



THE HONG KONG
POLYTECHNIC UNIVERSITY

香港理工大學

Pao Yue-kong Library

包玉剛圖書館

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

**DEVELOPING PUBLIC TRANSPORT INFORMATION
SYSTEMS WITH USE OF SMARTPHONE-BASED
HUMAN PROBE DATA**

PIYANIT WEPULANON

PhD

The Hong Kong Polytechnic University

2020

The Hong Kong Polytechnic University
Department of Civil and Environmental Engineering

**Developing Public Transport Information Systems with
Use of Smartphone-Based Human Probe Data**

PIYANIT WEPULANON

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

October, 2019

Certificate of originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Piyanit Wepulanon

Abstracts

This thesis contributes to the development of public transport information systems with use of human probe data. New methods for estimating three key performance indicators (KPIs) of bus transit systems are proposed in this study; namely, average bus passenger waiting times, real-time bus arrival times, and bus crowding levels. Without the data availability from in-vehicle sensing devices, smartphone-based human probe data are considered as the potential alternative data sources for estimating the three KPIs of bus transit system.

This thesis firstly focuses on the non-participatory sensing approach, passive Wi-Fi data. Through a detailed investigation, this thesis aims to understand the capabilities and limitations of using passive Wi-Fi data for deriving the bus transit information. The thesis proposes a new method for estimating the average bus passenger waiting times at a single bus stop. The proposed method is designed to handle massive noise and the temporal uncertainties of these data. In order to achieve this, generalized classification features are introduced for describing the attributes of individual Wi-Fi devices at a bus stop. The features can then be used for identifying waiting passengers' devices and estimating the average bus passenger waiting times at bus stop.

The thesis then proposes a novel framework for developing a real-time bus arrival time information system using participatory-based bus data contributed by bus passengers. This real-time information can be provided without the need for bus tracking devices and data provision from the bus operators. The proposed framework is developed to cope with the particular characteristics of participatory-based bus data such as inconsistencies in the bus data due to the participation of multiple passengers on the same bus, and the availability of bus data in the spatial and temporal dimensions.

Finally, the participatory-based bus data are used for developing a bus crowding prediction system. This system aims to predict the bus crowding levels of individual buses when they arrive at each bus stop. The crowding levels can be provided together with the bus arrival times so that passengers can make their boarding decisions with more relevant information when planning their journeys. Data mining techniques are adopted for the prediction of bus crowding levels based on several relevant factors, i.e. bus dwell times and bus headways at each bus stop. In the long run, the system developed in this thesis will be able to perform bus crowding prediction without the availability of passenger boarding and alighting information at bus stops.

Publications arising from the thesis

Journal papers

Wepulanon, P., Sumalee, A., & Lam, W. H. K. (2018). A real-time bus arrival time information system using crowd-sourced smartphone data: a novel framework and simulation experiments. *Transportmetrica B: Transport Dynamics*, 6(1), 34-53. (The content in Chapter 5 is mainly from this paper.)

Wepulanon, P., Sumalee, A., & Lam, W. H. K. (2019). Temporal signatures of passive Wi-Fi data for estimating bus passenger waiting time at a single bus stop. Paper accepted for publication in *IEEE Transactions on Intelligent Transportation Systems* (posted online: 12 July 2019). (The content in Chapter 4 is mainly from this paper.)

Wepulanon, P., Sumalee, A., & Lam, W. H. K. A bus crowding prediction system based on crowd-sourced smartphone data. To be submitted. (The content in Chapter 6 is mainly from this paper.)

Conference papers

Wepulanon, P., Lam, W. H. K., & Sumalee, A. (2017). Pedestrian facility usage monitoring using multiple sources of data. *Proceedings of the 22nd International Conference of Hong Kong Society for Transportation Studies*, 117-124.

Wepulanon, P., Lam, W. H. K., & Sumalee, A. (2016). Using multiple sources of data for monitoring facility usage on campus. *Proceedings of the 21st International Conference of Hong Kong Society for Transportation Studies*, 165-172.

Wepulanon, P., Sumalee, A., & Lam, W. H. K. (2015). Using in-campus Wi-Fi data for analysis of student activity sequence. *Proceedings of the 20th National Convention on Civil Engineering*.

Wepulanon, P., Sumalee, A., & Lam, W. H. K. (2014). Bus travel time estimation using crowd-source Wi-Fi Media Access Control addresses. *Proceedings of the 19th National Convention on Civil Engineering*.

Wepulanon, P., Sumalee, A., & Lam, W. H. K. (2013). Identification of public transit vehicle arrival with arrival time and dwell time estimation based on crowd-source Wi-Fi MAC address. *Proceedings of the 18th International Conference of Hong Kong Society for Transportation Studies*, 501-510.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, i.e. Prof. William H.K. Lam and Prof. Agachai Sumalee, for their supervision and patience throughout my PhD study. I have gained a lot of knowledge and great experiences from their kind guidance, particularly on how to be a good researcher and how to be successful in my future career.

I would like to acknowledge past and present fellow researchers in the transportation group of the Hong Kong Polytechnic University, including Dr. Karen Tam, Dr. Julio Ho, Dr. Renxin Zhong, Dr. Tianlu Pan, Dr. Jiankai Wang, and Dr. Teerapot Siripirote. I appreciate their support and still remember the very warm welcome when I started my PhD study. I would like to thank Thai students in Hong Kong (P'Tid, P'Volt, P'Neung, P'Jabby, P'Moo, Job, New, and Ohm) for sharing our experiences when we were far from our home.

Special thanks go to George, Ekalux Ua-areemitr, for the happiness you have brought into my life and all your support which got me through tough times. Also, I would like to express my gratitude to his family for their support.

Above all, I am deeply thankful to my parents for their unconditional love. I know I will always have their support no matter what I choose to pursue in my life.

Table of Contents

Certificate of originality	ii
Abstracts	iii
Publications arising from the thesis	iv
Acknowledgements	vi
Table of Contents	vii
List of Figures.....	xi
List of Tables	xiii
Chapter 1 Introduction and overview.....	1
1.1 Research motivation.....	1
1.2 Research objectives	3
1.3 Organization of the thesis.....	4
1.4 Research contributions	6
Chapter 2 Problem statements and literature review	10
2.1 Public transport information.....	10
2.1.1 Passenger waiting time	10
2.1.2 Bus arrival time prediction.....	13
2.1.3 Bus crowding	16
2.2 Smartphone-based human probe data	17
2.2.1 Wi-Fi.....	18
2.2.2 Bluetooth.....	21
2.2.3 Global Positioning System (GPS).....	22
2.2.4 Global System for Mobile Communications (GSM)	22
2.2.5 Motion sensors	23
2.3 Smartphone-based human probe data for public transport information systems	24
Chapter 3 Passive Wi-Fi monitoring: data characteristics and challenges	26
3.1 Introduction	26
3.2 Transformation of raw data	28

3.2.1 Raw data.....	28
3.2.2 Detection events.....	29
3.2.3 Session	30
3.2.4 Encounter	32
3.2.5 Trail.....	33
3.2.6 Dimensions of data analysis.....	34
3.3 Passive Wi-Fi data characteristics.....	35
3.4 Modeling uncertainties in passive Wi-Fi data based on a single Wi-Fi scanner.....	36
3.4.1 Uncertainties in device positioning.....	36
3.4.2 Uncertainties in detection periods.....	37
3.5 Other challenges in passive Wi-Fi monitoring.....	40
3.5.1 Determining configuration parameters for a Wi-Fi scanner	40
3.5.2 Long-term mobility tracking.....	41
3.5.3 System validation.....	41
3.5.4 Privacy	42
3.6 Summary of findings	48
Chapter 4 Passenger waiting time estimation at a single bus stop.....	49
4.1 Introduction	49
4.2 Background	51
4.3 System overviews.....	53
4.4 Data preparation	54
4.4.1 Primary Filtering.....	55
4.4.2 Activity Segmentation	55
4.4.3 Feature Extraction.....	55
4.4.4 Data Cleansing	57
4.5 AWT estimation	57
4.5.1 Classification features.....	58
4.5.2 Classifier	59
4.5.3 Two-stage classification.....	64
4.5.4 Average passenger waiting time	65
4.6 Empirical studies	65
4.6.1 Case study #1: Hong Kong	66
4.6.2 Case study #2: Bangkok.....	72

4.6.3	Limitations and suggestions for further development	75
4.7	Summary of findings	78
Chapter 5 A real-time bus arrival time information system based on crowd-sourced smartphone data.....		80
5.1	Introduction	80
5.2	System architecture	82
5.2.1	Smartphone application	83
5.2.2	Back-end processing	83
5.3	Bus data formulation	83
5.3.1	Bus route data	83
5.3.2	Crowd-sourced bus data.....	84
5.3.3	Historical information.....	84
5.4	Bus location filtering.....	86
5.4.1	Identifying bus running sequence	86
5.4.2	Bus location matching.....	87
5.5	Link travel time estimation	93
5.6	Bus arrival time prediction.....	94
5.6.1	Travel time prediction.....	94
5.7	Experimental studies	96
5.7.1	Case study #1: simulated bus services	96
5.7.2	Case study #2: real-world bus data	103
5.7.3	Limitations and suggestions for further development	105
5.8	Summary of findings.....	107
Chapter 6 A bus crowding prediction system based on crowd-sourced smartphone data		108
6.1	Introduction	108
6.2	Operational overviews.....	110
6.3	Problem formulation	111
6.4	Data preparation	112
6.4.1	Bus headways.....	112
6.4.2	Historical bus service pattern.....	114
6.5	Bus crowding estimation.....	114
6.5.1	Bus dwell time estimation.....	115
6.5.2	Bus crowding estimation based on bus dwell time.....	118

6.6	Bus crowding prediction	119
6.7	Experimental studies	121
6.7.1	Data description	121
6.7.2	Numerical results	123
6.8	Summary of findings	126
Chapter 7 Conclusions and recommendations for further research.....		129
7.1	Conclusions	129
7.2	Recommendations for further research	132
7.2.1	Extension of bus information systems based on passive Wi-Fi data	133
7.2.2	Extension of bus information systems based on crowd-sourced bus data	141
References.....		143
Appendix A Wi-Fi device discovery		155
A.1	Principle of Wi-Fi technology.....	155
A.2	Wi-Fi device discovery	158
A.2.1	Detection range	159
A.2.2	Monitoring cycle	159
Appendix B Experiments on Wi-Fi data characteristics.....		161
B.1	Equipment	161
B.2	Detection events	162
B.2.1	Experiment#1: observing detection events and event periods	163
B.3	Received Signal Strength Indicator	167
B.3.1	Experiment#2: observing RSSI.....	167
B.3.2	Experiment#3: observing RSSI with physical obstacles.....	171
B.4	Missed detection.....	174
B.4.1	Experiment#4: observing missed detection from various Wi-Fi devices	174
B.4.2	Experiment#5: observing missed detection from various distance	176
B.4.3	Experiment#6: observing missed detection based on different Wi-Fi scanner setups.....	178
B.4.4	Experiment#7: observing missed detection in a crowded space	180
Appendix C Examples of raw data.....		185
C.1	Passive Wi-Fi data.....	185
C.2	Observed passenger waiting times	186
C.3	GPS bus data	187
C.4	Passenger boarding and alighting at bus stops	188

List of Figures

Figure 1.1: Overall framework of the thesis	5
Figure 2.1: Bus trajectories in a time-space diagram.....	13
Figure 2.2: Differences between bus travel time estimation and prediction.....	14
Figure 3.1: Examples of captured Wi-Fi data.....	28
Figure 3.2: Detection events from a MAC-ID	29
Figure 3.3: An example of timeline visualization.....	32
Figure 3.4: The detection area divided into two circular zones.....	37
Figure 3.5: Time window constraints of activity duration.....	38
Figure 4.1: Overviews of the proposed system for AWT estimation at a single bus stop.....	53
Figure 4.2: Detection range of a Wi-Fi scanner installed at a bus stop area.....	54
Figure 4.3: Generalized classification features of the RSSI observations from a waiting passenger (MAC A), and a passing vehicle (MAC B).....	58
Figure 4.4: Time window constraints for identifying potential candidates	60
Figure 4.5: Bipartite graph representation	63
Figure 4.6: Data filtering performance (Case study #1)	67
Figure 4.7: AWT estimation accuracy for Case study #1	69
Figure 4.8: Data filtering performance (Case study #2)	73
Figure 4.9: AWT estimation accuracy for Case study #2.....	74
Figure 5.1: System architecture and operational overviews	82
Figure 5.2: Examples of candidate location determination	88
Figure 5.3: Example of a candidate graph	91
Figure 5.4: Examples of two overlapping bus routes on a simulated road network.....	97
Figure 5.5: The MAPE of bus arrival time prediction	102
Figure 6.1: Operational overviews.....	110
Figure 6.2: Candidate location formalization for dwell time estimation.....	116
Figure 6.3: A candidate graph for the bus dwell time estimation.....	117
Figure 7.1: Extension of device monitoring areas using two Wi-Fi scanners	133
Figure 7.2: Device outflows at a bus stop.....	135
Figure 7.3: A time-space diagram showing a vehicle trajectory and Wi-Fi signals from a passenger's device	137
Figure 7.4: Practical distance or spacing between two consecutive transit stops.....	137
Figure 7.5: A schematic map of the pedestrian tunnel.....	139
Figure 7.6: Available MAC-IDs for each 5-minute time period	139
Figure 7.7: Walking time estimation results	140
Figure A.1: A 48-bit MAC address.....	156
Figure A.2: Wi-Fi connection in (a) an infrastructure mode, and (b) an ad hoc mode.....	156

Figure A.3: The Wi-Fi discovery phase for (a) passive scanning and (b) active scanning ..	157
Figure B.1: A portable Wi-Fi scanner used in this study.....	162
Figure B.2: General equipment setup for the designed experiments.....	163
Figure B.3: Average event periods with an LPF.....	166
Figure B.4: The RSSI plot (an iPhone).....	168
Figure B.5: The RSSI plot (a Samsung smartphone).....	169
Figure B.6: The RSSI plot (a Samsung tablet)	169
Figure B.7: The RSSI plot (a Lenovo smartphone)	170
Figure B.8: The RSSI plots for the cases with and without obstacles (an iPhone).....	172
Figure B.9: The RSSI plots for the cases with and without obstacles (a Samsung smartphone).....	172
Figure B.10: The RSSI plots for the cases with and without obstacles (a Samsung tablet). 173	173
Figure B.11: The RSSI plots for the cases with and without obstacles (a Lenovo smartphone)	173
Figure B.12: Comparison of detection events based on the distance of Wi-Fi devices from a Wi-Fi scanner.....	176
Figure B.13: Wi-Fi monitoring using different channel-hopping velocities	177
Figure B.14: MAPE of occupancy time.....	180
Figure B.15: MAPE of occupancy time in a crowded space	182
Figure B.16: The proportion of detection events from Experiment#6 in each Wi-Fi channels.....	183
Figure B.17: The proportion of detection events from Experiment#7 in each Wi-Fi channels.....	183

List of Tables

Table 2.1: Summary of the previous studies using passive Wi-Fi data for public transport development.....	19
Table 3.1: The cases of time window constraints.....	39
Table 3.2: Examples of unencrypted SSIDs from a Wi-Fi device.....	45
Table 4.1: Summary of the previous studies on people dwell time/waiting time estimation using Wi-Fi data.....	52
Table 4.2: The information of a Wi-Fi record.....	56
Table 4.3: The case of time window constraints.....	60
Table 4.4: The conditions of two bus stops for system evaluation.....	66
Table 4.5: Probability distribution of passenger waiting times.....	71
Table 5.1: Candidate location formalization cases.....	90
Table 5.2: Bus location sampling methods.....	98
Table 5.3: Numerical results from the 17 bus datasets (Case study #1).....	100
Table 5.4: Numerical results from the 17 bus datasets (Case study #2).....	104
Table 6.1: Bus crowding by LOS categories.....	122
Table 6.2: Numerical results from the 11 bus datasets.....	124
Table 7.1: Device locations based on two Wi-Fi scanners.....	133
Table B.1: Summary of Wi-Fi device states during the experiment.....	164
Table B.2: Number of detection events during each device state.....	164
Table B.3: Summary of event periods.....	165
Table B.4: The mean and S.D. of RSSI (an iPhone).....	168
Table B.5: The mean and S.D. of RSSI (a Samsung smartphone).....	169
Table B.6: The mean and S.D. of RSSI (a Samsung tablet).....	169
Table B.7: The mean and S.D. of RSSI (a Lenovo smartphone).....	170
Table B.8: The mean and S.D. of the RSSI for the cases with and without obstacles (an iPhone).....	172
Table B.9: The mean and S.D. of the RSSI for the cases with and without obstacles (a Samsung smartphone).....	172
Table B.10: The mean and S.D. of the RSSI for the cases with and without obstacles (a Samsung tablet).....	173
Table B.11: The mean and S.D. of the RSSI for the cases with and without obstacles (a Lenovo smartphone).....	173
Table B.12: Evaluation of missed detection probability.....	174
Table B.13: An example of detection events from a Wi-Fi device.....	175
Table B.14: Wi-Fi monitoring capacity.....	178
Table B.15: Wi-Fi monitoring capacity in a crowded space.....	181

Chapter 1

Introduction and overview

1.1 Research motivation

Public transport is one of the sustainable solutions for improving accessibility and alleviating urban traffic congestion problems. In many developing and developed countries, the local governments have been attempting to encourage people to use public transport services for their daily travel within the urban areas by issuing supportive government policies and using advanced technologies (Buehler, 2009). The use of public transport is strongly affected by the quality of the public transport services, which can be evaluated in several dimensions such as information provision, service reliability, and comfort throughout the passengers' journeys (Redman et al., 2013).

Due to the advancement of sensing and communication technologies, Advanced Public Transportation Systems (APTS) have been implemented to improve the efficiency of public transport systems. The APTS applications have been developed based on six purposes: traveler information, fleet management, transit safety and security, transportation demand management, electronic fare payment, and intelligent vehicle systems (U.S. Department of Transportation, 2006).

Providing public transport information is one of the effective methods for improving passenger satisfaction, especially in Asian fast-growing cities where public transport is highly utilized and passenger journey lengths are increasing. In Hong Kong, public transport is responsible for about 90% of the total daily person trips (Transport Department, 2017). By providing timely public transport information, passengers can better plan their journeys resulting in a reduction of their total journey times.

Nevertheless, the provision of public transport information could be limited in some cities by various reasons. For instance, the Automatic Vehicle Location (AVL) systems which require a vehicle tracking device or sensor for tracking the real-time locations of each individual transit vehicle may not be implemented in some developing countries due to budget constraints. As a result of the densely populated development in Hong Kong, the bus transit system has been very successful in delivering the service while generating a profitable margin for its operation. The competitive nature of the bus market in Hong Kong seems to be a good case for bus service providers to improve their services. However, the current situation in Hong Kong discourages the bus operators from deploying the APTS and/or the integrated bus information systems. It is because each bus operator may run a risk of losing passengers to other bus companies if passengers would know in advance from the APTS which bus will arrive first at the bus stops. Moreover, the government sector has not contributed any direct financial subsidies for the operation of bus services in Hong Kong. As such, the Hong Kong SAR Government cannot force the bus operators to provide bus information in an environment as competitive as the bus market in Hong Kong. Therefore, integrated bus information may not be disseminated by the bus operators as they are reluctant to provide any information which could affect their revenues.

In addition to the provision of public transport information, service performance evaluation is also essential for transit operational improvement and planning. A number of key performance indicators (KPIs) have been proposed for defining the transit service quality in several dimensions. Some primary indicators can be estimated using the transit vehicle information derived from in-vehicle sensing devices in the AVL system. For example, service reliability can be evaluated based on transit vehicle location data. By installing a vehicle tracking device on each transit vehicle, the transit operation can then be observed practically via the AVL system over time. Furthermore, the AVL data can be used for estimating some performance indicators which cannot be directly estimated from observations such as the average passenger waiting times.

Improving public transport systems is challenging, particularly when AVL data from transit vehicles are not available. In view of this limitation, considerable attention has been paid to obtaining vehicle location data from smartphone devices (Guo et al., 2015). With the embedded sensors in smartphones, the ubiquitous mobile devices have been considered as human probes with the capability of sensing their surroundings during the users' activities. For transportation studies human probe data could offer opportunities for deriving mobility information that may not be available from the conventional stationary sensors.

Under the big data arena this thesis proposes new methods for deriving important public transport information based on human probe data. Both non-participatory sensing and participatory sensing approaches are used in this research for developing bus transit information systems. To a certain extent, the use of smartphone-based human probe data is

not very common for transportation studies. Hence, there is a need for investigating the opportunities and applications of using smartphone-based human probe data for developing public transport information systems. The new methods are proposed in this thesis for estimating three different KPIs; namely, average bus passenger waiting times, real-time bus arrival times, and bus crowding levels.

The three KPIs help improving bus services in terms of advanced traveler information systems and bus service performance evaluation. Firstly, the provision of real-time bus arrival time information together with the bus crowding information of the arriving buses broadens the benefits of the bus passenger information system. Passengers can make better decisions when choosing a departure time, travel mode choices, route choices, and bus choices. Secondly, bus-stop-based performance can be evaluated with use of the average passenger waiting times at bus stops. The real-time bus arrival time information can be further used for evaluating the bus service reliability and the bus service frequency at bus stops, while the bus crowding information could indicate the in-vehicle service performance and provide additional information for passengers' consideration in making their travel choices.

1.2 Research objectives

The ultimate goal of this research is to provide alternative methods of deriving public transport information with use of the smartphone-based human probe data when AVL data or in-vehicle sensor data from bus transit vehicles are not available. As described previously, both non-participatory sensing and participatory sensing approaches are incorporated for collecting data to estimate the three significant KPIs of bus services; namely, average passenger waiting times, real-time bus arrival times and bus crowding levels.

The first part of this study is dedicated to developing a passenger waiting time estimation system based on **non-participatory sensing approaches**. In particular, passive Wi-Fi data derived from waiting passengers are exploited for estimating the average passenger waiting time at a single bus stop. The specific objectives of this part of the thesis can be defined as follows:

- (1) To provide a comprehensive understanding of using passive Wi-Fi data, as a non-participatory based data source, for developing public transport information systems.
- (2) To develop an alternative approach for a bus passenger waiting time estimation system at a single bus stop based on the findings from the passive Wi-Fi data characteristics in (1).

Then the next part of the thesis focuses on **participatory sensing approaches** in which the relevant bus data can be contributed by participating bus passengers. The crowd-sourced bus data are used for providing two types of timely bus information to bus passengers. The primary objectives of this part can be summarized as follows:

- (3) To address the characteristics of the crowd-sourced bus data and the challenges of using the participatory-based data for developing traveler information systems.
- (4) To develop a real-time bus arrival time prediction system based on the crowd-sourced bus data characteristics in (3) without the need for in-vehicle sensing devices.
- (5) To develop a bus crowding prediction system based on the crowd-sourced bus data by integrating the bus arrival time information from (4) into the system.

1.3 Organization of the thesis

The overall framework of the thesis is illustrated in Figure 1.1 together with the inter-relationships between the chapters of the thesis. The thesis comprises four main parts. After the research introduction in Chapter 1, the first part provides a review of relevant literature in Chapter 2. The problem statements and relevant literature of the three KPIs (i.e. the average bus passenger waiting times, real-time bus arrival times, and bus crowding levels) are firstly described in Chapter 2 in order to provide a fundamental understanding of the previous methods, assumptions, and challenges of estimating the three KPIs of bus services. This is followed by a review of the previous studies on using smartphone-based human probe data for deriving beneficial information for transportation studies, so as to provide an overview of various sensing technologies and sensor data processing methodologies. Based on the foundations of this research study, the smartphone-based human probe data are exploited to tackle the challenges in estimating the three KPIs for developing bus transit information systems.

The second part of the thesis focuses on using non-participatory sensing approaches for estimating bus information. To be more specific, the main objective of this part is to develop an alternative method for estimating passenger waiting times. As a non-participatory sensing approach, passive Wi-Fi data are exploited for the waiting time estimation.

Chapter 3 firstly presents the insightful details of using passive Wi-Fi data for mobility analysis since Wi-Fi data has yet been widely used for deriving public transport information. The factors affecting the quality and quantity of these Wi-Fi data are investigated based on

designed experiments. Then, two uncertainty models are proposed for addressing the challenges in device positioning and activity duration estimation, followed by a discussion on the challenges of using passive Wi-Fi data for transportation studies.

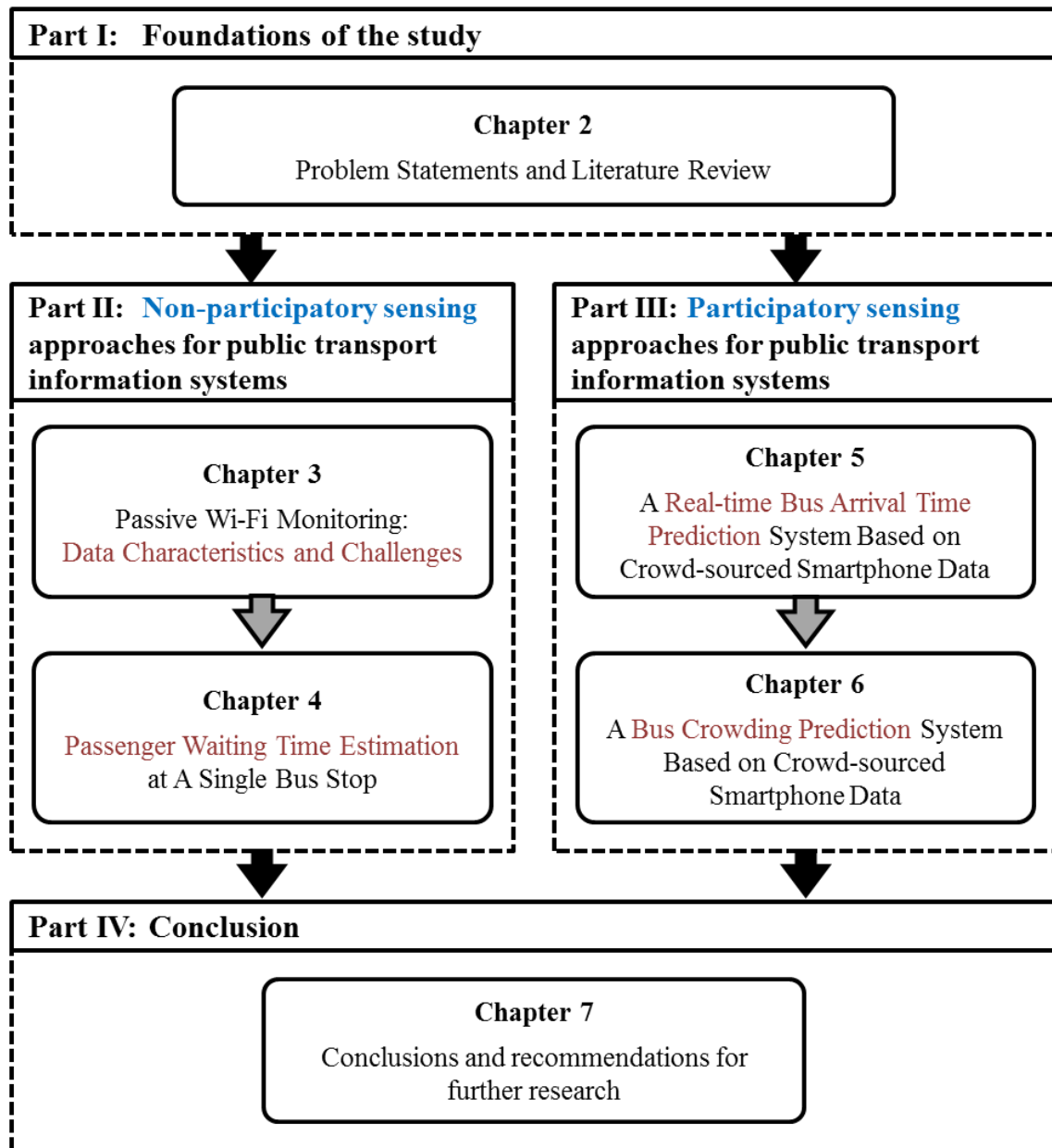


Figure 1.1: Overall framework of the thesis

Chapter 4 proposes an alternative method for bus passenger waiting time estimation at a single bus stop. The estimation method is developed to handle the massive noise data in the Wi-Fi data collected from bus stop environments. Generalized classification features of passive Wi-Fi data are introduced for facilitating the bus passenger waiting time estimation. Then a classifier is developed for identifying the Wi-Fi data derived from waiting passengers.

Multi-day observations were used for training the classifier in order to overcome the temporal uncertainties in the passive Wi-Fi data.

The third part of the thesis intends to develop a bus transit information system based on participatory sensing approaches in which bus passengers can contribute to the provision of the relevant bus data. The central goal of this part is to provide significant information to bus passengers so that they can make more informed travel choices based on the available information. The bus transit information system comprises two sub-systems for bus arrival time prediction and bus crowding prediction.

Chapter 5 presents the novel framework of a real-time bus arrival time prediction system based on the use of smartphone-based human probe data. As the in-vehicle sensor data for the bus mode is not available in Hong Kong, the relevant bus data can be contributed alternatively by participating passengers on the buses. The characteristics of the crowd-sourced bus data are different from the bus data from the AVL systems. Since data inconsistencies can be encountered in crowd-sourced data, bus location filtering methods are developed by taking into account the quality of the bus data both in spatial and temporal dimensions. Also, a bus arrival time prediction method is proposed to address the uncertainty in bus data availability.

Chapter 6 introduces a bus crowding prediction system as an extension of the real-time bus arrival time information system in Chapter 5. Firstly, a bus dwell time estimation method based on crowd-sourced bus data is proposed. Then, the bus dwell times together with the bus arrival time information derived from Chapter 5 are incorporated into bus crowding prediction so as to provide bus crowding information to bus passengers.

Finally, the fourth part (described in Chapter 7) summarizes the research findings with conclusions and recommendations for further research.

1.4 Research contributions

In order to clarify the research contribution, the main methodologies and results presented in each of the four core chapters are summarized as follows:

Chapter 3: Passive Wi-Fi monitoring: data characteristics and challenges

- Temporal uncertainties in passive Wi-Fi data are investigated based on designed experiments. Then quantitative analyses are conducted for identifying significant factors which could affect the probability of detecting Wi-Fi devices.
- Two uncertainty models for device positioning and activity duration estimation based on passive Wi-Fi data are proposed for further development of public transport information systems.

Chapter 4: Passenger waiting time estimation at a single bus stop

- A new method for estimating passenger waiting time at a single bus stop based on passive Wi-Fi data is proposed.
- Generalized classification features of passive Wi-Fi data are introduced in order to describe spatial attributes of each individual Wi-Fi device and to facilitate passenger waiting time estimation.
- A classifier is developed based on temporal uncertainties in passive Wi-Fi data. Without prior knowledge on waiting passengers' Wi-Fi devices, a modified bi-partite matching method is employed to match Wi-Fi records with potential waiting passengers.
- The proposed system is evaluated based on two case studies in which Wi-Fi data were collected from different bus stop environments. The results show that the proposed system can improve the accuracy of average passenger waiting time estimation at a single bus stop, compared to the baseline accuracy from the half-headway estimation method.

Chapter 5: A real-time bus arrival time information system based on crowd-sourced smartphone data

- A novel framework for developing a real-time bus arrival time information system based on crowd-sourced bus data is proposed in order to tackle challenges due to the need for in-vehicle sensing devices.
- Bus location filtering methods are developed to handle the bus data inconsistencies in the bus datasets reported by multiple passengers. Both spatial and temporal uncertainties are taken into account for bus location filtering.

- Without travel time data on some road segments, a bus travel time prediction method is developed based on historical traffic patterns of each road segment and its adjacent segments.
- The proposed system is tested using both simulated bus data and real-world bus data. The results show that the performance of bus arrival time prediction system relies on the levels of bus passenger participation.

Chapter 6: A bus crowding prediction system based on crowd-sourced smartphone data

- A bus crowding prediction system based on crowd-sourced bus data is proposed as an extension of the real-time bus arrival time information system.
- In addition to bus headways which can be derived from a real-time bus arrival time information system, a method for bus dwell time estimation based on crowd-sourced bus data is developed for improving bus crowding prediction performance.
- The system is evaluated using the same set of real-world bus data in Chapter 5. The results show that sufficient crowd-sourced bus data can be used for bus crowding prediction with comparable accuracy to AVL bus data.

In summary, the contribution of this thesis consists of the new methods for estimating the three different KPIs of bus services with use of smartphone-based human probe data only. The results can help improving the effectiveness of APTS in terms of advanced traveler information systems and bus service performance evaluation particularly in cities where AVL data from bus transit vehicles are not available. With the proposed new methods, the KPIs of bus services can be frequently updated in the long run regardless the availability of Global Positioning System (GPS) data from bus transit vehicles. Since the three KPIs have been considered as useful inputs in several public transport planning and operation models, the thesis output can be further contributed to the future-state development of public transport services in terms of transit operational management, system design, and planning. The availability of updated information could lead to new insights and new avenues of research for enhancing bus services. With this, more empirical evidence and valid assumptions can be used for improving the development of public transport models in future.

Part I
Foundations of the study

Chapter 2

Problem statements and literature review

This literature review is presented in two primary sections. Based on the main objectives of this thesis, the first section provides an insightful description of public transport information estimation/prediction problems (i.e. passenger waiting time estimation, bus arrival time prediction, and bus crowding prediction), and explores the literature on the estimation/prediction methods. The second section of this chapter presents an overview of smartphone-based human probe data. A brief review of sensing opportunities for transportation research is provided for each of these types of sensor-produced data.

2.1 Public transport information

2.1.1 Passenger waiting time

A direct method for deriving Average Waiting Time (AWT) of passengers at a bus stop is to observe individual waiting passengers and base the calculations on the results of those observations. In practice, however, it is costly and time-consuming to perform such direct measurement continuously over time. In most cases, several surveyors are required to achieve accurate direct measurement at a bus stop with overlapping or common bus routes. Conducting interview surveys is one of the estimation approaches which could reduce the excessive effort required for manual observation. Despite the advantages of this method, several studies have found that bus passengers tend to overestimate their waiting times (Fan et al., 2016). It was found that the provision of real-time transit information affects the passengers' perception of their waiting times. The waiting time perceived by passengers who can access real-time information is insignificantly different from the measured waiting time

(Watkins et al., 2011). Apart from bus arrival time and bus frequency information, it was reported in the literature that some other important factors would also have certain impacts on the perceived passenger waiting time as well. These factors included, for example, passengers' time-window constraints, passenger's walking time to the destination from the bus stop where they are waiting (Mishalani et al., 2006), weather conditions (Lam and Morrall, 1982), etc.

In the literature, indirect methods have also been developed to estimate the AWT at a bus stop as a function of the times between two consecutive bus arrivals; namely bus headways. This could be a viable solution for AWT estimation as observing bus headways is simple, especially when real-time bus arrival time information is available in some developed cities. Two major assumptions were adopted in the conventional models: a uniform passenger arrival rate and independent bus headways (Welding, 1957; Osuna and Newell, 1972). The expected value of passenger waiting time $E(WT)$ can be estimated by:

$$E(WT) = \frac{E(HW)}{2} \left(1 + \frac{Var(HW)}{(E(HW))^2} \right) \quad (2.1)$$

where $E(HW)$ is the expected value of the bus headways, and $Var(HW)$ is the variance of the bus headways. Suppose that there is no variation in bus headways. The average passenger waiting time can then be obtained by (2.1) as half of the bus headways. Due to the simplicity of AWT estimation, the half-headway approach has been widely adopted in many existing transportation models.

Based on the AWT estimation method in (2.1), several studies have reported that the major assumptions of constant and fixed headway could be invalid in actual bus operations. Firstly, the effects of bus arrival regularity were studied by Holroyd and Scraggs (1966). A constant variable was proposed for estimating a more accurate variance of the bus headways, as well as more accurate AWT estimation. Secondly, the half-headway approach could overestimate the AWT since passengers may time their arrivals at the bus stop particularly when the bus headways are relatively large (O'flaherty and Mancan, 1970; Seddon and Day, 1974; Bowman and Turnquist, 1981). The relationship between the expected passenger waiting time and the actual waiting time was also proposed. Jolliffe and Hutchinson (1975) described three types of passenger arrivals: coincident arrivals with no waiting time, scheduled arrivals, and random arrivals. Also, the AWT estimation model based on random and non-random passenger arrivals was proposed (Turnquist, 1978). Finally, the AWT could be affected by overlapping bus routes since passengers have multiple boarding choices (Marguier and Ceder, 1984).

Recently, more accurate AWT estimation models have been developed by relaxing the former assumptions and assuming other assumptions on passenger arrival patterns and bus

headway distribution. Amin-Naseri and Baradaran (2014) proved that the conventional estimation method in (2.1) overestimates the AWT. They proposed more accurate formulas for AWT estimation based on various assumptions on bus headways. In addition, it was suggested that passenger waiting behaviors are different for a multi-modal transit station which could result in the uniform passenger arrival assumption being inapplicable. In particular, some variables have a significant influence on the AWT of transferring passengers i.e. the reliability of feeder bus services (Chang and Hsu, 2001) and the capacities of the connecting and feeder services (Hsu, 2010). Furthermore, the passenger arrival distribution could be affected by some special conditions such as the availability of retail shops along the passenger transfer corridor. Since some passengers could make a stop during their transfers, the AWT estimation was developed based on two different passenger arrival distributions: direct transfer passengers and non-direct transfer passengers (Guo et al., 2011).

The advancement of bus tracking systems based on the Global Positioning System (GPS) has resulted in the availability of bus trajectory data, and the bus arrival time at each bus stop on the route. The technology enables a more convenient way of estimating passenger waiting times at a bus stop since bus headways can be estimated from the bus arrival times. However, the trajectory data alone may not be sufficient to provide complete bus arrival time data. Due to the lack of GPS signals, the system may be incapable of measuring bus locations, resulting in missing data in bus trajectories. In order to estimate AWT, the methods for handling the missing data were proposed by assuming the gaps in the missing data (McLeod, 2007).

It can be concluded that previous studies have focused on developing passenger waiting time estimation models based on two assumptions regarding passenger arrival patterns and regularity of bus headways. In this way, the estimation accuracy is dependent on how well the assumed distributions can represent the actual situation at a bus stop. One of the remaining challenges is that the passenger arrival patterns and the regularity of bus headways can be varied in both space and time dimensions. For instance, bus headways are usually affected by several factors during actual bus operations e.g. road traffic conditions and bus bunching. Therefore, a specific distribution may not be practical for representing bus headways at multiple bus stops, or at a single bus stop over different times of a day. In addition to the two major assumptions, passengers' boarding decisions could affect the AWT estimation accuracy. Individual passengers may have personal criteria for boarding a bus, especially when common or overlapping bus routes are operated at the same bus stop. In addition, some passengers may choose to wait for the next bus if the arriving bus is overcrowded.

2.1.2 Bus arrival time prediction

Suppose that bus n is dwelling at bus stop i . The bus arrival time at the next bus stop $i + 1$ can be predicted from (a) the bus travel time from bus stop i to the next stop $i + 1$, and (b) the bus dwell time at stop i . The bus arrival time prediction problem is generally defined by:

$$AR_{j,i+1} = DT_{j,i} + TT_{j,i,i+1} \quad (2.2)$$

where $AR_{j,i+1}$ is the predicted bus arrival time of bus j at stop $i + 1$, $DT_{j,i}$ is the bus dwell time of bus j at stop i , $TT_{j,i,i+1}$ is the predicted bus travel time from bus stop i to $i + 1$. Figure 2.1 illustrates two bus trajectories of the same bus route in a time-space diagram.

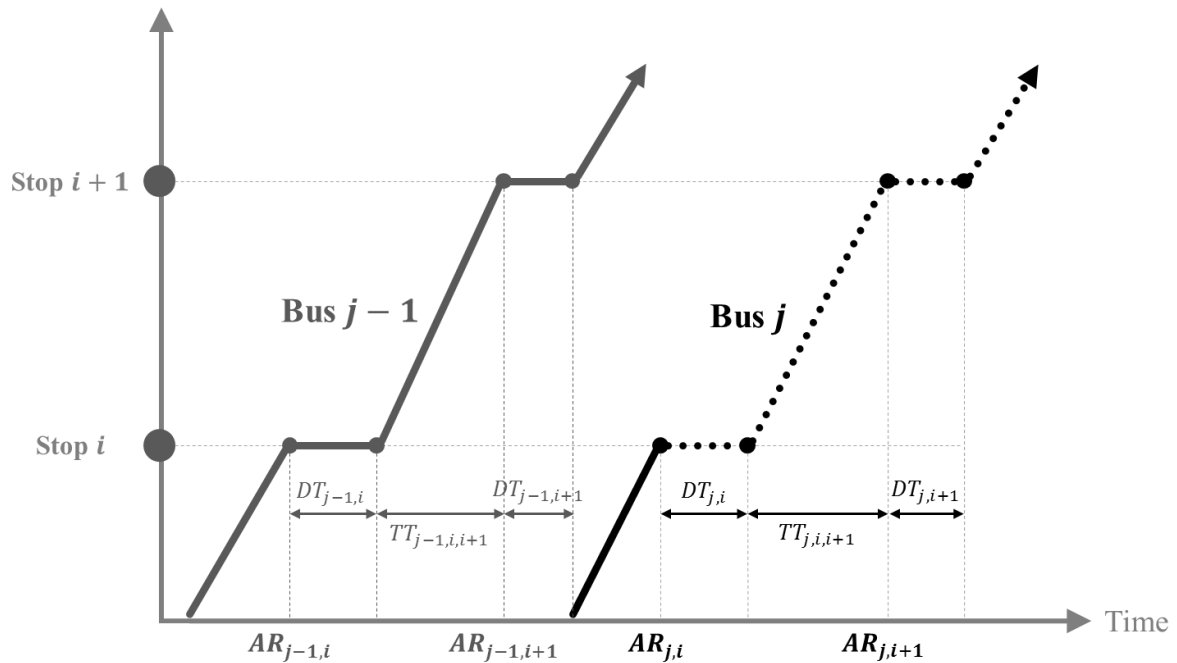


Figure 2.1: Bus trajectories in a time-space diagram

2.1.2.1 Bus dwell time

In the literature on bus arrival time prediction, various assumptions regarding bus dwell times have been employed in the prediction process. In general, bus dwell times constitute a small proportion of a bus trajectory compared to the bus travel times. Deriving bus dwell time information is challenging. Both sensing technologies and promising estimation methods are required for identifying the short time duration at a specific bus stop. Some earlier studies did not define bus dwell time as an explanatory variable for bus arrival time prediction, and assumed bus dwell times at be part of bus travel times (Lin and Zeng, 1999; Vanajakshi et al.,

2009), while some other studies estimated bus dwell time using AVL and Automatic Passenger Counting (APC) data (Shalaby and Farhan, 2004).

Bus dwell time estimation methods have been proposed in the previous related studies based on the major assumption that bus dwell times at a bus stop tend to increase when there are more boarding and alighting passengers at the bus stop. However, it has been suggested that other important factors can significantly affect bus dwell times such as fare payment types and bus crowding (Fernández et al., 2010; Fletcher and El-Geneidy, 2013; Tirachini, 2013). Several types of information are required for developing a precise bus dwell time estimation model (Li et al., 2006; Meng and Qu, 2013).

2.1.2.2 Bus travel time

Another variable for bus arrival time prediction is the bus travel time. Since bus travel times constitute the main part of a bus trajectory, considerable attention has been paid to the improvement of travel time estimation/prediction methods. In fact, the estimation methods and the prediction methods have been developed for different objectives.

Bus travel time estimation aims to reconstruct the travel times of bus trajectories completed in the past based on bus data collected during previous bus trajectories, whereas bus travel time prediction aims to forecast the travel time of bus trajectories that will start from the present time onwards (Mori et al., 2015). The differences are demonstrated in Figure 2.2.

