based schemes are not further discussed in this thesis.

### 2.6.2 Interpolation and Crowdsourcing

As explained previously, challenges for designing robust and efficient fingerprinting scheme do not only lay on the system itself but also other features, especially the construction of fingerprints datasets, which is very time consuming and requires dedicated
efforts. Traditionally training data collection is supposed to be done by robots in early schemes. In recent years, approaches focusing on eliminating the training phase or reducing dedicated efforts have been proposed and cause broad attentions.

An intuitive approach to reduce the training burden is fingerprint interpolation. Various methods for interpolation have been applied from the simple linear interpolation to more sophisticated methods like inverse distance weighting and kriging [48]. The hypothesis of such approaches is the spatial correlation of fingerprints.

Another method is reducing dedicated data collections by involving user collaboration. In Redpin [49], the user collaboration is conducted in both passive and active way. In the passive mode, when user walking around the place of interests, the device would periodically query GPS component to get locations feed associated with WLAN fingerprints. In the active mode however, when GPS is not available or reliable enough, the user will be prompted to identify his current location.

The same idea of user involvement in constructing fingerprints data is often referred as crowdsourcing in other contexts. The advantage of crowdsourcing is not only reducing initial training effort but also updating the dataset itself since users are involved in data collection at a running basis. An early attempt is presented in [50], where a year long experiment is conducted in a campus environment, involving 200 users with various devices, performing over 1,000,000 localizations.

One major challenge for crowdsourcing is to reduce user prompting and disturbance. In Zee [51], inertial navigations on mobile phones are used to feed location fingerprints to the system. And by using backward deduction techniques like particle filter, initial location for DR is not needed. In [52], the user feedback is simplified to positive or negative, indicating the accuracy of reported location.

In other articles, the term of organic localization is used [26, 53, 54] but with a focus on the design of efficient user prompting algorithms. While works in this area are promising and progressing rapidly, to further explore the space for improvements, we need to first examine and analyze more fundamental features regarding the engineering of location fingerprints and fingerprinting systems.
Chapter 3

Related Work

Through comprehensive examination of current schemes for LPS, we could get a deeper understanding of this area. As introduced in Chapter 2, approaches to LPS are highly diverse, designed for various scenarios. In this thesis, we focus on indoor localization which is a significant aspect of LPS because of its broad applications and greater demands for performance. And WLAN location fingerprinting is proved to be one of the most practical and adaptive methodology for indoor localization.

Although already well studied and developed, thorough analysis and systematic evaluation of WLAN location fingerprinting itself are merely seen in the community. The development of WLAN location fingerprinting has hit its bottleneck and is hard to progress without well rounded study on its fundamental aspects. Yet related efforts have only began very recently.

This chapter is organized as follows. Section 3.1 presents related articles and discusses the space of improvement for LPS, especially indoor localization. In Section 3.2, we introduce previous works on analysis of WLAN signals and location fingerprints. In Section 3.3, we examine efforts on evaluation and comparative study of existing WLAN fingerprinting systems.

3.1 Exploring the Space for Improvements

Although abundant approaches and articles have been presented in the area of WLAN location fingerprinting, competent solutions addressing the increasing demands for indoor localization are yet to be found. The difficulty to push this technology forward to full practicality leads to the research on exploring potential breakthroughs for localization itself, in order to guide and orientate further developments more with a better
vision. In [55, 56], comparative surveys of various indoor positioning systems have been presented where each considered system is summarized based on metrics including accuracy, computational complexity, robustness etc. However, both works lack a clear classification of existing systems and emphasize the importance of fingerprinting method for indoor localization.

In [57], a survey on wireless based indoor positioning techniques has been presented as well as discussions on progressing towards more adaptive schemes. Classifications introduced include signaturing, in their contexts referred as proximity, triangulation and fingerprinting. And a similar conclusion matching ours has been drawled that fingerprinting is the most suitable and practical approach for wireless based systems, in our contexts as WLAN, for following reasons.

- **Affordability**: Fingerprinting has little hardware defined requirements for AP as well as client’s devices, making it straightforward to deploy on off the shelf infrastructures.

- **Reliability**: Compared to signaturing and radio triangulation, fingerprinting provides relative better accuracy with less constraints like LOS paths.

At the same time, challenges and trends for further development of wireless based fingerprinting can be list as follows.

- **Facing environmental heterogeneity**: The increasing dynamic of WLAN spaces put challenges on the flexibility of localization.

- **Hybrid and novel positioning techniques**: To further improve the performance, novel techniques are to be proposed and we have witnessed a great trend to combine different techniques to compensate drawbacks of individual solutions.

Our study is a direct response to above challenges, since addressing environmental heterogeneity largely lays on construction of adaptive datasets and the design hybrid positioning techniques requires comprehensive understanding and evaluation of those existing techniques.

A similar review of WLAN based positioning system is presented in [48] also with a focus on fingerprinting. Several fingerprinting positioning techniques have been compared. As a conclusion, probabilistic method performs better than other algorithms and yet efforts to further improve it are needed. In this thesis, we also pay attentions on evaluations of popular probabilistic positioning algorithms. More details are given
in later sections. Those early works prove the potential of our work on fundamental study of WLAN fingerprinting methods.

### 3.2 Analysis of WLAN Location Fingerprints

As an important aspect for WLAN fingerprinting, the analysis of WLAN location fingerprints is a necessary area to investigate. Due to the proliferation of WLAN fingerprinting systems in recent decades, this topic causes growing attentions.

In [58], authors conduct WLAN measurements with a laptop equipped with an Orinoco WLAN card. At each of four locations included in the experiment, RSS signals are sampled at a frequency of 1 Hz and sampling vector contains at most three elements, in another word, at most three APs’ signals are recorded.

Their results show many interesting points. First, user’s presence, orientation could have a noticeable impact on fingerprints samples due to the signal attenuation caused by human body which consists of 70% water and absorbs RSS considerably. This observation matches those proposed in directional analysis based radio triangulation introduced in Chapter 2. More efforts on statistical properties of fingerprints are also made with following key observations.

- **Left Skewness**: The temporal distribution of RSS from an AP at the same location is usually not turned to be normal, or Gaussian, but left skewed, contradicting the assumption of Gaussian probabilistic modeling approach.

- **Signal variances**: The results demonstrate a negative correlation between RSS variance measured by standard deviation and the signal strength level or propagation distance. Strong signals tend to have larger variance.

- **Stationarity of RSS**: The ergodic theorem is applied according to the Wiener definition of stationarity based on two criteria, stable RSS mean and variance and shape of autocovariance function over separate time intervals. Results show that RSS is not stationary enough as supposed to.

In [59], RSS characteristics of five heterogeneous 802.11 interfaces have been investigated. The results also indicate that RSS distributions more tend to be left skewed, and different receive interfaces have different sensitivity to RSS variations over distance. The key observation over different interfaces is that RSS distributions received at more sensitive interface turn to be more left skewed. For device heterogeneity, here
different WLAN card models for PC are included, while in our study, both PC and smartphones are considered, providing a more comprehensive comparison.

The work of heterogeneous WLAN interfaces measuring is carried forward in [1] which conducted a further analysis over six PC WLAN card models. Observations mentioned previously have been confirmed. Compared to previous works, a relatively longer term of experiments have included and new observations have been made as following.

- The results of long term experiments indicate that RSS distributions are heavily time dependent, different periods of time may lead to different patterns in RSS readings, especially the variances.

- RSS at location without LOS to the AP have smaller standard deviations. Since RSS variances severely reduce the reliability of localization, this result suggests that, in contrast to common intuitions, for fingerprinting, locations without LOS path to AP may have better performance.

An area closely related to statistical analysis of WLAN fingerprints is the modeling of temporal and spatial dependencies of RSS values, which is preliminarily explored in [60]. In our work, long term experiments have also been carried out and interesting observations and analysis have been made, presented in Chapter 4.

### 3.3 Comparative Study on WLAN Fingerprinting Systems

Although being a specific branch of LPS, WLAN fingerprinting itself is a highly diverse and complex mechanism, which can be decomposed into various aspects and components. The work of comparative study on WLAN fingerprinting specifically has only began in recent years, due to the vast development and great potentials of it.

In [61], a taxonomy of radio location fingerprinting is made, shewed in Fig 3.1. The taxonomy is based on four aspects regarding the structure of fingerprinting systems. First is the Scale of localization from building level to city wide. Second is measurement methods where authors point out that various radio features can be used as fingerprints, which is related to a fundamental question of the definition of fingerprints and is further discussed in Chapter 4. Role identifies the dependency on specific infrastructures of the system. In this thesis, the signaturing and fingerprinting is not
Chapter 3. Related Work

Figure 3.1: A taxonomy of radio location fingerprinting

clearly differentiated although authors use terms of *descriptive* and *spatial* to refer sign-naturing and fingerprinting in our contexts respectively. In our work, the evaluation is focus on fingerprinting based on more accurate classifications.

In [62], an evaluation of city scale WLAN localization systems is presented, with conclusions on impacts of the density of AP, the size and refreshing frequency of datasets on the performance of the system. However, in our work, we focus on building level scale which is a more common scenario for WLAN fingerprinting system.

In [63], a set of tests in an indoor positioning scenario using WLAN signal strengths is performed to determine the influence of different positioning algorithms and parameters with differentiation of static and filtering methods. In [64], the emerging field of error estimation and optimization of fingerprinting based positioning engines has been explored.

*The topic of systematic evaluation and parametric optimization of WLAN location fingerprinting is very new. And in this thesis, we evaluate WLAN fingerprinting systems from a more comprehensive perspective, considering the totality of a fingerprinting system instead of only the positioning algorithms, and also influencing properties of WLAN fingerprint itself, making our work a unique contribution to this area.*
Chapter 4

Key Properties of WLAN Fingerprints

A fundamental aspect of WLAN location fingerprinting is the engineering of WLAN fingerprints. In this Chapter we introduce datasets we constructed in this study as well as analysis of WLAN location fingerprints from novel perspectives. Section 4.1 details the methodology and tools we used for experiments. In Section 4.2, we analysis visibilities of AP during measurements. In Section 4.3 we discuss properties RSSI series collected from measurements and their impacts on fingerprinting systems. In Section 4.4, we analysis correlations between important properties of WLAN location fingerprints. And in Section 4.5, we introduce impacts of VAP, an influencing feature of recent WLAN technologies that has not been discussed before.

4.1 Datasets design and structure

For a fingerprinting system, apart from the localization engine or algorithms, the construction of fingerprints itself is non-trivial. In our experiments we conducted measurements on both PC and smart phones. The objects of those experiments are listed as follows.

- Studying influencing aspects and important properties of the engineering of WLAN fingerprints.
- Providing datasets for evaluation of WLAN location fingerprinting.

4.1.1 Testbeds and Tools

The experiments are based on both Windows laptop and Android smart phone. Specific features of those devices are listed in Table 4.1. The reason for this particular
selection of devices is that mobile phones and laptops are the most important platforms for wireless indoor localization. In earlier works, experiments and systems are mostly deployed on laptops. However, in recent years, the proliferation of mobile phones has received great attentions for localization. Thus in our experiments laptops and smartphones are both included.

<table>
<thead>
<tr>
<th>Category</th>
<th>Model</th>
<th>System</th>
<th>WLAN interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smart-phone</td>
<td>HTC nexus one</td>
<td>Android 2.3.6</td>
<td>Broadcom BCM4329</td>
</tr>
<tr>
<td>PC Laptop</td>
<td>ASUS m51</td>
<td>Windows7</td>
<td>Intel WiFi Link 5100 AGN</td>
</tr>
</tbody>
</table>

Table 4.1: Features of measuring devices

All experiments are conducted in informatics forum, housing the school of Informatics, university of Edinburgh where most of the WLAN infrastructures are for academic and college purposes. The building mainly contains 8 individual sections which are listed below.

- **Basement**: The restricted accommodating several laboratories for perception and behavior research, referred as B

- **InSpace**: A separated area besides the ground floor for informatics related exhibitions and new medias, referred as I.

- **Ground Floor**: Containing large open areas and several broad meeting rooms, mostly used for reception and hosting conferences, referred as G.

- **Up floors**: The five floors are main functionalities of the building, housing majority of the offices and labs of the school, containing corridors and some open spaces, especially on the second floor, referred by the floor numbers.

For the ease of measurements and to capture impacts of crowds, all measurements are conducted on up floors. 802.11 AP in the area are mostly *Cisco Air* modules described in Table 4.2.

<table>
<thead>
<tr>
<th>Module names</th>
<th>protocol</th>
<th>No. of deployments</th>
<th>Existing sections</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP1131</td>
<td>802.11g</td>
<td>39</td>
<td>ALL</td>
</tr>
<tr>
<td>AP1142</td>
<td>802.11n</td>
<td>8</td>
<td>G, 1, 2, 4</td>
</tr>
</tbody>
</table>

Table 4.2: 802.11 AP modules deployed in informatics forum
Experiments can be categorized into two parts. For studying the temporal dependences of RSS readings which in our study is used to construct WLAN fingerprints, long term stationary measurements on both ASUS m51 laptop and HTC nexus one smart-phone have been conducted at a fixed location which is room 1.17 on the first floor.

For the smart phone measurement, data is collected on the HTC nexus one with a measuring application developed specifically for this project. Android API WiFi-manager is used to control the Broadcom BCM4329 interface. And BroadcastReceiver is leveraged in the application to interact with WiFimanager and receive the results. Features being retrieved from measurements include Basic service set identification (BSSID), Service set identification (SSID), RSSI, Channel and UNIX timestamp. The WiFi scanning is not set at a fixed frequency. The broadcastreceiver trigers the subsequent scan when results of the previous scan are received. A single scan on HTC nexus one would normally take about 1 second.

Datasets from the laptop are collected with inSSIDer [65] from metageek, which is a popular tool for WLAN infrastructure managements. The outputs of the software is stored in GPX files which is a file extension designed specifically for GPS related applications. The results of WLAN monitoring are associated with geographical co-ordinates which can be used for spatial management of the network. However in our project, all measurements are done in a fixed location.

Apart from geographical information, Features recorded in the GPX files include those captured by the phone as well as additional information like network type and signal quality. The reported Media Access Control (MAC) address is equivalent to BSSID in this context. A parser written in Java is developed to extract needed features from the raw GPX file and convert its timestamps into standard UNIX timestamps for the convenience of analysis.

The second part of the experiments is to collect RSS samples through corridors and open areas on the up floors of the building. Specific site-survey applications are developed on Android platform and all the measurements are based on grid planed on floors. The user interface of the application on HTC nexus one is shewed in Fig 4.1, which contains following components and functionalities.

- **Main display**: Displaying the ground truth of the selected floor with the yellow square identifying current position (cell).

- **Start and status display**: The Start button could trigger repeating scanning for
certain times. The status bar beside it shows the progress of measurements on current cell.

- **Set and cell control**: The *Set* button could initialize measurements by creating a new recording file on the external memory of the phone. Cell control buttons, *Next* and *Last*, allow user to navigate through cells.

- **Floor control**: The follow control buttons, *LF* and *HF*, allow user to choose from the five floors.

### 4.1.2 Long Term Stationary Measurements of WiFi Scanning

To conduct stationary experiments of WiFi scanning on both smart-phone and laptop, we carried out three different measurements described as follows.

- **SPDaug13**: Short for stationary RSS sampled data collected with HTC nexus one at the week August 13 to August 18.

- **SLDaug13**: RSS samples collected at a 7 hours measurement with ASUS m51 started at the midnight of August 13.

- **SPDaug24**: RSS samples collected at measurements on HTC nexus one from August 24 to August 29.

All datasets are stored in MySQL databases where each dataset is saved in an individual table. Data structures of all tables are identical, making it able to use a same
interface for data retrieving. The structure is described in Table 4.3. Each row of the table records one WiFi scan reading for a certain AP, one scan could capture several readings. Apart from inherent features, the LN variable, short for line number, acts as the primary key. The No. of readings identify how many readings have been recorded together with current reading.

<table>
<thead>
<tr>
<th>LN</th>
<th>BSSID</th>
<th>RSSI</th>
<th>Channel</th>
<th>SSID</th>
<th>Timestamps</th>
<th>No. of readings</th>
</tr>
</thead>
</table>

Table 4.3: Data structure of tables for stationary measurements

To build comparison between WiFi scanning on phone and laptop, we pick up a fraction from SPDaug13 which contains several hours of data and has overlap measuring periods with the laptop dataset SLDaug13. The selected overlapped fraction is stored in a separate table named SPDaug13P. Basic information of those tables are listed in Table 4.4. The Size stands for the total number of rows recorded in the table. Start time and End time indicates the exact duration of each measurement. Covered BSSID stands for number of distinct AP represented by BSSID being seen during the measurement.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Start Time</th>
<th>End Time</th>
<th>Size</th>
<th>No. of BSSID</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPDaug13</td>
<td>2012/Aug/13 01:30:23</td>
<td>2012/Aug/18 17:24:28</td>
<td>5314132</td>
<td>65</td>
</tr>
<tr>
<td>SLDaug13</td>
<td>2012/Aug/13 03:01:56</td>
<td>2012/Aug/13 10:47:04</td>
<td>285279</td>
<td>84</td>
</tr>
<tr>
<td>SPDaug24</td>
<td>2012/Aug/24 11:02:54</td>
<td>2012/Aug/29 14:04:44</td>
<td>4794787</td>
<td>61</td>
</tr>
</tbody>
</table>

Table 4.4: Basic information of datasets of stationary measurements

4.1.3 Constructing Fingerprint maps

A fundamental aspect of WLAN fingerprinting system is to construct a fingerprint map. Here we introduce our WLAN signals site-survey measurements on the five up floors of the building. As mentioned before, our measurements and evaluation are based on cells deployed mainly on corridors of each floor. Each cell is assigned with a unique ID.

As introduced before, data collections for a fingerprinting system contain two steps corresponding to Off line and On line phases respectively. In our experiments, for the ease of evaluation, data of off line and on line phases is collected at a same time. All
Measurements are collected from 18th to 26th September 2012, mostly during late mornings, using HTC nexus one, where at least 20 samples at each cell are collected. In the simulations of fingerprinting systems, the data is divided into two fractions used as training data for off line phase and testing data for on line phase respectively. Thus we minimize the impacts of temporal variations of WLAN fingerprints and make our evaluations of fingerprinting system more convincing. The site-survey data also provides materials for spatial analysis of WLAN fingerprints.

The basic information of collected data including cells deployments and covered WLAN networks are listed in Table 4.5. Cell deployments are on corridors except on the second floor where an large open space is also included, which is why more Cells are planed on this floor.

<table>
<thead>
<tr>
<th>Floor</th>
<th>No. of Cells</th>
<th>Covered BSSID</th>
<th>Covered SSID</th>
<th>Received Channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>100</td>
<td>76</td>
<td>13</td>
<td>1,5,6,9,11,13</td>
</tr>
<tr>
<td>2nd</td>
<td>144</td>
<td>88</td>
<td>10</td>
<td>1,5,6,9,11,13</td>
</tr>
<tr>
<td>3rd</td>
<td>98</td>
<td>79</td>
<td>12</td>
<td>1,5,6,8,9,11,13</td>
</tr>
<tr>
<td>4th</td>
<td>76</td>
<td>65</td>
<td>12</td>
<td>1,5,6,8,9,11,13</td>
</tr>
<tr>
<td>5th</td>
<td>78</td>
<td>67</td>
<td>18</td>
<td>1,4,5,6,8,9,11,13</td>
</tr>
</tbody>
</table>

Table 4.5: Basic information of sit survey datasets

Results are also stored in MySQL databases. Two different data structures are used. The first is for storing raw WiFi scanning readings where each row represents an individual reading, same with Table 4.3. The second is an aggregated data structure where readings from a same AP at a same cell are aggregated. For each aggregated record, various statistical features of aggregated readings are recorded, described in Table 4.6, where ID is the primary key, Num indicates the number of raw records being aggregated and Mean and Standard deviation are statistical features of aggregated RSSI series. The aggregated data structure is created for the convenience of analysis and information retrieval for probabilistic modeling based fingerprinting system, which is explored further in Chapter 5.

<table>
<thead>
<tr>
<th>ID</th>
<th>Floor</th>
<th>Cell</th>
<th>BSSID</th>
<th>SSID</th>
<th>Channel</th>
<th>Num</th>
<th>Deviation</th>
<th>Mean</th>
</tr>
</thead>
</table>

Table 4.6: Data structure of aggregated site-survey datasets
4.2 Analysis of AP Visibilities

A important feature of WLAN location fingerprints is the visibility of AP, which is rarely discussed before. Since WLAN fingerprinting is based on AP transmitting signals over spaces, the visibility of AP over locations is essential for constructing fingerprints. As discussed previously, fingerprinting methods usually depend on assumptions of rather stable AP visibilities. Some assume an identical set of AP is available for the whole place of interests and some suggest that AP visibilities should be constant over time. Yet those assumptions usually cannot meet the reality of WLAN enabled environments. Thus to further understand limits and performances of WLAN fingerprinting, detailed analysis of AP visibilities is needed.

4.2.1 Temporal Variations

Here we use stationary measurements data SPDaug24, SPDaug13, SPDaug13P and SLDaug13 to analyze temporal variations of AP visibilities. The main feature we retrieve from datasets is numbers of visible AP (NVA) time series, i.e., number of distinct AP captured by each scan. SPDaug24 and SPDaug13 both contain several days of data and are chosen to represent long term variations of NVA. SPDaug13P and SLDaug13 are relative shorter datasets conducted at the same period of time and are chosen to show short term variations and comparison between laptop and phone measurements. The WiFi scanning capabilities of the ASUS m51 laptop and HTC nexus one are different since the former could capture both 2.4GHz and 5GHz while the phone only support 2.4GHz. To provide more convincing comparison, we only consider AP transmitting at 2.4GHz when retrieving NVA time series from laptop.

Fig 4.2 and Fig 4.3 visualize NVA time series from four datasets respectively with red curves represent generalized additive models (GAM) regressions [66]. The GAM method is chosen for regression since temporal distributions of NVA are clearly not matching parametric models. Thus non-parametric methods like GAM is more suitable. Since the stationary NVA time series contain long periods of measurements, individual values in the series with large time interval can be treated as independent variables. Thus univariate series of NVA are better fitted to multivariate regression models like GAM, whose mathematical expression is presented in Eqn 4.1. Where $\mathbb{S}$ is the set of predictor variables and $Y$ is the response variable.

$$E(Y) = S_0 + S_1(X_1) + \ldots + S_p(X_p)$$ (4.1)
Figure 4.2: Short NVA time series and regressions from SLDaug13 and SPDaug13P
Also empirical cumulative distribution functions (ECDF) are generated based on for those NVA series showed in 4.4. Those figures give an intuitive observations of AP visibilities and differences in scanning capabilities between laptop and smart-phone. Surprisingly, results show that AP visibilities are not stable as we thought but show significant variations and temporal dependences.