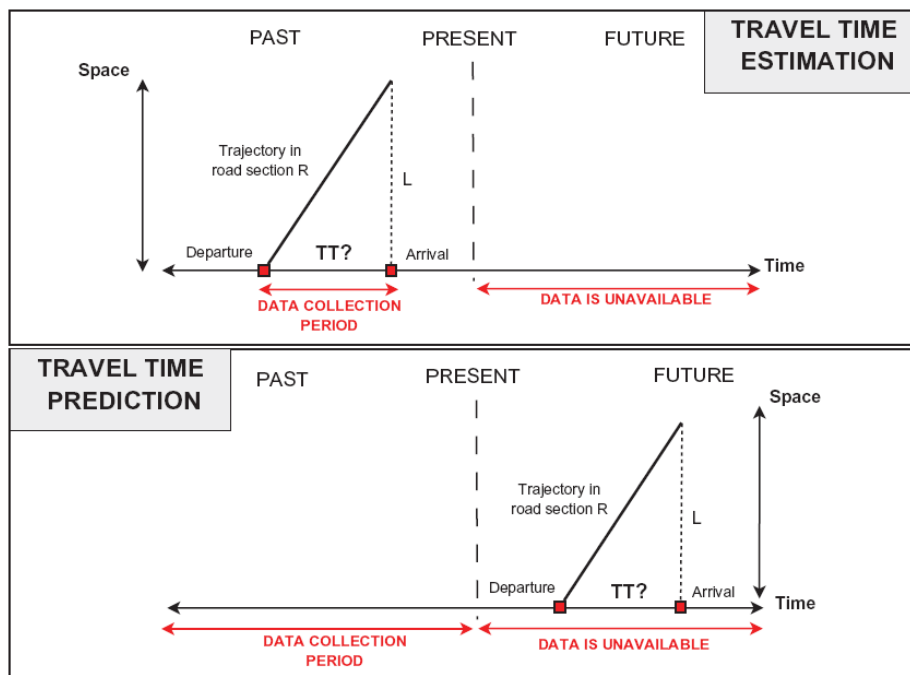


Figure 2.2: Differences between bus travel time estimation and prediction (source: (Mori et al., 2015))

For bus transit systems, bus data can be derived from various sensor types. Bus arrival time prediction systems using APC data and Automatic Vehicle Identification (AVI) data have been developed (Chen et al., 2007; Yu et al., 2011). Based on well-established sensing technologies, AVL systems have been widely implemented for vehicle tracking. Since GPS is generally employed in AVL systems, bus location data can be collected periodically and recorded in a database (data extraction approximately once every 5 to 30 seconds). The flexibility of an AVL system in obtaining vehicle location data from individual buses has motivated more studies on bus arrival time prediction based on AVL data.

Since bus location can be measured in every short time period, AVL data could be sufficient for estimating bus travel times on individual road sections in a bus trajectory. One of the challenges of AVL data, however, is GPS measurement errors. Each GPS data point may not indicate a location on the road network particularly on the sections surrounded by high-rise buildings. A few studies have described travel time estimation methods from AVL bus data (Lin and Zeng 1999; Jeong 2005). Firstly, a GPS location can be matched to the closest location on a road segment. Next, data filtering methods based on bus route information were proposed in order to remove the location data with significant errors. Then, remaining bus locations can be interpolated for estimating the travel time on a road section.

Of the bus arrival time prediction models based on AVL bus data that have been developed, the relevant existing methods can be classified as follows:

- **Historical-based models**

Early approaches adopted historical-based models for bus arrival time prediction (Jeong 2005). The models assume recurrent traffic conditions. Hence, the bus travel time at a certain time interval is assumed from the estimated travel times in the past during the given time period. Due to the model assumption, the historical-based prediction could be ineffective for non-recurrent traffic conditions.

- **Parametric models**

With these models, bus travel time prediction is based on a function of explicative variables. Linear regression models have been developed by formulating a mathematical function based on a set of independent variables. Previous studies investigated various independent variables for travel time prediction including travel distance, bus dwell time, and schedule adherence (Lin and Zheng, 1999; Jeong, 2005; Cats and Loutos, 2015). Linear regression models require a sufficient amount of datasets for parameter estimation. In addition to linear regression models, Kalman filter-based bus travel time predictors have been developed (Chien and Kuchipudi, 2003; Shalaby and Farhan, 2004; Vanajakshi et al., 2009). A Kalman filter predicts the future states of dependent variables based on the current state and the prediction

errors made in previous time steps. Therefore, the prediction is a recursive procedure and relies on regular updates from newly available bus travel times in each time step.

- **Non-parametric models**

Non-parametric estimation assumes that a function of explicative variables cannot be pre-defined for travel time prediction. Several methods have been developed for conducting non-parametric regression. For example, k-nearest neighbor (k-NN) is a method which identifies a neighborhood from historical bus data that is most similar to the bus data in the current time interval. Chang et al. (2010) showed that a k-NN prediction model provided accurate prediction results. However, the method required a longer execution time for the prediction when the size of the historical data was expanded. Due to the capability of providing a unique solution, Support Vector Machine (SVM) learning has been adopted for bus travel time prediction based on historical bus arrival information (Bin et al., 2006; Yu et al., 2011).

Artificial Neuron Network (ANN) models have been developed for solving complex and non-linear problems. An ANN imitates the mechanism of a brain by constructing a collection of neuron nodes. Jeong (2005) developed ANN-based bus travel time prediction using three inputs: bus arrival times at stops, bus dwell times, and schedule adherence. Yu et al. (2011) developed an ANN model for bus travel time prediction based on the travel times of the preceding buses and bus headways. Developing ANNs requires a training process for adjusting the weight of neuron nodes and the connections between nodes. The accuracy of ANN-based prediction is reliant on the ANN constructed after the training process.

- **Hybrid models**

Some studies integrated several models in order to improve bus travel time prediction performance. A model combining a Kalman filtering method with an SVM model was proposed by Yu et al. (2010), while the mixed model developed by Li et al. (2017) for bus arrival time prediction was supported by multiple methods, i.e. k-NN, K-means, Kalman filtering, and Markov chain. Yu et al. (2017) used a Random Forests model to enhance the bus travel time prediction performance based on a k-NN model. The results from previous studies showed that hybrid models improved bus travel time prediction accuracy compared to each of the participant models.

2.1.3 Bus crowding

In the related literature on bus crowding, the main research direction is to investigate the effects of bus crowding on passengers' satisfaction and passengers' decisions on mode choices and route choices (Tirachini et al., 2013; Yap et al., 2018). Both objective and subjective measures have been proposed for bus service evaluation. An objective standard

which is widely used for the evaluation of bus services is the number of standing passengers per square meter (Li and Hensher, 2013). In some cases, bus crowding is considered as one of the service-quality indicators for determining a level of service (LOS) scale for bus transit services. Das and Pandit (2015) proposed LOS benchmarks in which bus crowding levels are determined by the number of passengers per seat. However, a specific standard may not be applicable for all countries due to the differences in what is considered unacceptable crowding levels as perceived by the passengers in different cities. For subjective measures, several studies have investigated and quantified the values of bus crowding based on stated preference experiments. For example, pictures of crowding levels were used for valuation (Batarce et al., 2016; Tirachini et al., 2017).

There is limited literature on bus crowding estimation and prediction. Basically, bus crowding on a bus can be estimated from passenger boarding and alighting data at individual bus stops. The observations on boarding and alighting passengers have become practical with the availability of in-vehicle sensing systems for transit ridership tracking such as APC. Thus, the estimation of bus crowding can be straightforward with use of the APC data. The relationship between the number of on-board passengers and the number of boarding and alighting passengers is described by

$$NB_{b,s} = NB_{b,s-1} + LP_{b,s} - UP_{b,s} \quad (2.3)$$

where $NB_{b,s}$ is the number of passengers on bus b after the bus departs from stop s , $LP_{b,s}$ is the number of boarding passengers at stop s , and $UP_{b,s}$ is the number of alighting passengers at stop s .

Automatic Fare Collection (AFC) systems can also be a complementary data source for bus crowding estimation. Zhang et al. (2017) proposed a real-time passenger flow estimation and prediction method based on the integration of AVL and AFC data. Since AFC generally provides partial counts of boarding and alighting passengers, probability functions were proposed for estimating the total number of boarding and alighting passengers. A prediction method was also proposed by Zhang et al. in order to forecast bus crowding when buses arrive at the remaining bus stops on the bus route concerned.

2.2 Smartphone-based human probe data

Smartphone market penetration has been increasing rapidly over the past two decades, with smartphone subscriptions reaching 5.1 billion at the end of 2018 (Ericsson mobility report, 2019). The functions on these smartphones have also been improving based on the

advancements in mobile sensing and communications technologies. As a result, smartphones have begun to be considered as potential sensing devices for human mobility tracking since people tend to carry the devices with them almost everywhere and at all times. In addition, with the various sensors embedded in smartphones, the ubiquitous devices have been considered as human probes which are capable of sensing their surroundings during the users' mobility. For transportation studies, human probe data could offer unprecedented opportunities for deriving mobility information that may not be available from conventional sensing methods. Data mining techniques have been developed to extract the mobility information from various types of sensors. Essential mobility information consists of meaningful locations where people conducted activities and the time when the activities were conducted. In this section, significant sensors embedded in smartphones are introduced together with the use of sensor data in previous transportation studies.

2.2.1 Wi-Fi

Wi-Fi is a wireless communication technology for connecting Wi-Fi devices to each other and/or to the Internet. To facilitate the communication of Wi-Fi devices, a Media Access Control (MAC) address is assigned to each device as a unique identifier for identifying an individual device in Wi-Fi networks. According to the Wi-Fi communication standards, it could be assumed that Wi-Fi devices will continue broadcasting Wi-Fi packets with their MAC addresses as long as the Wi-Fi function is enabled. Therefore, the presence of such Wi-Fi enabled devices can be discovered using a Wi-Fi scanner which is responsible for listening to the Wi-Fi communication in nearby Wi-Fi networks. More details of the principle of Wi-Fi technology and Wi-Fi device discovery are provided in Appendix A.

By installing a Wi-Fi scanner at a fixed location, the captured Wi-Fi data can be exploited for analyzing human activity in a particular location such as statistics on space utilization (Abedi et al., 2014). The data can be further analyzed to classify the type of users based on their activities (Qin et al., 2013; Schauer and Linnhoff-Popien, 2017). Moreover, the analysis can be extended in the spatial dimension using multiple scanners installed at different points of interest (Prentow et al., 2015; Danalet et al., 2014).

In the previous research related to public transport systems, the use of passive Wi-Fi data has been categorized into two types based on Wi-Fi scanner installation. A set of Wi-Fi scanners can be installed either on each transit vehicle or in a transit station. The in-vehicle installation is suitable for capturing mobility information along the bus route, while the fixed-location installation is for capturing the mobility at transit stations or bus stops. Previous studies that used passive Wi-Fi data for transportation development are summarized in Table 2.1.

Table 2.1: Summary of the previous studies using passive Wi-Fi data for public transport development

Authors	Wi-Fi scanner installation	No. of installed Wi-Fi scanners	Complementary data	Transportation mode	Derived public transport information	Validation by manual observations
Handte et al. (2014)	In-vehicle	Single	Yes	Bus	- Bus crowd density	Yes
Myrvoll et al. (2017)	In-vehicle	Single	Yes	Bus	- Bus crowd density	Yes
Oransirikul et al. (2019)	In-vehicle	Single	Yes	Bus (campus)	- Bus crowd density	Yes
Liu et al. (2016)	In-vehicle	Single	No	Bus	- Real-time location - Arrival time	Yes
Shlayan et al. (2016)	Fixed-location	Multiple	No	Train/Bus Train Bus (intercity)	- Passenger count - Waiting time - Origin-Destination flow	No
Oransirikul et al. (2016)	Fixed-location	Single	No	Bus (rural area)	- Bus stop crowd density	Yes
Song and Wynter (2017)	Fixed-location	Multiple	No	Train	- Timetable - Arrival time	Yes

2.2.1.1 In-vehicle Wi-Fi scanner

Most of the previous studies related to the bus mode were developed based on an in-vehicle Wi-Fi scanner. The Wi-Fi data were considered as a complementary source of data to support GPS data in the early stage of these related studies. A crowd density estimation system was proposed by Handte et al. (2014). A Wi-Fi scanner was installed on each bus in order to estimate the number of on-board passengers over time. Since GPS data were taken into account for matching the Wi-Fi detection timestamps to a bus location, the system was able to provide crowd density information for each bus route segment. The results show that the estimated crowd density was approximately 20% of the on-board passengers. More accurate models for estimating the number of on-board passengers using both a deterministic method (Oransirikul et al., 2019) and a probabilistic method (Myrvoll et al., 2017) were proposed. The model included a scaling factor for adjusting the number of detected Wi-Fi devices. The manual passenger counting of passengers was required on individual bus trips for training purposes. The work concluded that the estimation results were promising.

In addition to the bus crowding estimation systems, a real-time bus tracking system using passive Wi-Fi data as a single data source was developed (Liu et al., 2016). Instead of identifying the Wi-Fi devices carried by the on-board passengers, the system focused on the densely distributed Wi-Fi access points along the bus routes in urban areas. To identify the real-time bus location, an access point fingerprint database was established for each road segment. A bus location was estimated on a real-time basis using the current detectable access points in comparison with the fingerprint database. Since the bus location can be estimated on a real-time basis, bus travel time estimation was performed as well as bus arrival time prediction. Furthermore, bus headways at individual bus stops can also be derived.

2.2.1.2 Fixed-location Wi-Fi scanner

The use of fixed-location Wi-Fi scanners for public transport information systems has been studied. Firstly, with a Wi-Fi scanner network in a transit station, i.e. a train station and an interstate bus terminal, passenger mobility information was investigated including passenger count at a gate, the passenger flow between two strategic locations by OD pair, and passenger waiting time (Shlayan et al., 2016). Second, Wi-Fi scanners were installed at the platform of individual train stations in order to estimate the train timetable including the number of trains and the arrival/departure times (Song and Wynter, 2017).

2.2.2 Bluetooth

Bluetooth is a type of wireless technology used for short-distance data communication. Bluetooth devices are connected in an ad hoc mode in which a master device is responsible for controlling device discovery and establishing connections. Similar to Wi-Fi devices, a unique Bluetooth device can be identified by a Bluetooth MAC address. The mobility of Bluetooth devices can be detected by a number of Bluetooth scanners which search for discoverable Bluetooth devices nearby. Kostakos et al. (2010) presented algorithms and visualization techniques for describing pedestrian mobility patterns in urban contexts based on discoverable Bluetooth devices captured by Bluetooth scanners in urban areas. Malinovskiy et al. (2012) used Bluetooth data for investigating pedestrian dwell time and travel time distributions over different times of a day. The study found that Bluetooth data provided low sample rates of between 2% and 5%. Therefore, Bluetooth data might not provide sufficient samples in cases of low-volume pedestrians. Weppner and Lukowicz (2013) developed a method for estimating crowd density in a mass event. The densities were identified as one of the seven discrete densities from a nearly empty space to an extremely highly-populated space. Bluetooth data were used for examining the behavioral patterns of visitors in mass events such as visitor flows and re-visiting patterns (Versichele et al., 2012; Delafontaine et al., 2012).

More studies have focused on using Bluetooth data for estimating vehicular traffic information: travel times (Martchouk et al., 2010; Barcelö et al., 2010; Porter and Arriaga, 2013; Araghi et al., 2014), intersection delays (Bhaskar and Chung, 2013), vehicle trajectories (Michau et al., 2017), and OD matrices (Barcelö et al., 2010; Laharotte et al., 2015). Although vehicle travel times between Bluetooth scanners can be estimated simply, previous studies have found that detection range of the Bluetooth scanner normally affects travel time estimation accuracy. It was suggested that a shorter distance between Bluetooth scanners results in lower estimation accuracy (Haghani et al., 2010; Bhaskar and Chung, 2013). Moreover, the accuracy of OD estimation relies on missed detection probabilities when a Bluetooth device cannot be captured by some of the Bluetooth scanners on the device's travel path (Araghi et al., 2014; Laharotte et al., 2015; Michau et al., 2017).

Bluetooth data were also integrated with other data sources. Bhaskar and Chung (2014) developed a vehicle trajectory estimation method which incorporated both Bluetooth data and loop detector data into the trajectory estimation along a section of a motorway. Kostakos (2008) proposed methods for estimating the number of on-board passengers and passengers' OD bus stops. By installing a Bluetooth scanner on a bus, Bluetooth data were collected from on-board passengers and integrated with GPS bus location data for the estimation.

Both Bluetooth and Wi-Fi MAC addresses have been used for human mobility analysis. However, previous studies have found that the proportion of Bluetooth and Wi-Fi MAC addresses could be different for various study cases. In the case of road traffic, more Bluetooth MAC addresses were captured as compared to Wi-Fi MAC addresses (Abbott-Jard et al., 2013). In contrast, the number of Wi-Fi MAC addresses was higher than the Bluetooth MAC addresses in cases of non-motorized traffic i.e. pedestrians and cyclists (Abedi et al., 2015).

2.2.3 Global Positioning System (GPS)

The GPS is a satellite-based system which can identify the location of a GPS-receiver with a timestamp. The system requires GPS signals from at least four satellites for identifying a device location. GPS sensors have been integrated into smartphones to provide location-based services such as navigation systems. In transportation studies, a GPS signal trace from an individual device is considered for describing the mobility of an individual. In the mid-1990s, GPS data were incorporated into travel surveys (Battelle Memorial Institute, 1997). It was found that GPS-based travel surveys could improve the quality of time and position data when compared to observation data, which was affected by human errors. In addition, the misreporting issue caused by respondents' manual records could be overcome since travel data are automatically recorded by GPS devices. However, the accuracy of travel survey data could be limited by GPS positioning accuracy, the battery life of GPS devices (e.g. smartphones), and GPS data transmission costs (Shen and Stopher, 2014).

For bus transit systems, a smartphone with GPS has been considered as a viable data collection tool for obtaining bus trajectories, speeds, and travel times. Biagioni et al. (2011) proposed an automatic transit tracking system in which a smartphone was placed in each vehicle for location tracking. The system made use of GPS vehicle trajectories to predict vehicle arrival times. The results showed that passenger waiting times can be significantly reduced with the availability of arrival time information. Furthermore, smartphone applications can be developed for establishing two-way data provision systems which gather necessary information from bus passengers and provide beneficial information in return. Lee and Yim (2014) proposed a system for providing the most up-to-date location of individual buses. Without the need for dedicated GPS devices, GPS bus data can be provided by on-board bus passengers.

2.2.4 Global System for Mobile Communications (GSM)

GSM is a standard for mobile communications. In order to establish a connection between mobile phones, the system needs to identify the location of the phones in the GSM network. Here a set of base stations is responsible for providing connections between mobile devices

and the GSM network in which each base station handles the mobile traffic within its coverage area. Therefore, the location of a mobile phone in the GSM network can be regularly updated when there is a change in its position to the base station areas due to the phone's movement.

Since mobile phones generate signal traces over multiple base stations in the GSM network, previous research has developed methods for deriving mobility information from cellular signals such as travel speeds (Birle and Wermuth, 2006), travel times (Bar-Gera, 2007), traffic flows (Astarita et al., 2006), densities (Ratti et al., 2006), route choices (Tettamanti et al., 2012), and trip OD (White and Wells, 2002; Caceres et al., 2007; Zhang et al., 2010).

Due to the long distances between base stations, typical location errors based on GSM data can vary from fifty meters to two kilometers (Chung and Kuwahara, 2007). These location errors could pose additional challenges in the estimation of transport information. In addition, mobile phone signals provide infrequent location data. The locations of a mobile phone are updated only when the phone is using GSM services, or when the phone movement results in its relocation to another base station area.

2.2.5 Motion sensors

Accelerometers are typical smartphone sensors used for motion detection. The sensors detect acceleration forces or the change of velocities by measuring the magnitude and direction of the acceleration. Accelerometers have been used for identifying users' activities since human movement generates acceleration and/or rotational motion. In addition, the sensors require lower power consumption compared to GPS sensors (Su et al., 2016). The measured signals can differ based on different human activities and when there are some variations in acceleration across the human body (Miluzzo et al., 2008; Hoseini-Tabatabaei et al., 2013).

In transportation studies, acceleration data have been used for traffic mode detection such as walking, running, bike, car, bus, subway (Yang, 2009; Siirtola and Röning, 2012; Shafique and Hato, 2015; Shin et al., 2015). Traffic mode detection can be improved by integrating acceleration data with gyroscope data and magnetometer data (Yu et al., 2014; Jahangiri and Rakha, 2015; Fang et al., 2016; Fang et al., 2017). Since gyroscopes measure the rotational motion of smartphones and magnetometers measure the change of a magnetic field, the motion of a smartphone can be described by direction, acceleration, orientation, and magnetism. Furthermore, motion-sensing data can be integrated with GPS data to improve the accuracy of traffic mode detection (Feng and Timmermans, 2013; Bedogni et al., 2016; Broach et al., 2019). Shen et al. (2019) found that the data sampling frequencies of smartphone sensors significantly affect the accuracy of travel mode extraction, especially for motorized modes such as buses and cars.

2.3 Smartphone-based human probe data for public transport information systems

Smartphone-based human probe data can be derived with and without the participation of the smartphone users. On the one hand, sensor data collection can be participative when the smartphone users are involved in the sensor data collection. The users' permission is required for activating most smartphone sensors such as GPS, accelerometers, gyroscopes, and magnetometers. In most cases, users need to use a smartphone application for such sensor data to be collected. The sensor data may be sent to a database server and/or be processed using smartphone resources. On the other hand, some types of sensor data can be collected in a non-participative way. For instance, passive Wi-Fi/Bluetooth signals and GSM signals are broadcast regularly when the sensors are enabled. Without direct participation from smartphone users, the passive signals can be collected by a number of sensing devices such as MAC address scanners.

In order to develop public transport information systems based on smartphone sensor data, it is necessary to determine which types of sensor data are suitable for the objectives of each information system. In the following parts of this thesis, both non-participatory sensing and participatory sensing methods are used for estimating bus transit information when sensor data from buses are not available.

Part II of the thesis aims to develop a new method for estimating AWT at a single bus stop. In such a case, non-participatory sensing approaches could be more feasible for collecting smartphone data from waiting passengers. Passengers may not consider AWT information to be useful for their journey planning, resulting in a lack of motivation for contributing related information for AWT estimation. In contrast, Part III of the thesis aims to develop traveler information systems in which related bus data can be contributed by bus passengers, and in return, the passengers can derive bus arrival time and bus crowding information to support their journey planning. Since bus data are collected from passengers' smartphones, the optimal data sampling frequencies are investigated in order to provide accurate bus information with minimum battery consumption.

Part II

**Non-participatory Sensing for
Public Transport Information Systems**

Chapter 3

Passive Wi-Fi monitoring: data characteristics and challenges

With reference to the study objective (1) shown in Chapter 1, this chapter provides insightful details of using passive Wi-Fi data for human mobility analysis. The use of passive Wi-Fi data for human mobility analysis is then demonstrated. As the foundation of this study, experiments were designed to address the knowledge gap regarding passive Wi-Fi data characteristics for the development of various public transport information systems. Furthermore, two stochastic models are proposed in this chapter to investigate the uncertainty issues for device positioning and activity duration estimation. As a guideline for practical development in the future, other challenges in the use of passive Wi-Fi data as well as privacy issues in relation to passive Wi-Fi data collection are discussed. Finally, the analysis of the Wi-Fi data and the subsequent discussion in this chapter will lead to the development of the proposed public transport information system which is to be presented in Chapter 4 of this thesis.

3.1 Introduction

As shown in the literature review in Chapter 2, the use of Wi-Fi data as a means of human mobility tracking has been gaining attention recently due to this method offering several advantages. With the use of Wi-Fi technology having now spread to almost every corner of the world, Wi-Fi access has become a necessity of modern life. To facilitate this ‘always-connected’ lifestyle, Wi-Fi technology has been embedded into personal mobile devices, especially smartphones. As the market penetration of smartphones has increased, Wi-Fi-enabled devices have become increasingly ubiquitous. This ubiquity has enabled alternative

approaches to be taken to human mobility analysis in various disciplines, such as business analytics (Zeng et al., 2015), tourism (Nunes et al., 2017), and the environment (Wang et al., 2018). Methodologies for Wi-Fi data acquisition have been proposed on both a participatory and an opportunistic basis, along with the methods for deriving meaningful information from the raw data.

In transportation research, Wi-Fi-enabled devices have been considered as a potential source of transport data. Apart from the previous research related to public transport systems presented in Chapter 2, the major focus is on using Wi-Fi data for human activity and behavior analysis since Wi-Fi devices provide opportunities for deriving digital footprints while the users are conducting their day-to-day activities. Data mining techniques have been applied to extract essential mobility information from the Wi-Fi data, including meaningful stay locations, to identify where people have conducted their activities, and the time when those activities were conducted. This information has also been utilized for developing activity models (Danalet, 2015). Wi-Fi data have been exploited in traffic fields for estimating vehicle travel times (Abbott-Jard et al., 2013), and estimating pedestrian destinations (Danalet et al., 2014). In previous related studies, Wi-Fi data have been obtained from outdoor environments where a great amount of noisy data is also present.

In order to develop transport information systems, Wi-Fi data should be collected in a non-intrusive way. Since direct participation from the users could raise concerns over their privacy, using passive Wi-Fi data could be more suitable for the system development. However, there are a number of challenges in collecting and using passive Wi-Fi data. The study of data characteristics is necessary as a fundamental first step to understand the nature and limitations of the data before further development.

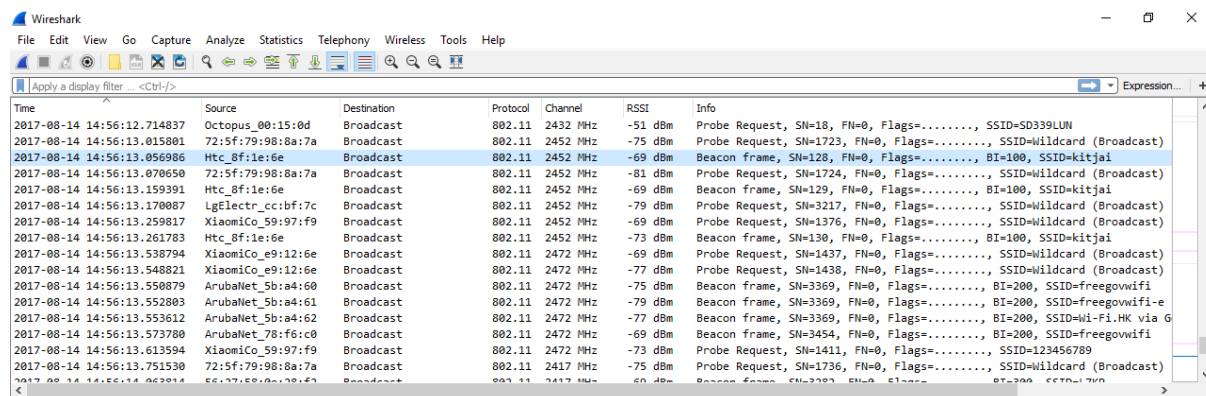
In what follows, the detailed procedures for deriving and processing passive Wi-Fi data will be described followed by a discussion on passive Wi-Fi data characteristics regarding each potential factor which could affect the data quality for human mobility analysis. In Section 3.2, an explanation on the transformation of raw data into meaningful information is provided. The characteristics of passive Wi-Fi data are summarized in Section 3.3. Section 3.4 introduces two uncertainty models based on the data characteristics. Other challenges which should be considered for system development are introduced in Section 3.5. Finally, this chapter closes with a summary of findings.

3.2 Transformation of raw data

This section shows the transformation of raw data derived from a fixed-location Wi-Fi scanner. The results are summarized as a set of basic terms in order to facilitate mobility analysis. As the foundation of other terms, detection events will be firstly described. Then several abstractions are discussed including session, encounter, and trail based on the concepts proposed by Kostakos et al. (2010). The raw data are further investigated in this thesis and additional indicators for each abstraction are proposed to cover more aspects of the mobility analysis. The potential approaches for utilizing the indicators are also discussed.

3.2.1 Raw data

A set of captured Wi-Fi packets from a Wi-Fi scanner operating in the 2.4 GHz radio band is shown in Figure 3.1. The captured packets are analyzed using a Windows Application for packet analysis called Wireshark. More examples of Wi-Fi data are provided in Appendix C



Time	Source	Destination	Protocol	Channel	RSSI	Info
2017-08-14 14:56:12.714837	Octopus_00:15:0d	Broadcast	802.11	2432 MHz	-51 dBm	Probe Request, SN=18, FN=0, Flags=....., SSID=SD339LUN
2017-08-14 14:56:13.015801	72:5f:79:98:8a:7a	Broadcast	802.11	2452 MHz	-75 dBm	Probe Request, SN=1723, FN=0, Flags=....., SSID=Hildcard (Broadcast)
2017-08-14 14:56:13.056986	Htc_Bf:1e:6e	Broadcast	802.11	2452 MHz	-69 dBm	Beacon frame, SN=128, FN=0, Flags=....., BI=100, SSID=kitjai
2017-08-14 14:56:13.070650	72:5f:79:98:8a:7a	Broadcast	802.11	2452 MHz	-81 dBm	Probe Request, SN=1724, FN=0, Flags=....., SSID=Hildcard (Broadcast)
2017-08-14 14:56:13.159391	Htc_Bf:1e:6e	Broadcast	802.11	2452 MHz	-69 dBm	Beacon frame, SN=129, FN=0, Flags=....., BI=100, SSID=kitjai
2017-08-14 14:56:13.170887	LgElectr_cc:bf:7c	Broadcast	802.11	2452 MHz	-79 dBm	Probe Request, SN=3217, FN=0, Flags=....., SSID=Hildcard (Broadcast)
2017-08-14 14:56:13.259817	XiaomiCo_59:97:f9	Broadcast	802.11	2452 MHz	-69 dBm	Probe Request, SN=1376, FN=0, Flags=....., SSID=Hildcard (Broadcast)
2017-08-14 14:56:13.261783	Htc_Bf:1e:6e	Broadcast	802.11	2452 MHz	-73 dBm	Beacon frame, SN=130, FN=0, Flags=....., BI=100, SSID=kitjai
2017-08-14 14:56:13.538794	XiaomiCo_e9:12:6e	Broadcast	802.11	2472 MHz	-69 dBm	Probe Request, SN=1437, FN=0, Flags=....., SSID=Hildcard (Broadcast)
2017-08-14 14:56:13.548821	XiaomiCo_e9:12:6e	Broadcast	802.11	2472 MHz	-77 dBm	Probe Request, SN=1438, FN=0, Flags=....., SSID=Hildcard (Broadcast)
2017-08-14 14:56:13.550879	ArubaNet_5b:a4:60	Broadcast	802.11	2472 MHz	-75 dBm	Beacon frame, SN=3369, FN=0, Flags=....., BI=200, SSID=freegowifi
2017-08-14 14:56:13.552803	ArubaNet_5b:a4:61	Broadcast	802.11	2472 MHz	-79 dBm	Beacon frame, SN=3369, FN=0, Flags=....., BI=200, SSID=freegowifi-e
2017-08-14 14:56:13.553612	ArubaNet_78:f6:c0	Broadcast	802.11	2472 MHz	-77 dBm	Beacon frame, SN=3369, FN=0, Flags=....., BI=200, SSID=Wi-Fi.HK via G
2017-08-14 14:56:13.573780	ArubaNet_78:f6:c0	Broadcast	802.11	2472 MHz	-69 dBm	Beacon frame, SN=3454, FN=0, Flags=....., BI=200, SSID=freegowifi
2017-08-14 14:56:13.613594	XiaomiCo_59:97:f9	Broadcast	802.11	2472 MHz	-73 dBm	Probe Request, SN=1411, FN=0, Flags=....., SSID=123456789
2017-08-14 14:56:13.751530	72:5f:79:98:8a:7a	Broadcast	802.11	2417 MHz	-75 dBm	Probe Request, SN=1736, FN=0, Flags=....., SSID=Hildcard (Broadcast)

Figure 3.1: Examples of captured Wi-Fi data

It can be observed that numerous Wi-Fi packets can be captured from several channels within a second. During one second, nine MAC addresses were discovered. The basic information was indicated in the packets for communication purposes including the detection timestamp, the transmitter MAC address, the receiver MAC address, the Received Signal Strength Indicator (RSSI), the frame subtype, and other essential information.

With the availability of Wi-Fi data, a number of studies consider a captured MAC address with its timestamp as data with the potential for conducting human mobility and activity analysis. However, the basic information in a Wi-Fi packet could be redundant for the analysis. The raw data need to be diminished so as to minimize memory usage and facilitate further data processing. Only vital data fields should be selected. Furthermore, the retained

information should not violate user privacy concerns. In general, two significant data fields are crucial for mobility analysis: the detection timestamp and the transmitter MAC address. Additional data fields may be required based on the objectives of each analysis.

3.2.2 Detection events

A detection event is defined as an occurrence when the scanner records a packet. The detection then contains the significant information of the packet e.g. the detection timestamp, and the MAC address (MAC-ID). A detection event i is denoted by:

$$event_i = \{timestamp_i, MAC_ID_i\} \quad (3.1)$$

where MAC_ID_i is the packet transmitter's MAC address.

Figure 3.2 shows the detection events of a MAC-ID captured by a fixed-location scanner during a 5-minute time window. The device was in the same state during the entire five-minute period.

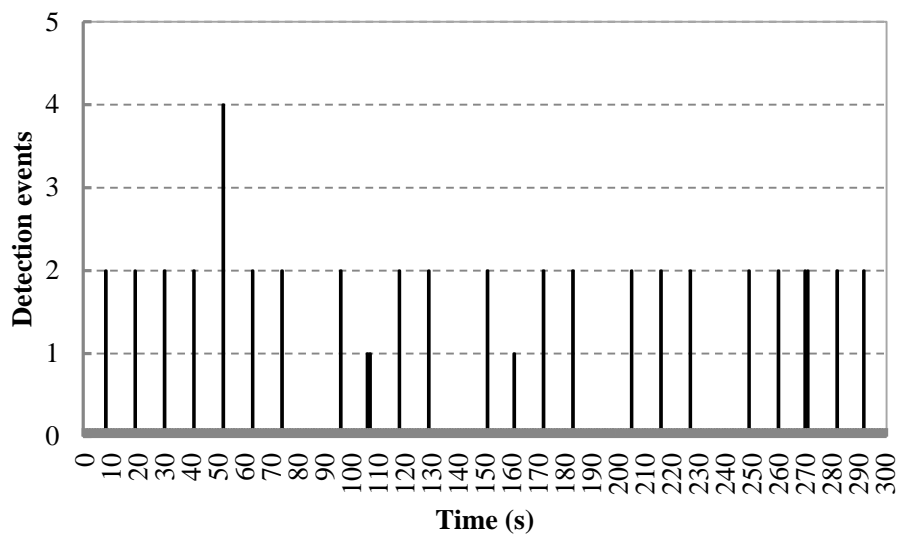


Figure 3.2: Detection events from a MAC-ID

It can be observed that the device generated multiple detection events during the 5-minute period. The packets were detected as a burst, whereby multiple detection events are detected in a short period. In particular, a burst of detection events is mainly influenced by active scanning. Since a Wi-Fi device can broadcast Probe Request packets (PRQs) to every channel during an active scan, a scanner may capture the PRQ packets within the scanning time window. Basically, an active scan is performed in a very short time period (i.e. milliseconds). Thereby, several PRQ packets can be captured within a second as can be seen in Figure 3.2. The general duration of an active scan was investigated by Hu et al. (2015). The study found

that the average time width for an active scan was up to 3 seconds in practice. This implies that PRQ packets from the same active scan could be captured within a 3-second time window.

3.2.3 Session

As is the nature of human activity, a person may conduct his or her activities at the same location several times within a time span (e.g. in a day). Therefore, a scanner can capture detection events which were generated from multiple activities during a monitoring period. To facilitate further analysis, the detection events captured during an activity should be isolated from the others. A series of detection events which were captured during the same activity is defined as a session.

Identifying different activities from the detection events of a MAC-ID is challenging. As can be seen from Figure 3.2, the time between consecutive detection events is varied. Since there is no certain rule for identifying distinct activity duration, a session can be defined empirically. Here, a session is defined as a series of detection events whereby the time between any consecutive detection events is less than a session threshold TH_{SS} . A set of assumptions for identifying distinct activity duration can be considered to determine the threshold. Firstly, the common period between consecutive detection events can be investigated for determining the threshold. A new session could be assumed if the time between consecutive detection events is greater than the common period. Also, the threshold can be empirically determined from manual observations of people's behavior in the study area, such as general activity durations. A session i is denoted by:

$$\mathit{session}_i = \{\vec{e}_{i,j}, j = 1, 2, \dots, J\} \quad (3.2)$$

where $\vec{e}_{i,j}$ is the j^{th} detection event of a MAC-ID, and J is the total number of detection events during the session. To simplify the mobility analysis, a session can also be represented using the information in the detection events of the session. For instance, a session can be denoted by:

$$\mathit{session}_i = \{MAC_{ID_i}, \mathit{start_time}_i, \mathit{end_time}_i\} \quad (3.3)$$

where $\mathit{start_time}_i$ and $\mathit{end_time}_i$ are the timestamp of the first and the last detection event of MAC_{ID_i} in the detection area respectively. Apart from the timestamps, other data fields can be included for additional analysis. For example, an RSSI attribute can be included for device positioning purposes.

Two additional indicators can be included to describe a session:

- **Event period:** the time between consecutive detection events. This indicator can imply the continuity of detection. If a device is frequently detected in every short period, further analysis based on the session will be more reliable. For example, the estimation of activity duration will be more accurate if the device is detected with regular event periods. The statistics of event periods in a session can be described in terms of mean and standard deviation.
- **Presence time:** the duration from *start_time* to *end_time*. The presence time of a session can be considered as the activity duration of the device's user.

Accordingly, a set of sessions introduces an array of indicators for mobility analysis:

- **Density:** the number of MAC-IDs present within the detection area at a given time t .
- **Device inflow:** the number of new MAC-IDs within a specified time window $[t_1 - t_2]$.
- **Device outflow:** the number of MAC-IDs leaving the detection area within a specified time window.
- **Retention:** the number of MAC-IDs dwelling in the detection area over a specified time window.
- **Device class:** the class of a MAC-ID based on its presence time. The device's class aims to classify the devices into several groups. In the literature on customer behavior analysis, the device's class was used for categorizing the customers based on the time spent in a commercial space (Yan et al., 2017). Furthermore, customer loyalty can be analyzed from the class over multi-day observations. The device's class was also used for determining persistent devices which are detected in the detection area for a considerable time period, as well as transient devices which are detected for only a short time period.

Kostakos et al. (2010) suggested data visualization for making sense of the raw data from a scanner. Figure 3.3 shows an example of the timeline visualization from a Wi-Fi scanner. The Wi-Fi scanner was installed in front of an elevator waiting area at the Hong Kong Polytechnic University campus. A set of Wi-Fi data from fifteen MAC-IDs during 8:00-8:30 was selected for data visualization. The x-axis is the detection time span, while each value on the y-axis is allocated for a discrete MAC-ID. Also, a detection event is represented by a circle.

Firstly, it can be observed that a session normally consists of multiple detection events in which the event periods can vary in a session. The presence time of a session implies the duration of activities conducted in the study area. In addition to individual sessions, a diagonal line (the red dotted line) was introduced to observe the device inflow over time. Moreover, the device's class can be initially classified as persistent devices or transient

devices. On the one hand, the sessions below the line tend to be persistent as the devices remain in the detection area for a significant time period such as the MAC-ID no.1. The device could be a nearby stationary device such as a Wi-Fi access point (AP). On the other hand, transient devices with high mobility are above the diagonal line.

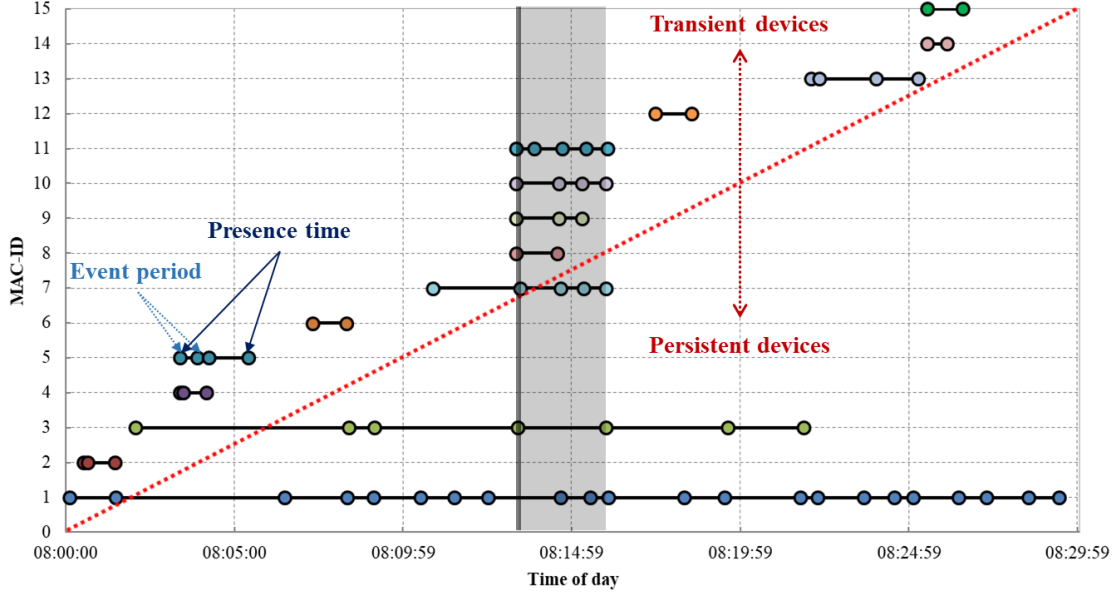


Figure 3.3: An example of timeline visualization

According to the visualization, we propose vertical lines for additional indicators including density and retention. With a vertical line (e.g. the vertical grey line at 08:13), the density can be measured as the number of MAC-IDs at a given time. Furthermore, the sessions which cover a pair of vertical lines are counted for retention over a specified time window. For instance, the sessions of MAC-ID no.1, 3, 7, 10, and 11 cover the gray area between 08:13 and 08:16 which means the devices occupied the study area over the three minutes.

3.2.4 Encounter

An encounter was introduced in order to capture the interaction among the devices in the study area. An encounter is defined as when any two MAC-IDs have an overlapping session. An encounter i is then denoted by:

$$encounter_i = \{MAC_{ID_{A_i}}, MAC_{ID_{B_i}}, s_start_time_i, s_end_time_i\} \quad (3.4)$$

where $s_start_time_i$ is the starting time of an overlapping session between $MAC_{ID_{A_i}}$ and $MAC_{ID_{B_i}}$, and $s_end_time_i$ is the ending time of the session.

Further analysis of group behavior benefits from encounter information since the co-presence of Wi-Fi devices is recorded. An example of an encounter can be seen from MAC-ID no.10 and no.11 in Figure 3.3 during 08:13-08:16

3.2.5 Trail

The concepts of session and encounter are based on the raw data from a single Wi-Fi scanner at a fixed location. Further analysis for device detection across multiple Wi-Fi scanners was also suggested. Here, a trail of a MAC-ID is defined as describing the sequence of device detection at different monitoring locations. A device can be re-identified across multiple Wi-Fi scanners using its unique MAC-ID. In general, the network of Wi-Fi scanners is represented by a directed weighted graph $G(N, E)$, in which N is a set of nodes representing the locations of Wi-Fi scanners and E is a set of edges linking between a pair of nodes. The edges usually represent the travel path between the two locations while the weight represents the travel distance. A trail of a MAC-ID is defined by a sequence of sessions associated with the scanners. A trail i of a MAC-ID is then denoted by:

$$\mathit{trail}_i^{MAC-ID} = \{ \mathit{node}_{i,j}, \mathit{s_start_time}_{i,j}, \mathit{s_end_time}_{i,j} \}, j = 1, 2, \dots, J \quad (3.5)$$

where $\mathit{node}_{i,j}$ is the j^{th} scanner's location represented by a node ID, $\mathit{s_start_time}_{i,j}$ and $\mathit{s_end_time}_{i,j}$ are the timestamp of the first and the last detection event at the scanner's location, and J is the total locations on a trail where the MAC-ID is re-identified.

Since a MAC-ID may generate multiple trails within a time window, different trails should be identified. A temporal threshold TH_{TR} was introduced for trail identification. A new trail is started when the time difference between a pair of consecutive sessions on any two nodes is greater than the threshold. In most cases, the maximum travel time between the two locations is estimated to determine the threshold. Therefore the difference between the starting time of the current session and the ending time of the previous session is taking into account, together with the distance between the locations represented by the linking edge.

With the topology of Wi-Fi scanner locations, a set of trails is beneficial for mobility analysis in the spatial dimension over a time span.

- **Path travel time:** the time spent traveling from one node to another one.
- **Trip travel time:** the time spent on a trail.
- **Attractive node:** the nodes where most MAC-IDs were re-identified. In other words, attractive nodes are the hotspots of the network. The popular nodes can be further classified as the origin nodes of most activities if the nodes are the first location of the trails. In the same way, destination nodes can be determined.