As we can see, the NVA series on laptop is much more disperse than that on phones,
Figure 4.4: ECDF of NVA time series from four datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Min</th>
<th>Max</th>
<th>Mode</th>
<th>Mean</th>
<th>Median</th>
<th>Deviation</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPDaug13</td>
<td>5</td>
<td>31</td>
<td>20</td>
<td>19.02</td>
<td>19</td>
<td>2.4292</td>
<td>12.77141%</td>
</tr>
<tr>
<td>SPDaug24</td>
<td>7</td>
<td>31</td>
<td>19</td>
<td>18.88</td>
<td>19</td>
<td>2.4434</td>
<td>12.99344%</td>
</tr>
<tr>
<td>SPDaug13P</td>
<td>10</td>
<td>27</td>
<td>19</td>
<td>18.97</td>
<td>19</td>
<td>2.03427</td>
<td>10.7256%</td>
</tr>
<tr>
<td>SLDaug13</td>
<td>1</td>
<td>30</td>
<td>21</td>
<td>19.87</td>
<td>21</td>
<td>4.146605</td>
<td>20.86858%</td>
</tr>
</tbody>
</table>

Table 4.7: Statistical features of NVA time series from four datasets

indicating larger variances. And AP visibilities seem to be more constant on the phone. An interesting finding is that different datasets collected on the phone during different periods of time show great similarities according to ECDF figures.

Those intuitive observations are verified with detailed in features in Table 4.7 where another statistical tool, coefficient of variation (CV) described in Eqn 4.2, is included to measure the dispersion specifically, where \( \sigma \) and \( \mu \) represent standard deviation and mean of the variable respectively. Those features clearly back our conclusions.

\[
C_v = \frac{\sigma}{\mu} \tag{4.2}
\]
4.2.2 Spatial Coverages

Spatial coverage of AP visibilities is also an important factor for fingerprinting systems. Many fingerprinting systems require coverages of an identical set of AP in the space of interest. However, in reality, this requirement usually cannot be met. Our site-survey data show that no AP can cover all cells on its located floor. Summary of AP coverages on five floors is presented in Fig 4.5. The values represented by charts are amount of AP with certain ranges of coverage, measured by number of covered cells.

<table>
<thead>
<tr>
<th></th>
<th>1st Floor</th>
<th>2nd Floor</th>
<th>3rd Floor</th>
<th>4th Floor</th>
<th>5th Floor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous coverage</td>
<td>11 (14.5%)</td>
<td>6 (7%)</td>
<td>6 (7.6%)</td>
<td>1 (1.5%)</td>
<td>6 (9%)</td>
</tr>
</tbody>
</table>

Table 4.8: Summary of AP with continuous coverages

Another interesting findings about AP coverage is that, opposite to normal assumptions, Coverage of AP signals is usually not continuous. Previous works often assume that an AP should have continuous coverages over cells. However, in our experiments, AP coverages are often interrupted. We count numbers of AP that has continuous coverage over more than one cells. The results are shewed in Table 4.8, which indicate AP with continuous coverages are very rare.
Chapter 4. Key Properties of WLAN Fingerprints

4.3 Properties of RSSI Series

As discussed earlier, various features of WLAN signals can be used for fingerprinting. Most of existing WLAN fingerprinting systems construct their fingerprints using RSSI.
In this work, evaluations and simulations of fingerprinting systems are also based on RSSI. In the context of WLAN fingerprinting, RSSI is usually considered series instead of individual readings since in both on line and off line phases, multiple readings are collected at each location. Thus it is important to further analyze RSSI series for WLAN location fingerprinting.

4.3.1 Temporal Distributions and Impacts of LOS

Here we examine long term temporal distributions of RSSI time series as well as impacts of LOS. We pick up representative AP from all AP can be see from the four stationary datasets SPDaug24, SPDaug13, SPDaug13P and SLDaug13 and extract its RSSI time series. In this analysis, two AP are selected. The first is an experimental WiFi board with BSSID ’00:80:48:57:C7:A3’ labeled as LOSap, which is deployed very close to scanning devices and has LOS path to both of them.

The second AP selected is part of the university WiFi network ‘central’ with BSSID ’00:24:97:83:3B:C0’, labeled as NLOSap, which is deployed beside room 1.12 on the first floor and has no LOS paths. They are chosen to analysis the impact of LOS paths on received RSSI because of their relatively higher visibilities during all measurements. Their number of appearances in four datasets are listed in Table 4.9. Both representative AP are among the most constant AP in all datasets and have similar number of appearances, making them fully comparable.

<table>
<thead>
<tr>
<th>AP Label</th>
<th>BSSID</th>
<th>SLDaug13</th>
<th>SPDaug13P</th>
<th>SPDaug13</th>
<th>SPDaug24</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOSap</td>
<td>00:80:48:57:C7:A3</td>
<td>7480</td>
<td>18445</td>
<td>273552</td>
<td>243446</td>
</tr>
<tr>
<td>NLOSap</td>
<td>00:24:97:83:3B:C0</td>
<td>7464</td>
<td>17285</td>
<td>269518</td>
<td>240314</td>
</tr>
</tbody>
</table>

Table 4.9: Number of appearances of representative AP in four datasets

Fig 4.6 to Fig 4.9 demonstrate RSSI time series as well as their histograms and GAM regressions from both representative AP in four datasets. Gaussian distribution curves are generated to fit empirical histograms. Clear different patterns between LOSap and NLOSap can be observed from all measurements on both laptop and smart phone. RSSI time series from NLOSap are clearly more normally distributed than those from LOSap.

Detailed statistical features of those series from LOSap and NLOSap are listed in Table 4.10 and Table 4.11 respectively. Here we introduce a new statistical tool, skewness test, to quantize the skewness of RSSI histograms. The test value can be
positive or negative, while negative values indicate the left tail of the density function is longer than the right tail and it is left skewed and positive values have opposite
Chapter 4. Key Properties of WLAN Fingerprints

(a) RSSI series from LOSap

(b) RSSI series from NLOSap

(c) Histogram of RSSI series from LOSap

(d) Histogram of RSSI series from NLOSap

Figure 4.7: RSSI time series and distributions from SPDaug13P

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Min</th>
<th>Max</th>
<th>Mode</th>
<th>Mean</th>
<th>Median</th>
<th>Deviation</th>
<th>CV</th>
<th>Skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLDaug13</td>
<td>-85</td>
<td>-54</td>
<td>-59</td>
<td>-60.91</td>
<td>-60</td>
<td>3.82</td>
<td>6.27%</td>
<td>-2.28</td>
</tr>
<tr>
<td>SPDaug13P</td>
<td>-88</td>
<td>-58</td>
<td>-68</td>
<td>-70.28</td>
<td>-70</td>
<td>3.89</td>
<td>5.53%</td>
<td>-0.204</td>
</tr>
<tr>
<td>SPDaug13</td>
<td>-94</td>
<td>-58</td>
<td>-74</td>
<td>-74.7</td>
<td>-75</td>
<td>3.84</td>
<td>5.15%</td>
<td>0.126</td>
</tr>
<tr>
<td>SPDaug24</td>
<td>-92</td>
<td>-57</td>
<td>-71</td>
<td>-71.52</td>
<td>-71</td>
<td>3.96</td>
<td>5.54%</td>
<td>-0.306</td>
</tr>
</tbody>
</table>

Table 4.11: Statistical features of RSSI series of NLOSap (00:24:97:83:3B:C0)

implications. The skewness of a variable is defined as the third standardized moment
Chapter 4. Key Properties of WLAN Fingerprints

(a) RSSI series from LOSap

(b) RSSI series from NLOSap

(c) Histogram of RSSI series from LOSap

(d) Histogram of RSSI series from NLOSap

Figure 4.8: RSSI time series and distributions from SPDataug13

described in Eqn 4.3. Where $\kappa_3$ and $\kappa_2$ are the third and the second cumulants.

$$\gamma_1 = \frac{\kappa_3}{\kappa_2^{3/2}}$$

(4.3)

Very interesting findings can be found from the two tables. We can clearly observe differences on scanning capabilities between laptop and smart-phone since for each representative AP, RSSI received by laptop is generally higher than phone. And unsurprisingly, RSSI from NLOSap is lower than that from LOSap. However, lower RSSI reduce the scale of differences RSSI received at heterogeneous devices since RSSI is a logarithmic scale, and also reduce the variance and coefficient of variation. Also from both figures and tables we can see that NLOSap RSSI series distributions are more
normal and less skewed. Thus the results back the conclusion presented in [1] that AP without LOS paths to the receiver is better for localization. From our analysis, more comprehensive evidences for this conclusion are found, listed as follows.

- **Patterns of distribution**: RSSI series from AP without LOS paths are more normally distributed and less skewed, making it more suitable for parametric modeling based fingerprinting methods like Gaussian modeling.

- **Scale of variations**: AP without LOS reduce the scale of RSSI temporal variations and impacts of devices heterogeneity, thus improving robustness of RSSI space distance based methods like NN.
4.3.2 Normality and Skewness

As mentioned previously, normality and skewness of RSSI series are important properties of WLAN fingerprints and have significant impacts on parametric modeling methods which are popular for WLAN fingerprinting. Thus it is necessary to analysis those two features of collected fingerprints. Here we also use skewness test equation introduced in Eqn 4.3. For normality, we use Shapiro-Wilk test described in Eqn 4.4, where $\mu$ is the mean of the variable $x$ and $x(i)$ is the $i$th order statistic. The constants $a_i$ are given by Eqn 4.5, where $\mathbb{M}$ is the expected values of order statistics of independent and identically distributed (i.i.d.) random variables sampled from the standard normal distribution, and $\mathbb{V}$ is the covariance matrix of them. We extract the P-value of test statistic as outputs.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x(i)\right)^2}{\sum_{i=1}^{n} (x_i - \mu)^2} \quad (4.4)$$

$$A = m^T \mathbb{V}^{-1} \left(m^T \mathbb{V}^{-1} \mathbb{V}^{-1} m\right)^{1/2} \quad (4.5)$$

![Figure 4.10: Normality and skewness test results of RSSI series on First Floor](image)

As we know, the normality test is meaningless for long time series since large samples are easily rejected by the model. Thus we did not include normality test in the analysis of stationary data. However, in site-survey measurements, short RSSI time series are collected on each cell. For small samples, normality test techniques
such like ShapiroWilk test are very sensible measures of how close the variable is to normal distribution. Here we have conducted both normality and skewness tests for all cells on each floor. The results are presented in Fig 4.10 to Fig 4.14 with right curves representing local polynomial regression (LOESS). Statistics of test results are
Chapter 4. Key Properties of WLAN Fingerprints

Figure 4.13: Normality and skewness test results of RSSI series on Fourth Floor

Figure 4.14: Normality and skewness test results of RSSI series on Fifth Floor

The results provide interesting implications on statistical properties of RSSI series at different locations. For the normality test, resulted P-values are very widely dispersed including several orders of magnitude, indicating very different relations to
normal distributions. However, the mean and median P-values on all floors show a constant scale. If we set the alpha level to be a typical figure as 0.5%, then those values are sufficient for the hypothesis of normal distributions. As a conclusion, normal distribution is suitable for most of the series, however, some series differ a lot and would cause severe impacts on fingerprinting method based on the hypothesis of normal distributed RSSI series.

For the skewness test, the results indicate very diverse patterns for distributions on all floors. The majority of RSSI series is left skewed and median values are all negative. Previous works like [58] suggest RSSI time series are mostly left skewed. Our results show that although the majority of RSSI series are left skewed, the dominance of left skewness is not significant, there is still large number of series that is right skewed or evenly distributed.

### 4.4 Correlations of Fingerprint Features

In this section we summarize influencing features of WLAN location fingerprints that have been discussed previously in this chapter and try to find correlations between them. Thus we could further understand relations of fundamental factors of the engineering of WLAN location fingerprints and explore the space for improvements. Main features of WLAN fingerprints are listed as follows.

- **Constancy**: The constancy of AP visibilities is essential feature for fingerprints. In our context, AP appearing more frequently during measurements at certain location are considered more constant.
• **Strength**: The signal strength of received signals indicated by RSSI is the main feature used to construct fingerprints.

• **Stability**: The stability of received signals, in our context, measured by a reverse index, the standard deviation of RSSI series.

• **Probability**: Considering RSSI series as a random variable, patterns of its probabilistic distributions have significant impacts on fingerprinting, especially probabilistic modeling based approaches. Various techniques can be used to analysis this factor. Here we mainly consider results of Shapiro-Wilk normality tests.

Here we introduce the statistical tool we used, Pearson correlation coefficients described in Eqn 4.6, where \( \mu_x, \mu_y \) and \( \sigma_x, \sigma_y \) are mean and standard deviation of the two variables respectively.

\[
\rho = \frac{\sum [(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y}
\]

(4.6)

<table>
<thead>
<tr>
<th>Floor</th>
<th>Constancy-Strength</th>
<th>Constancy-Stability</th>
<th>Strength-Stability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Floor</td>
<td>0.4050907</td>
<td>0.1883634</td>
<td>0.5228903</td>
</tr>
<tr>
<td>2nd Floor</td>
<td>0.6191411</td>
<td>0.3272367</td>
<td>0.4071382</td>
</tr>
<tr>
<td>3rd Floor</td>
<td>0.6430674</td>
<td>0.3887892</td>
<td>0.4136755</td>
</tr>
<tr>
<td>4th Floor</td>
<td>0.6001379</td>
<td>0.35482358</td>
<td>0.5356236</td>
</tr>
<tr>
<td>5th Floor</td>
<td>0.6507656</td>
<td>0.45762849</td>
<td>0.6038725</td>
</tr>
</tbody>
</table>

Table 4.13: Pair correlations between strength, constancy and stability

<table>
<thead>
<tr>
<th>Floor</th>
<th>Normality-Strength</th>
<th>Normality-Constancy</th>
<th>Normality-Stability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Floor</td>
<td>-0.2752412</td>
<td>-0.3455074</td>
<td>-0.0936583</td>
</tr>
<tr>
<td>2nd Floor</td>
<td>-0.2008754</td>
<td>-0.3113624</td>
<td>0.0035135</td>
</tr>
<tr>
<td>3rd Floor</td>
<td>-0.2828671</td>
<td>-0.327165</td>
<td>-0.02073372</td>
</tr>
<tr>
<td>4th Floor</td>
<td>-0.2545603</td>
<td>-0.3534598</td>
<td>-0.04899797</td>
</tr>
<tr>
<td>5th Floor</td>
<td>-0.2381025</td>
<td>-0.311509</td>
<td>-0.05791279</td>
</tr>
</tbody>
</table>

Table 4.14: Correlations between Shapiro-Wilk test values and other features

In Fig 4.13 we present pair correlations between strength, constancy and stability. One intuitive assumption about WLAN signaling is that AP with stronger signal also have higher constancy of visibilities, which is verified by our analysis since there
is a clear positive correlation between strength and constancy. And since higher signal level would increase the scale of standard deviation as well, positive correlations between strength and stability as well as constancy and stability are not surprising.

Correlations between normality test values and strength, constancy, stability respectively are presented in Fig 4.14. The negative correlation between normality and strength back our conclusions made in Section 4.3.1 that AP with lower signal level produce more normal distributed RSSI series. The correlations between normality and constancy maybe affected by impacts of sample sizes on ShapiroWilk test. And stability is proved to be irrelevant to other features.

4.5 Impacts of Virtual Access Points

Properties of WLAN location fingerprints are heavily dependent on WLAN technology itself which is developing rapidly. WLAN infrastructures leveraged in our works are all based on IEEE 802.11 standards. One interesting feature observed in our experiments that have not been noticed before is VAP, which are now widely deployed in WLAN infrastructures to boost capabilities and performances, especially in campus environments. We found that VAP is largely used in informatics forum. All WiFi fingerprinting schemes so far are only considering physical AP with distinct signaling patterns. However, VAP, which is based on time division and hosted on physical AP, behave quit differently. Since WLAN interfaces on consumer devices like smart phones and laptops can only give limited information, it is difficult to differentiate VAP and physical AP from scanning results. Intuitively the use of VAP could severely affect WLAN location fingerprinting in following aspects, which are verified by our experiments.

- **Discriminability**: WLAN fingerprints are constructed with physical layer features of the signal, in our context RSSI. Ideally fingerprints from different AP should have distinct patterns. However VAP hosted on an identical device share the similar signaling characters. Thus reduce the discriminability of fingerprints.

- **Computational Complexity**: The use of VAP could largely increase the total number of BSSID involved in a fingerprinting system, thus increasing the computational complexity of localization considerably.
4.5.1 Controlled VAP and Temporal Similarities

During the stationary measurements of \( SPDaug24 \) and its subset \( SPDaug24P \), we implemented controlled VAP on a Gateworks Avila transmitting board working under OpenWrt, which is able to set up at most four VAP at the same time. A bash script is written and installed on the board to change the number of enabled VAP from one to four repeatedly every 15 minutes. Each VAP is given a unique SSID as \( Vap1 \), \( Vap2 \), \( Vap3 \), \( Vap4 \) respectively. Basic information of readings from those VAP are listed in Table 4.15. Since the number of VAP is changing repeatedly, the four VAP have different number of readings during the measurements.

<table>
<thead>
<tr>
<th>VAP name</th>
<th>No. of Readings</th>
<th>Mean of RSSI</th>
<th>Deviation of RSSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vap1</td>
<td>18266</td>
<td>-47.2313</td>
<td>3.917474</td>
</tr>
<tr>
<td>Vap2</td>
<td>13676</td>
<td>-47.42088</td>
<td>4.079981</td>
</tr>
<tr>
<td>Vap3</td>
<td>9084</td>
<td>-47.71312</td>
<td>4.092947</td>
</tr>
<tr>
<td>Vap4</td>
<td>4476</td>
<td>-47.78575</td>
<td>4.418284</td>
</tr>
</tbody>
</table>

Table 4.15: Basic features of readings from VAP

<table>
<thead>
<tr>
<th></th>
<th>Vap2</th>
<th>Vap3</th>
<th>Vap4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vap1</td>
<td>0.6312198</td>
<td>0.4556699</td>
<td>0.4173344</td>
</tr>
<tr>
<td>Vap2</td>
<td>———</td>
<td>0.4806676</td>
<td>0.4542306</td>
</tr>
<tr>
<td>Vap3</td>
<td>———</td>
<td>———</td>
<td>0.4074805</td>
</tr>
</tbody>
</table>

Table 4.16: Pearson correlation coefficients between each pair of VAP

<table>
<thead>
<tr>
<th></th>
<th>Vap2</th>
<th>Vap3</th>
<th>Vap4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vap1</td>
<td>0.9975086</td>
<td>0.9962104</td>
<td>0.9949792</td>
</tr>
<tr>
<td>Vap2</td>
<td>———</td>
<td>0.996124</td>
<td>0.995826</td>
</tr>
<tr>
<td>Vap3</td>
<td>———</td>
<td>———</td>
<td>0.995837</td>
</tr>
</tbody>
</table>

Table 4.17: Cosine similarities between each pair of VAP

We analysis similarities of RSSI readings from VAP hosted on same board by computing Pearson correlation coefficients and cosine similarities between each pair of them, shewed in Table 4.16 and Table 4.17 respectively. The RSSI time series from VAP on same physical board are not identical but share great similarities. Pearson
<table>
<thead>
<tr>
<th>Floor</th>
<th>Representative Pair</th>
<th>Deployment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Floor</td>
<td>00:24:97:83:3b:c– (0/1)</td>
<td>1.12</td>
</tr>
<tr>
<td>2nd Floor</td>
<td>08:1f:f3:23:32:6– (0/1)</td>
<td>2.01</td>
</tr>
<tr>
<td>3rd Floor</td>
<td>00:24:97:0d:e5:f– (0/1)</td>
<td>3.01</td>
</tr>
<tr>
<td>4th Floor</td>
<td>00:24:97:2d:01:f– (0/1)</td>
<td>4.01</td>
</tr>
<tr>
<td>5th Floor</td>
<td>00:24:97:83:31:a– (0/1)</td>
<td>5.01</td>
</tr>
</tbody>
</table>

Table 4.18: Basic information of selected pairs of VAP

<table>
<thead>
<tr>
<th>Floor</th>
<th>Correlation Coefficients</th>
<th>Cosine Similarities</th>
<th>Differences in Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Floor</td>
<td>0.9977322</td>
<td>0.9999606</td>
<td>2/40</td>
</tr>
<tr>
<td>2nd Floor</td>
<td>0.972237</td>
<td>0.9998644</td>
<td>7/120</td>
</tr>
<tr>
<td>3rd Floor</td>
<td>0.9967735</td>
<td>0.9999709</td>
<td>7/61</td>
</tr>
<tr>
<td>4th Floor</td>
<td>0.9978508</td>
<td>0.999976</td>
<td>4/60</td>
</tr>
<tr>
<td>5th Floor</td>
<td>0.9983491</td>
<td>0.9999739</td>
<td>2/71</td>
</tr>
</tbody>
</table>

Table 4.19: Basic information of selected pairs of VAP

coefficients clear indicate positive correlations between them and the angular distance measured by cosine similarities are very small. As we can see all the similarity values are above 99%.

### 4.5.2 Spatial Similarities

An important impact of VAP on fingerprinting system is the location discriminability. If various AP produce similar patterns of signals over cells, it is more difficult to detect locations. Here we examine spatial similarities of representative pairs of VAP deployed by the university on each floor. An interesting character of VAP deployed in the forum is that VAP on same board share identical BSSID except the last digit, making it straightforward to recognize them. The information of representative AP on each floor is presented in Table 4.18.

The analysis is based on our site-survey data. For each AP, the mean of its RSSI recorded at visible cells are extracted. We also use Pearson correlation coefficients and cosine similarities to analysis similarities between RSSI vectors of pair of VAP. Results are presented in Table 4.19. The spatial patterns of VAP signals show even greater similarities and from both criteria, RSSI vectors from pair of VAP are almost
identical, which are visualized in Fig. 4.15. However, pairs of VAP do not have identical signal coverage over cells. The differences in signal coverage are presented in the last column, where numerators are number of differently covered cells and denominators are amount of cells covered by both VAP.

Figure 4.15: Spatial RSSI vectors of pairs of VAP on each floor
Chapter 5

Systematic Evaluation of WLAN Fingerprinting Systems

This chapter presents systematic evaluations of WLAN fingerprinting systems based on site-survey data introduced and analyzed in the last chapter. All evaluations are based on simulations implemented in R, which is chosen for its simplicity and functionality for data processing and visualization. We first define the composition of a WLAN fingerprinting systems with following key components.

- **Definitions of fingerprints**: WLAN location fingerprints are based on WLAN network signaling and varied signal features can be used such like RSSI, SNR, frequency offset etc. A fingerprinting system should first identify WLAN features it leveraged to construct fingerprints.

- **Fingerprints selections**: After constructing location fingerprints, proper fingerprints should be selected to feed localization systems to provide reliable location estimations.

- **Spatial clustering**: To improve efficiency and performance, spatial clustering methods are commonly leveraged in WLAN fingerprinting systems which divide the space of interest into subspaces based on fingerprints characters. In localization, the system would first detect the possible subspace the user is within before determining fine locations.

- **Localization engine**: The core component of fingerprinting systems is the localization engine which take collected fingerprints and run time readings as inputs and provide one or multiple location estimations as outputs.
• **Results aggregation:** Since localization engines could provide multiple location estimations, it is commonly used in fingerprinting system to aggregate multiple estimating results to give a combined estimation. For example, the popular method of k nearest neighbor (KNN) is to get multiple results from NN algorithm which could be aggregated by varied methods including centroid.

Those components summarize the construction of a typical WLAN fingerprinting system and each component could influence the behavior and performance of the whole system. To better study WLAN location fingerprinting and understand its limits and potentials, it is necessary to decompose the system and evaluate it systematically, considering impacts of individual components.

In this thesis, we focus on most influencing, practical and diverse components. Most of existing WLAN fingerprinting systems leverage RSSI values to construct WLAN fingerprints which is proved to be an efficient and straightforward method. In this work, we also focus on RSSI based WLAN fingerprints.