- **Node class:** the class of a node based on device presence times. The nodes where most devices spend a considerable amount of time tend to be the locations where people conduct major activities.
- **Attractive path:** the node sequence which is a segment of most trails. Popular paths are useful when there are multiple travel paths from one location to another location.
- **Attractive entry:** the preceding node which is most visited before a particular node. The attractive entry of a node implies the preferred entry to a location.
- **Attractive exit:** the following node which is most visited after a particular node. The attractive exit of a node implies the preferred exit from a location.
- **Attractive trail:** the most generated trails.

In practice, there are several challenges in the multiple-location based monitoring system. First, a pair of nodes may have multiple travel paths resulting in multiple edges linking the nodes. As a result, a certain travel path cannot be identified for a MAC-ID that is detected at the two nodes. Additional nodes might be required if path identification is necessary for analysis. Second, a scanner could miss a MAC-ID during the detection process. The missed detection problem results in imprecise trails since the detection sequence is different from the actual location sequence. To this end, these and other possible challenges should be carefully addressed based on various objectives of mobility analysis.

3.2.6 Dimensions of data analysis

To broaden the analytical capability, Wi-Fi data transformation can be performed in several dimensions. The primary dimensions for mobility analysis are summarized as follows:

- A temporal dimension is suitable for tracking the changes of information at a particular location over time. Also, the time resolution is dependence on the objectives of an individual study. For instance, a study may focus on the changes in mobility over different times of day, or over days of a week. With the continuous operation of a Wi-Fi scanner, the temporal scale can be extended for monthly and yearly analysis. Sessions and encounters are examples of temporal transformation.
- A spatial dimension is feasible when multiple scanners are installed at different locations. Hence, the spatial dimension is for observing differences in mobility over spaces within a time window. The activities can then be compared across different monitoring locations.
- A spatial-temporal dimension is valuable for mobility analysis since mobility is described over times and spaces. A trail is an obvious example of spatial-temporal transformation.

- The transformed data can be analyzed in terms of individual data or aggregate data. The raw data can be transformed to trace the mobility of individual MAC-IDs. Alternatively, the individual data can be combined at an aggregate level. Examples of the aggregate data described in this section consist of density, device inflow and outflow, retention, attractive nodes, paths, entries, exits, and trails.

3.3 Passive Wi-Fi data characteristics

Empirical studies have been conducted in previous research so as to understand the underlying characteristics of the Wi-Fi data. Firstly, it was found that the number of detection events captured from a Wi-Fi device could be varied based on the model and status of the Wi-Fi device (Freudiger, 2015; Hu et al., 2015). Furthermore, due to the potential of RSSI data for locating Wi-Fi devices, some studies investigated the characteristics of RSSI in Wi-Fi data. The results can be summarized into three findings. First, there was a relationship between RSSI and the distance between a Wi-Fi device and a Wi-Fi scanner. RSSI values were decreased when the distance between a Wi-Fi device and a Wi-Fi scanner was extended (Lui et al., 2011; Xu et al., 2011). Second, there was a variation in RSSIs among different Wi-Fi devices. Third, RSSI values were decreased when physical obstacles were presented between Wi-Fi devices and the Wi-Fi scanner (Garcia-Villalonga and Perez-Navarro, 2015; Nakatani et al., 2018).

In this study, three experiments were conducted for investigating the variation in the number of detection events and RSSI values. The experimental results are consistent with the findings from previous studies in the literature. The experimental results with the experimental details are provided in Appendices B.1-B.3. It is noteworthy that Appendix B can provide a comprehensive understanding for the readers who have no expertise in Wi-Fi technology. The readers may read Appendix B before continuing to the next section.

Apart from the variation in the number of detection events and RSSI values, missed detection is one of the critical factors which could affect the reliability of mobility analysis based on passive Wi-Fi data. Due to the mechanism of Wi-Fi communication, the number of Wi-Fi packets can be countless at a particular time especially in crowded areas with numerous Wi-Fi devices. Not all the Wi-Fi packets may be captured by a Wi-Fi scanner. This implies that a set of detection events in a specific time window is the subset of all Wi-Fi packets at the time. In such a case, missed detection also occurs.

Since the missed detection probability has yet been evaluated in previous studies, it is necessary to investigate the characteristics of temporal uncertainties in passive Wi-Fi data. In

order to observe the missed detection probability in different scenarios, four experiments were designed and carried out. The experimental details are provided in Appendix B.4. It can be summarized from the experimental results that the missed detection probability is uncertain for each individual Wi-Fi device. The missed detection probability of a Wi-Fi device can be affected by four main factors: the device's model and status, the distance between the Wi-Fi device and the Wi-Fi scanner, the Wi-Fi scanner setup, and the number of Wi-Fi signals in the study environments.

3.4 Modeling uncertainties in passive Wi-Fi data based on a single Wi-Fi scanner

According to the empirical analysis of passive Wi-Fi data, the data characteristics pose several challenges for human mobility analysis. This section provides a discussion on two major challenges which should be considered for developing a Wi-Fi data processing algorithm in order to derive accurate and reliable analytical results.

3.4.1 Uncertainties in device positioning

Since a Wi-Fi scanner has a detection range which can cover a wide area, a detection event could be derived from any Wi-Fi device within that range. Although RSSI information can be considered for representing the distance between the Wi-Fi device and the Wi-Fi scanner, it is still challenging to identify the precise location within the detection area using a single Wi-Fi scanner. Basically, there is a variation in RSSI values from a Wi-Fi device even when the device is in a stable state at a fixed location (See Experiment#2 in Appendix B for further results).

To tackle the device positioning challenges, several techniques using multiple Wi-Fi scanners have been developed. This method is suitable for spacious locations where the installation of multiple Wi-Fi scanners is feasible such as in an airport terminal or a shopping mall. Moreover, the method may require several RSSI observations from a Wi-Fi device. Then the device's RSSI can be compared with the prior RSSI measurements and the device position can be identified. This implies that the positioning accuracy could decrease at a location where people conduct short activities since the activity duration might be too short for deriving sufficient RSSI observations.

The installation of multiple Wi-Fi scanners might be unfeasible for some activity locations with a limited space. In general, a single Wi-Fi scanner could be adequate for identifying the

activity location where it is not necessary to identify a specific device position. For example, a system could aim to estimate customer activity duration in a small coffee shop regardless of which service section the customer occupied. However, the device position could be roughly identified as a zonal position instead of a precise location within the detection area. Figure 3.4 illustrates the detection area of a Wi-Fi scanner with an omnidirectional antenna. The detection area is divided into two circular zones with the Wi-Fi scanner in the center of both zones.

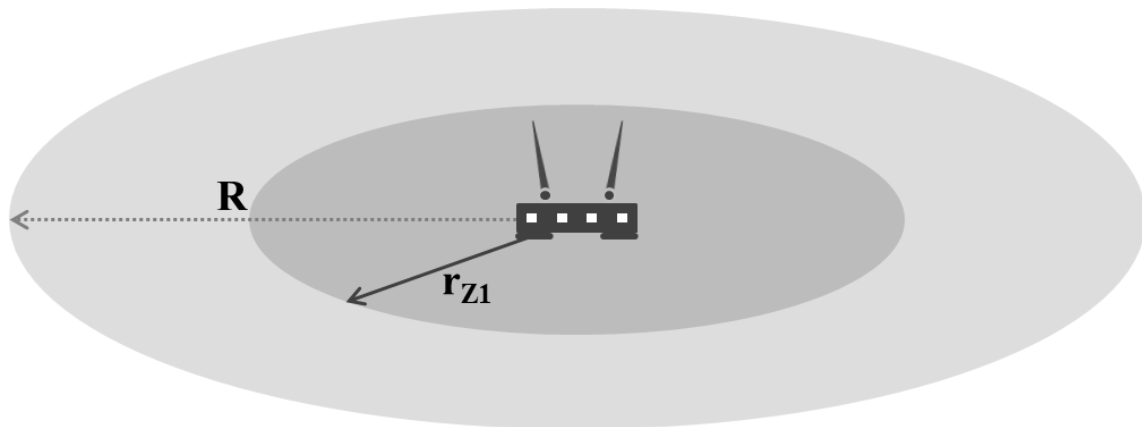


Figure 3.4: The detection area divided into two circular zones

Since the Wi-Fi scanner is in the center of both zones, the RSSI observations from a device in Zone A tend to be stronger than the RSSI observations from Zone B. It can be noted that the zonal positioning still involves additional challenges due to the variation in RSSI observations from various Wi-Fi devices, or even from the same device. A set of prior knowledge could be useful for developing zonal positioning techniques in future studies. Moreover, the positioning accuracy could be dependent on the number of zones. The positioning result might be inconclusive if there are multiple zones in such distance.

To model the monitoring zones, the Wi-Fi scanner detection area can be assumed to be a circular shape with a radius R . Suppose that the detection area is divided into Z zones and the distance from the scanner to the boundary of each zone z_i is denoted by r_{z_i} . It can be noted that the radius R can be a stochastic value due to the uncertainty in the scanner detection range which could be different for individual Wi-Fi devices.

3.4.2 Uncertainties in detection periods

The results from the experiments designed for observing missed detection (in Appendix B.4) show that the presence time of a Wi-Fi device calculated from passive Wi-Fi data is shorter than the actual occupancy duration. The occupancy time error results in a challenge for

estimating activity duration. In the related literature, the presence time at a particular location is usually used for representing the activity duration of the device user at the location. In such a case, the estimated activity duration involves some errors due to the sparse nature of passive Wi-Fi data. According to the experimental results, it can be concluded that the degree of occupancy time errors is varied based on the state of individual Wi-Fi devices which can be changed over time. It is challenging to estimate the degree of error without having detailed information about the current device state such as the Wi-Fi network association status.

To model the occupancy time errors in activity duration, time window constraints can be initially formulated based on the characteristics of passive Wi-Fi data. Suppose that the scanner detection range is divided into two zones as can be seen in Figure 3.4. The inner zone (Zone A) covers the study location (e.g. a coffee shop), while the outer zone (Zone B) is outside the area of interest. Given the observed activity duration of an individual as the period between the time of entry into the study location in Zone A and the exit time, an assumption is that the Wi-Fi record from the individual's device should be detected during the activity or during a similar time period. However, the scanner detection range also covers Zone B resulting in potential errors in estimating activity duration from Wi-Fi data. Figure 3.5 illustrates the time window constraints of an observed individual in space and time dimensions, while Table 3.1 summarizes the case of constraints with the factors for determining time windows.

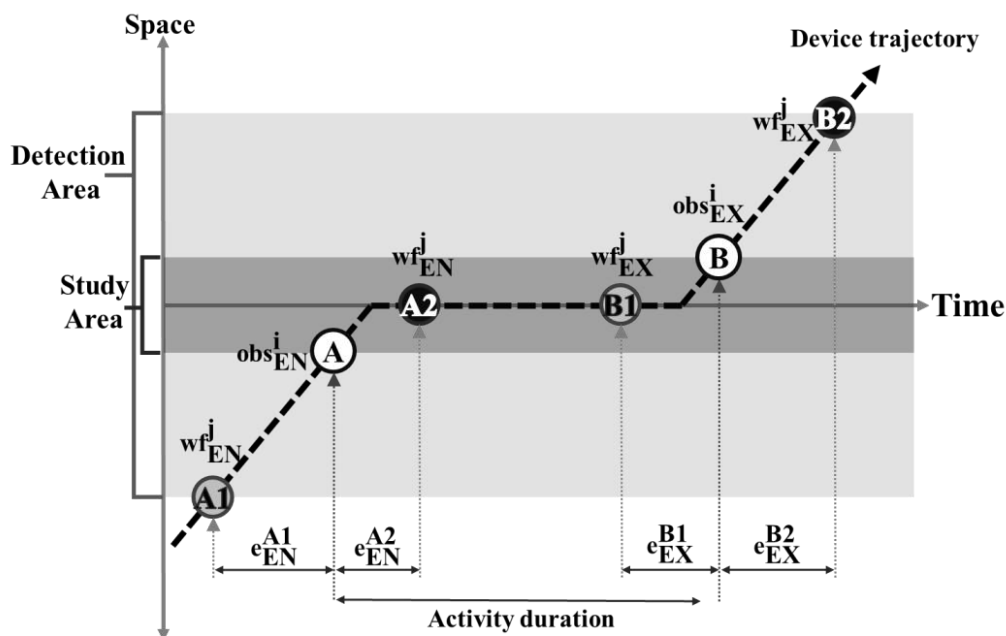


Figure 3.5: Time window constraints of activity duration

Table 3.1: The cases of time window constraints

Case	Target time window	MAC detection time	Factors for determining time windows
A1	Entry time	Before entry time	Travel speed (i.e. walking speed or vehicle speed)
A2	Entry time	After entry time	Detection frequency
B1	Exit time	Before exit time	Detection frequency
B2	Exit time	After exit time	Travel speed (i.e. walking speed or vehicle speed)

Basically, the entry time of an individual can be represented by the first Wi-Fi observation of a MAC-ID in the same way as the exit time can rely on the last observation. In the case of the A1 entry time wf_{EN}^j , the first Wi-Fi observation can be detected before the actual entry time to Zone A obs_{EN}^i since the scanner detection range covers a distance further than the study area. In contrast, the observation can also be detected after the exit time from Zone A due to the sparseness of Wi-Fi detection (case A2). Hence, the range of entry time error can be estimated using the travel time in Zone B and the detection period of the device. In the same way, the last Wi-Fi observation can be detected either before the exit time from Zone A (case B1) or after the exit time (case B2). The range of exit time error can be described as well.

Given the observed activity duration of an individual i as the entry time obs_{EN}^i and the exit time obs_{EX}^i and a Wi-Fi record j as the first observation time wf_{EN}^j and the last observation time wf_{EX}^j , the range of activity duration error based on the Wi-Fi record can be estimated.

$$RoE(j) = \left[- \left(\left(f^j - \frac{(R - r_{z1})}{v_{in}} \right) + \left(f^j - \frac{(R - r_{z1})}{v_{out}} \right) \right), \frac{R - r_{z1}}{v_{in}} + \frac{R - r_{z1}}{v_{out}} \right] \quad (3.6)$$

The notations are described as follows:

R : The radius of the scanner's detection area (m).

r_{z1} : The distance from the scanner to the end of the boundary of the study location (m).

v_{in} : Average travel speed when entering to the study area (m/s).

v_{out} : Average travel speed when exiting from the study area (m/s).

f^j : The estimated maximum detection period between Wi-Fi observations (s)

The equation can be simplified by assuming an equal travel speed v_t for v_{in} and v_{out} , as can be shown in (3.7).

$$RoE(j) = \left[- \left(2 \left(f^j - \frac{(R - r_{z1})}{v_t} \right) \right), 2 \left(\frac{R - r_{z1}}{v_t} \right) \right] \quad (3.7)$$

It can be noticed that the estimated maximum period between Wi-Fi observations f^j provides negative errors which result in underestimation of activity duration, whereas a long scanner detection range causes overestimation from positive errors. In addition, the range of error can be stochastic with the availability of statistical information for the stochastic parameter: the scanner detection range, travel speed, and detection period. Since the variation of other parameters may not be critical, a crucial parameter which strongly affects the estimated activity duration can be f^j .

The error may not be significant for a long activity duration. However, it is important to verify the estimation result for a system that aims to capture short activity durations especially for durations of less than one minute. If the activity duration is too short for detecting sufficient detection events, missed detection can occur. This could pose extra challenges for activity sequence estimation when the activity sequence of an individual has to be identified.

3.5 Other challenges in passive Wi-Fi monitoring

This section introduces three relevant challenges regarding the use of passive Wi-Fi data for human mobility analysis. The challenges should be taken into account when developing a system based on passive Wi-Fi data.

3.5.1 Determining configuration parameters for a Wi-Fi scanner

The first step before developing any software is to determine the two configuration parameters for a Wi-Fi scanner: the detection range and the monitoring mode. The parameters should be adjusted by considering their suitability for the system being developed. First, the antenna specification can be decided including the monitoring Wi-Fi frequency band (2.4 GHz and/or 5.0 GHz), the antenna type (an omnidirectional type or a directional type), and the antenna gain.

- A preliminary trial can be conducted for identifying the dominant Wi-Fi frequency in the study area.
- A suitable antenna type can be selected. An omnidirectional antenna is widely used for capturing human mobility, while a directional antenna could be more suitable for road traffic monitoring.

- The Wi-Fi scanner detection range can be determined. The detection range can be extended by a powerful antenna gain which results in a larger number of detected MAC-IDs (Abedi et al., 2015). However, a large number of noise data can be included in the Wi-Fi dataset as well as the higher activity duration errors.

Second, the Wi-Fi monitoring mode can be configured either in a channel-specific mode or in a channel hopping mode. For the channel-specific mode, the dominant Wi-Fi channels can also be identified by conducting a trial. Otherwise, specific channels of the target Wi-Fi network can be specified. For the channel hopping mode, a channel hopping velocity must be selected. Although the results from the experiment designed for observing missed detection show that the missed detection probability can be decreased by applying a lower channel hopping velocity, the use of high channel hopping velocities may be suitable for some specific objectives. For instance, a system may need to detect the presence of a specific Wi-Fi device promptly when Wi-Fi traffic is very low but the device's communication channel is unknown. Hence, it is necessary to set the Wi-Fi scanner to loop over all available channels rapidly.

3.5.2 Long-term mobility tracking

A MAC-ID is assumed to be unique for identifying an individual Wi-Fi device during wireless communication. In practice, some studies have found that a MAC-ID can be assigned to multiple Wi-Fi devices (Bhaskar et al., 2015). This duplication is rare but can happen due to several possibilities such as manufacturers reusing MAC-IDs, MAC spoofing tools being used for changing the factory-assigned MAC-ID, and MAC randomization being applied for enhancing device security.

A duplicate MAC-ID results in a challenge for long-term mobility tracking since it is uncertain whether Wi-Fi observations with the same MAC-ID are from the same device. Therefore, algorithms for identifying a unique device should be developed to fulfill the objectives of the system being developed.

3.5.3 System validation

Two issues could be involved when passive Wi-Fi data are used for mobility analysis. Firstly, it is important to justify the use of Wi-Fi data for representing the behavior of the entire population. The Wi-Fi data would only be partial as it is unlikely to capture the data from all existing devices. An evaluation of the Wi-Fi data penetration rate might be necessary. Second, validating Wi-Fi data filtering algorithms is an onerous task. In general, data filtering methods are necessary for identifying the target devices in noisy Wi-Fi data. If there is a need

to demonstrate that the algorithm can distinguish the target Wi-Fi devices from the noise data, the MAC-ID of the target devices in the study area is required.

To derive the Wi-Fi data penetration rate and validating Wi-Fi data filtering algorithms, a process for collecting a set of all target MAC-IDs is unavoidable. In such a case, MAC-ID surveys can be conducted at the study location. A group of volunteers is needed for approaching all target people and having them complete the survey process before they leave the study location. This method could be ineffective if all target MAC-IDs cannot be derived. Another more practical method could be conducting control experiments. In a control experiment, a group of volunteers can be asked to provide their MAC-IDs. Moreover, a set of scenarios can be designed to imitate a real-world situation. Then, the volunteers can be asked to behave according to the designed experiments.

Since the studies on Wi-Fi data penetration rates and the performance of data filtering algorithms require substantial resources, previous literature on using passive Wi-Fi data for mobility analysis has evaluated the proposed systems in terms of estimation accuracy for the final results e.g. activity duration and travel time. The evaluation could be considered sufficient since the main objective is to develop an information estimation system. For some system objectives, further justification might be required for providing a valid conclusion based on the analytical results from the passive Wi-Fi data

3.5.4 Privacy

Passive Wi-Fi data are considered anonymous since the only identification information is the MAC address. Although personal sensitive information cannot be revealed from the MAC address and be linked to individuals, the suitability of collecting passive Wi-Fi data has been questioned. On one hand, beneficial information can be provided to improve people daily life experience since people mobility can be analyzed from the available Wi-Fi data. On the other hand, deriving Wi-Fi data can be considered privacy intrusive from some points of view. To this end, the practicality of using passive Wi-Fi data should be discussed as a guideline based on the existing case studies.

In the related literature, similar concerns over the privacy of many developed applications have been discussed. The perspective on privacy can be varied based on what is considered personal information for an individual. In the current data-driven era, various data have been collected from users while they are connected to online services. For instance, location-based information can be collected via Google Maps when the navigation service is running. Also, the information can be linked to the user's Google account and be considered as a part of the user's profile. In this way, the application can provide more context-relevant information based on the locations the user has visited.

The trade-off between users' interests and their privacy is known as the personalization-privacy paradox and has been the subject of research (Aguirre et al., 2015). It has been suggested that collecting users' data may raise privacy concerns and result in a negative attitude towards the contribution of information (Ketelaar and van Balen, 2018), especially when the information is collected for the benefits of profit-oriented companies. In the following sections, two case studies on collecting passive Wi-Fi data are firstly discussed. Then, the possibility of deriving sensitive information from passive Wi-Fi data is summarized followed by a discussion on the applicable guidelines on developing a system based on passive Wi-Fi data.

3.5.4.1 Case studies

Two cases of collecting passive Wi-Fi data in London are worthwhile discussing here. The first case was conducted by an advertising firm in 2013. The main objective was to provide location-based, and targeted personal advertising based on pedestrians' behavior. A trial on passive Wi-Fi data collection was conducted for around two weeks in the Cheapside area of central London. Twelve data collection units were installed in recycle bins with digital advertising screens (Shubber, 2013).

The action raised privacy concerns resulting in further discussion about the suitability of passive Wi-Fi data collection. The advertising firm responded to the concerns by explaining that the data collection was legal since the system performed data encryption and focused on anonymized and aggregated Wi-Fi data (Miller, 2013). Therefore, personal information such as a person's name and home address could not be revealed. According to the Data Protection Act in the UK, the collected data could be valid if individual Wi-Fi devices were not specifically being tracked and the data were sufficiently anonymized. However, the City of London Corporation called a halt to the data collection based on a suggestion that such data collection should only be executed with public support (Battersby, 2013).

Another case involved a government section, Transport for London (TfL). Wi-Fi data were collected from 54 London underground stations for one month in 2016. The collected data were processed to provide benefits in terms of customer information, operations and safety information, transportation planning, and investment prioritization (Transport for London, 2017). Although the Wi-Fi data were anonymized similar to the first case study, the one-month trial by TfL was conducted with less public concerns over privacy issues. Recently, TfL has started collecting Wi-Fi data from the London underground network and has been doing so since the 8th July 2019 (O'Malley, 2019).

Even though the two cases focused on collecting the same data type, only one project was permitted to proceed to the next phase. The factors behind the different levels of success can

be discussed. According to a review of the TfL Wi-Fi pilot, it can be concluded that the key to its success was adherence to three major policies. Firstly, TfL followed the guidance from the Information Commissioner's Office in order to inform the customers about the pilot. Multiple communication methods were employed to disseminate the detailed information about the pilot, including the collected data, the data collection purposes, the plan for using the data, and the method used for preventing customers' data from being collected. Moreover, TfL clearly outlined its commitment to protecting customers' privacy. The data encryption and data storage methods applied were described, while the policy on data usage showed that TfL did not aim to identify individuals from the collected data. Finally, customer benefits were introduced to demonstrate how the collected data could be used to improve customers' travel experiences. For example, the availability of crowding information made it possible to offer the customers better journey planning assistance.

Whereas TfL disseminated the important details to the customer in advance, there was no apparent evidence that the advertising firm had announced the data collection details before its trial. It was not clear to the public about which data would be collected and the company's plan on how it would use their data. People could assume that the collected data would include their personal data. In addition, it seemed that the objective of the trial was mainly intended to benefit the firm, resulting in public criticism.

The key success factors from the TfL case are consistent with a previous study on mobile application privacy (Almuhimedi et al., 2015). The study suggested that users want to be in control when it comes to sharing their information. It is important for the users to understand and to acknowledge how their data are collected and processed by the third-party. Users should have the option of configuring their device so as not to provide the information such as toggling the Wi-Fi function.

3.5.4.2 The possibility of deriving sensitive information from passive Wi-Fi data

In the related literature, the possibility of using passive Wi-Fi data for deriving personal sensitive information and identifying an individual has been investigated. Several studies found that physical locations where a detected Wi-Fi MAC-ID has been visited could be assumed from the unencrypted Service Set Identifier (SSID identifying the name of a Wi-Fi network data in the broadcast probe requests). For instance, the Wi-Fi enabled smartphone of a student of the Hong Kong Polytechnic University may broadcast a probe request specifying 'PolyUWLAN' as the SSID. Table 3.2 shows the PRQs of a detected MAC-ID during one second from Experiment#2 in Appendix B.

Table 3.2: Examples of unencrypted SSIDs from a Wi-Fi device

PRQ Sequence No.	SSID
1529	freegovwifi
1549	Apple Store
1550	NETGEAR54
1551	Universities via Y5ZONE
1552	StarbucksHK WiFi
1553	PolyUWLAN
1554	CSL
1559	Universities via CSL
1560	-Y5ZONE-
1561	freegovwifi
1563	#####
1565	Apple Store
1567	Universities via Y5ZONE
1568	StarbucksHK WiFi

It can be observed that the device broadcast a set of PRQs with identified SSIDs in one second. Most of the SSIDs were public Wi-Fi services and at least three of them were from the in-campus Wi-Fi services i.e. ‘PolyUWLAN’, ‘Universities via Y5ZONE’, and ‘Universities via CSL’. Only one SSID of the PRQ sequence no. 1563 may not be a public network. The SSID was identified by 8 digits which were replaced by ‘#’ in the example for data privacy. Some studies have concluded that the retrieval of SSIDs can be used for identifying the potential interests of individuals (Chernyshev et al., 2015), or the individuals themselves (Cunche, 2014). However, a study suggested that the identification of individuals is rarely practical using Wi-Fi data alone. Moreover, it is uncertain whether the list of SSIDs is completely relevant to the device owner (Watts et al., 2011). The information presented in Table 3.2 could support the assumptions as most SSIDs were public Wi-Fi services and none of them could be linked to a specific location.

Apart from the SSID retrieval, the session of the detected MAC-IDs can be considered. For instance, the MAC-ID of a specific person could be assumed if his/her occupancy duration near a Wi-Fi scanner is known and there is only one detected MAC-ID that matches the duration (Cunche, 2014). The linked information can be used in both helpful cases and illegal cases such as tracking criminals or celebrities (Wilkinson, 2014). Next, techniques that force responses from Wi-Fi enabled devices were proposed (Musa and Eriksson, 2012). The proposed methods aim to increase the detection probability of a Wi-Fi device. One method is emulating an AP using a popular SSID. The AP then broadcasts query packets resulting in the response packets from the Wi-Fi devices which used to connect with the popular SSID.

Since a number of possible threats can be raised from MAC identification, an approach to preventing individual tracking has been developed. Considerable attention has been paid to MAC randomization since the method aims to replace the actual MAC-ID with random

MAC-IDs during the Wi-Fi communication. Here random MAC-IDs can be locally assigned to an individual Wi-Fi device and be used during the unassociated state when the device performs active scans. A random MAC-ID will be employed for a short duration. Therefore, MAC randomization should be sufficient for preventing MAC address tracking in the long run.

MAC randomization was firstly available for Apple mobile devices in 2014 with the release of iOS8. For Android devices, Google introduced the function in 2015 with the Android6, Marshmallow. However, it was found that not all Android manufacturers enable MAC randomization. In addition, researchers have suggested that some Wi-Fi devices can be detected even though MAC randomization techniques are implemented. Firstly, a content-based attack which leverages the information in PRQ packets has been introduced (Freudiger, 2015; Vanhoef et al., 2016; Martin et al., 2017). Next, a timing-based attack observes the temporal patterns of the network scan to isolate an individual device (Matte et al., 2016; Waltari and Kangasharju, 2016).

It can be concluded that the possibility of identifying an individual using passive Wi-Fi data is arguable. Any attempted identification using SSID data alone might be unsuccessful since the data could be too general, as can be seen in Table 3.2. One potential way to identify an individual's MAC-ID is to follow the target person with a Wi-Fi scanner. The MAC-ID of the target can be identified when only one MAC-ID remains during the presence time of the target. Also, a method for handling MAC randomization is required if the target device implements MAC randomization. Such an identification method is considerably intrusive without the target's consent.

To this end, there is clearly a need for further discussion on the use of passive Wi-Fi data. An extreme case could result in collecting passive Wi-Fi data that is illegal. In such a case, numerous applications and services that could benefit millions of people would be prohibited, even though the publicly broadcast data already exist in the air (Watts et al., 2011). To maintain the balance between user profits and user privacy, clear guidelines and policy on collecting and using passive Wi-Fi data should be defined.

3.5.4.3 Guidelines on developing a system based on passive Wi-Fi data

The fact remains that the issue of passive Wi-Fi data which can be collected without a smartphone user's consent raises a number of privacy concerns. Basically, the smartphone user can acknowledge and accept the terms and conditions of a smartphone application in order to use the provided service. By their acceptance, the users already provide their direct consent for any actions indicated in the complicated terms. Some applications provide a

better notification format as a pop-up message when the application needs to use the built-in sensors or access smartphone data.

Deriving the direct consent from users is challenging for some sensor devices since the data collection process can be done passively without any direct communication channel to notify the involved users. For instance, a video surveillance system may be installed in a shopping mall for security purposes. Without an obvious notification, some customers may not be aware that their activities have been recorded. In addition, customers do not have an option to avoid being recorded. Collecting passive Wi-Fi data is comparable to the video surveillance case. However, the people who acknowledge that their data are potentially being collected can choose not to share their Wi-Fi data by turning off the Wi-Fi function of their devices. Similar challenges could be raised in the future when the Internet of Things and ubiquitous computing technologies become universal. Data collection and processing will then be performed by any devices without any spatial or temporal constraints.

Due to the challenge of gaining public acceptance, any system that is based on passive Wi-Fi data should be carefully designed. Previous research and case studies should be considered in the system design and implementation. Here guidelines for developing a system based on passive Wi-Fi data can be proposed with two major key components: data security and customer notification.

First, the system has to provide data security in both data collection and data processing. Data depersonalization is a requirement to ensure that a MAC-ID will not be linked to an individual. The system should not store the Wi-Fi data with the captured MAC-ID. A function to transform the captured MAC-ID to cryptographic data should be provided regardless of the availability of MAC randomization in the user's device. An encrypted MAC-ID needs to be as unique as the device identifier and should not be decrypted to reveal the captured MAC-ID. Then, only necessary information should be retained for each Wi-Fi record. Other irrelevant information such as SSID data needs to be discarded in order to prevent context-based attacks on MAC randomization. Furthermore, security must be provided for data storage especially for the centralized system. Finally, the system should control data access from unauthorized users.

Second, an appropriate policy for keeping customers informed should be determined. The system implementation should be announced in detail in order to inform the customers in advance. As can be seen from the TfL study case, the announcement should consist of general information about the system such as the data collection method, the collected information, the data collection purposes, data usage, and customer benefits. While general information can introduce the basic concept, security information should also be provided to ensure user privacy. The methods for privacy protection should be elaborated. Moreover, the user should understand that they can opt out from data collection. In such cases, the device

configuration method should be described. Finally, the information should be disseminated using multiple types of media in order to ensure that significant information is well perceived by the customers.

3.6 Summary of findings

Wi-Fi technology has not yet been specifically developed for mobility tracking purposes particularly for the development of public transport information systems. Although passive Wi-Fi data have been used for human mobility analysis, the characteristics of the Wi-Fi data for mobility tracking have still to be adequately addressed in several dimensions. As the foundation for further development of public transport information systems, this chapter provided a detailed investigation of passive Wi-Fi data characteristics.

To be more specific, temporal uncertainties in passive Wi-Fi data are evaluated based on designed experiments. The results show that missed detection probabilities could be different for each individual Wi-Fi device. Also, the chance of missed detection is decreased when the device locations are close to the Wi-Fi scanner. In addition, the Wi-Fi scanners which were configured with lower channel hopping velocities (e.g. 1 ch/s) can capture more detection events from Wi-Fi devices resulting in a lower missed detection probability.

Based on the experimental results, two uncertainty models are proposed for describing the uncertainties in device positioning and detection periods. It can be concluded that the range of activity duration error can be affected by (a) the frequency of capturing a detection event from a Wi-Fi device, (b) the travel speed of Wi-Fi devices, and (c) the size of study area in the detection range of Wi-Fi scanner. In future research, such findings can be advantages for evaluating the reliability of mobility analysis based on passive Wi-Fi data.

With the abovementioned findings, the next chapter of this thesis will focus on developing a system for average passenger waiting time estimation at an urban bus stop. This is in fact one of the key performance indicators for estimating the quality of bus services at individual bus stops. This research direction could lead to a new method for deriving average passenger waiting times which have not been estimated automatically from individual passengers. The passenger waiting time estimation algorithms will be developed based on the passive Wi-Fi data characteristics in the following Chapter 4. It is noteworthy that the bus-stop-based waiting time can be considered as an initial study for further development i.e. the route-based passenger waiting time for individual bus routes which is a crucial parameter for improvements of bus route design and operation.

Chapter 4

Passenger waiting time estimation at a single bus stop

In connection to the study objective (2) in Chapter 1, this chapter proposes an alternative method for bus passenger waiting time estimation using passive Wi-Fi data. With the underlying mechanism for Wi-Fi communication, the presence of Wi-Fi enabled devices in a particular area can be passively discoverable. The mobile devices carried by bus passengers can be exploited to estimate passenger waiting time at a bus stop without the passengers' direct participation. Passenger waiting time estimation using such opportunistic data is challenging due to the particular characteristics of the Wi-Fi data described in Chapter 3, as well as the additional challenges from the Wi-Fi data collected from bus stop environments. This chapter proposes a methodology to handle massive noise in Wi-Fi data and identify the potential Wi-Fi records which are derived from passengers' devices. The filtered data can then be used to estimate passenger waiting time. Two bus stops were selected for system evaluation. To derive more justifiable results, the two case studies have different traffic conditions and passenger crowding at the bus stops. The practicality is investigated in terms of estimation accuracy and insightful analysis.

4.1 Introduction

Passenger waiting time constitutes the part of the total transit trip time which could be the most challenging factor to measure. For bus transport, the value of waiting time is crucial as it could represent the duration of the passengers' exposure to traffic emissions. Since passengers have direct experiences of waiting for buses, the waiting time could affect their perceived performance of bus services. Average passenger waiting time (AWT) has been suggested as one of the performance measures to evaluate the quality of bus services.

Furthermore, AWT has been considered in transit assignment models for operational improvement and planning (Guihaire and Hao, 2008; Liu et al., 2010).

In previous literature on this topic, indirect methods have been developed to estimate AWT at a bus stop as a function of bus headway. Although observing bus headway is simple, the estimation relies on two major assumptions regarding passenger arrival patterns and bus headway distribution which can be different for individual bus stops. The fact remains that estimating AWT at a bus stop is challenging in practice.

Unlike other modes of public transport, the AWT of bus passengers could be affected by a wider range of factors on supply and demand sides. As described in Chapter 2, the reliability of bus services can vary due to the effects of several factors on actual bus operations e.g. road traffic conditions. Also, individual passengers may have their own criteria for boarding a bus, especially when common bus routes are operating at the same bus stops. In addition, some passengers may choose to wait for the next bus if the arriving bus is overcrowded. Third, the availability of real-time bus arrival time information could affect the passengers' journey plan and their boarding decision. Several studies have found that the assumption regarding uniform passenger arrival can be violated, since passengers may time their arrival even for the frequent bus services (Luethi et al., 2007; Frumin and Zhao, 2012; Ingvardson et al., 2018).

For densely populated cities such as Hong Kong, making assumptions on passenger arrival patterns and bus services is more onerous. Firstly, passengers may arrive at a bus stop in a batch (Fernández, 2010). The batch arrival phenomenon could be encountered at multimodal transit stops, and/or in the peak time period of routine activities such as after office hours. Next, high passenger loading could affect passenger behavior. Passenger serving patterns may not be based on a first-in, first-out order as some passengers may prefer to wait for the next bus. Although high-frequency bus services are operated in Hong Kong, passenger waiting time could be longer than expected. Another challenge for Hong Kong is the accessibility of bus arrival time information. The bus system is competitive and bus operators are responsible for their own investment and operating costs. Therefore, bus agencies have the right to withhold any information (i.e. bus arrival time) which could affect their revenue.

To tackle these challenges, this chapter proposes an alternative solution for estimating bus passenger waiting time using opportunistic data. To be more specific, it is proposed that passive Wi-Fi data be considered for AWT estimation owing to the ubiquitous nature of the mobile devices carried by bus passengers (e.g. smartphones) as well as the general use of Wi-Fi technologies. The Wi-Fi data can be collected without the direct participation of the bus passengers in the data provision. Since the underlying mechanism for Wi-Fi communication is exploited by the proposed method for AWT estimation, the strong assumptions regarding passenger arrivals and bus headway could be unnecessary.

Despite the advantages of passive Wi-Fi data, a number of challenges pose extra difficulties in AWT estimation. First, massive noise is included in Wi-Fi data collected from bus stop environments. Second, the temporal resolution of Wi-Fi data is uncertain among different devices. Finally, the Wi-Fi data would only be partial as it is unlikely to capture the data from all waiting passengers. To this end, a methodological framework is proposed in this chapter for handling the challenges in AWT estimation based on the passive Wi-Fi data collected from bus stop environments. In this study, AWT is initially estimated for each bus stop regardless of individual bus lines. With the continuous availability of stop-based AWT, bus transit performance can be evaluated in the spatial aspect over the transit network and in the temporal dimension over time.

This chapter is organized as follows. Section 4.2 provides the essential background on acquiring passive Wi-Fi data, and thereafter reviews the related works. Section 4.3 presents an overview of the proposed system. The details of significant processes including data preparation and AWT estimation are described in Sections 4.4 and 4.5, respectively. Section 4.6 presents the results with insightful analysis performed on a case study. Finally, a summary of findings is given in Section 4.7.

4.2 Background

In the review of the previous related literature in Chapter 2, one study using methodology similar to the proposed system adopted passive Wi-Fi data to estimate people dwell time for indoor activities. The most challenging task is identifying the precise entry and exit time at a study location since the resolution of passive Wi-Fi data can be sparse over the temporal dimension. A device can broadcast a Wi-Fi packet anywhere between a few seconds and at periods of several minutes. Therefore, the estimation accuracy relies on the frequency with which Wi-Fi data can be detected from the device.

The findings from Chapter 3 indicate that the detection frequency of a Wi-Fi device at a particular location could be affected by various factors including the device characteristics (models and current operational states), the scanner characteristics (detection range), and interference of environmental/physical obstacles in wireless communication. Due to the variability in the detection frequency, three major approaches have been implemented for dwell time estimation. Table 4.1 summarizes the combination of approaches which have been adopted in previous studies.

Table 4.1: Summary of the previous studies on people dwell time/waiting time estimation using Wi-Fi data

Authors	Participation of Smartphone Users	No. of Wi-Fi Scanners	Waiting Time/Dwell Time Resolution	Study Location	Occupancy Duration
Wang et al. (2014)	Requisite	Single	Continuous	Indoor (a coffee shop, an airport)	Short-Intermediate
Manweiler et al. (2013)	Requisite	Single	Discrete	Indoor (a café, a library)	Short-Intermediate
Shu et al. (2016)	Non-requisite	Multiple	Continuous	Indoor (an airport)	Intermediate-Long
Yan et al. (2017)	Non-requisite	Multiple	Discrete	Indoor (a shopping mall)	Short-Long
Le et al. (2017)	Non-requisite	Single	Discrete	Indoor (an office)	Short-Long

To begin with participatory-based systems, participants can be asked to install a smartphone application which can improve the detection frequency by increasing Wi-Fi data broadcasting. The participatory-based systems are capable of estimating short occupancy durations with high accuracy (Wang et al., 2014; Manweiler et al., 2013), whereas non-participatory based systems are suitable for capturing the activities on which people tend to spend a considerable amount of time in the study locations.

Next, multiple-location based systems were introduced. Wi-Fi scanners can be installed at a point of interest as well as in its surrounding areas. In this way, people mobility can be tracked over space and time. Localization techniques were implemented to enhance the accuracy of spatial information on a human trace (Shu et al., 2016; Yan et al., 2017). Dwell time at the point of interest was estimated by considering the entry and exit time at the location. This method was implemented in large buildings with more open space e.g. an airport terminal and a shopping mall. In contrast, single-location based systems were employed for smaller study areas such as a coffee shop and an office. Using a single scanner substantially reduces the monitoring area and results in the lower accuracy of dwell time/waiting time estimation.

Finally, customer behaviors were classified based on the time spent at the study locations (Manweiler et al., 2013; Yan et al., 2017; Le et al., 2017). With this method, the time resolution is reduced into a discrete scale since the major objective is not estimating dwell time. As this chapter aims to estimate AWT at a bus stop, none of the three major approaches can be directly adopted. Firstly, the system should be passive. Direct participation in data collection should not be a requirement. Moreover, the multiple-location based system is not suitable since bus stop areas are limited. Lastly, the waiting time needs to be estimated as a continuous variable.

It should be highlighted that the proposed system involves two additional challenges. First, massive noise from non-potential data sources can be included in Wi-Fi data collected from bus stop environments. Second, there is a time limitation for detecting waiting passengers. Waiting time could be short due to the regularity of bus services at a bus stop. Consequently, the waiting time might be insufficient for detecting Wi-Fi data from passengers' devices. To overcome the difficulties of using passive Wi-Fi data for AWT estimation, this study proposed a methodological framework which exploits multi-day Wi-Fi data of the potential waiting passengers at a bus stop. Then a new classification feature is introduced in order to facilitate AWT estimation.

4.3 System overviews

The proposed system is comprised of three major processes: data collection, data preparation, and AWT estimation. The operational overviews are shown in Figure 4.1.

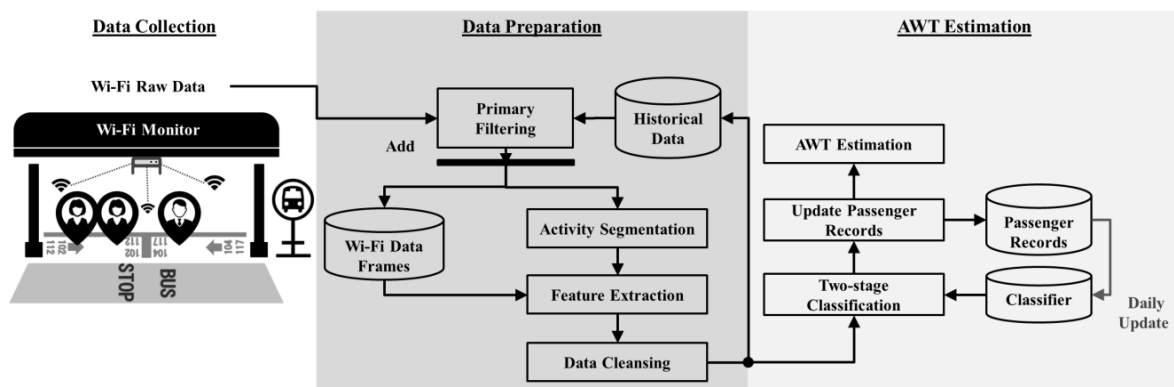


Figure 4.1: Overviews of the proposed system for AWT estimation at a single bus stop

- **Data collection**

Suppose that a Wi-Fi scanner is installed in the middle of the passenger waiting area of a bus stop, and an omnidirectional antenna is applied to the scanner. The detection area becomes a circular shape with a radius R . Then, the distance from the scanner to the end of the passenger waiting area, r_w , can be measured. Figure 4.2 shows the detection range of a scanner installed at a bus stop.

The Wi-Fi scanner is responsible for capturing the Wi-Fi data in the detection range. In general, a captured Wi-Fi packet consists of numerous data required for communication. To minimize database storage and lessen privacy concerns, only four data fields are retained as a detection event for AWT estimation:

- a) **MAC-ID:** An identifier for each Wi-Fi device.
- b) **Timestamp:** The timestamp of the detection event.
- c) **RSSI:** Received Signal Strength Indicator.
- d) **Frame subtype:** Identifying communication purposes.