Spatial clustering is commonly deployed in WLAN fingerprinting schemes like [38, 45] to provide additional constrains to location estimations. The basic methodology is to detect similarities of fingerprints collected from certain subspaces and first localize the user to a subspace before determining its fine grained location within. In some contexts like [45], the two steps are termed coarse and fine localizations respectively. The criterion for detecting spatial relations of fingerprints is commonly AP visibilities [38] based on the assumption that locations seeing a same set of AP should be spatial related. However, analysis shewed in the last chapter indicates that AP visibilities in WLAN enabled environments are not stable and constant as supposed, but highly dynamic and incoherent, which limit the feasibility of spatial clustering. Also, benefits from spatial clustering mainly lay on reducing computational costs while drawbacks of fingerprinting method mainly lay on performances and reliability. Thus in this work, spatial clustering is not included.

In this thesis, we mainly focus on fingerprint selection, localization engines and results aggregation which are key factors for fingerprinting systems, and thorough study and evaluations of them are rarely provided before. In Section 5.1, we present an overview of considered components and methods included for each component. In Section 5.2, simulated systems composed of varied methods for each component are evaluated and analyzed systematically from novel perspectives.
5.1 Components Overview

In this section we introduce methods included in fingerprint selection, localization engines and results aggregation components respectively. We construct our collection of fingerprints based on the site-survey data introduced in Chapter 4 with the aggregated data structure. For each cell, multiple readings from a same AP are combined as one fingerprint of that AP.

5.1.1 Fingerprints Selection

In our fingerprints datasets, usually a large amount of fingerprints is collected at each cell, demonstrated in Fig 5.1, which is a common phenomenon in fingerprinting environments due to the proliferation of WLAN infrastructures. To consider all fingerprints collected in off line for localization is infeasible for following reasons.

• **Computational costs**: More fingerprints to be considered by localization algorithms would significantly increase the computational burden of the system and reduce the speed for localization while efficiency is a key feature for positioning services.

• **Localization noises**: Obviously not all fingerprints are reliable since fingerprints are constructed based on WLAN signaling which itself can be noisy. And signals from varied AP have different feasibilities for positioning considering location discriminability. Thus leveraging all collected fingerprints without selection could have severe impacts on the system.

Thus, fingerprint selection is a key factor for fingerprinting system to pick up suitable fingerprints to feed the localization algorithm. In our context, fingerprint selection is to chose best fingerprints for each candidate cell. In our simulations, we include 7 different selection methods listed as follows.

• **Constancy**: As a major feature of WLAN fingerprints, constancy is very commonly used as a criterion for fingerprint selection. The method is to select fingerprints from AP which appear more frequently during the off line phase, or training measurements, i.e., AP have higher visibilities during the measurement. In our fingerprints datasets, for each aggregated fingerprint, we record the number of its appearance during site-survey measurements. For each candidate cell, we select fingerprints based on recorded number of appearances. Fingerprints with same constancy are selected randomly.
Figure 5.1: Summary of number of fingerprints collected at cells

- **Stability**: The stability of WLAN signals, in our context, measured by standard deviation of RSSI series from which a fingerprint is aggregated. Stability is an intuitive requirement of WLAN location fingerprints, especially for similarity based localization algorithms such like NN, where RSSI readings from same AP
are supposed to be stable and comparable. For each cell, we select fingerprints with less standard deviation.

- **Strength**: The strength of a fingerprint is measured by the mean of its RSSI series. For each cell, fingerprints with higher average RSSI value are selected.

- **Variance**: The spatial variance of fingerprints from a certain AP over cells. Ideal AP signals for fingerprinting should be varied at different cells in order to discriminate locations. If its fingerprints at different cells are highly similar, then the AP is not suitable for positioning. In this method, for each cell, we extract all fingerprints from every visible AP at that cell and measure its spatial variance by standard deviation. Fingerprints from AP with higher spatial variance are selected.

- **Coverage**: Another key property of AP used in fingerprinting system is spatial availability, or signal coverage. An ideal AP should have sufficient signal coverage over the space of interest. Here we select fingerprints with AP that is visible at larger number of cells, i.e., have better coverage.

- **Constancy+Strength**: We propose a novel hybrid selection method combing constancy and strength methods. In our measurements, we found that in my cells, large amount of fingerprints share the same constancy. Simple Constancy method would randomly select from fingerprints having same number of appearances. Here we select fingerprints based on two criteria at the time. First fingerprints with higher constancy are selected. Then within fingerprints with same constancy, those having higher strength are given priorities.

- **Constancy+Stability**: Another hybrid selection method is combing constancy and stability. Fingerprints are first ranked based on constancy. Then for fingerprints with same constancy, those more stable are given priorities.

### 5.1.2 Localization Engines

Localization engine, or positioning algorithms, is the core and most diverse component of a fingerprinting system. Various techniques and theories can be leveraged. As mentioned previously, our simulations are based on grid where the space of interest are divided into cells. In a grid based system, behaviors of varied localization engines can be summarized to a simple interface presented in Fig 5.2. The engine takes two
inputs. First is the off-line training datasets with selected fingerprints associated with each candidate cells and second is the on-line fingerprints readings, in our context, WiFi scanning results. Based on those inputs, the engine reports location estimations of the user, choosing most suitable candidate cells according to on-line readings.

![General interface of localization engines for fingerprinting](image)

In our work, we implemented 6 different localization engines which can be put into two categories listed as follows. By including various methods from each category, our evaluations are fully comprehensive and comparable.

- **Probability based methods**: By modeling RSSI time series as random variables based on varied distributions, the probability that user is at certain location could be given based on probabilistic models.

- **Similarity (distances) based methods**: Location estimations are given by comparing off-line readings with on-line readings and analyzing similarities or distances between them.

### 5.1.2.1 Parametric Modeling

The probabilistic modeling of RSSI series can be done in two ways, parametric modeling and more sophisticated non-parametric modeling techniques. For its popularity in fingerprinting and simplicity, parametric modeling is chosen in our works. As introduced previously, the task of probability based localization engine is to find a candidate cell $x$ that maximize the probability that the user is at that cell given on-line readings $S$, notated as $P(x/S)$. According to the equivalence as Eqn 5.1 proved by Bayesian
Chapter 5. Systematic Evaluation of WLAN Fingerprinting Systems

62

Theorem. The probability could be computed Eqn 5.2. \( P(s_i/x) \) represents the probability based on the \( i \)th selected fingerprint, which is computed based on its probabilistic model.

\[
\arg \max_x [P(x/S)] = \arg \max_x [P(S/x)]
\]

(5.1)

\[
P(S/x) = \prod_{i=1}^{k} P(s_i/x)
\]

(5.2)

**Gaussian modeling** is an important approach for probabilistic localization engine proposed in [67, 38]. The method is based on the intuitive assumption that RSSI time series are normal distributed, described in Eqn 5.3, where \( \mu \) and \( \sigma \) are mean and standard deviation of the variable respectively. Probabilities of user’s possible locations are computed based on Gaussian distribution functions with mean and standard deviation recorded in training data.

\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2}
\]

(5.3)

**Log-normal distribution** is closely related to Gaussian distribution which is distribution of a variable whose logarithm is normally distributed. In our measurements, we found RSSI series distributions sometimes more seem to be log-normal distributed which is also suggested in some articles [63, 68]. Thus we also include log-normal modeling method in our simulations. We retrieve mean and standard deviation of logarithms of RSSI series recorded in training data.

5.1.2.2 NN and Angular Similarity

NN is a very popular approach for localization engine in fingerprinting systems. The positioning algorithm receives two RSSI vectors \( S_t \) and \( S_o \) represents training data and on-line readings respectively. Each candidate cell in training datasets contains a RSSI vector \( S_t \) from selected fingerprints. NN is to determine user’s location by finding candidate cell whose training vector has shortest distance to the reading vector. And choosing a proper metric for measuring vector distances is non-trivial.

**Euclidean distance**, or \( L^2 \) distance, is the most common metric for NN described in Eqn 5.4, where \( S \) and \( P \) are the two vectors to be compared.

\[
d(S, P) = \sqrt{(s_1 - p_1)^2 + (s_2 - p_2)^2 + \ldots + (s_3 - p_3)^2}
\]

(5.4)

**Manhattan distance**, or \( L_1 \) distance, is a very popular non-euclidean distance metric which is defined as the sum of the absolute differences of values from two vectors, described in Eqn 5.5. Manhattan distance is very related to Euclidean distance with
which it can be generalized as Minkowski distance described in Eqn 5.6, where $t$ is the order of the distance.

$$d(S, P) = \sum_{i=1}^{n} |s_i - p_i|$$  \hspace{1cm} (5.5)

$$d(S, P) = \left( \sum_{i=1}^{n} |s_i - p_i|^t \right)^{1/t}$$  \hspace{1cm} (5.6)

**Mahalanobis distance** is a modern distance metric proposed in [69] which is much more sophisticated than former ones. The distinction of Mahalanobis distance is that it is not based on $L^p$ space and it takes into account the correlations between compared vectors. The mathematical expression of Mahalanobis distance is presented in Eqn 5.7, where $S$ is the covariance matrix of $S$ and $P$ of the same distribution. In our context, each RSSI vector $S$ is a multivariate vector whose elements are selected fingerprints. We assume RSSI signals from different AP are independent from each other, then we construct the covariance matrix $S$ as a diagonal matrix described in Eqn 5.8, where $\sigma_i$ is the standard deviation of $i$th selected fingerprints. An interesting feature of Mahalanobis distance is that it is based on assumptions of stable patterns of RSSI series distributions and it also takes into account signal variances which is similar to probability modeling. Some articles [1, 70] suggest Mahalanobis distance may outperform other metrics.

$$d(S, P) = \sqrt{(S - P)^T S^{-1} (S - P)}$$  \hspace{1cm} (5.7)

$$d_{i,j} = \begin{cases} 0, & \text{if } i \neq j \\ \sigma^2_i, & \text{if } i = j \end{cases}$$  \hspace{1cm} (5.8)

We implemented NN method based on all three distance metrics. As mentioned previously, NN based on distance metrics is to analysis similarities between *off-line* and *on-line* readings. However, similarity analysis of vectors can be done with varied techniques. Angular analysis of vectors is an important technique for vectors analysis, which is rarely mentioned in the context of WLAN fingerprinting.

To make our evaluations more comprehensive, here we include a novel approach using **cosine similarity** as the localization engine and compare it with other popular approaches. Cosine similarity measures the similarity between two vectors by computing the cosine of the angle between them, described in Eqn 5.9.

$$\text{Similarity} = \frac{S \cdot P}{||S|| \cdot ||P||}$$  \hspace{1cm} (5.9)
Chapter 5. Systematic Evaluation of WLAN Fingerprinting Systems

5.1.2.3 Tolerance for AP Invisibility

An issue about localization engines closely related to fingerprint selections is the tolerance for AP invisibility. As introduced previously, each candidate cell in datasets is associated with a selected set of fingerprints and the localization engine estimate user's location by comparing fingerprints vectors from candidate cells with readings vectors.

However, a common situation in WLAN environments, as proved in Chapter 4, is unstable visibilities of AP. Some AP recorded during the off-line phase may not be visible in on-line readings. In fingerprinting systems, fingerprint selection component would select most reliable and efficiency fingerprints for each cell. However, there is still a chance that some selected AP may not be captured during the on-line phase.

Ideally, for localization engines, variables in training and reading vectors should be identical. But due to AP invisibility, that requirement cannot always be met. And performances of varied localization algorithms are affected by this phenomenon at different levels.

In our works, for parametric modeling methods and NN based on Euclidean and Manhattan distances, if the \( i \)th selected AP for a candidate cell is not visible in reading vectors, an exceeding small probability \( P(s_i/x) \) or large distance \( d_i \) is given. Since those methods calculate final results by aggregating probabilities and distances for each fingerprints, they are still able to give estimations in that situation.