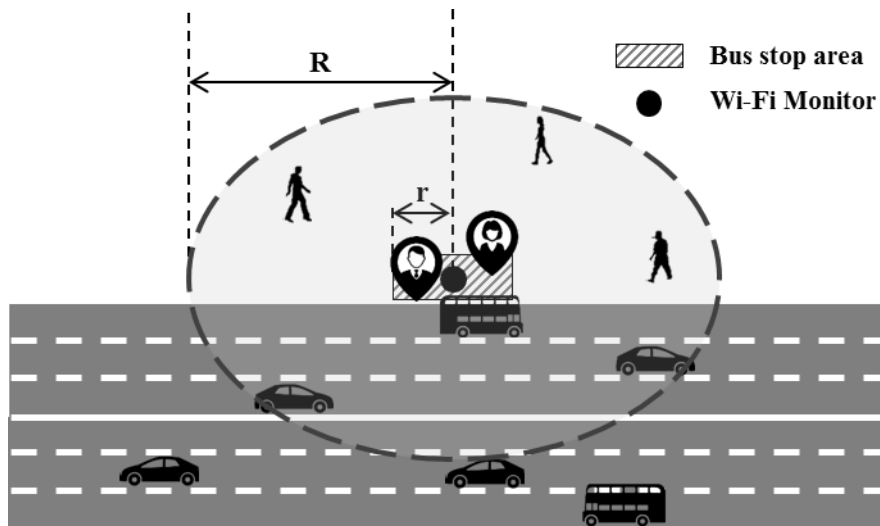


Figure 4.2: Detection range of a Wi-Fi scanner installed at a bus stop area

- **Data Preparation**

Meaningful information from detection events is extracted in this process. The information is used to facilitate data filtering which aims to remove noise from the Wi-Fi data.

- **AWT estimation**

The main objective is to distinguish the set of Wi-Fi data which belong to waiting passengers from the remaining noise. This is the most challenging task since the proportion of Wi-Fi data from waiting passengers is remarkably low. To overcome this difficulty, classification features are introduced and adopted with a machine learning technique.

4.4 Data preparation

This section describes the processes of preparing the raw data for AWT estimation: primary filtering, activity segmentation, feature extraction, and data cleansing. Most of the noise data will be removed through each process. Also, the raw data will be transformed in order to facilitate AWT estimation.

4.4.1 Primary Filtering

The primary sources of noise in Wi-Fi data are stationary devices nearby the bus stop, i.e. access points. The Wi-Fi data captured from access points can be promptly identified using frame subtype information in the detection events. Particular frame subtypes which only identify the action of the access points (e.g. probe response and beacon) can be recognized. Moreover, a dynamic lookup database recording a list of stationary devices nearby the bus stop can be established. The method for establishing the lookup database has been proposed (Hu et al., 2015). With the lookup database, frame subtype information may not be necessary for primary filtering. However, the data field is beneficial for instantaneous filtering which could minimize the computational time of the primary filtering process.

4.4.2 Activity Segmentation

Since a passenger may appear in a bus stop area several times in a day, the detection events of a MAC-ID can be captured from multiple sessions. It is essential that each detection event needs to be assigned to the proper session. The session sequence can be identified by considering the continuity of event detection over time. The basic idea is that consecutive events of the same session are supposed to be captured within a short time period and a session should be ceased when a MAC-ID is no longer detected within a pre-defined time window. The time window can be determined using the detection period between consecutive detection events of the same MAC-ID.

A ten-minute interval is identified in this study based on the results of Experiment#1 in Appendix B, assuming that most Wi-Fi devices can be captured within the ten-minute interval. The assumptions are consistent with the previous study (Freudiger, 2015) which showed that the average detection period of some smartphone models exceeds five minutes for an idle state but never reaches ten minutes.

It is noteworthy that the amount of Wi-Fi data is not reduced during this process. The outcome only assigns a session sequence to the existing detection events of individual MAC-IDs. Once the detection events are segmented into individual sessions, the next step is to extract essential information for each session.

4.4.3 Feature Extraction

A Wi-Fi record is defined as representing the meaningful information of a session. The information can be constructed from the detection events captured during the session. The

Wi-Fi record of a MAC-ID, MAC , during a session number, seq , is described by a feature set $f v_{MAC}^{seq}$. Table 4.2 summarizes the information attributes of a Wi-Fi record categorized into 3 groups to describe (i) time information, (ii) statistics of detection period, and (iii) statistics of RSSI.

Table 4.2: The information of a Wi-Fi record

Attribute	Description
entry_time	The timestamp of the first detection event
exit_time	The timestamp of the last detection event
observation_count	Total detection events
observation_period_mean	The average time period between consecutive detection events
observation_period_sd	The standard deviation of the period between detection events
observation_period_max	The maximum period between consecutive detection events
RSSI_mean	Average RSSI
RSSI_sd	The standard deviation of RSSI
RSSI_frequency [range_1]	Frequency of RSSI values in range 1
...	...
RSSI_frequency [range_L]	Frequency of RSSI values in range L

During Wi-Fi communication, a Wi-Fi device could broadcast Wi-Fi packets as a burst. A set of packets from a MAC-ID can be detected several times in just a few seconds. As a result, some features including the statistics of the detection period and RSSI need a low-pass filter to reduce potential bias in the statistical values. Herein, for each MAC-ID, a basic Wi-Fi packet is considered if the MAC-ID has not been detected within the previous 3-second time window. Otherwise, the packet will be omitted. The low-pass filter is based on an empirical finding on the detection period of regular Wi-Fi devices (Hu et al., 2015).

One attribute set that should be highlighted is the frequency of RSSI values in various ranges. Given a set of Wi-Fi observations and supposing that RSSI values are discretized into L ranges, there will be L attributes which summarize the RSSI frequency for each range. Empirical studies should be conducted to justify the number of ranges as well as the minimum and maximum RSSI values that could be detected by the Wi-Fi scanner at the study location. In general, RSSI observations from a Wi-Fi device can fluctuate even though the device remains in the same state (e.g. the same location). The range of values should be based on the general standard deviation of RSSI observations during a session. According to the results of the designed experiment no.2 and no.3 in Appendix B, RSSI values are discretized into ten ranges from -80 dBm (weak) to -40 dBm (strong). In cases where RSSI values exceed the defined range, the closest range is considered.

4.4.4 Data Cleansing

The final step of data preparation is to filter the Wi-Fi records which do not correspond to waiting passengers. Two basic characteristics are considered for filtering: (i) the presence time of the device, and (ii) the number of Wi-Fi observations.

4.4.4.1 Presence time

The presence time of a Wi-Fi record is derived from `entry_time` and `exit_time` attributes. Minimum and maximum duration thresholds are determined for filtering the devices which are unlikely to belong to waiting passengers. First, the minimum duration can be determined from the statistical information of bus dwell time at the bus stop whereas the maximum duration can be based on the manual observation of passenger waiting time. As a result, Wi-Fi records from on-board bus passengers and/or passing vehicles can be removed by the minimum duration threshold, while the records from other long-duration activities conducted nearby the bus stop can be filtered by the maximum duration threshold.

4.4.4.2 Unreliable observation

It is assumed that detection events will be detected regularly during a session. However, a Wi-Fi record may lack detection continuity during the session and the record can be considered insufficient for constructing reliable information and passenger waiting time estimation. In this study, the records which were observed only twice at the `entry_time` and `exit_time` are removed from the dataset.

4.5 AWT estimation

The majority of noise data are filtered out during data preparation. The remaining noise is assumed to be derived from mobile data sources e.g. passing vehicles, and passers-by. A promising way to distinguish between the data from waiting passengers and the data from other sources is identifying the device's position. If the device is in the bus stop area, it can be assumed to be a waiting passenger. Since the proposed system is non-participatory and based on a single Wi-Fi scanner, RSSI attributes could be a dominant indicator for such data classification.

This section firstly introduces the generalized classification features. Then, the methodologies for training a classifier are proposed based on the sparse nature of Wi-Fi detection in the

temporal dimension, followed by a two-stage classification which could reduce computational time.

4.5.1 Classification features

In order to perform data classification, the difference between the characteristics of the Wi-Fi records from waiting passengers and the noise data needs to be recognized. In spatial-temporal dimensions, waiting passengers tend to spend a considerable period at the bus stop area. Consequently, it can be assumed that most Wi-Fi data will be detected in the waiting area. On the other hand, passers-by who walk through the detection range of the Wi-Fi scanner with a constant walking speed should not spend a considerable amount of time at a particular location. Most of the walking time would be spent outside the bus stop.

Suppose that the relationship between the RSSI and the distance from the Wi-Fi scanner is a direct variation. The RSSI can represent the device location at the detection time. To this end, the statistics of RSSI observations are used for establishing the classification features for each Wi-Fi record. First, the frequency distribution of the RSSI values is generalized to simplify the classification. The generalized features consist of four data fields which describe RSSI distribution in terms of the distribution peak and range:

- **dist_min:** The range of minimum RSSI.
- **dist_max:** The range of maximum RSSI.
- **dist_width:** The number of ranges from dist_min to dist_max.
- **dist_mean:** The range of average RSSI based on the value of RSSI_mean in the Wi-Fi record.

Range Index	1	2	3	4	5	6	7	8	9	10	RSSI_mean
Upper Bound RSSI	-81	-76	-71	-66	-61	-56	-51	-46	-41	∞	
Lower Bound RSSI	$-\infty$	-80	-75	-70	-65	-60	-55	-50	-45	-40	
MAC A	0	0	0	0	1	1	2	2	1	0	52
MAC B	0	0	1	1	4	0	0	0	0	0	65

↓

Generalized value	dist min	dist max	dist width	dist mean
MAC A	5	9	5	7
MAC B	3	5	3	5

Figure 4.3: Generalized classification features of the RSSI observations from a waiting passenger (MAC A), and a passing vehicle (MAC B)

Figure 4.3 demonstrates the feature generalization of two Wi-Fi records. It can be seen that most of the RSSI observations from a waiting passenger (MAC A) are in the high range of values, whereas the observations from a passing vehicle (MAC B) lie in the lower ranges. This could imply that most observations from a waiting passenger were detected in the bus stop area near the location of the Wi-Fi scanner, while the observations from a passing vehicle were detected at a location away from the scanner.

4.5.2 Classifier

The classifier requires a learning phase for adjusting the relationships between the input features and classification results. In this study, labeling the features is challenging due to the lack of known MAC-IDs from bus passengers. A practical way to identify noise data is based on manual observation of individual waiting passengers at the bus stop.

4.5.2.1 Training assumptions

With the manual observation of passenger waiting times, the duration of time when there is no waiting passenger at the bus stop can be identified. Then, the Wi-Fi records captured during such duration can be assumed as noise data. In addition, repeated MAC-IDs can be discovered in multi-day observations. The MAC-IDs could be assumed as waiting passengers in the cases where the MAC-IDs are detected at a similar period for several days.

4.5.2.2 Data matching problem

Also, the Wi-Fi records from waiting passengers can be assumed by data matching. Let M , and N be the number of Wi-Fi records and total waiting passengers respectively. The objective is to match an observed waiting time N to a potential Wi-Fi record M . This becomes a matching problem and an assignment function μ can be defined:

$$\mu: \begin{cases} \{1, 2, \dots, N\} \rightarrow \{1, 2, \dots, M\} \\ \mathbf{a} \mapsto \mathbf{b}, \mathbf{a} = 1, 2, \dots, N, \mathbf{b} = 1, 2, \dots, M \end{cases} \quad (4.1)$$

where $\mu(a) = b$ indicates that the Wi-Fi record b is assumed to be detected from the observed passenger a .

The challenge is that a Wi-Fi record cannot be matched to an observed waiting time directly due to the sparseness of Wi-Fi data in the temporal dimension. To solve the matching problem, time window constraints are introduced for identifying potential Wi-Fi records.

4.5.2.3 Identifying potential candidates

The time window constraints can be formulated based on the observed passenger arrival time at the bus stop, and the departure time of the boarded bus. An assumption is that the Wi-Fi record from an observed passenger should be detected during the wait or a similar time period. Figure 4.4 illustrates the time window constraints of an observed passenger in space and time dimensions, while Table 4.3 summarizes the case of constraints with the factors for determining time windows.

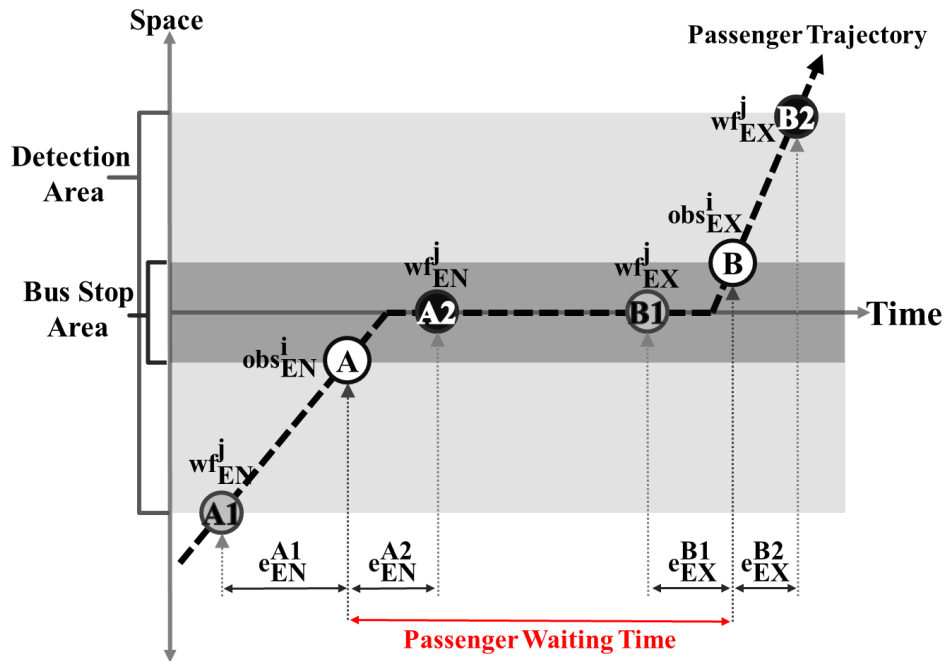


Figure 4.4: Time window constraints for identifying potential candidates

Table 4.3: The case of time window constraints

Case	Target Time Window	MAC Detection Time	Factors for Determining Time Windows
A1	Passenger arrival	Before passenger arrival	Walking speed
A2	Passenger arrival	After passenger arrival	Detection period
B1	Bus departure	Before bus departure	Detection period
B2	Bus departure	After bus departure	Bus speed

For the case of A1, the first Wi-Fi observation wf_{EN}^j can be detected before the actual passenger arrival at the bus stop obs_{EN}^i since the scanner detection range covers a distance further than the bus stop area. In contrast, the observation can also be detected after the passenger arrival due to the sparseness of Wi-Fi detection (case A2). Hence, the first time window is specified based on the observed passenger arrival time with a lower bound $obs_{EN}^i - e_{EN}^{A1}$ and an upper bound $obs_{EN}^i + e_{EN}^{A2}$. In the same way, the last Wi-Fi observation can be detected either before the bus departure (case B1) or after the departure (case B2).

Another time window based on bus departure time can be described as well from $obs_{EX}^i - e_{EX}^{B1}$ to $obs_{EX}^i + e_{EX}^{B2}$.

Given an observed passenger i with arrival time obs_A^i and departure time obs_D^i , a Wi-Fi record j will be considered as a potential candidate of the observed passenger when the entry_time and exit_time are within the following constraints:

$$obs_{EN}^i - \frac{R - r_w}{v_{ped}} \leq wf_{EN}^j \leq obs_{EN}^i + \left(f^j - \left(\frac{R - r_w}{v_{ped}} \right) \right) \quad (4.2)$$

$$obs_{EX}^i - \left(f^j - \left(\frac{R - r_w}{v_{bus}} \right) \right) \leq wf_{EX}^j \leq obs_{EX}^i + \frac{R - r_w}{v_{bus}} \quad (4.3)$$

The notations are described as follows:

obs_{EN}^i : The observed arrival time of a passenger i at the bus stop.

obs_{EX}^i : The observed departure time of the boarded bus.

wf_{EN}^j : The entry_time attribute of a Wi-Fi record j .

wf_{EX}^j : The exit_time attribute of a Wi-Fi record j .

R : The radius of the scanner's detection area (m).

r_w : The distance from the scanner to the end of the passenger waiting area (m).

v_{ped} : Average walking speed (m/s).

v_{bus} : Average bus speed from bus departure to the end of the detection area (m/s).

f^j : The estimated maximum detection period between Wi-Fi observations (s) which is determined by

$$f^j = \begin{cases} wf_{max}^j, & CV > 0.5 \\ wf_{mean}^j + 2wf_{sd}^j, & otherwise \end{cases} \quad (4.4)$$

where wf_{mean}^j , wf_{sd}^j , wf_{max}^j denote the attributes observation_period_mean, sd, and max of a Wi-Fi record j . The value of f^j is dependent on the coefficient of variation CV which is the ratio of the standard deviation to the mean of the period between observations. Identifying f^j is challenging due to the variability in the detection period among different Wi-Fi devices, or even in the same device but different states. On the one hand, the value should be sufficient to include a potential Wi-Fi record in the time windows. On the other hand, the time windows should not be too large to prevent the inclusion of excessive noise data in the candidate set.

In general, the value of f^j can be determined using the maximum period between Wi-Fi observations. In some cases, the maximum period may not be representable. The case A2 in Figure 4.4 can be used as an example. The first Wi-Fi record may be derived once the passenger used the device after passenger arrival, obs_{EN}^i . However, the device could be idle a significant of time from the arrival to the active state. Assuming that Wi-Fi records were only captured during the active state, the detection periods could be very short. Then the maximum period derived from the active device could be too short for including the Wi-Fi records in the time windows in (4.2) since the value of $f^j - \left(\frac{R-r_w}{v_{ped}}\right)$ is incomparable to e_{EN}^{A2} . A similar condition could occur for case B1 since the device could be idle a significant time before the bus departure. A way to solve this issue is extending the maximum period.

As only the mean and standard deviation are available, the distribution of detection periods could be assumed for estimating f^j . Firstly, it is assumed that the detection periods are not significantly varied when the mean is twice the standard deviation ($CV \leq 0.5$) and the detection periods are normally distributed. Consequently, f^j can be determined by adding twice the standard deviation to the mean and assuming that the maximum period is in the normal distribution.

It is noteworthy that subtracting the walking time to the bus stop or the bus travel time from the bus stop from the estimated maximum detection period could result in the negative values in such term with a small value of f^j . Therefore, this study assumes that the effect of walking time and bus travel time can be compromised in the time window constraints. The side effect is that more Wi-Fi records may be considered as the candidate since the time windows are slightly widened. The time window constraints can be simplified as follows:

$$obs_{EN}^i - \frac{R - r_w}{v_{ped}} \leq wf_{EN}^j \leq obs_{EN}^i + f^j \quad (4.5)$$

$$obs_{EX}^i - f^j \leq wf_{EX}^j \leq obs_{EX}^i + \frac{R - r_w}{v_{bus}} \quad (4.6)$$

The simplified constraints will be considered for system implementation and discussion from this point.

4.5.2.4 A modified bipartite matching method

A bipartite graph $G = (U, V, E)$ is demonstrated in Figure 4.5. The vertices are divided into two disjoint sets for observed data U , and Wi-Fi records V . Each edge $e(i, j), i = 1, 2, \dots, N; j = 1, 2, \dots, M$ corresponds to a potential match between an observed passenger and a Wi-Fi record. The weight associated with each edge is defined:

$$\varepsilon(i, j) = \frac{|wt_i - at_i|}{wt_i} \quad (4.7)$$

where a function $\varepsilon(i, j)$ evaluates the percentage error between the observed waiting time of a passenger i , wt_i and presence duration of a Wi-Fi record j , at_i . The presence duration is derived from the entry_time and exit_time attributes.

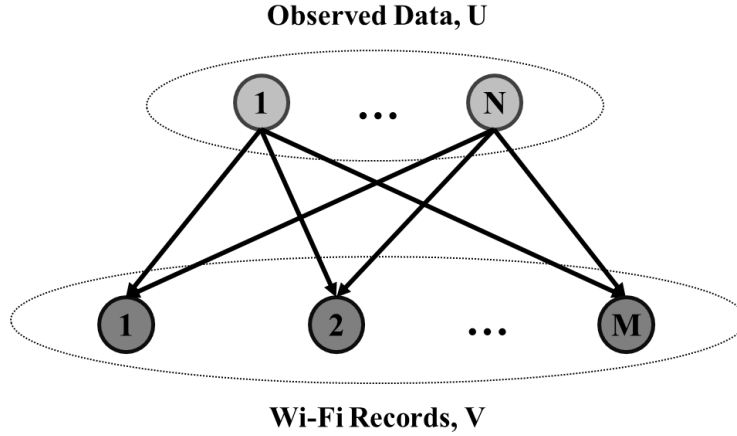


Figure 4.5: Bipartite graph representation

Two modifications can be applied to the bipartite graph. First, an edge $e(i, j)$ is removed from the graph if a Wi-Fi record is not a potential candidate for a waiting passenger. Second, a maximum weight threshold γ_{wg} can be applied as a filter to reduce the probability of misclassification. The threshold is used to ensure that the difference between actual waiting time and the presence duration of a Wi-Fi record is acceptable. In this study, the threshold is 0.4 based on empirical analysis of repeated MAC-IDs. All the edges with undesirable weight are removed from the graph, as well as the nodes without a linking edge. This implies that some observed passengers may not be assigned a Wi-Fi record since Wi-Fi data may not be captured from every passenger.

Finally, in order to find an optimum solution for the graph where each observed passenger node matches with the most potential Wi-Fi record, a minimum-weight bipartite matching can be formulated as follows:

$$\min_{\mu} \sum_{i=1}^N \sum_{j=1}^M \varepsilon(i, j) \delta(\mu(i) = j) \quad (4.8)$$

$$\text{s.t.} \quad \delta(\mu(i) = j) \in \{0, 1\}, \forall i \in \{1, 2, \dots, N\}, j \in \{1, 2, \dots, M\} \quad (4.9)$$

$$\sum_{j=1}^M \delta(\mu(i) = j) = 1, \forall i \in \{1, 2, \dots, N\} \quad (4.10)$$

$$\sum_{i=1}^N \delta(\mu(i) = j) \leq 1, \forall j \in \{1, 2, \dots, M\} \quad (4.11)$$

The matching solution is adopted from the vehicle matching problem proposed by Wang et al. (2013) as the nature of the matching problem is equivalent when the modified bipartite graph is applied. The objective function (4.8) aims to find a solution with the minimum of overall weight. Herein the minimum weight implies the minimum of overall percentage errors between observed waiting times and device presence times. The constraint (4.9) ensures that the $\delta(\cdot)$ provides binary integers. A retained passenger node is required to match with a Wi-Fi record by the constraint (4.10). Finally, duplicate matching is not allowed for a Wi-Fi record. Only one device is assumed to be detected from a passenger. A Wi-Fi record is allowed to have at most one match by the constraint (4.11).

The classification features of a Wi-Fi record p which is matched to an observed passenger is recognized as the classification profile X_{train}^p of the waiting passenger class ($y_{train}^p = passenger$), while the features of a Wi-Fi record q which is never considered as a candidate of any observed passenger are recognized as the profile X_{train}^q of the noise data class ($y_{train}^q = noise$). Moreover, another parameter which can be learnt during the training process is the minimum of RSSI_mean for the waiting passenger class, γ_{RM}^{WP} . This feature could facilitate the two-stage classification in the next sub-section.

4.5.3 Two-stage classification

Classification processes need to handle a large number of Wi-Fi records due to the large proportion of noise in Wi-Fi data. Here, a two-stage classification is therefore proposed to reduce computational time. First, the trained parameter γ_{RM}^{WP} initially classifies some records in which the RSSI_mean is noticeably low. It is assumed that the classification profile of such Wi-Fi records tends to represent noise data rather than waiting passengers.

Next, the k-NN classifier is adopted to classify a Wi-Fi record based on the classification feature by selecting k minimum distance values $d(B, C)$ for a majority vote. The distance function $d(\cdot)$ measures the similarity between two classification features: the classification features of a testing dataset B and a classification profile C . Euclidean distance can be applied as the function. The value of k is determined using an elbow method. A validation accuracy curve can be plotted against various values of k to find an elbow point where the smallest k gives the nearly highest accuracy.

The k-NN method is selected for classification due to the challenges in cross-validation processes. There is no evidence to ensure that the classification results are correct without the direct request of MAC-IDs from waiting passengers. Multi-day observations are necessary for assuming the repeated MAC-IDs to be waiting passengers. Since the number of observations is limited, AWT estimation is problematic in the early state of the system. However, training k-NN classifier can be performed with limited datasets. Also, the training process requires short computational time which allows the classifier to be trained regularly. The training can be performed daily at the end of the day when more repeated MAC-IDs are available.

4.5.4 Average passenger waiting time

AWT at a bus stop can be estimated from the potential Wi-Fi records of waiting passengers. The estimation can be performed to derive AWT during a time interval τ which is denoted by:

$$AWT^\tau = \frac{1}{N} \sum_{j=1}^N wf_{EX}^j - wf_{EN}^j; wf_{EX}^j \in \tau \quad (4.12)$$

where wf_{EN}^j and wf_{EX}^j are the entry_time and exit_time of a Wi-Fi record j , and N is the total number of records during the time interval.

4.6 Empirical studies

Two case studies were conducted in order to evaluate the proposed system. The first case study focuses on evaluating the system performance during evening peak time periods when more noise data were included in the collected Wi-Fi data due to congested road traffic conditions and considerable pedestrian flows nearby the bus stop. Then the second case study was conducted in order to evaluate the system performance in different bus stop environments. The conditions of the two bus stops are summarized in Table 4.4

During the observations, passenger loading on the buses was usually high due to travel demands during the evening peak periods. The road traffic was mainly congested. Also, considerable pedestrian flows were observed nearby the bus stop area. The average number of waiting passengers at the bus stop during a two-hour observation was 58. Manual observations of individual passengers were recorded for evaluation purposes including

passenger arrival times at the bus stop area, bus departure times, and the boarded bus numbers.

Table 4.4: The conditions of two bus stops for system evaluation

	Case study #01	Case study #02
City	Hong Kong	Bangkok
Time period	Evening peak (17:00-19:00)	Inter-peak (14:00-16:00)
No. of bus lines	9	11
Average no. of waiting passenger (a two-hour period)	<u>58</u>	<u>115</u>
Road traffic condition	<u>Mainly congested</u>	<u>Free-flow</u>
Pedestrian flows	<u>High</u>	<u>Low</u>

A group of volunteers was asked to conduct such manual observations by standing nearby the bus stop areas. For each individual passenger, the volunteers needed to record the passenger arrival time at the bus stop. Then the boarded bus route number and the departure time of each passenger were recorded by identifying the passengers who boarded an individual bus during the bus dwell time at the bus stop. This means the volunteers needed to remember individual passengers at the bus stop area in order to identify the boarded passengers. Hence, the volunteers needed to write down some noticeable appearances of the passengers such as the color of their clothes.

Examples of the manual observations of individual passengers are provided in Appendix C. The records from a volunteer were compared to the others in order to derive the most accurate information. Finally, the four-day dataset was separated for training the classifier and testing the system. By using 3-day datasets for training and another dataset for testing, the multi-day evaluation was performed.

4.6.1 Case study #1: Hong Kong

4.6.1.1 Data filtering

As vital processes for handling noise data, data filtering performance is evaluated. Figure 4.6 is plotted from a two-hour observation. The massive noise in Wi-Fi data was significantly reduced after performing four core processes. First, the data from stationary devices, about 75% of the raw data, were removed through primary filtering. Next, feature extraction compressed the output from the primary filtering process into 13%. Also, ninety percent of the Wi-Fi records were filtered out through data cleansing as they were considered unreliable

for AWT estimation. Finally, the classifier identified 35 Wi-Fi records which accounted for 60% of the observed passengers.

Even though the raw data were refined by data filtering processes, the final result shows that the system cannot identify all waiting passengers. A basic reason for this can be that the devices are undetectable; either the Wi-Fi function was disabled or the passengers did not carry any Wi-Fi devices. Furthermore, some valid data could be lost due to the filtering mechanisms. For example, the Wi-Fi data from the lucky passengers who were able to board the bus without having to wait tended to be filtered out during the data cleansing process since the duration of their presence at the bus stop was too short and/or the number of Wi-Fi observations is insufficient.

In contrast, some invalid observations might be classified as waiting passengers. According to manual observations, it was found that some people can behave like a waiting passenger e.g. people who waited for a bus but finally left the bus stop without boarding a bus. The Wi-Fi data from those people are then invalid. To validate the filtering results and make a solid conclusion on data classification performance, a survey of passengers' MAC-IDs is necessary. Since the survey has not been conducted at this stage, AWT estimation accuracy is considered based on the main objective of this study.

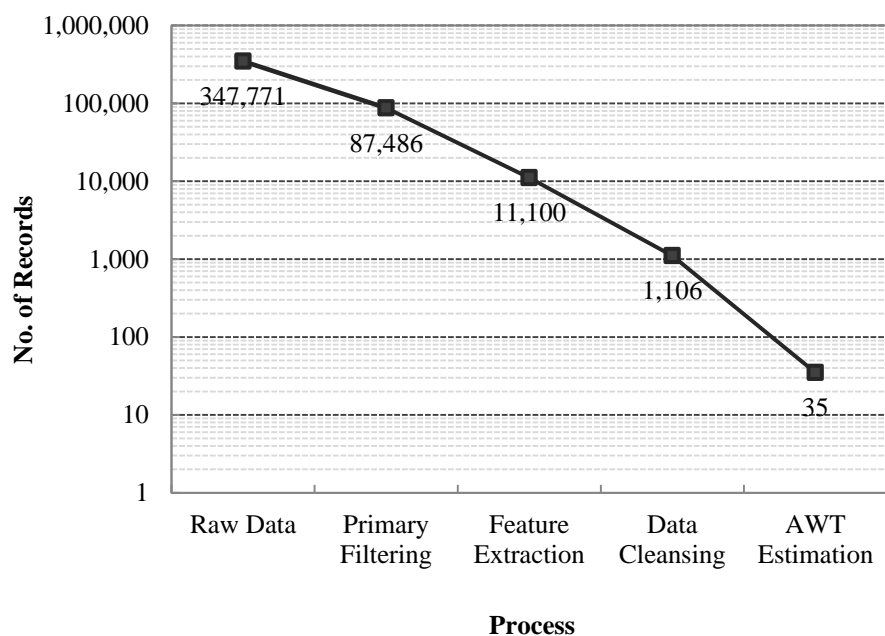


Figure 4.6: Data filtering performance (Case study #1)

4.6.1.2 Estimation results

The system performance is evaluated in terms of two measures: MAE and MAPE.

$$MAE = \frac{1}{N} \sum_{i=1}^N |AWT_{obs}^{\tau} - AWT_{est}^{\tau}| \quad (4.13)$$

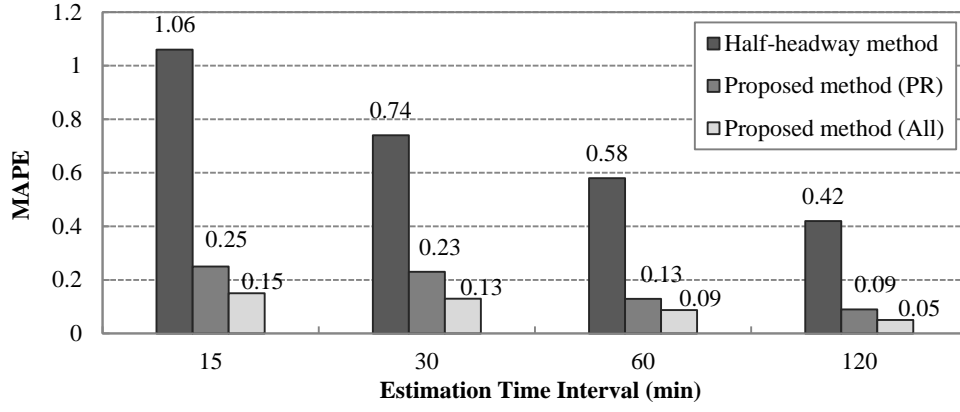
$$MAPE = \frac{1}{N} \sum_{i=1}^N \frac{|AWT_{obs}^{\tau} - AWT_{est}^{\tau}|}{AWT_{obs}^{\tau}} \quad (4.14)$$

where N is the number of data points, AWT_{obs}^{τ} and AWT_{est}^{τ} are the observed and estimated AWT at the bus stop during a time interval τ .

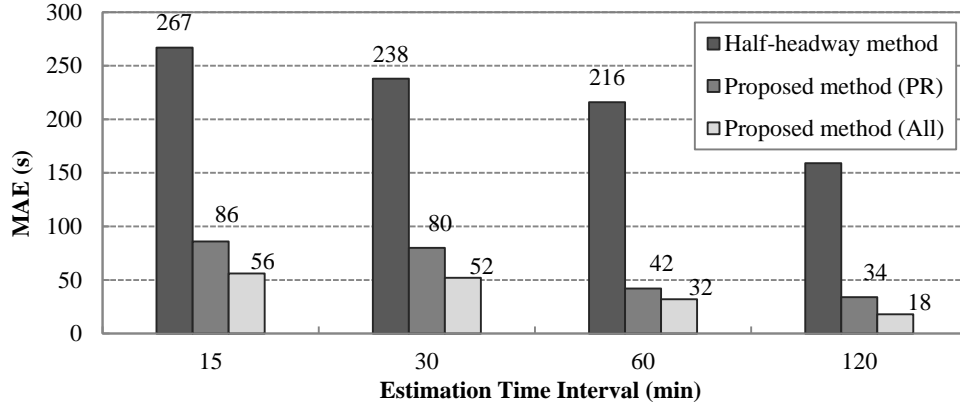
To analyze the system performance, three datasets are used for AWT estimation: (i) bus headway data, (ii) probe request (PR) packets in the raw data, and (iii) all Wi-Fi data. The first dataset is used for a half-headway method as the baseline performance. For the second dataset, only PR packets in the raw data are used for evaluation. Since the PRs are broadcast to all nearby devices, not to a specific one (e.g. a Wi-Fi access point), the effects of available Wi-Fi services nearby the bus stop can be investigated. The last dataset is raw Wi-Fi data which includes all captured packets.

To analyze data resolution effects on AWT accuracy, estimations were performed to provide AWT information for different estimation time intervals, i.e. every 15, 30, 60, and 120 minutes. Figure 4.7 shows the evaluation results in terms of (a) MAPE and (b) MAE based on different estimation intervals. The overall performance indicates that the proposed system improves AWT estimation from the baseline accuracy. The system can perform AWT estimation with 5-25% of MAPE and 18-86 seconds of MAE based on two significant factors: the detection period of Wi-Fi data, and the estimation interval.

There are two interesting points that should be highlighted here. First, the availability of Wi-Fi services nearby the bus stop improves the estimation accuracy. The proposed method performs better using all Wi-Fi data. The accuracy is decreased by up to 10% when only PR packets are considered. It should be noted that, in the PR dataset, not only are the Wi-Fi data from stationary devices removed but all the data involving nearby access points (i.e. data frames) are also discarded. Therefore, the detection period of passengers' devices in the PR dataset can be increased resulting in lower accuracy. Second, the accuracy is improved when the estimation interval is extended. The effects of the available Wi-Fi services and the estimation interval are further discussed in the following sub-sections.



(a)



(b)

Figure 4.7: AWT estimation accuracy for Case study #1 in terms of (a) MAPE, (b) MAE

4.6.1.3 Estimation errors in passenger waiting time

By assuming passenger waiting time from the presence time of a Wi-Fi record, a number of factors can affect the accuracy of individual waiting times. According to the time window constraints for identifying potential candidates (4.5) – (4.6), a Wi-Fi record can be detected at any point in the time windows. The range of estimation error for a Wi-Fi record j can be assumed based on the time windows.

$$RoE(j) = \left[-(2f^j), \frac{R-r}{v_{ped}} + \frac{R-r}{v_{bus}} \right] \quad (4.15)$$

The estimated maximum period between Wi-Fi observations f^j provides negative errors which result in underestimation of waiting time, whereas a long scanner detection range causes overestimation from positive errors. Since the variation of other parameters may not be critical, a crucial parameter here can be f^j .

In order to understand the typical detection period of a MAC-ID, the Wi-Fi records which are classified as waiting passengers were further analyzed to understand the typical detection period of a MAC-ID. Each Wi-Fi record consists of the average and standard deviation of the period between observations. Here, initial statistics of the typical detection period can be estimated. Firstly, the typical period could be assumed as the mean of the average values from passengers' Wi-Fi records. Also, the variation of the typical detection period is assumed from the mean of standard deviations.

For the dataset with only PR packets, the typical detection period is 59 seconds with 43 seconds of standard deviation. In other words, an observation can be detected from a MAC-ID every 59 seconds on average. The typical detection period is shorter for the dataset with all Wi-Fi data. An observation can be detected every 35 seconds with 32 seconds of standard deviation. This can explain the lower estimation accuracy when only PR data are considered. Filtering other packets results in a longer period between Wi-Fi observations and the range of error in (4.15) is extended. Moreover, the longer period could result in a higher probability of missed detection since the system requires more time to detect a waiting passenger.

Evaluation of the time for detecting a waiting passenger is valuable since it can describe the missed detection of short waiting times. Without the availability of passengers' MAC-IDs, one way to estimate the passenger detection time is to make assumptions based on Wi-Fi data characteristics. In the proposed system, at least three observations are needed so as not to filter out a valid Wi-Fi record during the data cleansing process. As a result, the average time duration for detecting a waiting passenger can be estimated using double the typical period between observations. To be more precise, the walking time to the bus stop area should be deducted since the scanner detection range is larger than the bus stop area. The relationship can be denoted by:

$$TD = 2TP - \left(\frac{R - r}{v_{ped}} \right) \quad (4.16)$$

where TD is the average time for detecting a waiting passenger, and TP is the typical period between observations of a Wi-Fi record.

4.6.1.4 Estimation errors in average passenger waiting time

Since AWT is calculated from individual Wi-Fi records during an estimation interval, the accuracy is reliant on the distribution of estimated waiting times. Table 4.5 presents a comparison of four passenger waiting time distributions during a two-hour observation. The distribution from manual observation shows that the majority of the waiting time is in the 1-10 minute range. According to the manual observation of the bus lines boarded, the passengers who waited for more than 10 minutes usually boarded the bus on uncommon routes with long headways. Also, most serving stops do not serve by other bus routes.

Table 4.5: Probability distribution of passenger waiting times

Dataset	Up to 1 minute	1-5 minutes	5-10 minutes	More than 10 minutes
Manual observation	0.11	0.49	0.25	0.15
Proposed method (all)	0.14	0.46	0.26	0.14
Proposed method (PR)	0.10	0.45	0.30	0.15
Half-headway method	0.07	0.22	0.41	0.30

Performing the proposed method using all Wi-Fi data provides waiting time distribution that is comparable to the observed one. The probability is slightly higher than the observation for the waiting times of up to 1 minute. The distribution could be shifted down due to the effects of the Wi-Fi detection period which results in the underestimation of waiting times. The distribution is even more different when the PR dataset is used. The probability of waiting times during 1-5 minutes is decreased while the probability of waiting times during 5-10 minutes is increased. It can be assumed that the lower accuracy is also affected by higher chances of missed detection, apart from the detection period. Since missed detection is generally encountered for short waiting times, the higher probability of the waiting times with a duration of 5-10 minutes could be reasonable.

For the half-headway method, the probability of waiting times with the duration of 1-5 minutes is lessened compared to the observed one whereas the probabilities are higher for the waiting times of more than 5 minutes. Based on the manual observation of passenger behavior, the accuracy of the half-headway method could be affected by the common bus routes. Since several bus lines are serving common stops, passengers can have multiple choices and they may decide to catch an alternative service depending on which bus arrives first. Also, the passenger arrival as a batch after office hours could violate the assumption of uniform passenger arrival.

The factors affecting the accuracy of AWT can be determined based on the waiting time distributions. Firstly, overestimation can be caused by missed detection. As discussed in the previous sub-section, the waiting time could be too short, especially for lucky passengers

who are able to catch their bus quickly. Otherwise, the presence duration of Wi-Fi records could be shorter than the bus dwell time and the records are excluded from AWT estimation.

Next, the accuracy can be decreased when the distribution of estimated waiting times is incomparable to the distribution of actual waiting times. Both overestimation and underestimation can be encountered. As can be seen from Table 4.5, using all Wi-Fi data provides more accurate waiting time distribution than is possible from the PR dataset. The result corresponds to the AWT estimation accuracy in Figure 4.7 which shows that using more Wi-Fi data results in better accuracy. In addition, as discussed in the data filtering results, misclassification can also occur. Invalid waiting times might be included in the distribution.

Finally, to estimate AWT during a time interval in (4.12), a Wi-Fi record might be included in an incorrect estimation interval due to the range of errors described in (4.15). In such a case, more errors are introduced to the distribution of estimated waiting times during the interval, and the accuracy of AWT during the interval is lessened. This implies that a shorter estimation interval can cause higher errors since a Wi-Fi record has a greater chance of being assigned to an incorrect interval. The results in Figure 4.7 show that extending the estimation interval can reduce the errors in AWT estimation. The accuracy is significantly improved for longer intervals (60 and 120 minutes).

It is worthwhile to notice that missed detection and unrepresentative distribution can also be encountered in the AWT estimation using manual observations, whereas the misclassification and incorrect estimation interval problems are introduced by Wi-Fi data. In addition, the performance of AWT estimation using the half-headway method should be investigated based on passenger behavior particularly for the bus stops with common bus routes.

4.6.2 Case study #2: Bangkok

A bus stop in Bangkok nearby Siam Technology College was selected for conducting the case study. The data collection was performed to derive four datasets in the same way as the first case study but the time duration was in the inter-peak time period (14:00-16:00). The bus stop is served by 11 bus lines which head towards different destinations in various directions. Some bus lines have common bus stops on the remaining routes. From the 11 bus lines, 32% of bus headways were less than 5 minutes, 25% between 5 and 10 minutes, and 43% more than 10 minutes.

The passenger loading on the buses was low-to-moderate and the passing vehicles traveled with free-flow speeds during the inter-peak time period. The pedestrian flows nearby the bus

stop area were insignificant. The average number of waiting passengers at the bus stop during a two-hour observation was 115.

4.6.2.1 Data filtering

Figure 4.8 shows the Wi-Fi data of a two-hour observation after performing the four core processes. In the same way as the first case study, the noise data were removed by the four processes. It can be noticed that the raw data captured during a two-hour observation were less than the Hong Kong case with a ratio of one to three approximately. The main reason is that the data collection in Bangkok was conducted during the inter-peak time period. Hence, there were fewer passing vehicles and passersby in the monitoring area during the time.

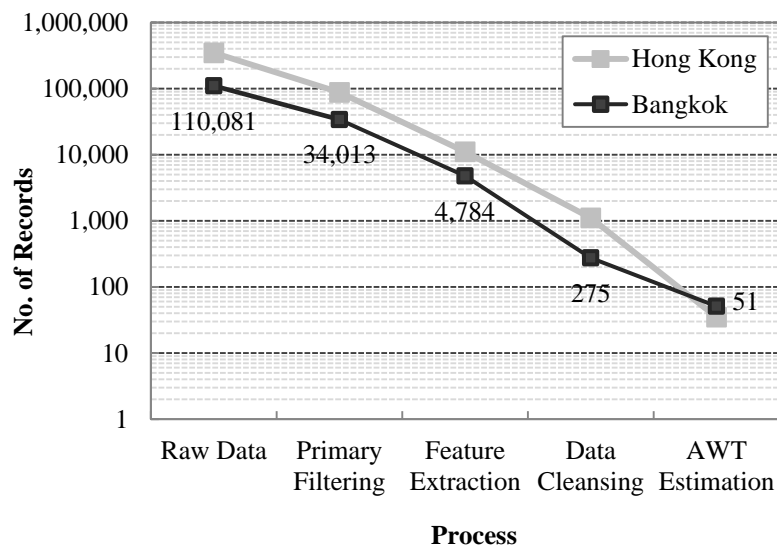


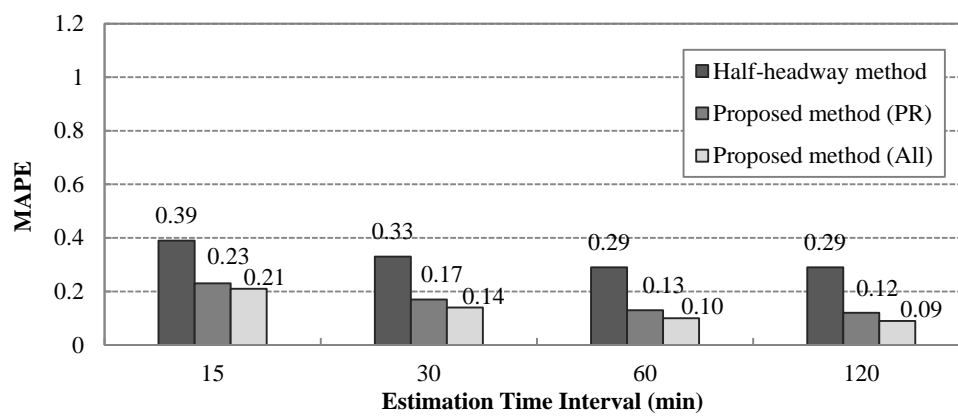
Figure 4.8: Data filtering performance (Case study #2)

The proportion of Wi-Fi records filtered by the primary filtering and the feature extraction processes was similar to the Hong Kong case, whereas the data cleansing process removed a larger proportion of non-potential Wi-Fi records. It can be assumed that the Bangkok case has a greater number of unreliable Wi-Fi records. This type of noise data could be captured from the passing vehicles. Since the vehicles were traveling with free-flow speeds during the inter-peak period, the vehicles spent only a few seconds in the Wi-Fi scanner detection range. As a result, one or two detection events could be captured from such vehicles and eventually filtered out by the data cleansing process. This implies that the data filtering process can perform better in the free-flow traffic conditions. The Wi-Fi records of passing vehicles may not be filtered out in the case of congested road traffic since more detection events could be derived from the vehicles. Then the Wi-Fi record cannot be identified as an unreliable record.