However, for NN based on Mahalanobis distance and cosine similarity methods, the requirement for constant AP visibilities is strict. And if not all selected fingerprints at a candidate cell are visible in the reading vector, that cell could not be given an estimated distance or similarity. And if all candidate cells cannot meet the requirement, the localization engine would report a NULL value, i.e., is not able to estimate the location.

Differences in tolerance for AP invisibility among varied localization engines are demonstrated in next section.

5.1.3 Results Aggregation

The idea of results aggregation is simple. Since localization engines could give multiple estimations, we may aggregate multiple estimated locations to give a better estimation. A typical approach leveraging results aggregation for location fingerprinting is KNN which aggregates multiple results got from NN, most commonly by finding the centroid. Although it is widely used in positioning systems, results aggregation is
rarely discussed separately while commonly regarded as an inherent functionality of localization algorithms. However, the impact of results aggregation is non-trivial and whether it could actually improve the overall performance of fingerprinting is still an open question.

In this thesis, we implemented results aggregation as an individual component and evaluate its influences. For the ease of analysis and explain, here we only consider double results aggregation. When results aggregation is enabled, the localization engine would report the two most likely locations of the user and the aggregated result is given by finding their midpoint.

## 5.2 Results and Analysis

![Cell deployments on the second floor for evaluations](image)

Evaluations are based on data collected from the second floor where 145 cells are deployed, presented in Fig 5.3. As shewed in the figure, measured cells on second floor are not only deployed on corridors but also a large open space, covered by 45 cells numbered from 10 to 54. Also the second floor has the highest number of visible AP during the measurement, thus it is the perfect choice to be the evaluation testbed.

At each cell, at least 20 samples are collected in our site-survey measurements. For simulations, the first 15 samples at each cell are used as **off-line** training data and the
rest samples are used as on-line readings. Raw data is aggregated to fingerprints from distinct AP.

If more than 5 fingerprints are recorded at a candidate cell, the fingerprint selection component would select 5 most suitable fingerprints by the chosen method. In system simulations, we include every combination of localization engines and fingerprint selection methods. And for every system we evaluate its performance with and without results aggregation respectively.

The size of each cell is $1.6 \times 1.6$ by meters. Coordinates of centroids of deployed cells are stored in databases. We measure positioning errors by meters and estimations with errors less than 0.8 meter are considered to be exact.

### 5.2.1 System Performances

<table>
<thead>
<tr>
<th>System</th>
<th>Min</th>
<th>Max</th>
<th>Median</th>
<th>Mean</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constancy</td>
<td>0</td>
<td>30.47</td>
<td>3.241</td>
<td>6.127</td>
<td>37 (26%)</td>
</tr>
<tr>
<td>Constancy.A</td>
<td>0</td>
<td>25.14</td>
<td>4</td>
<td>6.665</td>
<td>8 (6%)</td>
</tr>
<tr>
<td>Stability</td>
<td>0</td>
<td>50.11</td>
<td>16.32</td>
<td>16.82</td>
<td>10 (7%)</td>
</tr>
<tr>
<td>Stability.A</td>
<td>0.8</td>
<td>51.14</td>
<td>15.9</td>
<td>16.08</td>
<td>3 (2%)</td>
</tr>
<tr>
<td>Strength</td>
<td>0</td>
<td>28.62</td>
<td>4.663</td>
<td>6.99</td>
<td>33 (23%)</td>
</tr>
<tr>
<td>Strength.A</td>
<td>0</td>
<td>29.34</td>
<td>5.122</td>
<td>7.501</td>
<td>8 (6%)</td>
</tr>
<tr>
<td>Variance</td>
<td>0</td>
<td>62.01</td>
<td>9.86</td>
<td>14.1</td>
<td>16 (11%)</td>
</tr>
<tr>
<td>Variance.A</td>
<td>0.8</td>
<td>61.01</td>
<td>8.331</td>
<td>13.81</td>
<td>2 (1%)</td>
</tr>
<tr>
<td>Coverage</td>
<td>0</td>
<td>58.78</td>
<td>20.8</td>
<td>20.56</td>
<td>6 (4%)</td>
</tr>
<tr>
<td>Coverage.A</td>
<td>0</td>
<td>59.12</td>
<td>20.71</td>
<td>20.04</td>
<td>3 (2%)</td>
</tr>
<tr>
<td>Constancy+Strength</td>
<td>0</td>
<td>25.6</td>
<td>2.731</td>
<td>6.025</td>
<td>44 (31%)</td>
</tr>
<tr>
<td>Constancy+Strength.A</td>
<td>0</td>
<td>25.14</td>
<td>4.079</td>
<td>6.423</td>
<td>10 (7%)</td>
</tr>
<tr>
<td>Constancy+Stability</td>
<td>0</td>
<td>30.47</td>
<td>3.2</td>
<td>5.912</td>
<td>38 (26%)</td>
</tr>
<tr>
<td>Constancy+Stability.A</td>
<td>0</td>
<td>25.14</td>
<td>4.498</td>
<td>6.955</td>
<td>8 (6%)</td>
</tr>
</tbody>
</table>

Table 5.1: Statistics of performances of Gaussian Modeling

Fig 5.4 visualizes the performance of Gaussian modeling method with ECDF and spatial distributions of errors. We can clearly observe that fingerprint selection methods have significant impacts. Gaussian modeling working with Constancy, Strength, Constancy+Strength and Constancy+Stability selecting methods share very similar pat-
terns of performances and outperform the rest methods. Detailed statistics of those results are presented in Table 5.1, where the affix .A represents enabled results aggregation. Results show that Constancy+Strength method without results aggregation outperforms other systems by every criterion except the maximum error, where Constancy.A, Constancy+Strength.A and Constancy+Stability.A outperform it slightly.
Figure 5.5: Performances of Log-normal Modeling
<table>
<thead>
<tr>
<th>System</th>
<th>Min</th>
<th>Max</th>
<th>Median</th>
<th>Mean</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constancy</td>
<td>0</td>
<td>30.47</td>
<td>3.409</td>
<td>6.273</td>
<td>37 (26%)</td>
</tr>
<tr>
<td>Constancy.A</td>
<td>0</td>
<td>28.8</td>
<td>4.168</td>
<td>6.821</td>
<td>9 (6%)</td>
</tr>
<tr>
<td>Stability</td>
<td>0</td>
<td>50.11</td>
<td>16.77</td>
<td>17.23</td>
<td>9 (6%)</td>
</tr>
<tr>
<td>Stability.A</td>
<td>0.8</td>
<td>38.24</td>
<td>15.67</td>
<td>15.88</td>
<td>4 (3%)</td>
</tr>
<tr>
<td>Strength</td>
<td>0</td>
<td>29.61</td>
<td>4.052</td>
<td>7.015</td>
<td>34 (24%)</td>
</tr>
<tr>
<td>Strength.A</td>
<td>0</td>
<td>28.15</td>
<td>5.6</td>
<td>7.543</td>
<td>9 (6%)</td>
</tr>
<tr>
<td>Variance</td>
<td>0</td>
<td>62.01</td>
<td>9.86</td>
<td>14.21</td>
<td>16 (11%)</td>
</tr>
<tr>
<td>Variance.A</td>
<td>0.8</td>
<td>61.01</td>
<td>8.578</td>
<td>13.83</td>
<td>2 (1%)</td>
</tr>
<tr>
<td>Coverage</td>
<td>0</td>
<td>58.78</td>
<td>22.2</td>
<td>21.56</td>
<td>6 (4%)</td>
</tr>
<tr>
<td>Coverage.A</td>
<td>0</td>
<td>59.12</td>
<td>20.82</td>
<td>20.43</td>
<td>3 (2%)</td>
</tr>
<tr>
<td>Constancy+Strength</td>
<td>0</td>
<td>25.6</td>
<td>3.2</td>
<td>5.925</td>
<td>44 (31%)</td>
</tr>
<tr>
<td>Constancy+Strength.A</td>
<td>0</td>
<td>25.14</td>
<td>4.258</td>
<td>6.639</td>
<td>10 (7%)</td>
</tr>
<tr>
<td>Constancy+Stability</td>
<td>0</td>
<td>30.47</td>
<td>3.578</td>
<td>6.505</td>
<td>38 (26%)</td>
</tr>
<tr>
<td>Constancy+Stability.A</td>
<td>0</td>
<td>28.8</td>
<td>4.8</td>
<td>6.942</td>
<td>8 (6%)</td>
</tr>
</tbody>
</table>

**Table 5.2: Statistics of performances of Log-normal Modeling**

Fig 5.5 and Table 5.2 present performances of systems based on log-normal modeling. An interesting finding is that log-normal modeling systems have very similar behaviors and estimation results with Gaussian modeling systems. And Constancy+Strength also seems to be the best configuration for the system.
Figure 5.6: Performances of Nearest Neighbor Euclidean distance

Manhattan distance and Euclidean distance are closely related since they are both metrics for $L^p$ space and instances of Minkowski distance. Intuitive assumption it that they may have similar behaviors applied to fingerprinting systems.

Fig 5.6 and Table 5.3 present performances of systems based on NN with Euclidean distance. And Fig 5.7 and Table 5.4 show performances of systems based on NN with Manhattan distance. Figures show a clear similarity of patterns of errors from two methods. Statistics show that for Minkowski distances methods, constancy+Strength is the best configuration, providing best results in all features except maximum error which may be sightly outperformed by other configurations. The results also show that Manhattan distance has better performances than Euclidean distance, which is more commonly used as distance metric in fingerprinting.
Figure 5.7: Performances of Nearest Neighbor Manhattan distance
<table>
<thead>
<tr>
<th>System</th>
<th>Min</th>
<th>Max</th>
<th>Median</th>
<th>Mean</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constancy</td>
<td>0</td>
<td>30.4</td>
<td>2.731</td>
<td>5.684</td>
<td>43 (30%)</td>
</tr>
<tr>
<td>Constancy.A</td>
<td>0</td>
<td>29.18</td>
<td>3.697</td>
<td>5.974</td>
<td>13 (9%)</td>
</tr>
<tr>
<td>Stability</td>
<td>0</td>
<td>40.64</td>
<td>12.8</td>
<td>15.28</td>
<td>12 (8%)</td>
</tr>
<tr>
<td>Stability.A</td>
<td>0</td>
<td>33.14</td>
<td>12.05</td>
<td>13.07</td>
<td>3 (2%)</td>
</tr>
<tr>
<td>Strength</td>
<td>0</td>
<td>30.06</td>
<td>3.2</td>
<td>5.892</td>
<td>37 (26%)</td>
</tr>
<tr>
<td>Strength.A</td>
<td>0</td>
<td>29.34</td>
<td>4.025</td>
<td>6.015</td>
<td>14 (10%)</td>
</tr>
<tr>
<td>Variance</td>
<td>0</td>
<td>61.49</td>
<td>9.919</td>
<td>14.24</td>
<td>14 (10%)</td>
</tr>
<tr>
<td>Variance.A</td>
<td>0</td>
<td>61.24</td>
<td>9.515</td>
<td>13.89</td>
<td>4 (3%)</td>
</tr>
<tr>
<td>Coverage</td>
<td>0</td>
<td>53.07</td>
<td>21.92</td>
<td>20.62</td>
<td>8 (6%)</td>
</tr>
<tr>
<td>Coverage.A</td>
<td>0.8</td>
<td>53.86</td>
<td>22.25</td>
<td>20.77</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>Constancy+Strength</td>
<td>0</td>
<td>24.58</td>
<td>2.263</td>
<td>5.1</td>
<td>47 (33%)</td>
</tr>
<tr>
<td>Constancy+Strength.A</td>
<td>0</td>
<td>24</td>
<td>3.693</td>
<td>5.456</td>
<td>13 (9%)</td>
</tr>
<tr>
<td>Constancy+Stability</td>
<td>0</td>
<td>30.4</td>
<td>2.263</td>
<td>5.725</td>
<td>44 (31%)</td>
</tr>
<tr>
<td>Constancy+Stability.A</td>
<td>0</td>
<td>29.18</td>
<td>4.065</td>
<td>6.221</td>
<td>10 (7%)</td>
</tr>
</tbody>
</table>