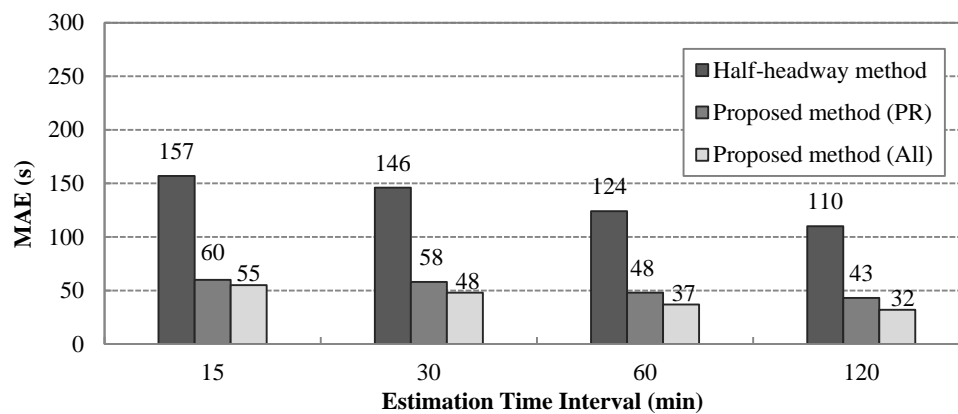
Lastly, fifty-one Wi-Fi records were identified as waiting passengers. The final results were accounted for approximately 50% of the observed passengers.

4.6.2.2 Estimation results and discussion

Figure 4.9 shows the AWT estimation accuracy in terms of (a) MAPE and (b) MAE. The figure is plotted based on the three datasets: bus headway data, probe request (PR) packets in the raw data, and all Wi-Fi data. It can be seen that the direction of accuracy results is similar to the Hong Kong case.



(a)



(b)

Figure 4.9: AWT estimation accuracy for Case study #2 in terms of (a) MAPE, (b) MAE

However, the results demonstrate a distinguishing point compared with the Hong Kong case. Firstly, the improvement of estimation accuracy using all Wi-Fi data is lower than the Hong

Kong case. The accuracy of the all Wi-Fi dataset is increased by 2-3% from the PR dataset for the Bangkok case, and 4-10% for the Hong Kong case. The main reason could be that there was no available Wi-Fi service at the bus stop in Bangkok. Although the bus stop is nearby a college, it was found that the college's Wi-Fi services may not cover the bus stop area which is located on the main road in front of the campus.

There was no data transmission between any Wi-Fi device and the college's access points found in the raw Wi-Fi data. Most detection events of a Wi-Fi device were PRQ packets. Only a few packets in the raw data identified the data transmission between Wi-Fi devices and personal hotspots. Therefore, without the availability of Wi-Fi services at the bus stop, the number of detection events of a MAC-ID in the all Wi-Fi dataset is not significantly different from the number in the PR dataset. This results in the near similarity of the typical detection period between the dataset with only PR packets and the dataset with all Wi-Fi data. It was found that the typical detection period of a MAC-ID in the PR dataset is 82 seconds with 54 seconds of standard deviation, while the period is slightly decreased to 80 seconds in the all Wi-Fi dataset with 55 seconds of standard deviation. As a result, the improvement of estimation accuracy using the all Wi-Fi dataset is insignificant compared with the Hong Kong case.

4.6.3 Limitations and suggestions for further development

The limitations of the proposed AWT estimation method based on passive Wi-Fi are discussed in this section, together with some suggestions for further development.

4.6.3.1 Data preparation

Four pre-defined thresholds are proposed in this study for facilitating the data preparation process. It should be noted that the values of the pre-defined thresholds are basically derived from empirical studies. First, a ten-minute interval is used for activity segmentation based on the experimental results. It was assumed in this study that most of the Wi-Fi devices in the detection area can be detected within the ten-minute interval. The assumption could result in overestimation of the passenger waiting time if the passenger re-visits the bus stop area within the pre-defined period since only the last visit should be considered for estimating the waiting time of passenger. In the future studies, this pre-defined threshold should be adjusted dynamically based on the real-time data collected in the recent previous time intervals at each individual bus stop.

Next, the minimum and maximum duration thresholds (i.e. the time windows) are used for data cleansing. These thresholds are derived from the manual observation of bus dwell times

and passenger waiting times at a bus stop. Also, these thresholds can be adjusted based on available bus dwell times and the passenger waiting times estimated in the previous time intervals.

The last pre-defined threshold is to ensure the reliability of information from a Wi-Fi record. In this study, only the Wi-Fi records which have more than two detection events are considered. This threshold can be examined further by conducting sensitivity analysis with consideration of prior information collected in the previous recent time intervals. The threshold value which provides the best AWT accuracy should be selected.

4.6.3.2 Training a classifier

The proposed system requires sufficient classification profiles from various Wi-Fi devices for reducing the misclassification rate due to the variation in RSSI of different Wi-Fi devices. The classification profile in this study is constructed based on the RSSI observations. Hence, the classification accuracy could be affected by the RSSI variation especially for the early stage of the system when the classifier database consists of only a few profile data. The misclassification can occur if the RSSI profiles in the classifier database are insufficient for distinguishing the Wi-Fi records of waiting passengers from other data sources. This implies that the classifier requires a learning period before the actual implementation to improve the classification accuracy. Then, the system can recognize more RSSI profiles of waiting passengers based on the repeated MAC-IDs in multi-day observations, assuming that some of the bus passengers will arrive at the bus stop in a certain time period for several days. However, the minor misclassification rate may not affect the AWT estimation accuracy due to the fact that very few valid data is needed for the AWT estimation in each time interval. The results in this chapter show that the accuracy of AWT estimation at a bus stop is satisfactory, although some MAC-IDs were incorrectly classified.

In addition, the effects of MAC randomization have not been addressed in the algorithm development. At the current stage, MAC randomization is available for the Wi-Fi devices which are operated by some iOS versions. Although the proportion of such Wi-Fi devices is insignificant, this could be a challenging problem for AWT estimation in the future if most Wi-Fi devices are applicable to MAC randomization. Hence, a thorough investigation of the Wi-Fi devices with MAC randomization needs to be conducted in order to address the device characteristics i.e. the time duration that a device will use the same MAC-ID before the randomization, the proportion of Wi-Fi devices which could be affected by MAC randomization, etc. For AWT estimation, overestimation of a passenger waiting time can occur if a MAC-ID is changed during the waiting period at bus stop. In such a case, the proposed AWT estimation method requires further improvements based on the new factors and/or prior information.

4.6.3.3 Lucky passengers

The most challenging issue of AWT estimation using passive Wi-Fi data is to estimate the actual waiting time of those lucky passengers who arrived the bus stop simultaneously with the arrival bus. Since this group of passengers spends a very short time period at a bus stop, the system has a lower chance to detect a valid Wi-Fi record from these passengers' devices. This could result in overestimation of AWT. Installing multiple Wi-Fi scanners at a bus stop area could increase the detection probability in both spatial and temporal dimensions. However, such improvement requires further investigation in future.

The proportion of lucky passengers at a bus stop could be varied over spaces and times particularly for the multi-route bus stops in densely populated urban areas in Hong Kong. In addition, the availability of real-time bus arrival time information could result in the larger proportion of lucky passengers since passengers may schedule their arrival at the bus stop. In the case that the proportion of lucky passengers becomes significant, there is a need to incorporate more factors and/or information in the proposed AWT estimation method. As a result, the system may require additional information from other data sources in order to model the effects of lucky passengers on AWT estimation. For instance, arrival time and waiting time information could be voluntarily provided by bus passengers.

4.6.3.4 Bus stop configurations and environments

The proposed AWT estimation method in this study was evaluated using the Wi-Fi data collected from two bus stop environments. However, the two bus stops have similar configurations in which bus passengers can randomly arrive and stand around the bus stop area without queuing. In Hong Kong, there are several types of bus stop configuration. For example, passengers may have to line up for boarding a bus at a stop with specific route number. In such cases, passengers may be queued up within the specified bus stop area. Furthermore, multiple bus stop signs can be used in a larger bus stop area in order to group the queuing lines of passengers waiting for the bus route numbers which operate on overlapping or common routes.

To implement the proposed AWT estimation method for Hong Kong bus transit systems, further research should be carried out to collect more field survey data for validation purposes. The accuracy of the proposed AWT estimation system should be justified based on more case studies on various types of bus stop in Hong Kong. In addition to the bus stop configurations, bus stop environments can be varied resulting in different magnitudes of noise in the Wi-Fi data collected from individual bus stops. In the future studies, AWT estimation methods can be developed based on multiple Wi-Fi scanners since a single Wi-Fi

scanner may be insufficient for detecting the Wi-Fi data from the waiting passengers at a bus stop in a densely-populated urban area.

4.7 Summary of findings

In this chapter, a new method is proposed for AWT estimation with use of Wi-Fi data collected at a single bus stop. Methodologies are developed in order to handle massive noise and spatial-temporal uncertainties in the passive Wi-Fi data captured from bus stop environments. The proposed system was evaluated using Wi-Fi datasets collected from two different bus stop environments. The results show that passive Wi-Fi data collected from bus stop environments are noisy. For the Hong Kong case, the number of waiting passengers' MAC-IDs identified by the proposed system is only 0.01% of the total detection events. It is found that the overall accuracy of AWT estimation is from 75-90% when only PRQs are considered for the estimation. The accuracy is improved to 80-95% when all types of Wi-Fi data are considered. The results imply that the availability of Wi-Fi services at bus stops can improve the accuracy of AWT estimation.

For the proposed AWT estimation system, the estimation error is affected by two main reasons. First, the frequency of capturing a detection event results in the estimation error in a passenger waiting time since the waiting time is assumed from the presence time of a Wi-Fi record. Second, the accuracy of AWT could be reduced if the distribution of passenger waiting times derived from the system is significantly different from the distribution of actual passenger waiting times. For instance, the system may not be able to capture any Wi-Fi signals from lucky passengers resulting in an overestimation of AWT at bus stops. It is noteworthy that the proposed system relies on the probability of detecting passive Wi-Fi data from waiting passengers even though the performance of AWT estimation is significantly improved from the conventional method. The penetration rate of deriving Wi-Fi data from waiting passengers should be analyzed before implementing the proposed system at a bus stop.

Future research can be focused in two directions. First, passenger classification algorithms can be improved when MAC-IDs of waiting passengers are available for validation purposes. Second, AWT estimation for individual bus routes can be developed so as to enhance the contribution to improvements in public transport services as well as transportation modeling. In order to identify the bus route from a MAC-ID, the MAC-ID is required to be re-identified by the Wi-Fi scanners, which are installed at multiple bus stops on the route.

Part III

**Participatory Sensing for
Public Transport Information Systems**

Chapter 5

A real-time bus arrival time information system based on crowd-sourced smartphone data

In Chapters 5 and 6 of this thesis, two bus information systems are proposed for providing important information to bus passengers. The two proposed systems are developed based on participatory sensing whereby bus data are collected from and contributed by the bus passengers. Firstly, this chapter proposes a novel framework for developing a real-time bus arrival time information system, using the crowd-sourced bus information contributed by bus passengers. On the one hand, passengers can access the real-time information via their smartphones. On the other hand, they can provide some bus data in return. In connection to the study objectives (3) and (4) in Chapter 1, particular characteristics of the crowd-sourced bus data are firstly introduced. Then, a number of data processing steps are presented and adopted in the proposed framework to handle the difficulties of real-time bus arrival time prediction. The real-time bus arrival time information system proposed in this chapter is evaluated using both simulated bus datasets and real-world datasets. The practicality of the proposed system is investigated in terms of prediction accuracy based on the different participation percentages of the bus passengers.

5.1 Introduction

As discussed in Chapter 2, real-time bus arrival time information systems have been broadly implemented in many modernized cities. The availability of timely information can improve passenger satisfaction since it enables them to better plan their journeys. Passengers can time

their arrival at a bus stop resulting in a reduction in their waiting times, especially when the bus stop is served by common bus routes and passengers have to queue up in a particular line to board a specific bus. Also, bus operators can use the real-time information to enhance their operational management.

In general, the real-time systems rely on a number of bus tracking devices. A device needs to be installed on each bus in order for the bus locations to be tracked periodically in real-time. Therefore, the implementation of the real-time systems could be limited by various reasons such as budget constraints preventing the installation of tracking devices on all buses. Moreover, the system requires the cooperation of bus operators in the provision of the data. In Hong Kong, bus operation is a competitive system and the government has not provided any direct financial subsidies to any of the bus service operators. Hence, the operators may not be willing to share any information that could affect their revenues. In addition, even after bus tracking systems have already implemented, bus arrival time information may not be properly disseminated to bus passengers. Without the installation of bus tracking devices, providing real-time bus arrival time information is challenging.

Zimmerman et al. (2011) proposed a bus information system that enables passengers to contribute to the transit service development by sharing some bus information using their smartphones. In the proposed system, passengers can manually record the in-vehicle travel time of their journeys. The travel times can then be used to predict the arrival times of the following buses. The two-way data provision system could be a viable solution since bus data can be gathered from a substantial amount of bus passengers.

The particular characteristics of crowd-sourced bus data have not been addressed in previous studies. First, unlike bus data from AVL systems, the bus data lack bus identification numbers. Second, inconsistencies in bus location data can also be encountered. Third, the frequency of acquiring bus location data is uncertain in both spatial and temporal dimensions. These limitations should be considered when developing a bus arrival time prediction system.

As an alternative method for deriving bus arrival time information, this chapter proposes a novel framework for developing a real-time bus arrival time prediction system based on crowd-sourced bus data. In this way, bus passengers can contribute the bus data instead of relying on bus tracking devices. The proposed system aims to overcome the challenges of real-time bus arrival time prediction caused by crowd-sourced bus data.

The rest of this chapter is organized as follows. Section 5.2 presents the system architecture and operational overviews. The bus database structure is described in Section 5.3. The details of the data processing steps, including bus location filtering, link travel time estimation, and bus arrival time prediction, are elaborated in Sections 5.4, 5.5, and 5.6, respectively. The

practicality of the proposed system is then evaluated in Section 5.7. Finally, Section 5.8 provides a summary of findings.

5.2 System architecture

The system architecture and operational overviews are demonstrated in Figure 5.1. The system comprises of two major parts: a smartphone application and a back-end processing server.

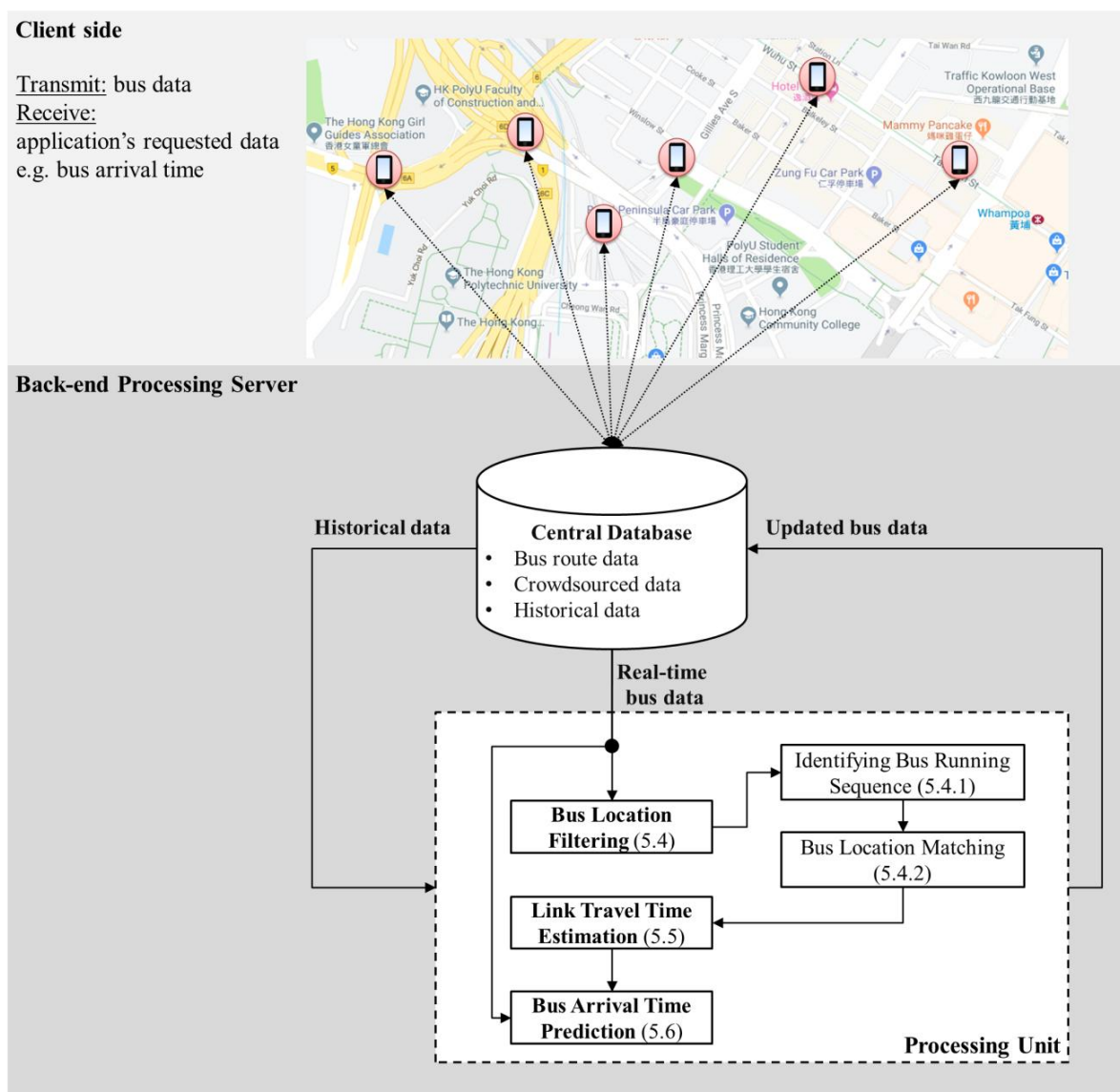


Figure 5.1: System architecture and operational overviews

5.2.1 Smartphone application

On the client-side, the crowd-sourced smartphones of bus passengers are considered as the data transmission tool. For this purpose, a smartphone application should be developed to fulfill two primary requirements. First, to gather bus data, a participating passenger will be requested to provide some information such as the serving bus line number and his or her destination bus stop. Bus location data with the instantaneous bus speed and the timestamp can be periodically identified by GPS after the user has pressed a start button to grant permission for data collection. The GPS operation should be stopped when the bus arrives at the passenger's specified destination. The second requirement of the smartphone application is to disseminate predicted bus arrival times to bus passengers. The users can request bus arrival time information by selecting a bus stop or a bus line number.

5.2.2 Back-end processing

On the server-side, a central database is responsible for recording the relevant data required for bus arrival time prediction such as the bus data provided by participating passengers. The data will be processed in accordance with a number of data processing steps which are considered as the core components of the framework. Practical algorithms can be applied to each core component. This study adopts several methods from related literature and adjusts some solution algorithms to tackle the challenges in crowd-sourced bus data. The data processing results will also be recorded in the database. Finally, the bus arrival time information in the server can be disseminated to the clients upon their smartphone application requests.

5.3 Bus data formulation

This section provides an introduction and definition of the related bus data in the system, including bus route data, crowd-sourced bus data, and historical bus data.

5.3.1 Bus route data

First of all, a road network can be represented by a link-node map, wherein the nodes represent as intersections or bus stops and the links represent the roadways in between. Herein, a road network consists of a set of nodes denoted by $ND = \{nd_1, \dots, nd_N\}$, where nd_i is the location of the node identification number (node ID) i indicated by its two geographical

coordinates in latitude and longitude, $nd_i = \{nd_{i,x}, nd_{i,y}\}$. Then a link between two node IDs a and b is denoted by $l(a, b)$.

Moreover, a set of relative positions on a link can be identified in order to facilitate the bus location matching process, which aims to match GPS observations with potential locations on the road network. A set of relative positions on a link $l(a, b)$ is denoted by $C_{a,b} = \{c_{a,b,1}, \dots, c_{a,b,N}\}$, where $c_{a,b,i}$ is the location of the i^{th} relative location on the link $l(a, b)$. The relative locations could be measured approximately at equal distances along the link, such as every 10 meters.

Next, individual bus routes can be defined by a sequence of traversed nodes along the route. The service route of a bus line number bn operating between an origin station (node x) and its terminus (node y) is denoted by $RN_{bn,A,B} = \{rn_{bn,A,B,1}, \dots, rn_{bn,A,B,N}\}$ where $rn_{bn,A,B,i}$ describes by the node ID of the node sequence i^{th} on the bus route, nid , and a binary variable indicating whether the node is a bus stop of the bus line, st ($rn_{bn,A,B,i} = \{nid, st\}$).

5.3.2 Crowd-sourced bus data

The time-ordered bus data of a bus line reported by participating passengers is represented by a collection of bus data $P_{bn} = \{p_{bn,1}, p_{bn,2}, \dots\}$. Each reported bus dataset $p_{bn,i}$ is denoted by $p_{bn,i} = \{rc_x, rc_y, v, t, ds\}$, where rc_x and rc_y indicate the GPS bus location in latitude and longitude, v is the instantaneous speed (km/hr), t is the timestamp, and ds is the node ID of the passenger's destination bus stop.

5.3.3 Historical information

A set of historical bus information is also required as the necessary input for the real-time bus arrival time prediction system. At the initial stage of the system, a number of GPS bus traces can be collected by a group of volunteers in order to derive the historical bus information before implementing the real-time system. During this stage, the volunteers can be asked to record bus datasets while they are riding on buses. The bus datasets can be recorded using a smartphone application that can continuously record a GPS bus location every few seconds (e.g. once every 5-second period). Since the fine-grained GPS bus traces can describe the bus trajectories in detail, historical bus information on the entire road network can be extracted.

Here, five types of historical bus information are recorded separately for each traversed link $l(a, b)$ and for each time interval τ .

- **Bus delay zone**

The bus delay zone is a road section where buses tend to travel at low speeds due to the presence of an intersection or bus stop. The delay zone, $\bar{q}_{a,b}^\tau$, can be recognized by a relative location $c_{a,b,i}$ which is the location where most buses start to travel at low speeds.

- **Bus delay time**

The average bus delay time at the intersection or the bus stop on the traversed link, $\bar{dt}_{a,b}^\tau$ can be estimated. The delay time of a bus can be estimated from the time when the bus started to travel at low speeds until it passed the intersection or the bus stop.

- **Bus travel time and speed**

The average bus travel time on an individual traversed link, $\bar{tt}_{a,b}^\tau$, can be estimated together with the average link bus speed, $\bar{v}_{a,b}^\tau$.

- **Traffic pattern**

To facilitate the bus arrival time prediction process, traffic patterns on the road network are recorded in the database. The traffic pattern on an individual traversed link during a time interval is denoted by $TP_{a,b}^\tau = \{tp_{a,b,1}^\tau, \dots, tp_{a,b,N}^\tau\}$. A feature $tp_{a,b,i}^\tau$ identifies the observing date, the average bus speed, and the speed range ($tp_{a,b,i}^\tau = \{date, tt, vr\}$). Herein, the speed range is calculated based on the average link travel time as follows:

$$vr = \left\lceil \frac{\bar{v}_{a,b}^\tau}{L} \right\rceil \quad (5.1)$$

where $\bar{v}_{a,b}^\tau$ is the average bus speed of the traversed link $l(a,b)$, and L is a pre-defined speed range. For example, the speed range is equal to 5 if the average speed is 25 km/hr and the pre-determined speed range is 5 km/hr.

In this study, the speed range is considered as an indicator for representing the traffic pattern of each link instead of using the average bus speed. It can help to reduce the computational time of bus travel time prediction for the entire road network. It is noted that the average bus speed is a continuous variable while the value of speed range is discrete and more finite. Therefore, using the speed range can accelerate the traffic pattern matching in the bus arrival time prediction process.

- **Time headway**

Let X be the node ID which is a service bus stop of a bus route. The average time headway of a bus line at a bus stop, $\bar{ht}_{X,bn,A,B}^\tau$, can be estimated from the difference in consecutive bus arrival times at the bus stop. In cases where the bus data are

insufficient for tracking all operating buses, the time headway from public bus schedules can be used instead.

5.4 Bus location filtering

A major difference between the crowd-sourced bus data and the bus data from in-vehicle tracking devices is the availability of bus identification number in the GPS data. The lack of bus identification in the crowd-sourced bus data may result in the bus data inconsistency when two datasets are reported by the passengers on several buses of the same route. Without the bus identification number, a bus dataset may refer to any operating bus.

Another data inconsistency caused by GPS measurement errors could be encountered when two bus datasets reported by multiple passengers on the same bus may indicate different bus locations at a time. The system has no simple way of identifying the more reliable dataset.

As some inconsistency of data from multiple participants could occur in different spatial and temporal dimensions, methods for handling these possible inconsistencies are proposed in this section based on an assumption that the reliability of reported bus data from multiple participating passengers is unknown. Firstly, a bus running sequence can identify each bus dataset to address the data inconsistency caused by the lack of bus identification number. Then, a bus location matching method is proposed to overcome the data inconsistency caused by GPS errors.

5.4.1 Identifying bus running sequence

Due to the regularity of bus services, there are several buses operating on the same route at a time. Therefore, it is necessary to know from which particular bus a bus dataset \vec{p}_i is reported. Suppose that each operating bus on a bus route can be recognized by its running sequence number r for the day. This process aims to identify the bus running sequence for each reported bus dataset \vec{p}_i .

A method of identifying the bus running sequence is to compare the reported bus location with the locations of the operating buses on the route. Let R be the most updated bus running sequence of a bus route for the day. The minimum between the reported bus location and the location of all operating buses can be calculated:

$$D = \min \sum_{j=k}^R \delta(rc_i, ul_j); 0 \leq k \leq R \quad (5.2)$$

where rc_i is the location of a bus route identified by the GPS coordinates in a reported bus dataset, ul_j denotes the location representing the most updated location of a bus running sequence j of the same route, $\delta(m, n)$ is the distance function measuring the distance between two GPS locations m and n , and k is the minimum bus running sequence which has not completed the service route at the terminus.

A bus running sequence r of the bus data can be identified:

$$r = \begin{cases} J; D < \theta_r \\ R + 1; D \geq \theta_r \vee R = 0 \end{cases} \quad (5.3)$$

where J is the bus running sequence which provides the minimum distance D , and θ_r is a predefined distance threshold. The solution function identifies the bus running sequence J as the source of the reported bus dataset $p_{bn,i}$ if the distance between the reported bus location and the location of the bus running sequence J is the shortest distance compared to other buses. However, the bus dataset may be reported from a new bus running sequence which the bus location has not been reported by any passengers. In such a case, the distance threshold is used for determining whether a new bus running sequence should be identified for the bus dataset.

With the bus running sequence, the time-ordered bus datasets of the bus line number bn , running sequence r can be denoted by $P_{bn,r,A,B} = \{p_{bn,r,A,B,1}, p_{bn,r,A,B,2}, \dots\}$ where $p_{bn,r,A,B,i} = \{rc_x, rc_y, v, t\}$.

5.4.2 Bus location matching

This process aims to handle the bus data inconsistencies caused by GPS measurement errors. To do so, potential candidate locations are firstly determined based on each GPS bus location since the GPS location may not indicate the actual bus location on the route. Then, spatial and temporal factors are incorporated in order to identify the more reliable bus locations in the crowd-sourced bus location data.

5.4.2.1 Candidate location determination

Given a GPS bus location rc_i , a set of candidate locations can be determined based on the range of GPS errors within a radius of ϱ . For each link in the error range, a candidate is defined as a relative location $c_{\alpha,\beta,\gamma}$ whereby the distance between the relative location and the GPS bus location, $\delta(rc_i, c_{\alpha,\beta,\gamma})$ is minimum compared with other relative locations. With the availability of bus route data provided by the passengers, only the links on the specific bus route are considered for determining candidates. It is noteworthy that a GPS location may have multiple candidates if the error range comprises more than one link. Here, the candidate set is denoted by $CL^i = \{cl_1^i, \dots, cl_N^i\}$.

Figure 5.2a illustrates the candidate locations of a GPS location rc_i in the GPS error range (represented by a dotted circle). Since there are three links in the error range, three candidates are identified: cl_1^i , cl_2^i , and cl_3^i . Also, Figure 5.2b shows a case of bus data inconsistency due to GPS errors. Suppose that two bus locations, rc_j and rc_k , are reported from two passengers on a bus at the same time. It can be seen that the two GPS locations may indicate different bus locations from cl_1^j and cl_1^k at the same time.

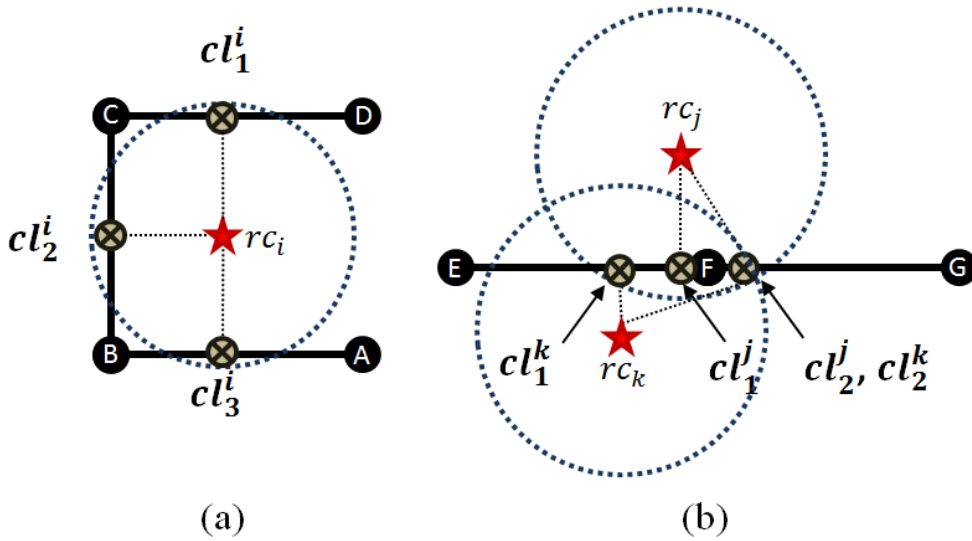


Figure 5.2: Examples of candidate location determination of (a) GPS location rc_i , and (b) GPS locations rc_j and rc_k .

The candidate point cl_j^i is denoted by $cl_j^i = \{\gamma, \alpha, \beta, v, t, \varepsilon\}$ where γ is the identification number of a relative location $c_{\alpha,\beta,\gamma}$ indicating the location on a link between node ID α and β , v is the instantaneous speed from the GPS data, t is the GPS timestamp, and ε is the GPS errors estimated by $\delta(rc_i, c_{\alpha,\beta,\gamma})$.

5.4.2.2 Candidate location formalization

Location formalization aims to arrange the scattered candidate locations on the road segments into a number of reference locations i.e. node locations. Given a candidate location in the middle of a link, the location can be relocated to one of the nodes either in the forward direction or the backward direction.

The relocating direction can be determined based on the candidate location on a link. Suppose that a link is divided into a regular zone and a delay zone with respect to the bus delay zone, $\bar{q}_{a,b}^t$, of the link. A candidate location in the regular zone will be relocated to the node in the backward direction, while the candidate location in the delay zone will be relocated to the node in the forward direction.

Next, the time when the bus is at the node location can be estimated based on the instantaneous bus speed specified in the bus dataset. It is assumed that a bus should travel with free-flow speeds on the regular zone and with congested speeds on the delay zone. Thus, the node arrival time can be estimated in four possible cases:

- (1) If a candidate location is in the regular zone and the instantaneous bus speed is a free-flow speed, the instantaneous bus speed can be used for the time estimation.
- (2) If the instantaneous bus speed indicates a congested speed for a candidate on the regular zone, the average bus speed on the link during the same time interval, $\bar{v}_{a,b}^t$, can be used for the estimation instead.
- (3) If a candidate location is in the delay zone and the instantaneous bus speed is a congested speed, the average bus delay time on the link, $\bar{dt}_{a,b}^t$, can be used for the estimation.
- (4) A bus could travel at a non-congested speed on a delay zone and pass an intersection/a bus stop without any deceleration. In such a case, the average bus speed on the link can be used for the estimation.

Table 5.1 summarizes the four bus location formalization cases with the bus speed parameters used for estimating the time when the bus is at the formalized location.

As a result of candidate location formalization, the time when the bus is at a node could be varied due to the relocation of several candidates to the same node location. The next step is to identify a number of reliable bus data from the existing candidates.

Table 5.1: Candidate location formalization cases

Case No.	Candidate location	Instantaneous speed	Relocation direction	Bus speed parameter for time estimation
1	Regular section	Free-flow	Backward	Instantaneous bus speed
2	Regular section	Congested	Backward	Average link bus speed
3	Delay section	Congested	Forward	Average link delay time
4	Delay section	Free-flow	Forward	Average link bus speed

5.4.2.3 Candidate graph

To identify the reliable bus datasets, spatial and temporal factors can be considered. In this chapter, a bus data identification method is developed based on the concept of the Spatio-Temporal Matching algorithm proposed by Lou et al. (2009). In the previous study, the algorithm was developed for GPS location matching based on a low-sampling-rate trajectory derived from an in-vehicle tracking device. Thus, considerable modifications to the method are necessary for handling the data inconsistency problem that occurs in crowd-sourced bus data.

To facilitate further data processing steps, the results of candidate location formalization during a processing time interval are represented by a direct graph. Each vertex denotes an estimated time when the bus is at a node location while each edge represents the bus travel time between a pair of nodes. The vertices are initially grouped by the node ID (e.g. a, b, \dots), and the groups of vertices are arranged in order by the node ID on the bus route $RN_{bn,A,B}$. Then, the sequence of the arranged group is denoted by a sequence number sn .

Each vertex in a group sequence is denoted by $NC^{sn} = \{nc_1^{sn}, \dots, nc_N^{sn}\}$ where N is the total vertices in the group. The vertex nc_w^{sn} is described by $nc_w^{sn} = \{\ell, t', \rho, \varepsilon\}$ where ℓ is the node ID, t' is the estimated time at the node, ρ is the identification number i of the reported dataset which is the source of the candidate cl_j^i , and ε is the GPS error inherited from the error of candidate location $cl_j^i \cdot \varepsilon$ before performing location formalization. Finally, each edge represents the transmission between a pair of vertices in two consecutive group sequences. Figure 5.3 shows a graphical presentation of a candidate graph.

In the next steps, each vertex and each edge on the graph will be associated with observation probability and transmission probability so as to determine the most reliable candidate for individual node locations.

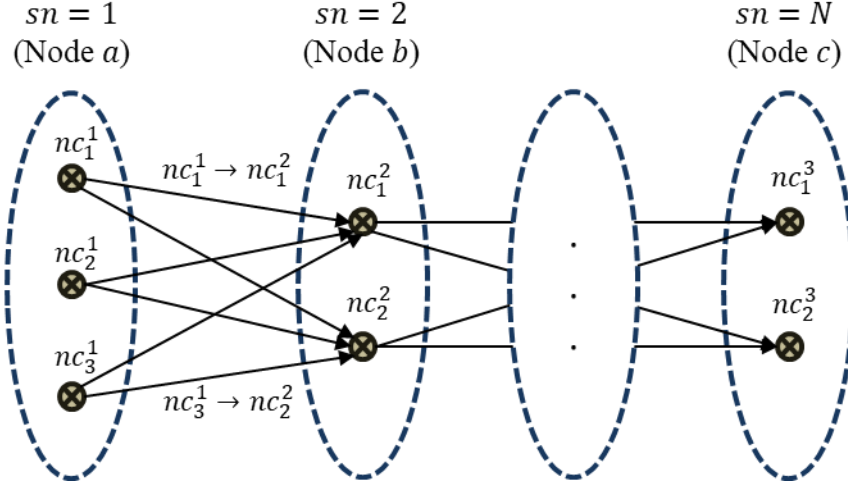


Figure 5.3: Example of a candidate graph

5.4.2.4 Observation probability

Observation probability aims to evaluate the spatial reliability of each vertex. The probability can be calculated based on the degree of GPS errors $\overline{nc}_w^{sn} \cdot \varepsilon$, and the statistical distribution of GPS errors indicated by a mean μ and a standard deviation σ :

$$N(nc_w^{sn} \cdot \varepsilon) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(nc_w^{sn} \cdot \varepsilon - \mu)^2}{2\sigma^2}} \quad (5.4)$$

5.4.2.5 Transmission probability

Each edge of the candidate graph represents the travel time between two node locations calculated from a pair of candidates. The objective of transmission analysis is to evaluate the temporal reliability of the travel time. Given the m^{th} vertex of the node sequence $i - 1$ and the n^{th} vertex of the node sequence i , the travel time between node sequences can be calculated:

$$ptt_{i-1,i} = nc_n^i \cdot t' - nc_m^{i-1} \cdot t' \quad (5.5)$$

Let $\overline{ptt}_{i-1,i}^\tau$ be the average travel time from the node sequence $i - 1$ to i during the time interval τ , and the average path travel time can then be calculated from the summation of average link travel times, $\overline{tt}_{a,b}^\tau$, along the path during the same time interval τ .

Then, the transmission probability can be defined by the likelihood between the candidate travel time, and the historical travel time:

$$F_t(\mathbf{nc}_m^{i-1} \rightarrow \mathbf{nc}_n^i) = 1 - \frac{|\mathbf{ptt}_{i-1,i} - \overline{\mathbf{ptt}}_{i-1,i}^t|}{\overline{\mathbf{ptt}}_{i-1,i}^t} \quad (5.6)$$

5.4.2.6 Result matching

The final transmission function between a pair of vertices can be denoted by:

$$\mathbf{F}(\mathbf{nc}_m^{i-1} \rightarrow \mathbf{nc}_n^i) = N(\mathbf{nc}_n^i, \varepsilon) \times F_t(\mathbf{nc}_m^{i-1} \rightarrow \mathbf{nc}_n^i) \quad (5.7)$$

The transmission function is used for evaluating the transmission probability of each edge in the candidate graph.

Suppose that the transmission function is the reliability score of a pair of vertices. To identify a set of reliable datasets from the candidate graph, the reliability score has to be calculated for a chain of vertices from the first node sequence to the last node sequence. A chain of vertices is denoted by $CP_C: \mathbf{nc}_{e_1}^1 \rightarrow \mathbf{nc}_{e_2}^2 \rightarrow \dots \rightarrow \mathbf{nc}_{e_N}^N$ where $\mathbf{nc}_{e_i}^{sn_i} \cdot \rho \neq \mathbf{nc}_{e_j}^{sn_j} \cdot \rho; \forall i = 1, 2, \dots, N - 1, \forall j = i + 1, \dots, N$. Each vertex $\mathbf{nc}_{e_i}^i$ on the chain is the vertex number e_i of a node sequence i , and N is the total node sequences in the graph.

Suppose that the transmission function is the reliability score of a pair of vertices. The reliability score of a chain of vertices, CP_C , from the first group sequence to the last group sequence can be calculated:

$$\mathbf{F}(CP_C) = \sum_{i=2}^N \mathbf{F}(\mathbf{nc}_{e_{i-1}}^{i-1} \rightarrow \mathbf{nc}_{e_i}^i) \quad (5.8)$$

Here the reliability score of a chain $\mathbf{F}(CP_C)$ can be used to represent the likelihood that a bus is at the node locations at the times t' indicated by the vertices in the chain.

Finally, the chain with the highest score is considered as the best solution for identifying the most reliable bus datasets:

$$CP = \operatorname{argmax}_{CP_C} \mathbf{F}(CP_C) \quad (5.9)$$

Given a solution $CP: nc_{e_1}^1 \rightarrow nc_{e_2}^2 \rightarrow \dots \rightarrow nc_{e_N}^N$, path travel time between consecutive node sequences can be obtained. Let a and b be the node IDs of a pair of node sequences $i - 1$ and i , respectively. The path travel time between the nodes during the time interval τ can be calculated:

$$ptt_{a,b}^\tau = nc_{e_i}^i \cdot t' - nc_{e_{i-1}}^{i-1} \cdot t' \quad (5.10)$$

where the time interval τ is identified by $nc_{e_i}^i \cdot t'$.

It can be noted that the last vertex $nc_{e_N}^N$ is considered as the most updated dataset of the bus. The bus information is recorded as a data point $uc_{bn,A,B,r} = \{cl, t', st\}$ where cl is the node ID of the node sequence N , t' is the time at the node location, and st is a binary variable indicating the bus operational status, either active or terminated. For instance, the status can indicate the service termination when the bus arrives at the bus terminus.

5.5 Link travel time estimation

Since a path of travel between two node sequences may consist of multiple links, link travel time estimation aims to decompose the path travel time into the travel time of individual links on the path. The link travel times can be estimated using historical travel time information.

Given a travel path from the node ID a to b , the link travel time between node ID c to the next node d on the travel path can be estimated by:

$$ltt_{c,d}^\tau = \frac{\bar{tt}_{c,d}^\tau}{\bar{ptt}_{a,b}^\tau} \times ptt_{a,b}^\tau \quad (5.11)$$

where $\bar{tt}_{c,d}^\tau$ is the historical link travel time between node m and n during time interval τ , and $\bar{ptt}_{a,b}^\tau$ is the historical path travel time during time interval τ calculated from the summation of average link travel times along the path during the same time interval τ .

For each time interval, reported bus datasets from several operating buses on the same road section could provide a set of travel times on an individual link, $LTT_{c,d}^\tau = \{ltt_{c,d,1}^\tau, \dots, ltt_{c,d,n}^\tau\}$. Since the link travel times could be varied during a time interval, the average of link travel times, $tt_{c,d}^\tau$, estimated in the current time interval can be derived by applying the stratified sampling technique to the dataset. In addition, the traffic pattern,

$tp_{c,d,i}^\tau$, is derived based on the average link travel time. It is noteworthy that the travel time information of some links may be unavailable in an individual time interval, since there may be no bus service on the links during the time.

5.6 Bus arrival time prediction

Given the updated bus information $uc_{bn,A,B,r}$, the bus arrival time at the remaining stops on the route can be predicted. Three types of information are essential for predicting the bus arrival time at a bus stop: (a) the time at the updated bus location, (b) the predicted travel time between the updated location and the bus stop, and (c) bus delay time at the bus stop. Suppose that the updated bus location is on the node ID u , and the bus stop is on the link $l(w, z)$. The relationship can be denoted by:

$$ar_{bn,r,z}^{\tau+1} = uc_{bn,A,B,r} \cdot t' + ptt_{u,z}^{\tau+1} - \overline{dt}_{w,z}^{\tau+1} \quad (5.12)$$

where $ar_{bn,r,z}^{\tau+1}$ is the predicted bus arrival time for the next time interval $\tau + 1$ of the bus line bn at the bus stop represented by the node ID z .

The updated bus location with the time at the location is derived from the bus location filtering process, while the other two parameters require the prediction. The bus dwell time at the bus stop will be subtracted from the predicted travel time since the delay time on a link is already included as a part of the link travel time.

In this study, the crowd-sourced bus data may not be sufficient for estimating bus dwell times at bus stops. Thus, historical bus delay times at the bus stop $\overline{dt}_{w,z}^{\tau+1}$ can be used to represent the predicted bus dwell time. In such a case, the use of historical dwell times is based on an assumption that the variation in bus dwell times of the operating bus lines at a bus stop is insignificant during a time interval. To relax the assumption, historical bus delay times can be recorded individually for each bus line. To this end, a vital process is performing the path travel time prediction.