Table 5.3: Statistics of performances of Nearest Neighbor Euclidean distance

<table>
<thead>
<tr>
<th>System</th>
<th>Min</th>
<th>Max</th>
<th>Median</th>
<th>Mean</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constancy</td>
<td>0</td>
<td>27.38</td>
<td>1.6</td>
<td>4.705</td>
<td>50 (35%)</td>
</tr>
<tr>
<td>Constancy.A</td>
<td>0</td>
<td>29.18</td>
<td>3.2</td>
<td>5.309</td>
<td>17 (12%)</td>
</tr>
<tr>
<td>Stability</td>
<td>0</td>
<td>50.79</td>
<td>11.65</td>
<td>14.33</td>
<td>14 (10%)</td>
</tr>
<tr>
<td>Stability.A</td>
<td>0.8</td>
<td>50.44</td>
<td>11.31</td>
<td>12.23</td>
<td>3 (2%)</td>
</tr>
<tr>
<td>Strength</td>
<td>0</td>
<td>30.06</td>
<td>3.2</td>
<td>5.678</td>
<td>38 (26%)</td>
</tr>
<tr>
<td>Strength.A</td>
<td>0</td>
<td>29.34</td>
<td>4.194</td>
<td>5.915</td>
<td>18 (13%)</td>
</tr>
<tr>
<td>Variance</td>
<td>0</td>
<td>61.49</td>
<td>9.919</td>
<td>14.24</td>
<td>14 (10%)</td>
</tr>
<tr>
<td>Variance.A</td>
<td>0</td>
<td>61.24</td>
<td>9.515</td>
<td>13.86</td>
<td>4 (3%)</td>
</tr>
<tr>
<td>Coverage</td>
<td>0</td>
<td>53.07</td>
<td>21.92</td>
<td>20.62</td>
<td>8 (6%)</td>
</tr>
<tr>
<td>Coverage.A</td>
<td>0.8</td>
<td>53.86</td>
<td>22.25</td>
<td>20.77</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>Constancy+Strength</td>
<td>0</td>
<td>24.58</td>
<td>1.6</td>
<td>5.16</td>
<td>51 (35%)</td>
</tr>
<tr>
<td>Constancy+Strength.A</td>
<td>0</td>
<td>25.13</td>
<td>2.707</td>
<td>4.898</td>
<td>16 (11%)</td>
</tr>
<tr>
<td>Constancy+Stability</td>
<td>0</td>
<td>27.38</td>
<td>1.931</td>
<td>5.135</td>
<td>48 (33%)</td>
</tr>
<tr>
<td>Constancy+Stability.A</td>
<td>0.8</td>
<td>29.18</td>
<td>3.394</td>
<td>5.503</td>
<td>12 (8%)</td>
</tr>
</tbody>
</table>

Table 5.4: Statistics of performances of Nearest Neighbor Manhattan distance
As explained previously, compared to other methods, NN Mahalanobis distance and cosine similarity are less tolerate for AP invisibility and when encountering such situations, the system may report NULL results. For systems based on those two engines, we record the number of NULL results reported. Obviously, NULL estimation is directly related to the chosen fingerprint selection method, thus it is a useful feature to evaluate fingerprint selection components as well.

Fig 5.8 and Table 5.5 show results from systems based on NN Mahalanobis distance. And Fig 5.8 with Table 5.5 show results from systems based on cosine similarity. Interrupted lines shewed in figures of errors distributions indicate NULL estimations. Statistics show that NULL is directly affected by fingerprint selections. Systems with Constancy, Stability and hybrid selection methods do not report NULL values. The rest
configurations would cause *NULL* reports, especially *Coverage* and *Variance* methods.

Here, we can observe benefits from results aggregation for both engines. For majority of configurations, results aggregation could reduce the maximum errors as shewed in the tables, which are not valid for other engines.

Figures of NN Mahalanobis distance based systems show similar patterns with former systems, while behaviors of cosine similarity systems are very different from others. For other algorithms, differences in performance working with varied configurations are significant where *Constancy*, *Strength* and hybrid methods clearly outperform the rest methods. However, in cosine similarity systems, differences are reduced especially in maximum errors.
Table 5.5: Statistics of performances of Nearest Neighbor Mahalanobis distance

<table>
<thead>
<tr>
<th>System</th>
<th>Min</th>
<th>Max</th>
<th>Median</th>
<th>Mean</th>
<th>Exact</th>
<th>NULL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constancy</td>
<td>0</td>
<td>30.4</td>
<td>3.2</td>
<td>5.034</td>
<td>38 (26%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Constancy.A</td>
<td>0</td>
<td>23.2</td>
<td>4</td>
<td>6.115</td>
<td>9 (6%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Stability</td>
<td>0</td>
<td>50.11</td>
<td>13.67</td>
<td>15.51</td>
<td>14 (10%)</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>Stability.A</td>
<td>0</td>
<td>51.14</td>
<td>13.77</td>
<td>14.52</td>
<td>6 (4%)</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>Strength</td>
<td>0</td>
<td>28.62</td>
<td>3.22</td>
<td>6.156</td>
<td>36 (25%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Strength.A</td>
<td>0</td>
<td>29.34</td>
<td>5.483</td>
<td>6.851</td>
<td>8 (6%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Variance</td>
<td>0</td>
<td>62.01</td>
<td>8</td>
<td>11.7</td>
<td>32 (22%)</td>
<td>19 (13%)</td>
</tr>
<tr>
<td>Variance.A</td>
<td>0</td>
<td>61.01</td>
<td>7.376</td>
<td>11.12</td>
<td>22 (15%)</td>
<td>19 (13%)</td>
</tr>
<tr>
<td>Coverage</td>
<td>0</td>
<td>53.07</td>
<td>20.35</td>
<td>19.33</td>
<td>31 (22%)</td>
<td>24 (17%)</td>
</tr>
<tr>
<td>Coverage.A</td>
<td>0.8</td>
<td>53.86</td>
<td>19.01</td>
<td>17.67</td>
<td>26 (18%)</td>
<td>24 (17%)</td>
</tr>
<tr>
<td>Constancy+Strength</td>
<td>0</td>
<td>31.44</td>
<td>2.187</td>
<td>5.149</td>
<td>45 (31%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Constancy+Strength.A</td>
<td>0.8</td>
<td>24</td>
<td>4</td>
<td>5.863</td>
<td>10 (7%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Constancy+Stability</td>
<td>0</td>
<td>30.4</td>
<td>2.925</td>
<td>5.342</td>
<td>40 (28%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Constancy+Stability.A</td>
<td>0</td>
<td>26.74</td>
<td>4.079</td>
<td>5.902</td>
<td>10 (7%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

Table 5.6: Statistics of performances of cosine similarity

<table>
<thead>
<tr>
<th>System</th>
<th>Min</th>
<th>Max</th>
<th>Median</th>
<th>Mean</th>
<th>Exact</th>
<th>NULL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constancy</td>
<td>0</td>
<td>33.35</td>
<td>13.19</td>
<td>13.6</td>
<td>6 (4%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Constancy.A</td>
<td>0</td>
<td>33</td>
<td>14.79</td>
<td>13.7</td>
<td>6 (4%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Stability</td>
<td>0</td>
<td>40.31</td>
<td>13.67</td>
<td>14.5</td>
<td>9 (6%)</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>Stability.A</td>
<td>0</td>
<td>40.31</td>
<td>13.02</td>
<td>13.5</td>
<td>9 (6%)</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>Strength</td>
<td>0</td>
<td>33.35</td>
<td>13.6</td>
<td>14.07</td>
<td>4 (3%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Strength.A</td>
<td>0</td>
<td>31.46</td>
<td>15.38</td>
<td>14.16</td>
<td>8 (6%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Variance</td>
<td>0</td>
<td>41.6</td>
<td>16</td>
<td>18.82</td>
<td>21 (15%)</td>
<td>19 (13%)</td>
</tr>
<tr>
<td>Variance.A</td>
<td>0.8</td>
<td>44.8</td>
<td>16.93</td>
<td>17.58</td>
<td>20 (14%)</td>
<td>19 (13%)</td>
</tr>
<tr>
<td>Coverage</td>
<td>0</td>
<td>43.2</td>
<td>26.24</td>
<td>24.29</td>
<td>26 (18%)</td>
<td>24 (17%)</td>
</tr>
<tr>
<td>Coverage.A</td>
<td>0.8</td>
<td>38.7</td>
<td>25.85</td>
<td>24.2</td>
<td>25 (17%)</td>
<td>24 (17%)</td>
</tr>
<tr>
<td>Constancy+Strength</td>
<td>0</td>
<td>33.35</td>
<td>13.25</td>
<td>13.8</td>
<td>4 (3%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Constancy+Strength.A</td>
<td>0</td>
<td>33</td>
<td>14.13</td>
<td>13.76</td>
<td>4 (3%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Constancy+Stability</td>
<td>0</td>
<td>33.35</td>
<td>13.42</td>
<td>13.85</td>
<td>5 (3%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Constancy+Stability.A</td>
<td>0</td>
<td>33</td>
<td>14.9</td>
<td>13.71</td>
<td>6 (4%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>
Another distinction of cosine similarity algorithm is that \textit{Constancy}+\textit{Strength} is not the best configuration for it as for other algorithms. Surprisingly, fingerprint selection based on \textit{Coverage}, which performs badly in other systems, reports the most exact estimations working with cosine similarity. And \textit{Constancy} seems to be the best selection method regarding other features.

Figure 5.10: Performances of varied localization engines with optimized configurations
Now we compare performances of localization engines with their optimized configurations. Results aggregation is not enabled in all systems. Based on former results, we choose Constancy to be the fingerprint selections method for cosine similarity and choose Constancy+Strength for other engines. The comparison is presented in Fig 5.10 with ECDF and spatial distributions of errors. Results show that cosine similarity has poorer performance compared to others whose performances are much similar. We retrieve key statistics features of positioning performances, maximum, median and mean errors of each systems and compare them in Fig 5.11.

5.2.2 Correlations of Localization Performances

Here we analysis results of varied fingerprinting systems from novel perspectives. In previous contexts, comparison between different fingerprinting systems are mainly conducted by analyzing probability distributions of errors. However, a question rarely answered is whether those errors are inherent in localization systems, or caused by collected data of fingerprints. And another related question is whether positioning performances are related to physical locations, i.e., whether different locations in WLAN environments have different feasibilities for fingerprinting.

Those questions could be addressed by analyzing correlations of localization per-
performances from varied systems, through which we can verify whether positioning er-
rors are more dependent on systems or environments. Here single estimations from
localization engines under optimized configurations are considered.