5.6.1 Travel time prediction

In general, a path travel time can be predicted by calculating the summation of predicted travel times of the links along the path. Since the crowd-sourced bus data may not always

provide a series of link travel times in every time step, the link travel time prediction methods which require a series of information cannot be applied for the prediction.

This study adopted the traffic pattern matching algorithm proposed by Vanitchakornpong et al. (2013) in order to perform the link travel time prediction. The objective function was developed to predict a link travel time by searching for the historical traffic pattern which is most similar to the current one. Without the availability of link travel time information in the current time step, the algorithm can predict the link travel time by considering (a) the spatial correlation between the traffic pattern of the link and its adjacent links as well as (b) the temporal correlation between the traffic patterns of the link in several time steps.

Let $\vartheta_{l,\tau}^\omega$ be the current traffic pattern of a link $l = L(a, b)$ on the current date ω and time interval τ , represented by the current link speed range $tp_{c,d,i}^\tau \cdot vr$. The difference between a historical traffic pattern on a date φ during the same time interval and the current traffic pattern can be calculated in terms of a mismatch value:

$$m_{l,\tau}^\varphi = |\vartheta_{l,\tau}^\varphi - \vartheta_{l,\tau}^\omega| \quad (5.13)$$

where $\vartheta_{l,\tau}^\varphi$ is the historical traffic pattern represented by the link speed range $tp_{c,d,j}^\tau \cdot vr$.

To predict the link travel times for the time interval $\tau + 1$, the traffic patterns of the adjacent links, as well as the patterns of the link in previous time intervals, are also considered. The average mismatch value can be calculated:

$$M_{l,\tau}^\varphi = \frac{1}{N} \sum_{p \in P} \sum_{l' \in A(a,b)} |\vartheta_{l',\tau-p}^\varphi - \vartheta_{l',\tau-p}^\omega| \quad \forall \varphi \in \Phi \quad (5.14)$$

where the searching space is determined by the set of considered dates, Φ , the previous time intervals P , and the set of adjacent links $A(a, b)$. In this study, the two parameters were set as $\Phi = \{\omega - 1, \dots, \omega - 14\}$, and $P = \{0, 1, 2\}$.

The historical traffic pattern $\vartheta_{l',\tau-p}^\varphi$ is assumed for the prediction when the average mismatch value $M_{l,\tau}^\varphi$ is the minimum value in the searching space. Then, link travel times are predicted using historical travel times. It is noteworthy that the link travel times derived from the crowd-sourced data during a time interval could be insufficient for predicting the link travel times for the entire road network. In this case, the average link travel time $\bar{tt}_{a,b}^{\tau+1}$ will be assumed to represent the travel time for the next time interval.

5.7 Experimental studies

The proposed system was evaluated using both simulated bus datasets and the real-world bus datasets. In order to derive the crowd-sourced bus datasets, two types of information are required: GPS bus trajectories from the operating buses, the number of boarding and alighting passengers at each bus stop.

5.7.1 Case study #1: simulated bus services

5.7.1.1 Simulated bus data

(1) Bus trajectory data

The microscopic simulation software called VISSIM was used for simulating a road network with virtual environments. Firstly, private vehicles were included in the road network. The traffic flows from private vehicles were simulated by determining the vehicle desired speeds and the traffic volume. The parameters were adjusted to include the variation in traffic conditions over periods of time. Traffic data from the Transport Department of Hong Kong were used as the reference for the parameter calibration. Moreover, the vehicle movement was based on a car-following model and a lane-changing model provided by VISSIM.

Next, bus services were integrated into the road network including bus stops, bus routes, bus frequencies, and dwell times at bus stops. The distribution of bus frequencies applied the bus information provided by Kowloon Motor Bus Company. The dwell time distribution was calibrated using observation data from field surveys. In the simulation, a bus had to stop at every bus stop on the route. At this stage, bus trajectory data could be obtained for every second.

The simulated road network was represented by 74 nodes and 89 links. The average link distance was 132 meters. Twenty bus lines were assigned to operate on the road network. Figure 5.4 shows the simulated road network with examples of two bus routes. The bus routes are partially overlapped on three links.

(2) Crowd-sourced bus data

Two additional modifications of the simulated bus trajectories were required in order to obtain the crowd-sourced bus datasets reported by participating passengers.

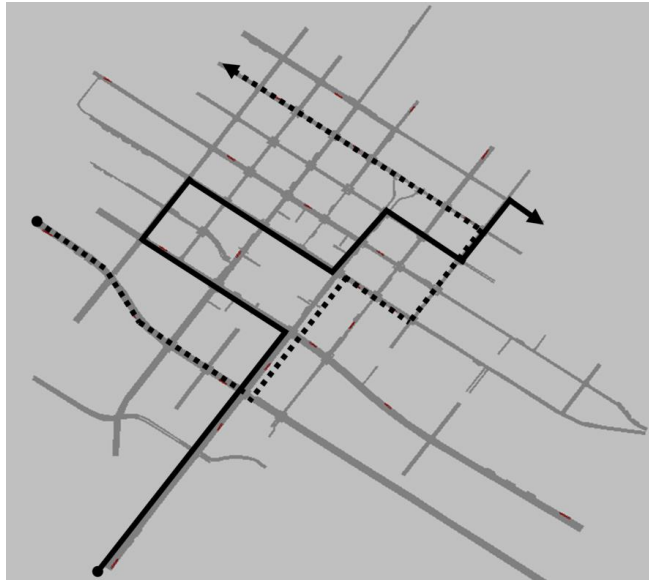


Figure 5.4: Examples of two overlapping bus routes on a simulated road network

- **Passenger participation**

The amount of reported bus datasets is dependent on the level of passenger participation. Since only some of the passengers on a bus could contribute to bus data collection, data sampling methods were applied to obtain the crowd-sourced bus data from a bus trajectory.

The first step was to simulate the total number of boarding and alighting passengers at each bus stop. Field observations were conducted on several bus routes in Hong Kong. A group of volunteers were asked to ride on various bus routes during evening peak periods (i.e. 17:00-19:00). For each bus stop, the numbers of boarding and alighting passengers were observed by the volunteers. In addition, the origin and destination bus stops of some random passengers were also recorded. Then, the distribution of passengers' journey lengths was calibrated using the observed data. As a result, a destination bus stop can be determined for each boarding passenger.

Next, the pattern of bus data derived from participating passengers was determined by indicating the passenger participation percentage, as well as the bus data sampling method. First, four crowd-sourced bus datasets were generated from 1%, 3%, 5%, and 10% of the total passengers. Second, four sampling methods were applied to each dataset. The sampling methods consisted of both continuous sampling using different sampling frequencies, and one-time sampling after the passengers boarded a bus. The sampling methods are summarized in Table 5.2. To sum up, sixteen crowd-sourced bus datasets were generated from each simulated bus trajectory.

Table 5.2: Bus location sampling methods

No.	Sampling method	Sampling frequency
1	Continuous sampling	Every 1 minute
2	Continuous sampling	Every 2 minutes
3	Continuous sampling	Every 3 minutes
4	One-time sampling	-

In addition to the crowd-sourced bus data, the AVL data were also simulated to compare the performance of the proposed system against the benchmark system in which a bus tracking device is installed on each individual bus. An AVL bus dataset was generated for every 30 seconds.

- **Smartphone GPS errors**

The bus locations provided by VISSIM were the actual locations on the road. Therefore, smartphone GPS errors were included in the location data to simulate the bus locations collected by crowd-sourced smartphones.

For each bus location, the degree of error was determined based on the GPS error distribution. To determine this distribution, a data collection survey was carried out by a group of volunteers. The GPS bus traces of several bus routes in Hong Kong were recorded by their smartphones. In this study, the GPS error is assumed as the perpendicular distance from a GPS location to the closest road section. The distribution of errors can be summarized as follows: 83% of total GPS locations have errors within 0-30 meters, 10% within 30-50 meters, and the remaining 7% over 50 meters.

The statistical distribution was used for modifying the actual bus location, as well as for identifying the GPS error region for identifying candidate locations (Section 5.4.2.1). In this study, the radius of the error range is 50 meters assuming that at least 90% of the reported bus data are retained for bus arrival time prediction. Finally, the mean and the variance of GPS errors can be derived for the observation probability function (Section 5.4.2.4)

For the AVL bus data, the degree of GPS errors was assumed according to the previous studies on GPS errors in AVL bus locations (Jagadeesh et al., 2004; Jeong 2005; Chen et al., 2013). The mean error was assumed to be 15 meters with 7 meters for the standard deviation.

5.7.1.2 Evaluation results

The system performance was evaluated using two performance measures: MAE and MAPE:

$$MAE = \frac{\sum |AR_z - \widehat{AR}_z|}{N} \quad (5.15)$$

$$MAPE(\%) = \frac{1}{N} \sum \frac{|AR_z - \widehat{AR}_z|}{tt_{u,z}} \times 100 \quad (5.16)$$

where AR_z is the observed bus arrival time at a node or a bus stop represented by the node ID z , \widehat{AR}_z is the estimated/predicted bus arrival time at the node or the bus stop, $tt_{u,z}$ is the observed travel time between the bus location at the node ID u and the bus stop location z , and N is the number of the observed bus arrival times.

The simulated bus datasets were independently used as the system inputs. In this case study, the system was developed to update bus arrival time information at every 1-minute interval based on the most available bus location. Bus location matching and link travel time prediction were performed at every 5-minute interval. The simulation was replicated several times in order to generate historical bus data from 14 weekdays.

This section provides a performance evaluation of two core processes. Firstly, the performance of bus location matching was analyzed based on

- (a) The MAE of the time estimation at a node location after performing location matching,
- (b) The average number of links between consecutive bus locations, and
- (c) The availability of link travel time per time interval.

Then the accuracy of bus arrival time prediction was evaluated by

- (d) The MAE of the bus arrival time prediction, and
- (e) The MAPE of the bus arrival time prediction.

Table 5.3 provides the numerical results from the 17 simulated bus datasets.

(1) Bus location matching performance

Bus location matching determines the reliable datasets in the crowd-sourced bus data. The process results in a number of relocated bus locations with the time at the locations.

Table 5.3: Numerical results from the 17 bus datasets (Case study #1)

Passenger participation	Sampling method	Bus location matching			Bus arrival time prediction	
		(a) MAE of the time at a node location (s)	(b) Average no. of links between bus locations	(c) Available link travel time/interval (%)	(d) MAE of the BAT prediction (s)	(e) MAPE of the BAT prediction (%)
AVL	-	10.34	1.14	67.98	23.14	23.61
10%	1	12.35	1.38	58.68	26.58	26.90
	2	13.38	1.64	55.71	26.91	27.13
	3	14.13	2.01	50.89	27.51	27.79
	4	15.99	2.74	41.66	30.69	28.35
5%	1	13.62	1.77	53.84	27.38	27.98
	2	14.76	2.50	46.27	28.55	28.44
	3	15.35	2.97	40.07	30.89	28.60
	4	16.03	3.57	29.23	31.25	29.88
3%	1	14.84	2.19	48.45	28.63	28.66
	2	15.76	3.49	38.29	31.30	29.49
	3	16.23	4.38	28.76	32.40	32.22
	4	N/A	N/A	<8	N/A	N/A
1%	1	15.48	2.44	34.36	31.56	29.31
	2	16.30	3.92	25.23	32.66	31.18
	3	17.34	5.97	17.53	34.21	34.69
	4	N/A	N/A	<1	N/A	N/A

The numerical results show that the MAE of the time estimation at node locations (a) is varied from 12 to 18 seconds. It can be observed that the accuracy is improved when more bus datasets are available due to the greater percentage of participating passengers and the higher data sampling frequency. Also, the greater number of reported bus datasets results in the shorter distance between the relocated bus locations. The average number of links between consecutive bus locations (b) shows a decreasing trend when more bus datasets are available.

The errors in the estimated time could be incorporated during several processes.

- First, during the location formalization process, the reported bus location is relocated to a node location using the historical link speed, the historical delay time, or the instantaneous speed. The considered speed may not result in accurate time estimation, especially when the bus is delayed at an intersection or a bus stop.
- Second, GPS measurement errors could be one of the underlying errors originally included in the crowd-sourced bus locations. It can be seen that the MAE of the time estimation from the AVL bus dataset is approximately 10 seconds, even though the average number of links between consecutive bus locations is close to one link.
- Third, the reliability of transmission probability could be decreased for a pair of bus locations with a longer distance. Since the path travel time is compared to the historical path travel time, the historical travel time could be less reliable for representing the current traffic conditions on a long road section.

In addition to the time accuracy at node locations, one of the results of bus location matching is a set of link travel times during the current time interval. The numerical results show that the percentage of available link travel times on the road network per time interval (c) is increased when more bus datasets from participating passengers are available. This will further affect the bus arrival time prediction performance.

(2) Bus arrival time prediction performance

The accuracy of the crowd-sourced datasets can be compared with the AVL bus dataset. The prediction accuracy of the AVL bus data is better than the crowd-sourced dataset with the most reported bus data (10% passenger participation, 1-minute sampling) for approximately 3% of MAPE, since the AVL bus data have an updated bus location in every 30-second time period. It can be observed that the AVL data provide more link travel times on 9% of the total links for each time interval.

For the crowd-sourced bus datasets, the prediction accuracy of the crowd-sourced bus data is improved when more reported bus datasets are available. The MAE of bus arrival time

prediction (d) is from 27 to 34 seconds, while the MAPE (e) is from 27% to 35%. Figure 5.5 shows the MAPE of the 17 datasets categorized by passenger participation percentages and data sampling methods.

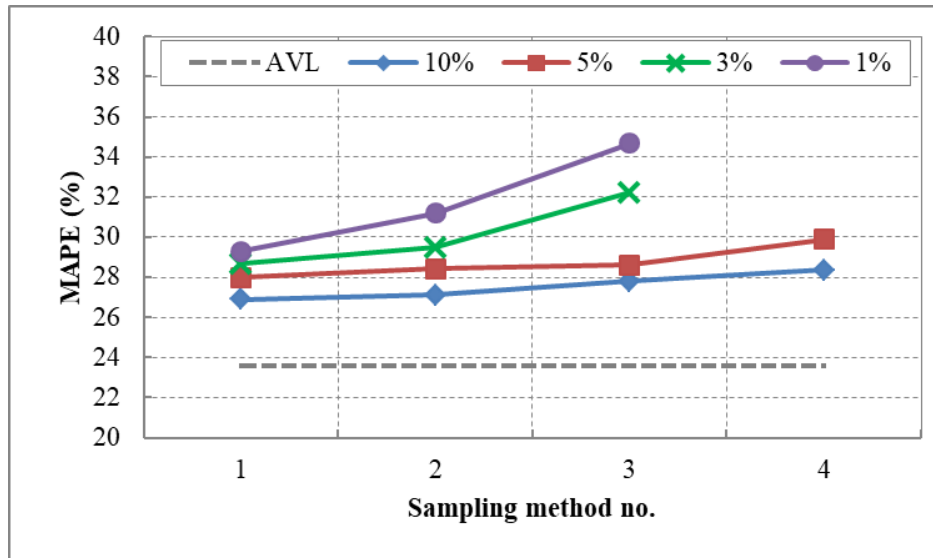


Figure 5.5: The MAPE of bus arrival time prediction

For the datasets from 5% and 10% passenger participation, the MAPEs are slightly increased when the lower sampling frequencies are applied from methods no.1 to no.4. The prediction accuracy is lower than 70% when some sampling methods are applied to the datasets from 1% and 3% passenger participation. In such a case, it can be assumed that the reported bus data are insufficient to provide updated bus data for real-time prediction.

The insufficiency of crowd-sourced bus data affects the prediction accuracy for three reasons.

- The number of links between consecutive bus locations (b) could be extended. The longer distance could result in the lower accuracy of link travel time estimation and prediction since the link travel times during a time interval are taken from the long path travel time based on the historical link travel times.
- The prediction algorithm could rely on historical data which may not represent the current traffic condition. The numerical results show that the datasets provide bus arrival time prediction accuracy of more than 70% when the availability of the link travel time in the current time interval (c) is greater than 30% on average.
- During a time interval, the reported bus data could be inadequate for providing the most updated locations of individual operating buses on the road network. In such a case, bus arrival time prediction is based on the historical time headway resulting in higher prediction errors.

It is noteworthy that the insufficiency of crowd-sourced bus data could be encountered when there are few passengers on each operating bus, such as during off-peak hours or in the high-frequency bus services.

5.7.2 Case study #2: real-world bus data

5.7.2.1 Data description

In this section, the proposed system is evaluated based on the real-world bus data collected from a number of buses in Bangkok. The data collection was conducted on a 14-kilometer road section where bus services were provided as part of eight bus routes. The study road section was represented by 39 nodes and 38 links. The average link distance was 358 meters. In the same way as the Case study #1, two types of information were recorded by volunteers: the GPS bus trace of individual buses, and the number of boarding/alighting passengers at each bus stop. Examples of GPS raw data and passenger boarding/alight observations are provided in Appendix C. The data were collected for five weekdays during 7:00-9:00. Here the datasets were separated so as to derive historical bus data from four-day datasets and to evaluate the proposed system using one-day datasets.

Since GPS errors were already included in the GPS bus locations, only the methods for generating crowd-sourced bus datasets were implemented based on the GPS bus trajectories. For this case study, the same passenger participating percentages and data sampling methods were implemented for generating 17 crowd-sourced bus datasets.

5.7.2.2 Evaluation results

The proposed system was evaluated using the same performance evaluation as the first case study. Table 5.4 provides numerical results from the 17 bus datasets.

For the bus location matching accuracy, the results show that the MAE of the time estimation at node locations (a) was in a range from 17 to 27 seconds based on the availability of crowd-sourced bus data. Also, a pair of bus locations covered approximately 1-2 links during each time interval. It could be assumed that the average link distance in this case study is longer than in the simulated case. A bus trajectory could cover only a few links during each time interval (i.e. a 5-minute interval), especially when considerable bus delay times affected the bus travel time during the morning peak time period.

Table 5.4: Numerical results from the 17 bus datasets (Case study #2)

Passenger participation	Sampling method	Bus location matching			Bus arrival time prediction	
		(a) MAE of the time at a node location (s)	(b) Average no. of links between bus locations	(c) Available link travel time/interval (%)	(d) MAE of the BAT prediction (s)	(e) MAPE of the BAT prediction (%)
AVL	-	14.52	1.04	68.57	65.45	23.12
10%	1	16.79	1.05	62.88	70.17	24.36
	2	17.13	1.24	57.10	76.88	25.71
	3	18.76	1.57	51.59	80.41	26.23
	4	20.25	2.86	32.85	84.72	27.10
5%	1	18.82	1.18	56.21	76.97	25.88
	2	19.70	1.38	49.05	81.56	26.61
	3	24.31	1.76	40.75	86.80	27.89
	4	N/A	N/A	<14	N/A	N/A
3%	1	21.78	1.24	50.43	86.44	27.65
	2	22.26	1.55	38.65	90.67	28.15
	3	26.52	2.10	27.52	95.31	29.06
	4	N/A	N/A	<5	N/A	N/A
1%	1	N/A	N/A	N/A	N/A	N/A
	2	N/A	N/A	N/A	N/A	N/A
	3	N/A	N/A	N/A	N/A	N/A
	4	N/A	N/A	<1	N/A	N/A

For the overall performance, the MAPE of bus arrival time prediction was 24-29% with 70-95 seconds of MAE. It can be observed that some bus datasets were insufficient for performing the bus arrival time prediction. The one-time data sampling method was practical only for the 10% passenger participation scenario. Moreover, the crowd-sourced bus data from 1% passenger participation were insufficient regardless of the data sampling methods. The main reason could be due to the Bangkok bus capacity being less than the bus capacity in the simulation case (the Hong Kong bus system). Single-decker buses serve the passengers in Bangkok, while double-decker buses are generally operated in Hong Kong. As a result, 1% of the passengers in a Bangkok bus could be impractical for the system implementation.

5.7.3 Limitations and suggestions for further development

The limitations of the real-time bus arrival time prediction system based on crowd-sourced bus data are discussed in this section, together with some suggestions for further development.

5.7.3.1 Bus running sequence identification

In the bus running sequence identification process, each reported bus data is assigned by a bus running sequence. Here, the running sequence of the closet bus from the same bus line number is assigned to the reported bus data. This method may cause some errors when multiple buses from the same bus line number are running on a similar road section (due to bus bunching or bus overtaking). Therefore, additional information should be considered in solving the problem. For instance, users' login data can be used since the smartphone application may request the users to log in before using the application. In such a case, the system can promptly identify the bus datasets reported by the same passengers. As a result, the group of participating passengers on an individual bus can be identified by the users' login data.

5.7.3.2 GPS measurement errors

GPS measurement errors in bus location data could affect the bus arrival time prediction accuracy. This type of error is originally included in the bus location data and the magnitude of the error is generally unknown. This study assumes that the bus locations within a pre-defined error range from any location on the road could be the potential candidate locations. In Case study #1, a 50-meter radius of the error range is assumed based on the distribution of GPS bus location errors collected by smartphones. However, GPS errors can be varied for different road sections. To improve the bus arrival time prediction accuracy, the error range

should be further adjusted for each link on the road network when sufficient bus locations are available.

5.7.3.3 Bus dwell time estimation

It is challenging to derive updated bus dwell time information based on crowd-sourced bus data since the frequency of bus location data may not be sufficient for bus dwell time estimation. In this case, the real-time bus arrival time prediction is reliant on historical bus dwell times. To update bus dwell time information, the system may ask some participating bus passengers to report a set of bus data every few seconds.

5.7.3.4 Method validation and system deployment

The practicality of the proposed system should be further investigated before the actual implementation. Sufficient survey data from the majority of bus transit network are required, especially from the congested urban areas. In Hong Kong, the GPS locations measured on some main roads may have larger errors due to the high-rise buildings in urban areas. The GPS signal quality could be inadequate for identifying a precise location using the built-in GPS receiver of smartphones. In such a case, the reported bus data from participating passengers could be unreliable. The road sections with significant GPS errors should be identified to improve the accuracy of bus arrival time prediction. The integration between the use of crowd-sourced bus data and other fixed-location sensors (e.g. Wi-Fi scanners and Radio Frequency Identification readers) can be considered.

In addition, the surveys should be sufficient for determining the expected passenger participation and the bus data sampling method since the bus arrival time prediction accuracy is mainly dependent on the quantity of crowd-sourced bus data. However, the number of participating passengers after the deployment could be different from the expected number. In such cases, the sampling frequency could be flexible based on the actual number of participating passengers after the deployment. For instance, in the early stage of deployment, the smartphone application may collect the bus data using a moderate sampling frequency. Then, the application may reduce the frequency when a more substantial number of participating passengers are involved. Moreover, increasing user acceptance after the deployment is important. The smartphone application should be developed to provide more benefits to the users and to persuade the users to share bus information.

In addition to the system evaluation in this chapter, the effects of different processing time intervals should be investigated in future studies. On the one hand, bus data insufficiency could be overcome by extending the processing time interval since more bus datasets can be

obtained during longer intervals. On the other hand, the extension of the time intervals could reduce the accuracy of bus arrival time prediction since timely updates on bus location data may not be provided.

5.8 Summary of findings

This chapter presents a novel framework for proposing a real-time bus arrival time information system. The proposed system is based on the two-way data provision concept in which a smartphone application is considered as a tool for both disseminating real-time bus information and gathering bus data from participating passengers. Without the need for in-vehicle tracking devices, the proposed system can provide an alternative solution for deriving bus arrival time information.

Unlike AVL bus data, crowd-sourced bus data pose extra challenges for bus arrival time prediction. The characteristics of crowd-sourced bus data are addressed in this chapter i.e. the lack of bus identification data, the bus data inconsistencies, and the uncertain availability of bus location data. A number of data processing steps are adopted as core components of the proposed framework to handle the particular challenges of this system. The practicality of the proposed system is investigated based on both simulated bus data and real-world bus data. The results show that crowd-sourced bus data have the potential to be a viable source of data for developing a bus arrival time information system. The prediction accuracy can be improved when more bus data are contributed by the participating passengers. It can be observed from the results that the prediction accuracy relies on the number of links between two consecutive bus locations and the percentage of available link travel times on the road network. With a larger crowd-sourced bus dataset, the number of links between two consecutive bus locations can be decreased resulting in more accurate link travel times. In addition, the percentage of available link travel times on the road network is increased and this could facilitate bus arrival time prediction.

Apart from bus arrival time information, bus crowding information could also be one of the significant factors which affect passenger boarding decisions. Even though bus arrival time information is disseminated to bus passengers, they may not choose to board the first bus that arrives if it is overcrowded. In the next chapter, the proposed bus arrival time prediction system will be extended in order to develop a bus crowding prediction system based on the crowd-sourced bus data. In this way, the public transport information system can provide more complete information to enable passengers to make the best decision.

Chapter 6

A bus crowding prediction system based on crowd-sourced smartphone data

In addition to bus arrival time information for study objective (4), bus crowding is one of the essential information for passengers' journey planning with long-distance travel. With the challenges of obtaining bus data from in-vehicle sensing devices, this chapter proposes a new method for bus crowding prediction based on the crowd-sourced bus data contributed by participating bus passengers. In connection to study objective (5), the results from bus arrival time estimation/prediction in Chapter 5 are incorporated into the method proposed in Chapter 6 for bus crowding prediction in which bus headways are considered as the primary factor. Moreover, bus dwell time estimation based on crowd-sourced bus data is also proposed in order to estimate the bus crowding level at a bus stop based on the bus dwell time and the headway. Then the bus crowding prediction process can take the estimated crowding information into account. The proposed bus crowding prediction system is evaluated using real-world bus datasets. The results show that when there is sufficient crowd-sourced bus data, the potential for bus crowding prediction is good with accuracy levels that are comparable to predictions made from the AVL bus data.

6.1 Introduction

Providing bus crowding information can improve passengers' travel experiences and the attractiveness of using the bus service. Bus crowding is one of the performance indices applied for evaluating the on-board comfort level. Previous studies have suggested that passengers' travel decisions are affected by crowding in the transit vehicles, especially for long-distance travel in highly-populated cities where passenger demand can outstrip the

available seating capacity of the transit vehicle. With the availability of bus crowding information, passengers can choose better mode choices, route choices, and bus choices. Furthermore, bus crowding affects the values of travel time due to the different perceptions of the values of in-vehicle travel times, passenger waiting times, and travel time reliability.

In Chapter 2, there is limited literature on bus crowding estimation and prediction since crowding on a bus can be derived from the passenger boarding and alighting data at individual bus stops. The observation of passenger boarding and alighting can be facilitated using in-vehicle sensing systems for transit ridership tracking such as APC, or closed-circuit television (CCTV) cameras. With the availability of the in-vehicle sensing systems, the estimation of bus crowding can be straightforward. On the other hand, the estimation involves significant challenges without in-vehicle sensing devices. Furthermore, bus crowding prediction poses extra difficulties in providing accurate crowding information. For each running bus, the system needs to predict the bus crowding at the remaining bus stops for the bus route concerned while the number of boarding and alighting passengers at the remaining bus stops is however unavailable at the current time interval.

The need for in-vehicle sensor data can be overcome since bus passengers can contribute to the provision of the bus data. In this chapter, the bus arrival time estimation/prediction results in Chapter 5 will be utilized for the prediction of bus crowding. Since bus headways can be available for all bus stops of a bus route, bus crowding prediction can be initially based on the bus headways. An initial assumption is that passenger demand for a bus route at a bus stop could be accumulated when bus headways are varied. However, the relationship between bus headways and bus crowding may not be stationary. For example, common or overlapping bus routes could be operated at the same bus stop, and passengers may therefore have several boarding options. In such cases, passengers may choose to board alternative bus lines resulting in the invalidation of the initial assumption.

Apart from bus headways, this study investigates the practicality of using bus dwell time information for estimating bus crowding levels. Firstly a bus dwell time estimation method based on crowd-sourced bus data is proposed. Then the estimated dwell times together with bus headways can be incorporated into bus crowding estimation. With the updated bus crowding information, the accuracy of bus crowding prediction for the remaining bus stops could be improved. Finally, bus crowding information can be disseminated to bus passengers in return.

The remainder of this chapter is organized as follows. A brief overview of the proposed bus crowding prediction system is presented in Section 6.2. The formulation of the bus crowding prediction problem is then provided in Section 6.3, followed by a description of data preparation based on the crowd-sourced bus data given in Section 6.4. The bus crowding estimation process is elaborated with a heuristic method for bus dwell time estimation in

Section 6.5, while Section 6.6 proposes a bus crowding prediction method. In Section 6.7, the proposed system is evaluated using real-world bus data. Finally, a summary of this chapter is presented with a summary of the findings in Section 6.8.

6.2 Operational overviews

Figure 6.1 depicts the operational overview of the proposed bus crowding prediction system.

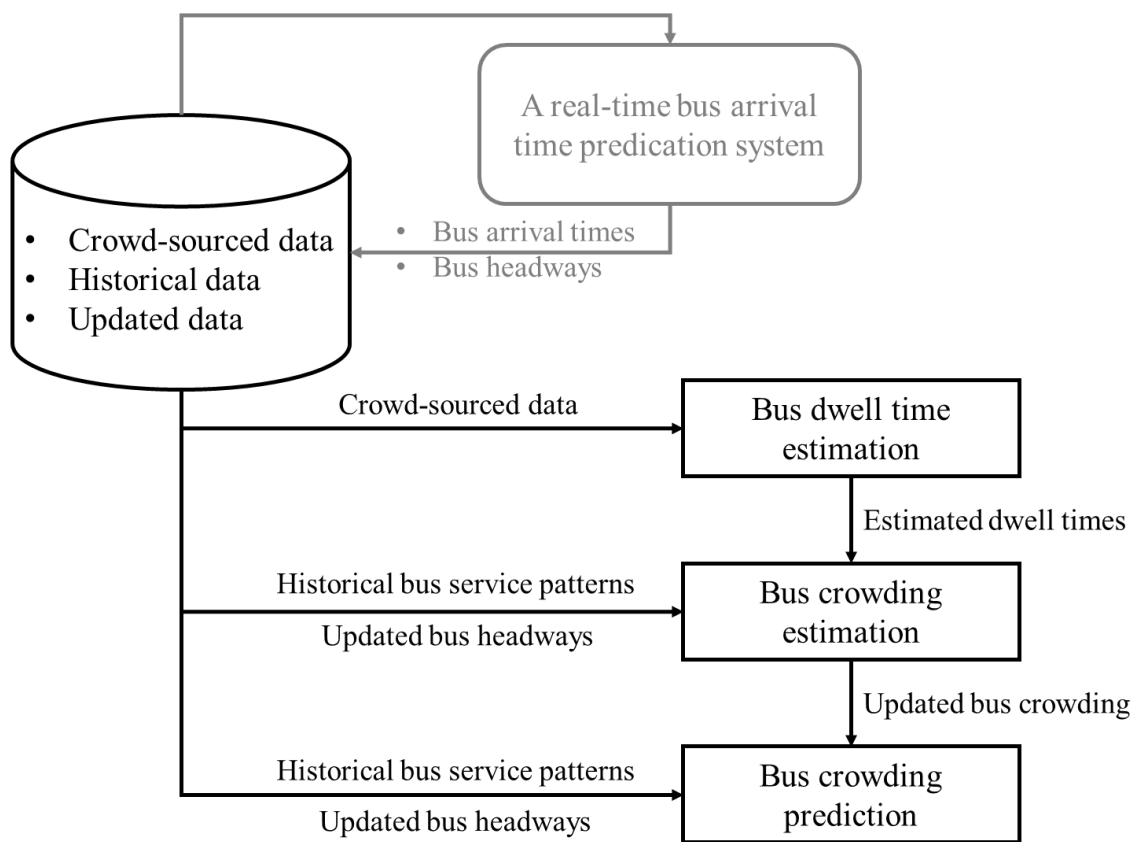


Figure 6.1: Operational overviews

In the same way as the proposed bus arrival time prediction system presented in Chapter 5, crowd-sourced bus data can be derived from the participating passengers. In addition to the primary bus data in a bus dataset, the smartphone application can be designed to request bus crowding data from the participating passengers. The reported bus crowding data can be used for estimating bus crowding and updating the estimation/prediction results. However, this study aims to investigate the potential of crowd-sourced bus data which requires minimum

manual inputs from the bus passengers. The reported bus crowding data will not be addressed in this study. The use of reported bus crowding will be investigated in future works.

At the back-end system, one of the system inputs is the bus arrival time results from the real-time bus arrival time information system. The bus arrival times can be processed to derive bus headways at individual bus stops. Next, in parallel with the real-time bus arrival time information, the reported bus data are processed for estimating the bus dwell times. Then, both the bus headways and bus dwell times are used for the bus crowding estimation. For an individual bus, the estimated bus crowding and headways are incorporated into the bus crowding predictions for the remaining bus stops of the bus route concerned.

6.3 Problem formulation

In traveler information systems, bus crowding information should be easily perceived by bus passengers. The number of on-board passengers can be discretized into several crowding levels such as Level of Services (LOS) categories LOS A to LOS F. The bus crowding information of an individual bus route or bus line can be denoted by an $R \times S$ matrix:

$$W_{bn,A,B,\tau} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1,S} \\ w_{21} & w_{22} & \dots & w_{2,S} \\ \vdots & \vdots & \ddots & \vdots \\ w_{r,1} & \dots & \dots & w_{r,S} \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix}_{R \times S} \quad (6.1)$$

The matrix describes the bus crowding information of a bus line bn which provides bus services from an origin bus stop on node ID A , to the bus terminus on node ID B . The information can be updated in every time interval τ based on the newly available bus data gathered during the time interval. Suppose that a bus route consists of S bus stops, and R bus running sequences have been operated on the route in a given day. The bus crowding level of a bus running number i after passenger boarding/alighting at a bus stop j is denoted by each element $w_{i,j}$ where $i \leq R$ and $j \leq S$.

The objective of bus crowding prediction is to provide a solution matrix $\tilde{W}_{bn,A,B,\tau}$ which can describe the information that is most similar to the actual data in $W_{bn,A,B,\tau}$.

$$\widetilde{W}_{bn,A,B,\tau} = \begin{bmatrix} \widetilde{w}_{11} & \widetilde{w}_{12} & \dots & \widetilde{w}_{1,k} & \widehat{w}_{1,k+1} & \dots & \widehat{w}_{1,S} \\ \widetilde{w}_{21} & \widetilde{w}_{22} & \dots & \widehat{w}_{2,k} & \widehat{w}_{2,k+1} & \dots & \widehat{w}_{2,S} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \widetilde{w}_{i,1} & \dots & \widetilde{w}_{i,j} & \widehat{w}_{i,j+1} & \vdots & \vdots & \widehat{w}_{i,S} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}_{R \times S} \quad (6.2)$$

For each processing time interval τ , the bus crowding level of a bus running sequence i can be estimated based on the current bus location. Suppose that a bus i has passed the j^{th} bus stop on the route at a time interval. The bus crowding level of the bus after the service at the k^{th} bus stop can be estimated and/or predicted:

$$w_{i,k} = \begin{cases} \widetilde{w}_{i,k}; k \leq j \\ \widehat{w}_{i,k}; j \leq k \leq S \end{cases} \quad (6.3)$$

where $\widetilde{w}_{i,k}$ is the estimated bus crowding level, and $\widehat{w}_{i,k}$ is the predicted bus crowding level.

6.4 Data preparation

The notation and definition of relevant bus information is provided in this section; namely, bus arrival times, bus headways, bus dwell times, and bus service patterns.

6.4.1 Bus headways

In this study, bus headways at a bus stop can be defined as the time difference between the arrival times of the same bus line at the bus stop. Firstly, the bus arrival time information of an individual bus route can be denoted by:

$$\widetilde{AR}_{bn,A,B,\tau} = \begin{bmatrix} \widetilde{ar}_{11} & \widetilde{ar}_{12} & \dots & \widehat{ar}_{1,k} & \widehat{ar}_{1,k+1} & \dots & \dots \\ \widetilde{ar}_{21} & \widetilde{ar}_{22} & \dots & \widehat{ar}_{2,k} & \widehat{ar}_{2,k+1} & \dots & \widehat{ar}_{2,S} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \widetilde{ar}_{i,1} & \dots & \widetilde{ar}_{i,j} & \widehat{ar}_{i,j+1} & \vdots & \vdots & \widehat{ar}_{i,S} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}_{R \times S} \quad (6.4)$$

where each element is the estimated bus arrival time $\widetilde{ar}_{i,j}$, or the predicted bus arrival time $\widehat{ar}_{i,j}$ based on the bus location at the current time interval τ . Here, the bus arrival times are derived from the real-time bus arrival time information system proposed in Chapter 5. The bus location matching and link travel time estimation processes result in the estimated bus arrival times, while the bus arrival time prediction process provides the predicted bus arrival times.

Given the bus arrival time information matrix, bus headways can be denoted by:

$$\widetilde{HT}_{bn,A,B,\tau} = \begin{bmatrix} \overline{\widetilde{ht}}_{21} & \overline{\widetilde{ht}}_{22} & \dots & \overline{\widetilde{ht}}_{2,k} & \overline{\widetilde{ht}}_{2,k+1} & \dots & \overline{\widetilde{ht}}_{2,S} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \overline{\widetilde{ht}}_{i,1} & \dots & \overline{\widetilde{ht}}_{i,j} & \overline{\widetilde{ht}}_{i,j+1} & \vdots & \vdots & \overline{\widetilde{ht}}_{i,S} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \end{bmatrix}_{R \times S} \quad (6.5)$$

where $\widetilde{ht}_{i,j}$ is the headway estimated from the difference between the bus arrival time $\widetilde{ar}_{i,j}$ and the arrival time of the previous bus at the bus stop j , and $\widehat{ht}_{i,j}$ is the predicted bus headway. It can be noted that bus headways are unavailable for the first bus of the given day.

Next, the bus headway can be represented by a discrete range which represents the deviation of the headway from the expected headway at the bus stop $E(ht_{i,j})$. The bus schedule can be used for determining the expected bus headway. The relative deviation of bus headway $\phi(ht_{i,j})$ is calculated by

$$\phi(ht_{i,j}) = \frac{E(ht_{i,j}) - \widetilde{ht}_{i,j}}{E(ht_{i,j})} \quad (6.6)$$

It is noteworthy that the bus headway deviation can result in a negative value which indicates that the bus arrival is earlier than the expected arrival, whereas a positive value indicates a late bus arrival.

The range of bus headway deviation is denoted by:

$$\widetilde{hd}_{i,j} = \begin{cases} \mathbf{1} & ; \phi(ht_{i,j}) < \min(\phi(ht)) \\ \mathbf{L} & ; \phi(ht_{i,j}) > \max(\phi(ht)) \\ \left\lceil \frac{\phi(ht_{i,j}) + \max(\phi(ht))}{L_h} \right\rceil & ; \textit{otherwise} \end{cases} \quad (6.7)$$

where the bus headway deviation can be discretized into L ranges with a range width L_h , $\min(\phi(ht))$ and $\max(\phi(ht))$ is the pre-defined minimum and maximum deviation of bus headway, respectively. Finally, a matrix of bus headway deviation $\widetilde{HD}_{bn,A,B,\tau}$ can be derived based on $\widetilde{HT}_{bn,A,B,\tau}$.

6.4.2 Historical bus service pattern

Since a bus crowding level $\tilde{w}_{i,j}$ can be estimated based on the bus headway and the bus dwell time patterns at a bus stop i , a set of historical patterns of the three types of bus information needs to be recorded in the database for the bus crowding estimation. Before the system deployment, high-resolution GPS bus traces and bus crowding information after the buses have departed each bus stop can be recorded by a group of volunteers. As described in Chapter 5, a GPS trace can be processed to derive the bus delay time on the delayed zone of each traversed link $dt_{a,b}^\tau$. Suppose that node ID b is the bus stop j on the bus route concerned. The bus delay time is assumed to be the bus dwell time at the bus stop

As a result, the bus service pattern of bus at a bus stop j during time interval τ is denoted by $sp_{bn,A,B,i,j,\tau}$ where i is the bus running sequence of the bus route in a given day. The pattern describes the date of bus operation, the estimated range of bus headway, the estimated range of bus dwell time, and the observed bus crowding level; $sp_{bn,A,B,i,j,\tau} = \{date, \tilde{hd}_{i,j}, \tilde{dw}_{i,j}, \tilde{w}_{i,j}\}$.

The range of bus dwell time can be calculated as follow:

$$\tilde{dv}_{i,j} = \left\lfloor \frac{dt_{a,b}^\tau}{L_d} \right\rfloor \quad (6.8)$$

where L_d is a pre-defined range width for the bus dwell times.

The data set SP including all historical bus service patterns will be used for further used for estimating bus crowding for each time interval.

6.5 Bus crowding estimation

Without in-vehicle sensor data, the number of boarding and alighting passengers cannot be adopted directly for the bus crowding estimation. In view of this, a new method for estimating bus crowding levels is proposed in this chapter by considering the patterns of bus dwell times and bus headways at a bus stop. This section firstly describes the bus dwell time estimation based on crowd-sourced bus data. Then the estimated dwell times of buses can be used for the bus crowding estimation as presented below.

6.5.1 Bus dwell time estimation

One of the challenges of using crowd-sourced bus data is data inconsistency. In Chapter 5, a bus location matching process was proposed in order to handle the bus data inconsistencies and to identify the more reliable datasets in the crowd-sourced data. The results from the bus location matching include a sequence of bus locations at nodes (i.e. intersections and bus stops) including the time spent at the node locations. Also, the bus travel times between pairs of nodes are estimated for the bus arrival time prediction purposes.

Estimating the bus dwell times at bus stops based on crowd-sourced bus data is more challenging. The resolution of bus location data could be insufficient for estimating the time spent at a particular location on a bus stop node. Due to these difficulties, this section proposes a bus dwell time estimation method for deriving bus dwell time information at some potential bus stops from the crowd-sourced bus data. To be more specific, some of the proposed bus location matching processes can be modified for the purpose of estimation.

In this chapter, the bus running sequence identification and candidate location determination processes can be implemented in the same way as the proposed bus arrival time prediction system in Chapter 5, while the details of other processes can be modified accordingly as shown below.

6.5.1.1 Candidate location formalization

Given the time-ordered bus datasets of the bus line number bn , running sequence r , $P_{bn,r,A,B}$, and set of candidate locations CL^i for each dataset $p_{bn,r,A,B,i}$, some candidate locations can be relocated to one of the node locations. Since this process aims to estimate bus dwell times at bus stops, only bus stop nodes on the bus route in the set $RN_{bn,A,B}$ are considered for relocating the bus data. Suppose that a bus stop is represented by node ID b , nd_b , linking to the preceding node on the route nd_a . The bus dwell time at the bus stop is defined by the time spent on the delay zone between the starting point of the zone (i.e. $\bar{q}_{a,b}^\tau$), and the bus stop node nd_b .

Relocating the candidates to node locations can provide the path travel times between a pair of nodes. Here, bus dwell times at the bus stop between the nodes constitute a part of the path travel time. Therefore, the distance between nodes could be too long for dwell time estimation. In order to limit the distance between a pair of nodes, a virtual node vd_a on a link $l(a,b)$ can be defined for each bus stop node nd_b on the bus route. The virtual node represents the starting location of the delay zone during the processing time interval τ , $vd_a = \bar{q}_{a,b}^\tau$, where buses tend to travel at low speeds in order to dwell at the bus stop.

Accordingly, candidate locations can be relocated to the bus stop nodes on the route and/or the virtual nodes of the bus stops.

Then, candidate location formalization can be performed based on the assumption that the variability in the travel times on the regular zone is less than the variability in delay times on the delay zone. Providing a bus data point on a delay zone, it is difficult to assume how long the bus has been delayed on the delay zone. In contrast, a free-flow bus speed can be assumed for relocating a bus data point on a regular zone to any location on the section.

Thus, a candidate location on the regular zone can result in two relocated locations on both the node in the backward direction and the virtual node in the forward direction. Figure 6.2 displays the candidate location formalization of a candidate on the regular zone of a link. Furthermore, the time at the new locations is estimated using (a) the instantaneous bus speed if the speed is a free-flow speed or (b) the average bus speed on the link during the same time interval, $\bar{v}_{a,b}^T$, if the instantaneous bus speed is a congested speed.