<table>
<thead>
<tr>
<th></th>
<th>Log-normal</th>
<th>NN-Euclidean</th>
<th>NN-Manhattan</th>
<th>NN-Mahalanobis</th>
<th>Cosine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>0.959</td>
<td>0.748</td>
<td>0.805</td>
<td>0.650</td>
<td>-0.037</td>
</tr>
<tr>
<td>Log-normal</td>
<td>———</td>
<td>0.785</td>
<td>0.840</td>
<td>0.602</td>
<td>0.019</td>
</tr>
<tr>
<td>NN-Euclidean</td>
<td>———</td>
<td>———</td>
<td>0.758</td>
<td>0.667</td>
<td>0.023</td>
</tr>
<tr>
<td>NN-Manhattan</td>
<td>———</td>
<td>———</td>
<td>———</td>
<td>0.542</td>
<td>0.034</td>
</tr>
<tr>
<td>NN-Mahalanobis</td>
<td>———</td>
<td>———</td>
<td>———</td>
<td>———</td>
<td>-0.123</td>
</tr>
</tbody>
</table>

Table 5.7: Pearson correlation coefficients of errors for varied systems

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Coincide</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian–Lognormal</td>
<td>134 (93%)</td>
<td>0</td>
<td>0.674</td>
<td>25.6</td>
</tr>
<tr>
<td>Gaussian–Euclidean</td>
<td>94 (65%)</td>
<td>0</td>
<td>2.527</td>
<td>24</td>
</tr>
<tr>
<td>Gaussian–Manhattan</td>
<td>93 (65%)</td>
<td>0</td>
<td>2.542</td>
<td>24.98</td>
</tr>
<tr>
<td>Gaussian–Mahalanobis</td>
<td>89 (62%)</td>
<td>0</td>
<td>2.942</td>
<td>31.07</td>
</tr>
<tr>
<td>Gaussian–Cosine</td>
<td>5 (3%)</td>
<td>13.67</td>
<td>14.13</td>
<td>36.42</td>
</tr>
<tr>
<td>Lognormal–Euclidean</td>
<td>96 (67%)</td>
<td>0</td>
<td>2.173</td>
<td>22.18</td>
</tr>
<tr>
<td>Lognormal–Manhattan</td>
<td>93 (65%)</td>
<td>0</td>
<td>2.489</td>
<td>24.98</td>
</tr>
<tr>
<td>Lognormal–Mahalanobis</td>
<td>87 (60%)</td>
<td>0</td>
<td>3.389</td>
<td>31.07</td>
</tr>
<tr>
<td>Lognormal–Cosine</td>
<td>5 (3%)</td>
<td>13.49</td>
<td>13.9</td>
<td>36.42</td>
</tr>
<tr>
<td>Euclidean–Manhattan</td>
<td>113 (78%)</td>
<td>0</td>
<td>2.442</td>
<td>24.98</td>
</tr>
<tr>
<td>Euclidean–Mahalanobis</td>
<td>82 (27%)</td>
<td>0</td>
<td>3.352</td>
<td>34.35</td>
</tr>
<tr>
<td>Euclidean–Cosine</td>
<td>6 (4%)</td>
<td>13.6</td>
<td>14.02</td>
<td>36.42</td>
</tr>
<tr>
<td>Manhattan–Mahalanobis</td>
<td>80 (56%)</td>
<td>0</td>
<td>3.751</td>
<td>31.44</td>
</tr>
<tr>
<td>Manhattan–Cosine</td>
<td>6 (4%)</td>
<td>12.9</td>
<td>13.66</td>
<td>36.42</td>
</tr>
<tr>
<td>Mahalanobis–Cosine</td>
<td>3 (2%)</td>
<td>14.4</td>
<td>14.64</td>
<td>36.42</td>
</tr>
</tbody>
</table>

Table 5.8: Comparisons of location estimations from varied systems

The analysis is conducted in two ways. First we analysis positioning errors mea-
sured by meters from varied systems and compute Pearson correlation coefficients be-
tween them. Thus we can verify whether accuracy of localization from different sys-
tems are related and dependent on the environment. Results are shewed in Table 5.7,
which back our conclusions in previous sections. Performances of Gaussian and log-normal modeling are closely related. And cosine similarity performs quite differently from all other systems, showing irrelevancy at every comparisons. Surprisingly, Mahalanobis distance metric which shares many similarities with parametric modeling methods, actually shows least correlations to them and poorer performance compared to other distance metrics.

Apart from the accuracy of positioning, it is also necessary to compare the location estimations from different systems themselves. Different estimations may lead to same level of errors. To study relations between systems, we need to explore how differently those systems estimate. To address this question, we extract raw estimation results from varied systems and compare their raw estimations for same user’s inputs. Results are presented in Table 5.8, where the Coincide column records numbers of coincide estimations, and Median, Mean and Max columns presents statistical features of distances between estimations from different systems based on same off-line readings. Those results cohere with results from correlation analysis. All systems except cosine similarity give coincide estimations most of the time. However, for each pair of systems, there are situations when systems estimate quit differently.

### 5.2.3 Cross Results Aggregation

Based on observations and analysis made in former sections, a question is raised that how to take advantages of multiple localization engines which may provide different location estimations. An intuitive answer to this question is cross results aggregation. Since individual systems may report huge errors and systems sometimes may estimate quite differently, it maybe useful to aggregate results from different systems to compensate errors.

Following this intuition, we implement cross results aggregation in our simulations. For the ease of analysis, we use the same double results aggregation method used in former evaluations, which aggregate two estimated locations by finding their midpoint. In cross results aggregation, we retrieve the most likely location from each chosen system.

We test performances of cross results aggregation with every combination of the 6 optimized systems. Results are showed in Table 5.9, which show there is no significant improvement. Overall performance still remains at the same level as individual systems.
<table>
<thead>
<tr>
<th>Combination</th>
<th>Min</th>
<th>Max</th>
<th>Median</th>
<th>Mean</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian–Lognormal</td>
<td>0</td>
<td>25.6</td>
<td>2.331</td>
<td>5.849</td>
<td>43 (30%)</td>
</tr>
<tr>
<td>Gaussian–Euclidean</td>
<td>0</td>
<td>24.58</td>
<td>2.263</td>
<td>5.406</td>
<td>41 (28%)</td>
</tr>
<tr>
<td>Gaussian–Manhattan</td>
<td>0</td>
<td>24.58</td>
<td>2.263</td>
<td>5.346</td>
<td>44 (31%)</td>
</tr>
<tr>
<td>Gaussian–Mahalanobis</td>
<td>0</td>
<td>24.8</td>
<td>2.465</td>
<td>5.444</td>
<td>35 (24%)</td>
</tr>
<tr>
<td>Gaussian–Cosine</td>
<td>0</td>
<td>23.04</td>
<td>8.708</td>
<td>8.778</td>
<td>3 (2%)</td>
</tr>
<tr>
<td>Lognormal–Euclidean</td>
<td>0</td>
<td>24.58</td>
<td>2.4</td>
<td>5.419</td>
<td>42 (29%)</td>
</tr>
<tr>
<td>Lognormal–Manhattan</td>
<td>0</td>
<td>24.58</td>
<td>2.263</td>
<td>5.235</td>
<td>46 (32%)</td>
</tr>
<tr>
<td>Lognormal–Mahalanobis</td>
<td>0</td>
<td>24.8</td>
<td>2.4</td>
<td>5.267</td>
<td>38 (26%)</td>
</tr>
<tr>
<td>Lognormal–Cosine</td>
<td>0</td>
<td>23.04</td>
<td>8.099</td>
<td>8.715</td>
<td>3 (2%)</td>
</tr>
<tr>
<td>Euclidean–Manhattan</td>
<td>0</td>
<td>24.58</td>
<td>2.187</td>
<td>4.796</td>
<td>47 (33%)</td>
</tr>
<tr>
<td>Euclidean–Mahalanobis</td>
<td>0</td>
<td>24.58</td>
<td>4.802</td>
<td>6.803</td>
<td>41 (28%)</td>
</tr>
<tr>
<td>Euclidean–Cosine</td>
<td>0</td>
<td>22.13</td>
<td>7.96</td>
<td>8.256</td>
<td>5 (3%)</td>
</tr>
<tr>
<td>Manhattan–Mahalanobis</td>
<td>0</td>
<td>24.58</td>
<td>2.263</td>
<td>4.832</td>
<td>40 (28%)</td>
</tr>
<tr>
<td>Manhattan–Cosine</td>
<td>0</td>
<td>23.04</td>
<td>8.099</td>
<td>8.39</td>
<td>8 (6%)</td>
</tr>
<tr>
<td>Mahalanobis–Cosine</td>
<td>0</td>
<td>24.01</td>
<td>8.099</td>
<td>8.358</td>
<td>3 (2%)</td>
</tr>
</tbody>
</table>

Table 5.9: Summary of performances of cross results aggregation

Marginal Benefits from cross results aggregation can be found on maximum errors since combinations like Euclidean–Cosine achieve maximum errors slightly lower than individual systems. However, this enhancement seems trivial and better schemes to take advantages of varied systems and further optimization of performances are yet to be found.
This study has deepened our understandings of WLAN location fingerprinting and its limits by analyzing the engineering of fingerprints itself as well as evaluation of various fingerprinting systems and approaches. This study is based on both stationary and site survey WLAN scanning measurements collected in an office/university environment.

It includes analysis of previously ignored features of WLAN fingerprints that have significant influences on fingerprinting systems. Contrary to common assumptions made by previous works, AP visibilities are found to be instable, showing significant temporal and spatial incoherence. The time series for number of visible APs (NVA) is extracted from stationary measurements and the results show that AP visibilities are heavily time dependent and affected by hardware heterogeneity, and NVA time series received at same devices, even from different periods of time, demonstrate similar patterns. Statistics clearly show that AP visibilities received on phone are more constant than that on laptop. By analyzing spatial distributions of AP visibilities from site-survey data, we find that continuous coverage is not common. Negative effects of VAP on location discriminability are validated by analyzing spatial and temporal similarities of signals from VAP hosted on same physical AP platforms. The results show that VAP on same devices provide different but very similar signals, which would hurt localization performance.

Interesting findings are also obtained from analyzing RSSI time series. LOS is shown to have negative influences since RSSI from AP without LOS paths to receivers tend to be more normally distributed and have less variations — the ideal properties for constructing location fingerprints. Normality and skewness tests are applied on RSSI series received at different locations and the results show that a greater proportion of locations witness skewed RSSI distributions. We study correlations of different
features of WLAN fingerprints and find signals with higher strength tend to have more constant visibilities and are less normally distributed.

Based on site-survey measurements and emulations in R, systematic evaluation of fingerprinting systems is conducted with focus on three main functionalities: fingerprint selections, localization engines and results aggregation. All combinations of the methods chosen for these functionalities are included to identify optimized systems. The results show that fingerprint selections have significant impact with the proposed Constancy+Strength method outperforming the rest, and results aggregation is not beneficial for fingerprinting unlike with other positioning systems.

Popular methods including parametric modeling (Gaussian and log-normal models) and NN (Euclidean, Manhattan, Mahalanobis distances) as well as cosine similarity are considered, where NN with Manhattan distance provides best performance. We find popular methods show similar behaviors and their location estimates are heavily correlated. The intuitive idea of cross results aggregation to compensate errors from individual systems is found to be ineffective. The results demonstrate that localization engines play an important but limited role in fingerprinting. However, positioning errors are dependent on the environment and constructed fingerprints. Thus the refined engineering of fingerprints is equally important as advanced localization systems for further development of WLAN location fingerprinting.

### 6.1 Future Work

Future work is needed to complement and support results of this study.

- Additional measurements collected at different spaces and environments are needed to validate observations from our experiments for generalizing our conclusions.

- Emulations of fingerprinting systems can be improved to include added functionalities and methods. Spatial clustering is a popular component in modern schemes which divides the space of interests into subspaces before fine grained positioning and its feasibility and practicality need to be verified. In recent years, sophisticated localization algorithms based on machine learning and compressed sensing have been proposed and it is helpful to test their behaviors.

- To further validate impacts of VAP, comparative evaluations should be conducted
with modified fingerprints datasets where VAP hosted on same physical platforms are filtered out.

Based on conclusions and observations made in this study, we suggest following ideas to address remaining challenges in this area.

- **Fingerprints clustering**: AP with similar patterns of signaling could reduce location discriminability, such like VAP hosted on same devices. Methods that cluster fingerprints could detect VAP as well as AP with similar patterns, thus selecting AP with distinct features and increasing efficiency and performance of localization.

- **Errors modeling and prediction**: Since positioning errors in fingerprinting systems are dependent on the environment and collected fingerprints, it is possible to model the errors based on data characteristics or spatial features, thus predicting localization accuracy given the sampled location or collected measurements, making fingerprinting systems more adaptive and flexible.
Bibliography


[65] Metageek LLC. inSSIDer homepage and description.