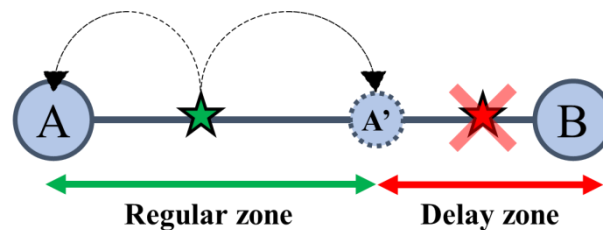


Figure 6.2: Candidate location formalization for dwell time estimation

The candidate location formalization process results in a number of bus locations at the bus stops with the virtual nodes representing the delay zones. The next step is to identify the representative data points for estimating bus the dwell time at individual bus stops.

6.5.1.2 Identifying a representative dwell time

To estimate the bus dwell time at bus stop node nd_b , the datasets which are relocated on the bus stop node and the virtual node vd_a on link $l(a,b)$ of the bus stop are considered. The datasets can be represented by a direct graph where each vertex denotes a relocated dataset. The vertices are divided into two groups: the datasets for the virtual node and the datasets for the bus stop node. Then each edge represents the bus dwell time on the delay zone based on the transition of bus data from the virtual node to the bus stop node. Assuming that there are M datasets for the virtual node and N datasets for the bus stop node, there will be $M \times N$ candidate dwell times in the graph.

Figure 6.3 illustrates a candidate graph for estimating the bus dwell time at a bus stop during a processing time interval. The virtual node and the bus stop node are represented by the node sequences $sn = 1$ and $sn = 2$, respectively. A vertex is denoted by nc_w^{sn} ; $nc_w^{sn} = \{\ell, t', \rho, \varepsilon\}$, where ℓ is the node ID or the virtual node ID, t' is the estimated time at the node, ρ is the identification number i of the reported dataset which is the source of the candidate cl_j^i , and ε is the GPS error inherited from the error of candidate location cl_j^i . ε before performing the location formalization.

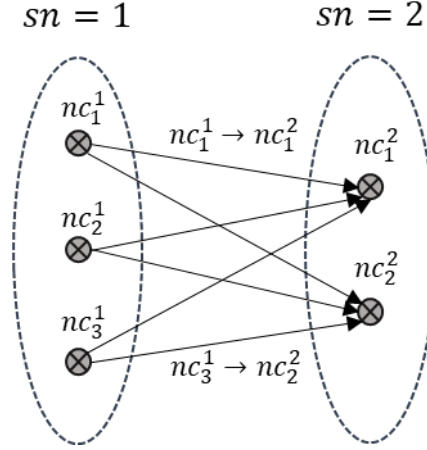


Figure 6.3: A candidate graph for the bus dwell time estimation

Each vertex is associated with an observation probability:

$$N(nc_w^{sn}, \varepsilon) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(nc_w^{sn}, \varepsilon - \mu)^2}{2\sigma^2}} \quad (6.9)$$

where μ and σ are the mean and the standard deviation of smartphone GPS errors, respectively.

Moreover, each edge is associated with a transmission probability which determines the likelihood between the candidate dwell time and the historical dwell time:

$$G_t(nc_m^1 \rightarrow nc_n^2) = 1 - \frac{|dw_2 - \overline{dw}_2^\tau|}{\overline{dw}_2^\tau} \quad (6.10)$$

where dw_2 is the candidate dwell time calculated from $nc_m^1 \cdot t' - nc_n^2 \cdot t'$, and \overline{dw}_2^τ is the average dwell time at the bus stop ($sn = 2$) during the time interval τ . The average bus dwell time can be derived from the historical delay time $\overline{dt}_{a,b}^\tau$, where b is the node ID representing the bus stop node sequence, and a is the node ID of the prior node of b on the bus route.

Hence, the bus dwell time derived from a candidate of the virtual node and a candidate of the bus stop node can be evaluated by the transmission function:

$$G(nc_m^1 \rightarrow nc_n^2) = N(nc_m^1, \epsilon) + \left(N(nc_n^2, \epsilon) \times G_t(nc_m^1 \rightarrow nc_n^2) \right) \quad (6.11)$$

Let $VP_C: nc_{e_1}^1 \rightarrow nc_{e_2}^2$ represents a pair of vertices where e_i is the vertex number of the node sequence i . The bus dwell time can be estimated from the vertex pair, VP , with the maximum transmission score:

$$VP = \operatorname{argmax}_{VP_C} G(VP_C) \quad (6.12)$$

Finally, the bus dwell time at bus stop node ID b during the time interval τ is denoted by:

$$\widetilde{dt}_{r,b}^\tau = nc_{e_1}^1 \cdot t' - nc_{e_2}^2 \cdot t' \quad (6.13)$$

Also, the range of dwell time $\widetilde{dv}_{r,b}^\tau$ can be calculated from L_d .

6.5.2 Bus crowding estimation based on bus dwell time

Since the variation of bus dwell time at a bus stop could be affected by several factors such as the number of boarding/alighting passengers and the number of on-board passengers, it could be assumed that some patterns in bus dwell times at a bus stop can describe the bus crowding level after a bus has departed from the bus stop. In addition, the relationship between bus dwell times and bus crowding levels can be varied due to the effects of bus headway deviation. On the one hand, passenger demand at a bus stop can be increased when the bus headway is extended resulting in a longer bus dwell time. On the other hand, the demand can change if passengers have several boarding options if there are overlapping or common bus routes at the bus stop. In such cases, passengers can choose to board an alternative bus if the bus headway of the expecting route is excessive.

In this chapter, both the bus dwell times and bus headways are incorporated for the bus crowding estimation. At this stage, the estimated bus dwell time $\widetilde{dv}_{r,b}^\tau$ and the estimated headway deviation range $\widetilde{hd}_{r,b}^\tau$ are available for a bus stop b . As a result of bus dwell time estimation, bus crowding can be approximately estimated from the historical bus service patterns (i.e. the pattern of bus dwell time and bus headway at the bus stop) which are most similar to the current pattern.

Let vector A denotes the current bus service pattern of a bus described by the bus dwell time and the bus headway at a time interval:

$$\mathbf{A} = [\widetilde{hd}_{bn,A,B,r,b}^{\tau}, \widetilde{dv}_{bn,A,B,r,b}^{\tau}]^T \quad (6.14)$$

where $\widetilde{dv}_{bn,A,B,r,b}^{\tau}$ is the range of bus dwell time $\widetilde{dt}_{r,b}^{\tau}$.

Vector B denotes a historical bus service pattern in the set SP which identifies the bus dwell time and the bus headway at a bus stop $j = b$. All datasets M indicating the bus service of the same bus route during the same time interval are considered as vector B_M in order to calculate the cosine similarity $\pi(A, B)$ between the historical data and the current bus data.

$$\pi_i(\mathbf{A}, \mathbf{B}_i) = \frac{\mathbf{A} \cdot \mathbf{B}_i}{\|\mathbf{A}\| \|\mathbf{B}_i\|}; i = 1, 2, \dots, M \quad (6.15)$$

The bus crowding level at the bus stop is assumed to be similar to the historical bus crowding level when the pattern of bus dwell times and the bus headways are comparable. From all datasets M , the maximum cosine similarity may indicate several possibilities of historical bus dwell times and headway patterns which match the current pattern. Therefore, the estimation may provide a set of estimated crowding levels from several historical patterns $W'_{bn,A,B,r,b,\tau} = \{W'_{bn,A,B,r,b,\tau,1}, \dots, W'_{bn,A,B,r,b,\tau,N}\}$.

Then the availability of bus dwell times can be updated in a matrix:

$$DS_{bn,A,B} = \begin{bmatrix} ds_{11} & \cdots & ds_{1S} \\ \vdots & \ddots & \vdots \\ ds_{i1} & ds_{ij} & ds_{iS} \\ \vdots & \vdots & \vdots \end{bmatrix}_{R \times S} \quad (6.16)$$

where $ds_{i,j}$ is a binary variable indicating the availability of bus dwell time at a bus stop j estimated from a bus running sequence i .

6.6 Bus crowding prediction

Since bus headways can be derived from every processing time interval, the bus headway deviation matrix can be considered as the primary variable for the bus crowding prediction. Therefore, an initial assumption is that the bus crowding level after a bus has departed a bus stop is affected by the deviation of the bus headway at the bus stop over the time periods of a

day. For instance, a longer time headway could result in a higher crowding level on the bus since the number of waiting passengers at the bus stop has been accumulated during the headway.

However, as described earlier, passenger demand could be affected by the overlapping or common bus routes at the bus stop, and the bus crowding level could be varied for a specific deviation of headways at a time interval. Therefore, the estimated bus crowding based on bus dwell times can be incorporated into the prediction process. It is noteworthy that the estimated bus dwell times may be available only for some bus stops due to the uncertainty of crowd-sourced bus data.

The bus crowding levels of an individual bus i at bus stops on the route can be predicted based on the headway deviation matrix of the bus. To simplify the notations, the bus crowding levels of an individual bus i in the matrix $\tilde{W}_{d,bn,A,B,\tau}$ is denoted by:

$$\tilde{W}_i = [\tilde{w}_{i,1} \quad \dots \quad \tilde{w}_{i,j} \quad \hat{w}_{i,j+1} \quad \dots \quad \hat{w}_{i,S}] \quad (6.17)$$

In a similar way, the bus arrival time matrix, the headway deviation matrix, and the dwell time availability matrix can be denoted by $\tilde{A}R_i$, $\tilde{H}D_i$, and DS_i , respectively.

The bus crowding can be estimated/predicted by searching for the historical pattern of bus headway deviation that is most similar to the current one. The searching space can be determined by setting the number of searching dates Φ and time intervals P . For example, all headway deviation matrices of the buses on the dates $\omega - \Phi$ during the time intervals $\tau - P$ will be considered for bus crowding prediction, where ω is the current date, and τ is the current processing time interval.

Next, a scoring coefficient can be defined based on the potential of each headway deviation matrix in the searching space. Given the dwell time availability matrix DS_i and the sets of estimated bus crowding levels of the bus i at bus stops during the current time interval $W'_{i,j}$, the scoring coefficient can be calculated:

$$\varpi_k = \sum_{j=1}^S ds_{i,j} \quad ; \forall (ds_{i,j} \times \tilde{w}_{k,j}) \in W'_{i,j}, ds_{i,j} \in DS_i \quad (6.18)$$

where i is the current bus, k is a bus from the search space, S is the total bus stops on the route, $ds_{i,j}$ is the dwell time availability of the current bus i at a stop j , and $\tilde{w}_{k,j}$ is the historical bus crowding level of the bus k at a stop j .

The scoring coefficient can be normalized to facilitate headway deviation matching:

$$\boldsymbol{\omega}'_k = \begin{cases} \mathbf{1} & ; \boldsymbol{\omega}_k = \mathbf{0} \\ \frac{\boldsymbol{\omega}_k}{\sum_{j=1}^S ds_{i,j}} & ; \textit{otherwise} \end{cases} \quad (6.19)$$

Then, a historical headway deviation matrix \widetilde{HD}_k will be included for bus crowding estimation. The difference between a historical headway deviation matrix from a bus k and the current headway matrix deviation from a bus i can be calculated

$$\eta_k = \boldsymbol{\omega}'_k \times D(\widetilde{HD}_i, \widetilde{HD}_k) \quad (6.20)$$

where η_k is the mismatch score indicating the degree of difference between the two matrices, and $D(\cdot)$ is the Euclidean distance function. Suppose that there are N headway deviation matrices in the searching space. The bus crowding level matrix \widetilde{W}_i predicted at the current time interval τ can be estimated from the bus crowding level matrix \widetilde{W}_k which η_k is the minimum score for all mismatch scores η_1, \dots, η_N .

It is noteworthy that the headway deviation matrix is unavailable for the first running bus of the day. In such a case, the bus arrival time matrices $\widetilde{AR}_i, \widetilde{AR}_k$ can be used instead.

6.7 Experimental studies

6.7.1 Data description

Further to Case study #2 in Chapter 5, real-world bus data were used for evaluating the proposed bus crowding prediction system. There are 33 bus stops from 39 nodes on the road section, and the eight bus lines share some of the bus stops. Based on the bus arrival time prediction results in Chapter 5, some crowd-sourced datasets were insufficient for providing adequate bus arrival time information. Therefore, eleven crowd-sourced datasets were used for system evaluation in this chapter:

- An AVL bus dataset,
- Four sets of 10% participating passengers with the four sampling methods,
- Three sets of 5% participating passengers without the dataset of the one-time sampling method, and
- Three sets of 3% participating passengers without the dataset of the one-time sampling method.

In order to evaluate the feasibility of using bus dwell times and bus headways for estimation of the bus crowding level in practice, the bus crowding data reported by participating passengers are not involved in the system evaluation at this stage.

Furthermore, the GPS data were used for creating historical bus service patterns *SP*. The estimated bus dwell times were discretized into a range with a width set at 10 seconds. Also, the range of a bus headway deviation was estimated based on the expected bus headway at each bus stop. Since there is no bus schedule provided to the public, the average bus headway was considered as the expected bus headway. Moreover, the bus headway deviation was classified into twenty ranges by setting the minimum deviation as -1.0, the maximum deviation as +1.0, and the range width as 0.1.

Next, bus crowding levels on an individual bus were derived from the number of boarding and alighting passengers at individual bus stops which were recorded by the volunteers as described in Case study #2 in Section 5.7.2. The number of on-board passengers after the bus departed each bus stop was calculated and further discretized into the six LOS categories. In this chapter, the LOS categories were determined based on the bus capacity. The bus capacity of the eight bus lines has 34 seats and 46 standing capacity. Hence, the total capacity is 80 passengers per bus. In the previous studies, the number of passengers per seat has been suggested for determining the LOS of a bus (Katz and Rahman, 2010; Das and Pandit, 2015). However, the number of passengers per seat may not be well-represented in bus passengers' perception, especially when bus passengers need to determine an LOS to describe the bus crowding. This chapter defines the LOS categories based on percentages of available seats and standings compared to the bus capacity. The percentages were determined based on the average number of passengers per seat indices suggested in the above two previous related studies. Table 6.1 shows the criteria of the six LOS categories.

Table 6.1: Bus crowding by LOS categories

LOS category	Available seats	Available standings	Crowding level (passengers per seat)
A	> 50%	100%	0.50
B	< 20%	100%	0.79
C	0%	> 80%	1.27
D	0%	51-80%	1.67
E	0%	20-50%	2.09
F	0%	< 20%	2.35

It is noteworthy that the datasets were separated for creating the historical data and testing the proposed system. The historical data were derived from four-day datasets. The testing datasets were independently used as the system inputs. In this case study, the bus crowding prediction was performed at every 5-minute interval.

6.7.2 Numerical results

This section provides an overall performance evaluation of the proposed system for which the bus crowding estimation and bus crowding prediction processes were analyzed based on

- (a) Availability of bus dwell times calculated by the number estimated dwell times and the number of available bus arrival times per time interval

$$AD(\%) = \frac{\text{No. of estimated dwell times}}{\text{No. of estimated bus arrival times}} \quad (6.21)$$

- (b) The MAE of the estimated dwell times (seconds), and
- (c) The MAE of the estimated bus crowding levels (LOS).

Then the overall accuracy of bus crowding prediction was evaluated by

- (d) The MAE of the predicted bus crowding levels (LOS).

The MAEs in (b), (c), and (d) were calculated from:

$$MAE = \frac{\sum_{i=1}^N |y_i - \hat{y}_i|}{N} \quad (6.22)$$

where y_i is the actual value, \hat{y}_i is the estimated/predicted value, and N is the number of observations in each time interval.

Table 6.2 provides the numerical results from the 11 simulated bus datasets. The results are used for data analyses in the following two sub-sections for the illustration of the proposed system.

6.7.2.1 Bus crowding estimation

Since an estimated bus crowding level is assumed based on the pattern of bus dwell times and the bus headways, the estimation process may indicate several possibilities of the patterns in the database. As a result, the estimation may provide a set of estimated crowding levels $W'_{bn,A,B,r,b,\tau}$ from several historical patterns that will be further incorporated into the process for bus crowding prediction.

Table 6.2: Numerical results from the 11 bus datasets

Passenger participation	Sampling method	Bus crowding estimation			Bus crowding prediction	
		(a) Dwell times/ arrival times (%)	(b) MAE of bus dwell time estimation(s)	(c) MAE of bus crowding estimation (LOS)	(d) MAE of bus crowding prediction without updated bus crowding (LOS)	(e) MAE of bus crowding prediction with updated bus crowding (LOS)
AVL	-	34.67	24.83	0.73	0.74	0.66 (+10.81%)
10%	1	34.37	34.15	0.88	0.97	0.88 (+9.28%)
	2	23.68	39.84	1.05	1.12	1.04 (+7.14%)
	3	20.43	73.67	1.28	1.36	1.31 (+3.68%)
	4	N/A	N/A	N/A	1.82	N/A
5%	1	30.30	40.71	0.94	1.11	1.02 (+8.11%)
	2	12.06	54.40	1.41	1.45	1.43 (+1.38%)
	3	4.25	N/A	N/A	1.91	N/A
	4	N/A	N/A	N/A	N/A	N/A
3%	1	15.30	42.82	1.17	1.80	1.77 (+1.67%)
	2	2.44	N/A	N/A	2.13	N/A
	3	1.92	N/A	N/A	2.48	N/A
	4	N/A	N/A	N/A	N/A	N/A

To calculate the MAE of bus crowding estimation, the estimated bus crowding level in the set $W'_{bn,A,B,r,b,\tau}$ which provides minimum MAE is selected. Therefore, the MAE implies the accuracy of the most accurate result from all possible answers. The results show that the MAE of bus crowding estimation based on crowd-sourced bus data varies from 1.00-1.97 crowding levels depending on the availability of the crowd-sourced data. This means the most accurate bus crowding level in a set $W'_{bn,A,B,r,b,\tau}$ is different from the actual level on average.

The degree of bus crowding error is consistent with the bus dwell time estimation accuracy, in which the MAE of bus dwell time estimation is in a range of 34-74 seconds. Since the bus crowding estimation is dependent on the pattern of bus dwell times and the bus headways, the bus dwell time error directly affects bus crowding estimation accuracy. It can be noted that a bus dwell time is discretized into a dwell time range before performing bus crowding estimation. Thus, the large dwell time errors can be compromised.

In addition, the bus crowding may not be estimated for all traversed bus stops. The availability of bus dwell times (%) shows that crowd-sourced bus data could be insufficient for bus dwell time estimation. There are six out of the twelve datasets providing sufficient data for the estimation:

- The datasets from 10% participating passengers applying sampling methods 1-3,
- The datasets from 5% participating passengers applying sampling methods 1-2, and
- The datasets from 3% participating passengers applying sampling method 1.

The main reason for this is that a larger proportion of the crowd-sourced bus datasets may not be able to be used for the bus dwell time estimation. In the candidate location formalization for dwell time estimation, only bus locations on the regular zone of a road link are considered, while buses usually spend a longer duration of time on the delay zone.

6.7.2.2 Bus crowding prediction

The prediction performance is evaluated in two separate cases based on the incorporation of the estimated bus crowding data into the prediction process. The six datasets which are sufficient for bus crowding estimation are considered for validation of the bus crowding prediction results. The MAE of bus crowding prediction without considering the estimated bus crowding varies from 0.97-1.80, while the MAE of bus crowding prediction with the estimated bus crowding is in the range of 0.88-1.77 crowding levels. The prediction accuracy can be compared to the baseline accuracy from the AVL bus data which provides the predicted bus crowding level with 0.66 MAE. It is noteworthy that the historical data in this case study are derived from a four-day dataset. The bus crowding prediction could involve significant errors when the irregular bus services are included in the testing datasets without prior knowledge of the same service patterns in the historical data.

The results show that the accuracy of bus crowding prediction can be improved with the availability of bus dwell times and the estimated bus crowding levels. The degree of improvement is varied for different datasets. It can be observed that greater availability of bus dwell times can better improve the bus crowding prediction accuracy.

It can be concluded that crowd-sourced bus data can be used for bus crowding prediction. The crowd-sourced bus data can provide bus crowding prediction accuracy that is comparable with the AVL bus data when passenger participation is greater than 5%, and bus data sampling should be performed every one minute to maintain the prediction accuracy. However, the implication of errors requires further investigation based on passengers' perceptions. For instance, the difference between LOS B and LOS C could affect passengers' boarding decisions since some passengers may expect to find a seat under the LOS B. Providing inaccurate bus crowding information could affect the level of user acceptance and reduce the number of participating passengers. Furthermore, the optimum number of LOS categories should be identified based on bus passengers' perceptions. The prediction accuracy can be improved when bus crowding is described by a few LOS categories.

6.8 Summary of findings

This chapter proposed a bus crowding estimation/prediction system based on crowd-sourced smartphone data. Without the need for in-vehicle tracking devices, the proposed system can provide a new solution for deriving the bus crowding information. Since bus headways at bus stops are derived from the bus arrival time estimation/prediction results in Chapter 5, the system firstly incorporates bus headways into bus crowding prediction. However, the relationship between bus headways and bus crowding levels may not be stationary. Therefore, a method for bus dwell time estimation is also proposed in order to update properly the estimation of bus crowding levels. The case study results show that bus dwell times can improve bus crowding prediction accuracy. However, bus dwell times may not be available for every bus stop due to the sparseness of crowd-sourced bus data. It can be observed from the results that the prediction accuracy can be further improved when more bus dwell times are available. Although bus dwell time estimation may not provide accurate results, the estimated dwell times are still useful for bus crowding prediction.

For the overall performance, the proposed system can predict the bus crowding level with average errors of approximately within one LOS but can be up to two LOS depending on the availability of crowd-sourced bus datasets. The system can maintain the errors within one LOS when 5-10% of bus passengers can contribute to the provision of the relevant bus data. In addition, the system requires a bus dataset to be obtained every one minute.

In future studies, more accurate bus dwell time estimation methods can be developed to enhance the accuracy of bus crowding prediction as well as the accuracy of bus arrival time prediction. A method for updating bus crowding information is to request the information from on-board passengers. Participating bus passengers may directly provide bus crowding information on the bus they are riding. In the same way as bus dwell time information, the reported bus crowding data may not be regularly available for all bus stops. Although the information cannot be considered as the primary factor for bus crowding prediction, it can be used in the same way as bus dwell times (i.e. updating bus crowding information and being incorporated into bus crowding prediction). In such a case, bus crowding levels are estimated from the passengers' perception, which could be varied by each person under different environments, meaning inconsistencies in the reported bus crowding data could be encountered. The investigation of the potential of using reported bus crowding information for bus crowding estimation should be further investigated in future studies.

Part IV
Conclusions and Future Works

Chapter 7

Conclusions and recommendations for further research

7.1 Conclusions

This thesis proposed new methods for deriving important public transport information with use of human probe data and advanced data mining techniques, particularly when sensor data from transit vehicles are not available. Bus transit information systems based on human probe data were proposed for deriving three significant KPIs for bus service evaluation i.e. average bus passenger waiting times, real-time bus arrival times, and bus crowding levels. A brief summary of the key findings is given as follows.

Chapter 2 firstly gave the basic problem statement and relevant literature review of bus passenger waiting time estimation, bus arrival time prediction, and bus crowding prediction. As this thesis focused on the use of smartphone-based human probe data, Chapter 2 provides previous studies on the estimation and prediction of the three KPIs, followed with an overview of important smartphone sensing technologies and the applications for deriving mobility information for transportation studies. At the end of the chapter, previous related studies on using smartphone-based human probe data for deriving the three KPIs were summarized in order to provide the fundamental understanding for the development of bus transit information systems in the next following chapters.

In Hong Kong, the challenges for estimating bus information are due to the fact that the in-vehicle sensor data are not available to the public sector for the development of the bus information systems. In view of this, a bus passenger waiting time estimation system based

on a non-participatory sensing approach was therefore proposed in Part II of this thesis. Then Part III of the thesis presented the use of participatory sensing approaches for developing a real-time bus arrival time information system and a bus crowding prediction system.

With reference to Part II of this thesis, Chapter 3 focused on investigating possible uncertainties in passive Wi-Fi data. It was found that mobility information derived from passive Wi-Fi data could involve due to missing detection data. The results from designed experiments showed that a Wi-Fi scanner cannot capture every Wi-Fi data broadcast by a Wi-Fi device within its detection range. The probability of missed detection could vary based on the state of each individual Wi-Fi device and configurations of the Wi-Fi scanner. As a foundation for further development of public transport information systems, two uncertainty models were proposed for describing the uncertainties in device positioning and activity duration estimation based on passive Wi-Fi data. The models showed that the errors in activity duration could be affected by the Wi-Fi data detection frequency, the travel speed of Wi-Fi devices, and the size of study area.

Chapter 4 introduced a new method for bus passenger waiting time estimation at a single bus stop. The estimation method was developed based on the foundation understanding of passive Wi-Fi data in Chapter 3. Firstly, generalized classification features of passive Wi-Fi data were introduced in order to describe the spatial attributes of each Wi-Fi device. In order to identify waiting passengers from noisy Wi-Fi data, a classifier was developed based on the spatial and temporal uncertainties in the passive Wi-Fi data. The proposed system was evaluated using Wi-Fi datasets collected from two different bus stop environments. The results showed that the proposed system could provide 80-95% of AWT estimation accuracy which was more accurate than the baseline accuracy from the half-headway estimation method. Since passenger waiting times were assumed from the presence time of Wi-Fi devices, the Wi-Fi data detection frequency could strongly affect the estimation accuracy. In addition, the system might miss Wi-Fi signals from some groups of passengers who were able to board the bus quickly resulting in overestimation of AWT.

In connection to Part III of the thesis, Chapters 5 and 6 developed two bus information systems based on participatory sensing approaches in which bus passengers can contribute to the provision of the relevant bus data. First, a novel framework for developing a real-time bus arrival time information system based on crowd-sourced bus data was proposed in Chapter 5. Chapter 5 firstly addressed the characteristics of crowd-sourced bus data. Then the proposed framework was introduced including bus location filtering and bus arrival time prediction methods which were developed to overcome the challenges in crowd-sourced bus data. The proposed system was tested using both simulated bus data and real-world bus data. The results showed that the bus arrival time prediction accuracy is dependent on bus passenger participation in which more participation can provide comparable prediction accuracy with the AVL bus data. The MAPE of bus arrival time prediction derived from the largest bus

dataset (from 10% participating passengers using one-minute sampling frequency) was less than the MAPE from AVL bus data from 2-4% approximately.

As an extension of the real-time bus arrival time information system proposed in Chapter 5, a bus crowding prediction system based on crowd-sourced bus data was presented in Chapter 6. A bus dwell time estimation method was developed in order to estimate the bus crowding levels at some upstream bus stops. The updated bus crowding levels together with bus headways from the bus arrival time prediction system in Chapter 5 were incorporated into bus crowding prediction, in which the predicted bus crowding levels were based on the patterns of bus services represented by the two parameters. The system was initially evaluated using the same set of real-world bus data in Chapter 5. The results showed that the availability of bus dwell times could improve the accuracy of bus crowding prediction. With sufficient crowd-sourced bus data from 5-10% participating passengers, the proposed system could predict the bus crowding level with average error of 0.88 LOS while the average error from AVL bus data is 0.66 LOS.

To this end, the new methods proposed for estimating the three KPIs investigated in this thesis can improve the efficiency of APTS, especially when AVL data from bus transit vehicles are not available. At this stage, bus transit information systems can provide timely information (i.e. real-time bus arrival times and bus crowding levels) that helps bus passengers to make better their travel decision on which bus they should get on. Also, the quality of bus services can be evaluated based on the three KPIs concerned. Therefore, the government can continuously monitor and regulate the quality standard of bus services.

The limitations of applying the proposed methods were discussed in this thesis. It can be summarized that some limitations can be encountered in practice due to the characteristics of the related bus data that are opportunistically derived from bus passengers. First, passive Wi-Fi data may not be captured from all waiting passengers at a bus stop. Such missed detection could affect the accuracy of AWT estimation. Moreover, sufficient field surveys are necessary before implementing the proposed AWT estimation method since the characteristics of Wi-Fi data collected from various bus stop environments could be different. Second, the performance of the proposed methods based on crowd-sourced bus data is reliant on the number of participating passengers. Sufficient crowd-sourced bus data are required for providing accurate bus information.

With the current limitations of the proposed methods, multiple data sources can be considered for developing the future-state methods. For example, the integration between passive Wi-Fi data and crowd-sourced bus data can be advantageous to the estimation of the three KPIs. In addition to the smartphone-based human probe data, there are several data types from existing traffic sensors in a road network. Most sensors are installed at a fixed location to collect vehicle data at a particular location. In some cases, the stationary data could be more

reliable than probe vehicle data from mobile devices particularly for the road sections with high GPS or AVL measurement errors. However, the stationary sensors are installed only at some partial locations on the bus transit network. The limitation of stationary data in the spatial dimension could be one of the disadvantages compared to the probe vehicle data. Therefore, data fusion methods should be developed in future to maximize the use of multiple data sources which can provide more related information for estimation purposes.

To improve the efficiency of APTS in the future state, the estimation of three KPIs can be valuable for developing more advanced models (e.g. optimization models for transit scheduling and planning). This will benefit both public and private stakeholders since these model results are useful for public transit operational improvement, public transport system design, and planning. Basically, the previous related models have been developed based on a set of assumptions regarding each operational factor such as an average passenger waiting time, an average passenger arrival rate, etc. The model results can be significantly improved if more accurate and more updated input data can be provided. In the big data era, the estimation of the input data for public transport models can be performed continuously over time. These data can be used for validating the former assumptions adopted in the existing models. In this way, some assumptions in the past may be compromised and more realistic assumptions can be made. For instance, assuming an average passenger waiting time as half of the bus headway might be impractical for bus stop with common bus routes. The experimental results in Chapter 4 showed that the AWT estimation results from the half-headway method had significant errors. For the bus stops with common bus route services, the assumption regarding uniform passenger arrival could be violated since a passenger may choose to board the first arriving bus of the common routes.

In order to derive more updated input data, the participatory sensing approaches proposed in this thesis can be further developed. To a certain extent, the reported data from participating passengers can be further analyzed to derive more meaningful information such as passengers' route choices as well as their boarding decisions. Also, the behavioral changes of public transit passengers can be observed over time under the complex conditions of public transport services in Hong Kong. With the availability of updated information, more valid assumptions can be made for improving the accuracy of public transport models. Moreover, the newly available information could lead to new insights resulting in the new theories for modeling public transport services in future.

7.2 Recommendations for further research

Future research works can be developed in several aspects so as to improve the accuracy of bus passenger waiting times, bus arrival times, and bus crowding levels. In addition, the

practicality of deriving other related bus information from smartphone-based human probe data can be further investigated. The recommendations for further research based on passive Wi-Fi data and crowd-sourced bus data are provided in this section below.

7.2.1 Extension of bus information systems based on passive Wi-Fi data

7.2.1.1 Localization accuracy

As reported in the literature that the variation in RSSIs poses spatial uncertainties in mobility analysis, the use of multiple Wi-Fi scanners has been adopted in several previous studies for estimating more accurate device locations. The concept is to compare the RSSIs of a single Wi-Fi device at a time period from multiple Wi-Fi scanners. Figure 7.1 illustrates an example of two Wi-Fi scanners in which the detection range of each Wi-Fi scanner is partially overlapped with another range.

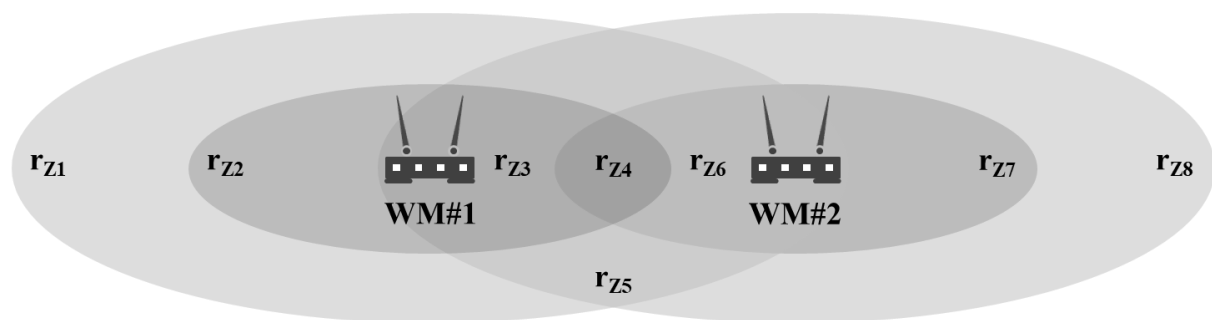


Figure 7.1: Extension of device monitoring areas using two Wi-Fi scanners

Suppose that an RSSI is classified as a strong signal or a weak signal. The device location can be roughly identified as one of the eight zones demonstrated in Table 7.1.

Table 7.1: Device locations based on two Wi-Fi scanners

Zone	RSSI from WM#1	RSSI from WM#2
r_{Z1}	Weak	-
r_{Z2}	Strong	-
r_{Z3}	Strong	Weak
r_{Z4}	Strong	Strong
r_{Z5}	Weak	Weak
r_{Z6}	Weak	Strong
r_{Z7}	Strong	-
r_{Z8}	Weak	-

For bus passenger waiting time estimation, the availability of multiple Wi-Fi scanners can enhance the efficiency of data filtering algorithms since the MAC-IDs outside passenger waiting areas at a bus stop can be discarded. Moreover, temporal uncertainties due to missed detection can be overcome. First, the extended detection range offers more time for detecting a Wi-Fi device. Second, the probability of missed detection can be decreased since the scanners may scan for different Wi-Fi channels at a specified time period.

7.2.1.2 Additional bus information

- **Bus passenger waiting time**

In Chapter 4, the proposed passenger waiting time estimation method provides estimation of AWT at an individual bus stop. However, the proposed method can be enhanced to estimate the AWT for individual bus routes without the AVL bus data. In order to do so, more Wi-Fi scanners can be installed at strategic bus stops on the bus transit network. Then, the routing of each bus can be identified when each MAC-ID is re-identified by other Wi-Fi scanners at the bus stops along the same bus route. However, the probability of a Wi-Fi scanner detecting on-board passengers and/or alighting passengers should be investigated for identifying the locations of strategic bus stops at which the Wi-Fi scanners should be installed at bus stops along the bus route concerned.

Based on the possible transformation of the raw data presented in Chapter 3, more factors can be considered for passenger waiting time estimation such as device outflows. In Chapter 3, a device outflow is defined as the number of MAC-IDs leaving the detection area at a specific time. Here, a Wi-Fi data set can be used to demonstrate device outflows at a bus stop. The Wi-Fi data were collected from a large bus stop area. To simplify the data analysis, the Wi-Fi scanner used for the data collection was configured for capturing the Wi-Fi signals within a smaller area i.e. one of the passenger queuing lines of a specific bus route number at the bus stop. Figure 7.2 shows the passenger device outflows by the time of a day compared to four bus arrival times of a bus route number (represented by four green dotted lines) during the data collection periods. It can be noticed that the graph has four peaks of device outflow values in which the time at the peaks are closely matched with the four bus arrival times observed at the bus stop. This implies that most of the waiting passengers have boarded the arrival buses and the Wi-Fi signals of these passengers' devices were no longer detected after the buses departed from the bus stop.

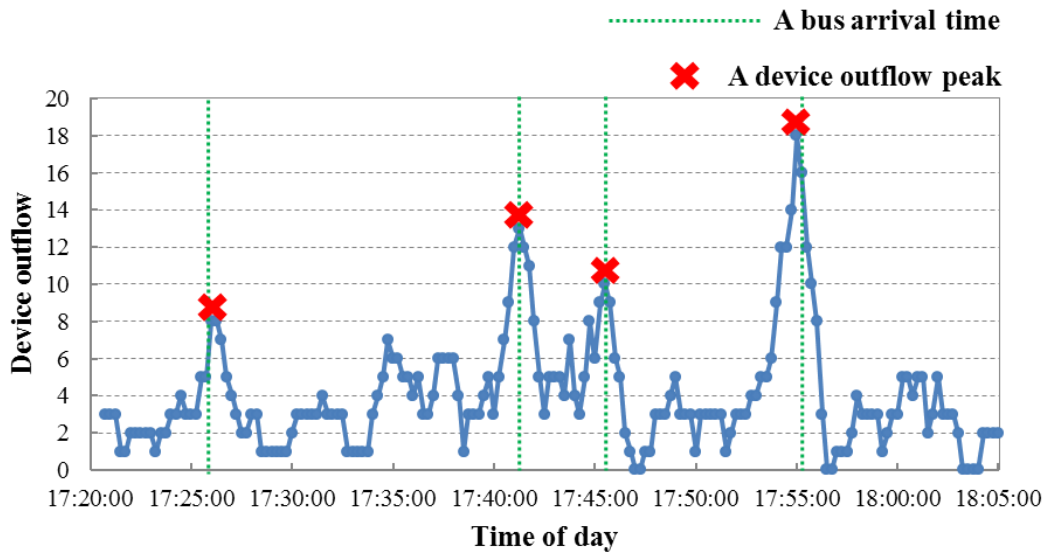


Figure 7.2: Device outflows at a bus stop

- **Bus arrival times and bus departure times**
 Further research is required for investigating the possibility of estimating bus arrival times based on a single Wi-Fi scanner at a bus stop with common bus routes. Two main challenges are involved in the bus arrival time estimation for this scenario: bus bunching and temporal uncertainties within the Wi-Fi data. In Hong Kong, several buses can arrive at a bus stop simultaneously, queue up, and depart from the bus stop within a similar time period. Then, a number of passengers may board several buses at a similar time period and the Wi-Fi signal from those passengers' devices will disappear simultaneously when the buses departed from the same bus stop. Although the example in Figure 7.2 shows the correlation between the peaks of device outflow and bus arrival times, bus bunching may result in only one peak of device outflow when several buses departed from the same stop at a similar time period. In order to address this issue, bus arrival time and bus departure time estimation may require prior information to identify the number of arrival buses from a single Wi-Fi scanner only. Further studies should be carried out to investigate this issue.

Moreover, the practicality of estimating bus travel times, bus arrival times, and bus departure times should be further investigated based on the availability of re-identified MAC-IDs across multiple Wi-Fi scanners in the bus transit network.

- **Passenger counting**
 One of the common challenges regardless of Wi-Fi scanner installation types is the accuracy of passenger counting data. The number of transit passengers and the number of detected Wi-Fi devices may not be a one-to-one relationship since some passenger-carried devices could be undetectable, and some passengers may carry more than one device. A scaling factor is necessary if the main objective is to estimate

the number of passengers i.e. on-board passengers, boarding passengers, alighting passengers, and waiting passengers at a transit station.

The system may require a large set of manual observation data for training the scaling function. Also, the method for handling the difficulties arising from an insufficient sample size could be necessary when there are only a few passengers in the study area. The performance of passenger crowding estimation should be evaluated in various passenger crowding levels.

For bus transit systems, estimating the number of waiting passengers at a bus stop with use of the Wi-Fi data only is more challenging. In Hong Kong, a bus stop is usually a shared facility for several bus routes with overlapping or common route sections. Therefore, during a given time interval at the same bus stop, the number of passengers waiting for each of these bus routes could be different and difficult to be estimated. As such, modeling the function for estimating the scaling factor of the detected passengers' Wi-Fi data would require additional and/or prior information particularly when there are very few passengers waiting at the bus stop and some of their Wi-Fi devices cannot be detected.

- **Passenger demands**

Identifying passengers' OD could be an uncomplicated task since the first and the last device location can be observed from the MAC-ID's session and/or trail. However, OD estimation may involve some errors due to the uncertainties in detection periods when the distance between two observing nodes is considerably short. As a consequence, the first detection event of a Wi-Fi device could be detected after the passenger's origin transit station. In the same way, the last detection event could be detected before the actual destination. The complications are different for the two Wi-Fi scanner installation modes.

In-vehicle-based monitoring system

The passenger OD matrix $m \times m$ of a transit route can be estimated where m is the number of transit stops on the route. The case of origin transit stop estimation can be described. Suppose that a passenger boards the transit vehicle at stop i , time t_i and the first detection event from the passenger's device is detected at time t_e . The origin transit stop can be incorrectly identified if the transit vehicle arrives the next stop $i + 1$ at time t_{i+1} where $t_{i+1} \geq t_e$. Figure 7.3 shows the time-space diagram of a transit vehicle trajectory with the detection events from a passenger's Wi-Fi device.

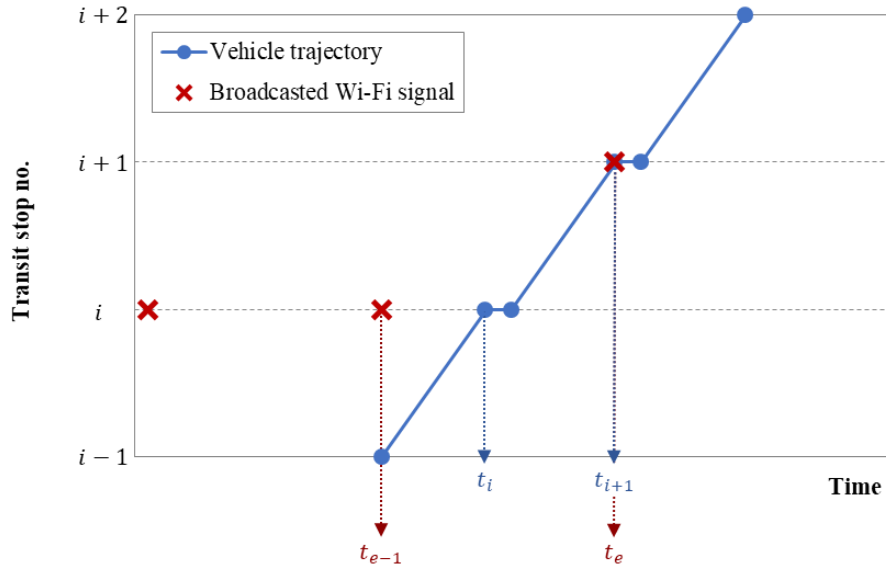


Figure 7.3: A time-space diagram showing a vehicle trajectory and Wi-Fi signals from a passenger's device

This case can occur when the travel time between two stops $t_{i+1} - t_i$ is shorter than the general detection period of the device, $E(t_e - t_{e-1})$. The same logic can be used to describe the incorrect inference of the destination transit stop when the last detection event is captured before the actual destination. Therefore, the probability of incorrect OD estimation is reliant on the distance between two consecutive transit stops on the route and the vehicle speed on the link. Figure 7.4 demonstrates the practical distance between two transit stops for OD estimation based on various average vehicle speeds (30-80 km/hr), and general detection periods (60-180 seconds).

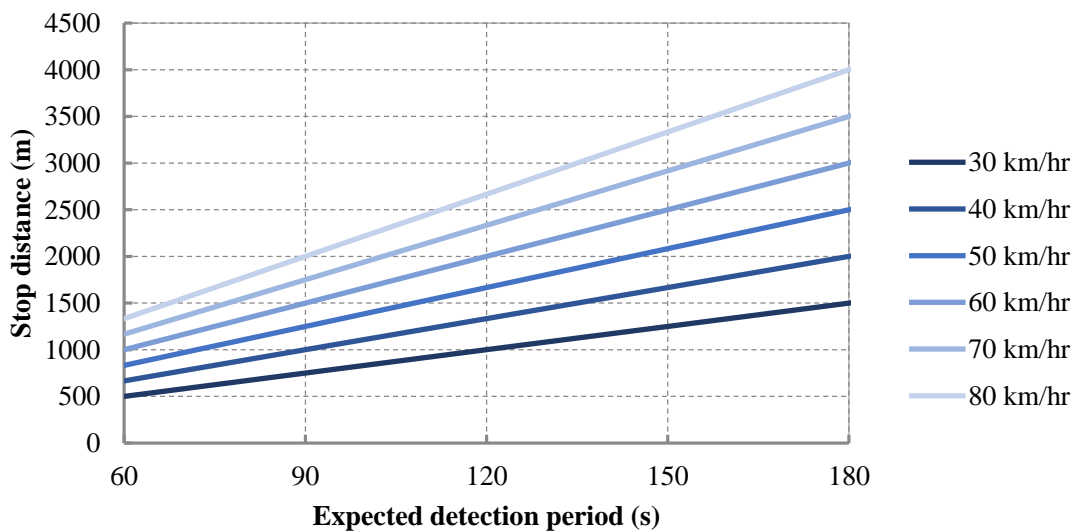


Figure 7.4: Practical distance or spacing between two consecutive transit stops

From the above example of average vehicle speeds and detection periods, it can be noted in Figure 7.4 that the practical stop distance or stop spacing is in a range of 500 meters to 4,000 meters. The results could imply challenges in passenger OD estimation, especially for urban public transit networks such as bus systems in which the distance or spacing between two consecutive stops tends to be shorter than 1 kilometer. However, the estimation accuracy could be promising for inter-city bus systems and rural-area bus networks.

Fixed-location-based monitoring system

There are two types of OD estimation based on the fixed-location installation. First, the Wi-Fi scanners can be installed at each transit station to capture the passenger OD transit stops. In such a case, the accuracy is affected by the capability of the scanners to capture Wi-Fi signals at the OD transit stops before passenger boarding and after passenger alighting. Missed detection can be encountered when passengers only spend a short time duration at the transit stops. Moreover, some detection events may be occasionally detected at the non-OD transit stops during the vehicle dwell time, resulting in the incorrect inference of the passenger OD.

Second, a set of Wi-Fi scanners can be installed in a transit station to capture passenger mobility in the station. Basically, the scanners are installed at significant locations such as exits, gates, and platforms. The challenge of OD estimation is similar to the first case but the chance of missed detection is lower. This is because passenger mobility is based on walking speeds rather than vehicle speeds. Therefore, the system has more time for Wi-Fi device discovery. In addition to the OD estimation, some studies may need to identify passenger travel paths in the station. In the cases that there could be several pedestrian paths between an OD pair, additional Wi-Fi scanners could be necessary for observing the individual paths. The number of required sensors can be determined by solving the sensor location optimization problem. One possible solution is considering the path travel time when the travel time distribution of each path is distinct from the others.

7.2.1.3 Passenger walking time information

Apart from the three significant KPIs in this thesis, passenger walking time is also one of the essential indicators used for bus service evaluation. Passengers may choose other modes of public transport if it takes too much time to walk to and from the bus stops. By installing a number of Wi-Fi scanners in the walking facilities linking to bus stops, the walking times can be estimated from partial data. However, the passenger walking time estimation requires a sufficient sample size for each estimation time interval.

A case study of walking time in a pedestrian crossing tunnel nearby a bus stop can be used as an example. Figure 7.5 shows a location map of the pedestrian tunnel.

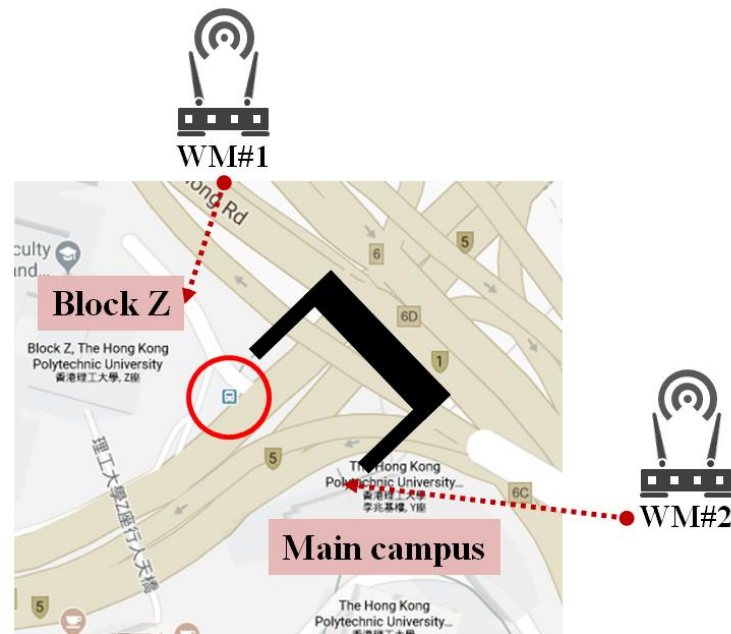


Figure 7.5: A schematic map of the pedestrian tunnel

The tunnel links Block Z building of the Hong Kong Polytechnic University with the main campus and there is a bus stop in front of Block Z. The walking distance is 165 meters. The walking times through the tunnel were estimated from Wi-Fi MAC-IDs captured by two Wi-Fi scanners installed at both ends of the tunnel. Walking time estimation is validated using observed walking times during a 90-minute period in the morning peak time. Figure 7.6 presents the number of MAC-IDs which were re-identified by the two Wi-Fi scanners. Figure 7.7 demonstrates the accuracy of walking time estimation by plotting the estimated walking times against the observed walking times.

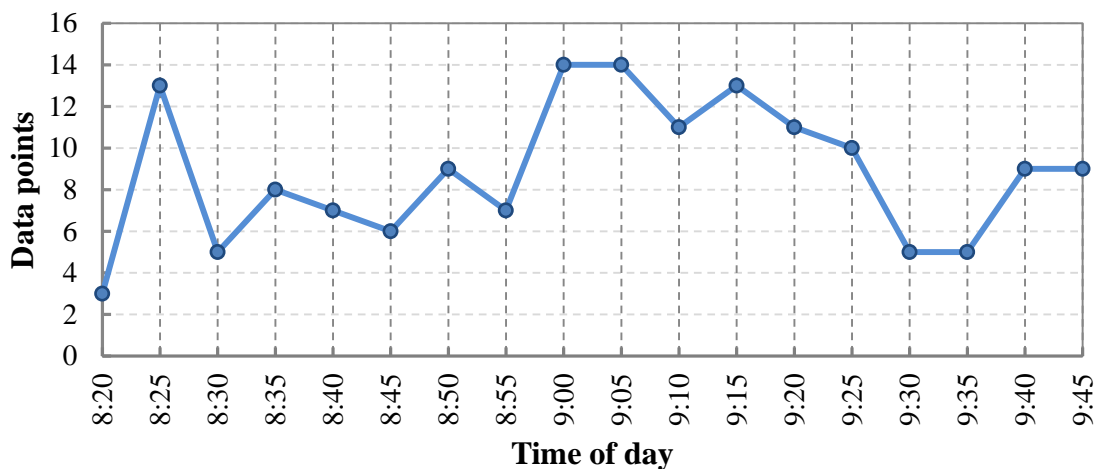


Figure 7.6: Available MAC-IDs for each 5-minute time period

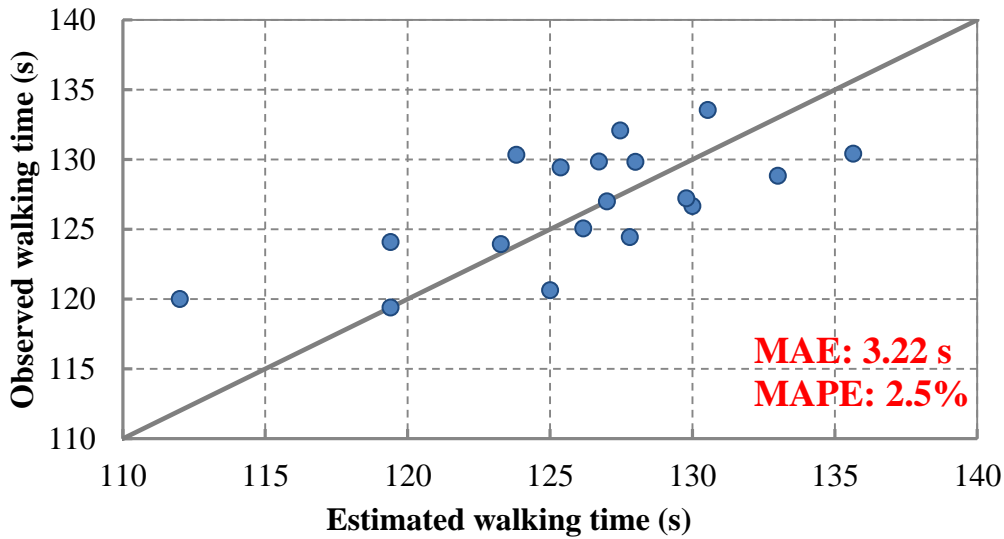


Figure 7.7: Walking time estimation results

7.2.1.4 Research direction

To provide insightful mobility analysis over a public transit network, future works could investigate the possibilities of integrating the use of in-vehicle Wi-Fi scanners on transit vehicles and the use of fixed-location Wi-Fi scanners at transit stops. As an example of the bus mode, the in-vehicle Wi-Fi scanner can only provide the condition of the operating buses such as the bus location and the bus crowd density. In the case that a bus is full, the crowd density information will remain stable during the time of full capacity. However, the bus operator may need to know the condition at bus stops when the operating buses are overcrowded in order to make decisions promptly for better operational management. With the integration, the passenger crowd density can be estimated for both the on-board passengers and the waiting passengers.

Moreover, the integration can be the foundation for passenger behavior studies. It is possible to capture the decision-making behavior of a waiting passenger since identifying the boarding bus is feasible. As a result, the integrated system could trace back to analyze the passenger decision under various conditions e.g. the number of buses that a passenger decided not to board due to the arriving buses being full. At the same time, the information can be used for bus service evaluation such as estimating the number of passengers left behind after a departure.

In terms of public transport information, the performance of passenger counting could be further improved and investigated. For several transportation models, there is a great value in receiving accurate results, including the crowd density on transit vehicles, the crowd density

at transit stations, passengers' OD stations, passenger flow on individual vehicles, boarding and alighting passengers. Due to the uncertainties in Wi-Fi device detection periods, it may not be possible to estimate passenger boarding and alighting times using passive data. Next, there is a limitation of passenger waiting time estimation studies, especially for bus passenger waiting times. A challenging research direction, therefore, is to exploit the passive Wi-Fi data for passenger waiting time estimation. In addition, the estimation results should be evaluated using manual observation of passenger waiting times.

7.2.2 Extension of bus information systems based on crowd-sourced bus data

The accuracy of bus information based on crowd-sourced bus data can be improved by developing prediction models and encouraging more bus passengers to participate. Firstly, advance prediction methods can be applied such as artificial neuron networks and convolutional neuron networks. Such prediction models require sufficient and accurate training data. Therefore, at the initial stage before system deployment, there is a need for collecting high-resolution bus data which cover the entire bus transit network. Moreover, the model development should consider both spatial and temporal uncertainties in bus travel time and/or bus dwell time information. For each time interval, crowd-sourced bus data may provide bus travel times for only some links on the road network and bus dwell times for only some bus stops.

It should be noted in this thesis that the proposed crowd-sourced systems rely on the number of participating passengers. Technology acceptance is one of the key components for the successful development of the proposed public transport information systems. Apart from providing timely and accurate bus information, it is important to design a system that requires only minimal input from the bus passengers and can provide more benefits for them in return. Hence, the factors which could attract more participating passengers should be studied (e.g. user benefits and battery consumption) in order to design an applicable smartphone application. Furthermore, further studies on computational times required for the proposed system should be carried out when more passengers contribute a large volume of bus data. Efficient solution methods should be developed particularly for real-time public transport information systems. With a well-designed application, passengers can provide additional bus information to the system e.g. passenger waiting time and bus crowding level.

Given the diversity of research and development (R&D) in the field of various data mining methods to short-term and long-term transit service planning, the methods presented in this thesis are by no means exhaustive. However, they do provide basic coverage of various important directions of R&D in this field, with particular attention to the estimation of three different KPIs for bus transit information systems to aid transit operation and service

improvements. It is hoped that this thesis will provide a state-of-the-art methodology for the development of multi-modal public transport information systems by both practitioners and researchers, as well as indicating new R&D opportunities and inspiring further efforts in this field. The new approach of using Wi-Fi data for the development of public transport information systems is expected to improve the planning, design, and operation of multi-modal transit systems and thereby increase the efficiency and reliability of transportation networks in our cities.

References

- Abbott-Jard, M., Shah, H., & Bhaskar, A. (2013). *Empirical evaluation of Bluetooth and Wifi scanning for road transport*. Paper presented at the 36th Australasian Transport Research Forum (ATRF). Retrieved from https://www.researchgate.net/profile/Ashish_Bhaskar/publication/290558910_Empirical_evaluation_of_Bluetooth_and_Wifi_scanning_for_road_transport/links/58040b8208ae310e0d9f511b/Empirical-evaluation-of-Bluetooth-and-Wifi-scanning-for-road-transport.pdf
- Abedi, N., Bhaskar, A., & Chung, E. (2014). Tracking spatio-temporal movement of human in terms of space utilization using Media-Access-Control address data. *Applied Geography, 51*, 72-81.
- Abedi, N., Bhaskar, A., Chung, E., & Miska, M. (2015). Assessment of antenna characteristic effects on pedestrian and cyclists travel-time estimation based on Bluetooth and WiFi MAC addresses. *Transportation Research Part C: Emerging Technologies, 60*, 124-141.
- Aguirre, E., Mahr, D., Grewal, D., de Ruyter, K., & Wetzels, M. (2015). Unraveling the personalization paradox: The effect of information collection and trust-building strategies on online advertisement effectiveness. *Journal of Retailing, 91*(1), 34-49.
- Almuhimedi, H., Schaub, F., Sadeh, N., Adjerid, I., Acquisti, A., Gluck, J., ... Agarwal, Y. (2015). Your location has been shared 5,398 times!: a field study on mobile app privacy nudging. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 787-796.
- Amin-Naseri, M. R., & Baradaran, V. (2014). Accurate estimation of average waiting time in public transportation systems. *Transportation Sciences, 49*(2), 213-222.
- Araghi, B. N., Olesen, J. H., Krishnan, R., Christensen, L. T., & Lahrmann, H. (2015). Reliability of Bluetooth technology for travel time estimation. *Journal of Intelligent Transportation Systems, 19*(3), 240-255.
- Astarita, V., Bertini, R. L., d'Elia, S., & Guido, G. (2006). Motorway traffic parameter estimation from mobile phone counts. *European Journal of Operational Research, 175*(3), 1435-1446.

- Barcelö, J., Montero, L., Marqués, L., & Carmona, C. (2010). Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring. *Transportation Research Record*, 2175(1), 19-27.
- Bar-Gera, H. (2007). Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. *Transportation Research Part C: Emerging Technologies*, 15(6), 380-391.
- Batarce, M., Munoz, J. C., & Ortuzar, J. (2016). Valuing crowding in public transport: implications for cost-benefit analysis. *Transportation Research Part A : Policy and Practice*, 91, 358-378.
- Battelle Transportation Division. (1997). *Global Positioning Systems for personal travel surveys, Lexington Area travel data collection test*. Retrieved from <https://www.fhwa.dot.gov/ohim/lextrav.pdf>.
- Battersby, L. (2013, August 24). Tracked from the moment you wake. *The Sydney Morning Herald*. Retrieved from <https://www.smh.com.au/technology/tracked-from-the-moment-you-wake-20130824-2shwq.html>.
- Bedogni, L., Felice, M. D., & Bononi, L. (2016). Context-aware Android applications through transportation mode detection techniques. *Wireless Communications and Mobile Computing*, 16(16), 2523–2541.
- Bhaskar, A., & Chung, E. (2013). Fundamental understanding on the use of Bluetooth scanner as a complementary transport data. *Transportation Research Part C: Emerging Technologies*, 37, 42-72.
- Bhaskar, A., Qu, M., & Chung, E. (2014). Bluetooth vehicle trajectory by fusing Bluetooth and loops: Motorway travel time statistics. *IEEE Transactions on Intelligent Transportation Systems*, 16(1), 113-122.
- Bhaskar, A., Qu, M., Nantes, A., Miska, M. & Chung, E., (2015). Is bus overrepresented in Bluetooth MAC scanner data? Is MAC-ID really unique?. *International Journal of Intelligent Transportation Systems Research*, 13(2), 119-130.
- Biagioni, J., Gerlich, T., Merrifield, T., & Eriksson, J. (2011). Easytracker: automatic transit tracking, mapping, and arrival time prediction using smartphones. *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, 68-81.
- Bin, Y., Zhongzhen, Y., & Baozhen, Y. (2006). Bus arrival time prediction using support vector machines. *Journal of Intelligent Transportation Systems*, 10(4), 151–158.
- Birle, C., and Wermuth, M. (2006). *The traffic online project*. Paper presented at the 13th ITS World Congress, Special Session: Cellular-based Traffic Data Collection. London, UK, October 8-12.
- Bowman, L. A., & Turnquist, M. A. (1981). Service frequency, schedule reliability and passenger wait times at transit stops. *Transportation Research Part A: General*, 15(6), 465-471.

- Broach, J., Dill, J., & McNeil, N. W. (2019). Travel mode imputation using GPS and accelerometer data from a multi-day travel survey. *Journal of Transport Geography*, 78, 194-204.
- Buehler, R. (2009). Promoting public transportation: Comparison of passengers and policies in Germany and the United States. *Transportation Research Record*, 2110(1), 60-68.
- Bulut, M. F., Demirbas, M., & Ferhatosmanoglu, H. (2014). Lineking: Coffee shop wait-time monitoring using smartphones. *IEEE Transactions on Mobile Computing*, 14(10), 2045-2058.
- Caceres, N., Wideberg, J. P., & Benitez, F. G. (2007). Deriving origin-destination data from a mobile phone network. *IET Intelligent Transport Systems*, 1(1), 15-26.
- Cats, O., & Loutos, G., 2015. Real-time bus arrival information system: an empirical evaluation. *Journal of Intelligent Transportation Systems*, 20(2), 138-151.
- Chang, H., Park, D., Lee, S., Lee, H., & Baek, S. (2010). Dynamic multi-interval bus travel time prediction using bus transit data. *Transportmetrica*, 6(1), 19–38.
- Chen, B. Y., Yuan, H., Li, Q., Lam, W. H. K., Shaw, S. L., & Yan, K. (2014). Map-matching algorithm for large-scale low-frequency floating car data. *International Journal of Geographical Information Science*, 28(1), 22-38.
- Chen, M., Yaw, J., Chien, S. I., & Liu, X. (2007). Using automatic passenger counter data in bus arrival time prediction. *Journal of Advanced Transportation*, 41(3), 267-283.
- Chernyshev, M., Valli, C., & Hannay, P. (2015). On 802.11 access point locatability and named entity recognition in service set identifiers. *IEEE Transactions on Information Forensics and Security*, 11(3), 584-593.
- Chien, S. I. J., & Kuchipudi, C. M. (2003). Dynamic travel time prediction with real-time and historic data. *Journal of Transportation Engineering*, 129(6), 608–616.
- Chung, E., & Kuwahara, M. (2007). Mapping personal trip OD from probe data. *International Journal of Intelligent Transportation Systems Research*, 5(1), 1-6.
- Cunche, M. (2014). I know your MAC Address: Targeted tracking of individual using Wi-Fi. *Journal of Computer Virology and Hacking Techniques*, 10(4), 219-227.
- Danalet, A. (2015). *Activity choice modeling for pedestrian facilities*. (Doctoral dissertation, École Polytechnique Fédérale de Lausanne). Retrieved from <http://dx.doi.org/10.5075/epfl-thesis-6806>
- Danalet, A., Farooq, B., & Bierlaire, M. (2014). A Bayesian approach to detect pedestrian destination-sequences from WiFi signatures. *Transportation Research Part C: Emerging Technologies*, 44, 146-170.
- Das, S., & Pandit, D. (2013). Importance of user perception in evaluating level of service for bus transit for a developing country like India: a review. *Transport Reviews*, 33(4), 402–420.

- Das, S., & Pandit, D. (2015). Determination of level-of-service scale values for quantitative bus transit service attributes based on user perception. *Transportmetrica A: Transport Science*, *11*(1), 1-21.
- Delafontaine, M., Versichele, M., Neutens, T., & Van de Weghe, N. (2012). Analysing spatiotemporal sequences in Bluetooth tracking data. *Applied Geography*, *34*, 659-668.
- Dell'Olio, L., Ibeas, A., & Cecin, P. (2011). The quality of service desired by public transport users. *Transport Policy*, *18*(1), 217-227.
- Fan, Y., Guthrie, A., & Levinson, D. (2016). Waiting time perceptions at transit stops and stations: Effects of basic amenities, gender, and security. *Transportation Research Part A: Policy and Practice*, *88*, 251-264.
- Fang, S. H., Fei, Y. X., Xu, Z., & Tsao, Y. (2017). Learning transportation modes from smartphone sensors based on deep neural network. *IEEE Sensors Journal*, *17*(18), 6111-6118.
- Fang, S. H., Liao, H. H., Fei, Y. X., Chen, K. H., Huang, J. W., Lu, Y. D., & Tsao, Y. (2016). Transportation modes classification using sensors on smartphones. *Sensors*, *16*(8), 1324.
- Feng, T., & Timmermans, H. J. (2013). Transportation mode recognition using GPS and accelerometer data. *Transportation Research Part C: Emerging Technologies*, *37*, 118-130.
- Fernández, R. (2010). Modelling public transport stops by microscopic simulation. *Transportation Research Part C: Emerging Technologies*, *18*(6), 856-868.
- Fernández, R., Zegers, P., Weber, G., & Tyler, N. (2010). Influence of platform height, door width, and fare collection on bus dwell time: laboratory evidence for Santiago de Chile. *Transportation Research Record*, *2143*(1), 59-66.
- Fletcher, G., & El-Geneidy, A. (2013). Effects of fare payment types and crowding on dwell time: fine-grained analysis. *Transportation Research Record*, *2351*(1), 124-132.
- Freudiger, J. (2015). How talkative is your mobile device? an experimental study of Wi-Fi probe requests. *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, 1-6.
- Frumin, M., and Zhao, J. (2012). Analyzing passenger incidence behavior in heterogeneous transit services using smartcard data and schedule-based assignment. *Transportation Research Record*, *2274*(1), 52-60.
- Garcia-Villalonga, S., & Perez-Navarro, A. (2015). Influence of human absorption of Wi-Fi signal in indoor positioning with Wi-Fi fingerprinting. *2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 1-10.
- Guihaire, V., & Hao, J. K. (2008). Transit network design and scheduling: A global review. *Transportation Research Part A: Policy and Practice*, *42*(10), 1251-1273.

- Guo, B., Wang, Z., Yu, Z., Wang, Y., Yen, N. Y., Huang, R., & Zhou, X. (2015). Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm. *ACM Computing Surveys (CSUR)*, 48(1), 1-31.
- Guo, S., Yu, L., Chen, X., & Zhang, Y. (2011). Modelling waiting time for passengers transferring from rail to buses. *Transportation Planning and Technology*, 34(8), 795–809.
- Haghani, A., Hamed, M., Sadabadi, K. F., Young, S., & Tarnoff, P. (2010). Data collection of freeway travel time ground truth with bluetooth sensors. *Transportation Research Record*, 2160(1), 60-68.
- Handte, M., Iqbal, M. U., Wagner, S., Apolinarski, W., Marrón, P. J., Navarro, E. M., ... Fernández, M. G. (2014). Crowd density estimation for public transport vehicles. *Proceedings of the EDBT/ICDT 2014 Joint Conference*, 315-322.
- Holroyd, E. M., & Scraggs, D. A. (1966). Waiting times for buses in central London. *Traffic Engineering and Control*, 8(3), 158-160.
- Hoseini-Tabatabaei, S. A., Gluhak, A., & Tafazolli, R. (2013). A survey on smartphone-based systems for opportunistic user context recognition. *ACM Computing Surveys (CSUR)*, 45(3), 1-51.
- Hsu, S. C. (2010). Determinants of passenger transfer waiting time at multi-modal connecting stations. *Transportation Research Part E: Logistics and Transportation Review*, 46(3), 404–413.
- Hu, X., Song, L., Van Bruggen, D., & Striegel, A. (2015). Is there wifi yet? How aggressive wifi probe requests deteriorate energy and throughput. *Proceedings of the 2015 Internet Measurement Conference*, 317-323.
- Ingvardson, J. B., Nielsen, O. A., Raveau, S., & Nielsen, B. F. (2018). Passenger arrival and waiting time distributions dependent on train service frequency and station characteristics: A smart card data analysis. *Transportation Research Part C: Emerging Technologies*, 90, 292-306.
- Jagadeesh, G. R., Srikanthan, T., & Zhang, X. D. (2004). A map matching method for GPS based real-time vehicle location. *Journal of Navigation*, 57(3), 429-440.
- Jahangiri, A., & Rakha, H. A. (2015). Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. *IEEE Transactions on Intelligent Transportation Systems*, 16(5), 2406-2417.
- Jason Chang, S. K., & Hsu, C. L. (2001). Modeling passenger waiting time for intermodal transit stations. *Transportation Research Record*, 1753(1), 69-75.
- Jeong, R. H. (2005). *The prediction of bus arrival time using automatic vehicle location systems data* (Doctoral dissertation, Texas A&M University). Retrieved from <https://oaktrust.library.tamu.edu/bitstream/handle/1969.1/1458/etd-tamu-2004C-CVEN-Jeong.pdf?sequence=1&isAllowed=y>

- Jolliffe, K., & Hutchinson, T. P. (1975). A behavioural explanation of the association between bus and passenger arrivals at a bus stop. *Transportation Science*, 9(3), 248-282.
- Ketelaar, P. E., & van Balen, M. (2018). The smartphone as your follower: The role of smartphone literacy in the relation between privacy concerns, attitude and behaviour towards phone-embedded tracking. *Computers in Human Behavior*, 78, 174-182.
- Kostakos, V. (2008). Using Bluetooth to capture passenger trips on public transport buses. arXiv:0806.0874. Retrieved from https://www.researchgate.net/profile/Vassilis_Kostakos/publication/220484433_Using_Bluetooth_to_capture_passenger_trips_on_public_transport_buses/links/0046352266cbc780b7000000/Using-Bluetooth-to-capture-passenger-trips-on-public-transport-buses.pdf
- Kostakos, V., O'Neill, E., Penn, A., Roussos, G., & Papadongonas, D. (2010). Brief encounters: sensing, modeling and visualizing urban mobility and copresence networks. *ACT Transaction on Computer-Human Interaction (TOCHI)*, 17(1), 2.
- Laharotte, P. A., Billot, R., Come, E., Oukhellou, L., Nantes, A., & Faouzi, N. E. (2015). Spatiotemporal analysis of Bluetooth data: Application to a large urban network. *IEEE Transactions on Intelligent Transportation Systems*, 16(3), 1439–1448.
- Lam, W. H. K., & Morrall, J. (1982). Bus passenger walking distances and waiting times: a summer-winter comparison. *Traffic Quarterly*, 36(3), 407-421.
- Le, T. V., Song, B., & Wynter, L. (2017). Real-time prediction of length of stay using passive Wi-Fi sensing. *2017 IEEE International Conference on Communications (ICC)*, 1-6.
- Lee, G., & Yim, J. (2014). Design of an Android Real-Time Bus Location Provider. *Life Science Journal*, 11(7), 619-625.
- Li, M. T., Zhao, F., Chow, L. F., Zhang, H., & Li, S. C. (2006). Simulation model for estimating bus dwell time by simultaneously considering numbers of disembarking and boarding passengers. *Transportation Research Record*, 1971(1), 59-65.
- Li, J., Gao, J., Yang, Y., & Wei, H. (2017). Bus arrival time prediction based on mixed model. *China Communications*, 14(5), 38–47.
- Li, Z., & Hensher, D. A. (2013). Crowding in public transport: a review of objective and subjective measures. *Journal of Public Transportation*, 16(2), 107-134.
- Lin, W. H., & Zeng, J. (1999). Experimental study of real-time bus arrival time prediction with GPS data. *Transportation Research Record*, 1666(1), 101-109.
- Liu, W., Liu, J., Jiang, H., Xu, B., Lin, H., Jiang, G., & Xing, J. (2016). WiLocator: WiFi-sensing based real-time bus tracking and arrival time prediction in urban environments. *2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*, 529-538.
- Liu, Y., Bunker, J., & Ferreira, L. (2010). Transit users' route choice modelling in transit assignment: A review. *Transport Reviews*, 30(6), 753-769.

- Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., & Huang, Y. (2009). Map-matching for low-sampling-rate GPS trajectories. *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 352-361.
- Luethi, M., Weidmann, U., & Nash, A. (2007). *Passenger arrival rates at public transport stations*. Paper presented at the TRB 86th Annual Meeting Compendium of Papers. Transportation Research Board. Washington DC, United States. Retrieved from <https://www.research-collection.ethz.ch/handle/20.500.11850/64815>
- Lui, G., Gallagher, T., Li, B., Dempster, A. G., & Rizos, C. (2011). Differences in RSSI readings made by different Wi-Fi chipsets: A limitation of WLAN localization. *2011 International Conference on Localization and GNSS (ICL-GNSS)*, 53-57.
- Malinovskiy, Y., Saunier, N., & Wang, Y. (2012). Analysis of pedestrian travel with static Bluetooth sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 2299(1), 137-149.
- Manweiler, J., Santhapuri, N., Choudhury, R. R., & Nelakuditi, S. (2013). Predicting length of stay at wifi hotspots. *2013 Proceedings IEEE INFOCOM*, 3102-3110.
- Marguier, P. H., & Ceder, A. (1984). Passenger waiting strategies for overlapping bus routes. *Transportation Science*, 18(3), 207-230.
- Martchouk, M., Mannering, F., & Bullock, D. (2010). Analysis of freeway travel time variability using Bluetooth detection. *Journal of Transportation Engineering*, 137(10), 697-704.
- Martin, J., Mayberry, T., Donahue, C., Foppe, L., Brown, L., Riggins, C., ... Brown, D. (2017). A study of MAC address randomization in mobile devices and when it fails. *Proceedings on Privacy Enhancing Technologies*, 2017(4), 365-383.
- Matte, C., Cunche, M., Rousseau, F., & Vanhoef, M. (2016). Defeating MAC address randomization through timing attacks. *Proceedings of the 9th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 15-20.
- McLeod, F. (2007). Estimating bus passenger waiting times from incomplete bus arrivals data. *Journal of the Operational Research Society*, 58(11), 1518-1525.
- Meng, Q., & Qu, X. (2013). Bus dwell time estimation at bus bays: A probabilistic approach. *Transportation Research Part C: Emerging Technologies*, 36, 61-71.
- Michau, G., Nantes, A., Bhaskar, A., Chung, E., Abry, P., & Borgnat, P. (2017). Bluetooth data in an urban context: Retrieving vehicle trajectories. *IEEE Transactions on Intelligent Transportation Systems*, 18(9), 2377-2386.
- Miller, J. (2013), August 12. City of London calls halt to smartphone tracking bins. *BBC News*, Retrieved from <https://www.bbc.com/news/technology-23665490>.
- Miluzzo, E., Lane, N. D., Fodor, K., Peterson, R., Lu, H., Musolesi, M., ..., Campbell, A. T. (2008). Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application. *Proceedings of the 6th ACM conference on Embedded Network Sensor Systems*, 337-350.

- Mishalani, R. G., McCord, M. M., & Wirtz, J. (2006). Passenger wait time perceptions at bus stops: Empirical results and impact on evaluating real-time bus arrival information. *Journal of Public Transportation*, 9(2), 89-106.
- Mori, U., Mendiburu, A., Álvarez, M., & Lozano, J. A. (2015). A review of travel time estimation and forecasting for Advanced Traveller Information Systems. *Transportmetrica A: Transport Science*, 11(2), 119-157.
- Musa, A. B., & Eriksson, J. (2012). Tracking unmodified smartphones using wi-fi monitors. *Proceedings of the 10th ACM conference on embedded network sensor systems*, 281-294.
- Myrvoll, T. A., Håkegård, J. E., Matsui, T., & Septier, F. (2017). Counting public transport passenger using WiFi signatures of mobile devices. *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 1-6.
- Nakatani, T., Maekawa, T., Shirakawa, M., & Hara, T. (2018). Estimating the physical distance between two locations with Wi-Fi received signal strength information using obstacle-aware approach. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3), 130.
- Nunes, N., Ribeiro, M., Prandi, C., & Nisi, V. (2017). Beanstalk: a community based passive Wi-Fi tracking system for analysing tourism dynamics. *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, 93-98.
- O'flaherty, C. A., & Mancan, D. O. (1970). Bus passenger waiting times in central areas. *Traffic Engineering and Control*, 11(9), 419-421.
- O'Malley, J. (2019), May 22. TfL is going to track all London Underground users using Wi-Fi. *Wired*. Retrieved from <https://www.wired.co.uk/article/london-underground-wifi-tracking>.
- Oransirikul, T., Nishide, R., Piumarta, I., & Takada, H. (2016). Feasibility of analyzing Wi-Fi activity to estimate transit passenger population, *2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, 362-369.
- Oransirikul, T., Piumarta, I., & Takada, H. (2019). Classifying passenger and non-passenger signals in public transportation by analysing mobile device Wi-Fi activity. *Journal of Information Processing*, 27, 25-32.
- Osuna, E. E., & Newell, G. F. (1972). Control strategies for an idealized public transportation system. *Transportation Science*, 6(1), 52-72.
- Porter, J. D., Kim, D. S., Magaña, M. E., Poocharoen, P., & Arriaga, C. A. G. (2013). Antenna characterization for Bluetooth-based travel time data collection. *Journal of Intelligent Transportation Systems*, 17(2), 142-151.
- Prentow, T. S., Ruiz-Ruiz, A. J., Blunck, H., Stisen, A., & Kjærgaard, M. B. (2015). Spatio-temporal facility utilization analysis from exhaustive wifi monitoring. *Pervasive and Mobile Computing*, 16, 305-316.

- Qin, W., Zhang, J., Li, B., & Sun, L. (2013). Discovering human presence activities with smartphones using nonintrusive wi-fi sniffer sensors: the big data prospective. *International Journal of Distributed Sensor Networks*, 9(12), 927-940.
- Ratti, C., Frenchman, D., Pulselli, R. M., & Williams, S. (2006). Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5), 727-748.
- Redman, L., Friman, M., Gärling, T., & Hartig, T. (2013). Quality attributes of public transport that attract car users: A research review. *Transport Policy*, 25, 119-127.
- Schauer, L., & Linnhoff-Popien, C. (2017). Extracting context information from Wi-Fi captures. *Proceedings of the 10th International Conference on Pervasive Technologies Related to Assistive Environments*, 128530, 123-130.
- Seddon, P. A., & Day, M. P. (1974). Bus passenger waiting times in Greater Manchester. *Traffic Engineering and Control*, 15(9), 442-445.
- Shafique, M. A., & Hato, E. (2015). Use of acceleration data for transportation mode prediction. *Transportation*, 42(1), 163-188.
- Shalaby, A., & Farhan, A. (2004). Prediction model of bus arrival and departure times using AVL and APC data. *Journal of Public Transportation*, 7(1), 41-61.
- Shen, J., Jiang, H., Yang, F., & Yao, Z. (2019). Trip mode recognition using smartphone sensor data under different sampling frequencies. *Web Intelligence*, 17(2), 151-160.
- Shen, L., & Stopher, P. R. (2014). Review of GPS travel survey and GPS data-processing methods. *Transport Reviews*, 34(3), 316-334.
- Shin, D., Aliaga, D., Tunçer, B., Arisona, S. M., Kim, S., Zünd, D., & Schmitt, G. (2015). Urban sensing: Using smartphones for transportation mode classification. *Computers, Environment and Urban Systems*, 53, 76-86.
- Shlayan, N., Kurkcu, A., & Ozbay, K. (2016). Exploring pedestrian Bluetooth and WiFi detection at public transportation terminals. *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 229-234.
- Shu, H., Song, C., Pei, T., Xu, L., Ou, Y., Zhang, L., & Li, T. (2016). Queuing time prediction using WiFi positioning data in an indoor scenario. *Sensors*, 16(11), 1958.
- Shubber, K. (2013, August 9). Tracking devices hidden in London's recycling bins are stalking your smartphone. *Wired*. Retrieved from <https://www.wired.co.uk/article/recycling-bins-are-watching-you>.
- Siirtola, P., & Röning, J. (2012). Recognizing human activities user-independently on smartphones based on accelerometer data. *International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI)*, 1(5), 38-45.
- Song, B., & Wynter, L. (2017). *Real-time public transport service-level monitoring using passive WiFi: A spectral clustering approach for train timetable estimation*. Retrieved from *arXiv preprint arXiv:1703.00759*.

- Su, X., Caceres, H., Tong, H., & He, Q. (2016). Online travel mode identification using smartphones with battery saving considerations. *IEEE Transaction on Intelligent Transportation Systems*, 17(10), 2921–2934.
- Transport and Housing Bureau (2017). *Public transport strategy study*. Retrieved from https://www.td.gov.hk/filemanager/en/publication/ptss_final_report_eng.pdf.
- Tettamanti, T., Demeter, H., & Varga, I. (2012). Route choice estimation based on cellular signaling data. *Acta Polytechnica Hungarica*, 9(4), 207-220.
- The Ericsson Company. (2019), *Ericsson mobility report* [PDF file].. Retrieved from <https://www.ericsson.com/49d1d9/assets/local/mobility-report/documents/2019/ericsson-mobility-report-june-2019.pdf>.
- Tirachini, A. (2013). Bus dwell time: the effect of different fare collection systems, bus floor level and age of passengers. *Transportmetrica A: Transport Science*, 9(1), 28-4.
- Tirachini, A., Hensher, D. A., & Rose, J. M. (2013). Crowding in public transport systems: effects on users, operation and implications for the estimation of demand. *Transportation Research Part A: Policy and Practice*, 53, 36-52.
- Tirachini, A., Hurtubia, R., Dekker, T., & Daziano, R. A. (2017). Estimation of crowding discomfort in public transport: results from Santiago de Chile. *Transportation Research Part A: Policy and Practice*, 103, 311-326.
- Transport for London. (2017). *Review of the TfL WiFi pilot*. Retrieved from <http://content.tfl.gov.uk/review-tfl-wifi-pilot.pdf>.
- Turnquist, M. A., (1978). A model for investigating the effects of service frequency and reliability on bus passenger waiting times. *Transportation Research Record*, 663, 70-73.
- U.S. Department of Transportation. (2006). *Advanced public transportation systems: the state of the art update 2006*. (No. FTA-NJ-26-7062-06.1). Retrieved from <https://rosap.ntl.bts.gov/view/dot/16448>
- Vanajakshi, L., Subramanian, S. C., & Sivanandan, R. (2009). Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses. *IET Intelligent Transport Systems*, 3(1), 1-9.
- Vanhoef, M., Matte, C., Cunche, M., Cardoso, L. S., & Piessens, F. (2016). Why MAC address randomization is not enough: An analysis of Wi-Fi network discovery mechanisms. *Proceedings of the 11th ACM on Asia Conference on Computer and Communication*, 413-424.
- Vanitchakornpong, K., Indra-Payoong, N., & Sumalee, A. (2013). Siamtraffic2.0: Traffic pattern search for travel time prediction in Bangkok road network. *Journal of Information Science and Technology*, 4(1), 1-10.
- Velayos, H., & Karlsson, G. (2003). Techniques to reduce IEEE 802.11 b MAC layer handover time. Retrieved from

- <https://cs.uwaterloo.ca/~brecht/courses/856-802.11-Network-Performance-2014/readings/handoff/reduce-handoff-time.pdf>
- Versichele, M., Neutens, T., Delafontaine, M., & Van de Weghe, N. (2012). The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: A case study of the Ghent festivities. *Applied Geography*, 32(2), 208-220.
- Waltari, O., & Kangasharju, J. (2016). The wireless shark: Identifying wifi devices based on probe fingerprints. *Proceedings of the First Workshop on Mobile Data*. Singapore, 1-6.
- Wang, J., Indra-Payoong, N., Sumalee, A., & Panwai, S. (2013). Vehicle reidentification with self-adaptive time windows for real-time travel time estimation. *IEEE Transactions on Intelligent Transportation Systems*, 15(2), 540-552.
- Wang, W., Chen, J., Hong, T., & Zhu, N. (2018). Occupancy prediction through Markov based feedback recurrent neural network (M-FRNN) algorithm with WiFi probe technology. *Building and Environment*, 138, 160-170.
- Wang, Y., Yang, J., Chen, Y., Liu, H., Gruteser, M., & Martin, R. P. (2014). *Tracking human queues using single-point signal monitoring*. *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, 42-54.
- Watkins, K. E., Ferris, B., Borning, A., Rutherford, G. S., and Layton, D. (2011). Where is my bus? Impact of mobile real-time information on the perceived and actual wait time of transit riders. *Transportation Research Part A: Policy and Practice*, 45(8), 839-848.
- Watts, M., Brunger, J., & Shires, K. (2011). Do European data protection laws apply to the collection of WiFi network data for use in geolocation look-up services?. *International Data Privacy Law*, 1(3), 149-160.
- Welding, P. I. (1957). The instability of a close-interval service. *Journal of the Operational Research Society*, 8(3), 133-142.
- Weppner, J., & Lukowicz, P. (2013). Bluetooth based collaborative crowd density estimation with mobile phones. *2013 IEEE International conference on pervasive computing and communications (PerCom)*, 193-200.
- White, J., & Wells, I. (2002). Extracting origin destination information from mobile phone data. *Proceedings of the 11th International Conference on Road Transport Information and Control*, 30-34.
- Wilkinson, G. (2014). *Digital terrestrial tracking: The future of surveillance*. Paper presented at the DEF CON 22. Las Vegas, Nevada, USA. Retrieved from <http://bofh.nikhef.nl/events/defcon/DEF%20CON%2022/DEF%20CON%2022%20presentations/Glenn%20Wilkinson%20-%20Updated/DEFCON-22-Glenn-Wilkinson-GRW-WP.pdf>
- Xu, L., Yang, F., Jiang, Y., Zhang, L., Feng, C., & Bao, N. (2011). Variation of received signal strength in wireless sensor network. *2011 the 3rd International Conference on Advanced Computer Control*, 151-154.

- Yan, C., Wang, P., Pang, H., Sun, L., & Yang, S. (2017). *Proceedings of the International Conference Multimedia Modelling*, 503-514.
- Yang, J. (2009). Toward physical activity diary: motion recognition using simple acceleration features with mobile phones. *Proceedings of the 1st International Workshop on Interactive Multimedia for Consumer Electronics*, 1-10.
- Yap, M., Cats, O., & van Arem, B. (2018). Crowding valuation in urban tram and bus transportation based on smart card data. *Transportmetrica A: Transport Science*, 16(1), 23-42.
- Yoo, S., Shin, Y., Kim, S., & Choi, S. (2014). Toward realistic WiFi simulation with smartphone "Physics". *Proceedings of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks 2014*, 1-6.
- Yu, B., Lam, W. H. K., & Tam, M. L. (2011). Bus arrival time prediction at bus stop with multiple routes. *Transportation Research Part C: Emerging Technologies*, 19(6), 1157-1170.
- Yu, B., Wang, H., Shan, W., & Yao, B. (2017). Prediction of bus travel time using random Forests based on near neighbors. *Computer-Aided Civil and Infrastructure Engineering*, 33(4), 333-350.
- Yu, B., Yang, Z., Chen, K., & Yu, B. (2010). Hybrid model for prediction of bus arrival times at next station. *Journal of Advanced Transportation*, 44(3), 193– 204.
- Yu, M. C., Yu, T., Wang, S. C., Lin, C. J., & Chang, E. Y. (2014). Big data small footprint: the design of a low-power classifier for detecting transportation modes. *Proceedings of the VLDB Endowment*, 7(13), 1429-1440.
- Zeng, Y., Pathak, P. P., & Mohapatra, P. (2015). Analyzing shopper's behavior through wifi signals. *Proceedings of the 2nd Workshop on Workshop on Physical Analytics*, 13-18.
- Zhang, J., Shen, D., Tu, L., Zhang, F., Xu, C., Wang, Y., ..., Li, Z. (2017). A real-time passenger flow estimation and prediction method for urban bus transit systems. *IEEE Transactions on Intelligent Transportation Systems*, 18(11), 3168-3178.
- Zhang, R., Liu, W., Jia, Y., Jiang, G., Xing, J., Jiang, H., & Liu, J. (2018). WiFi sensing-based real-time bus tracking and arrival time prediction in urban environments. *IEEE Sensors Journal*, 18(11), 4746–4760.
- Zhang, Y., Qin, X., Dong, S., & Ran, B. (2010). *Daily OD matrix estimation using cellular probe data*. Paper presented at the 89th Annual Meeting Transportation Research Board, 9. Washington DC, USA. Retrieved from <https://pdfs.semanticscholar.org/1fdc/79bcbfa16c2960025293fb760b5d919201d3.pdf>
- Zimmerman, J., Tomasic, A., Garrod, C., Yoo, D., Hiruncharoenvate, C., Aziz, R., ... Steinfeld, A. (2011). Field trial of tiramisu: crowd-sourcing bus arrival times to spur co-design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1677-1686.

Appendix A

Wi-Fi device discovery

A.1 Principle of Wi-Fi technology

Wireless Fidelity (Wi-Fi) is the name given to the technology used for wireless local area network connections according to the IEEE¹ 802.11 standard. Two radio frequency bands are generally used for Wi-Fi communication, including 2.4 GHz and 5.0 GHz spectrums. Each band is divided into multiple channels so as to enhance the communication performance and to facilitate Wi-Fi communication for several networks. The number of channels is different for each country depending on the countrywide allocation of radio spectrums.

To facilitate the communication of Wi-Fi devices, a set of 802.11 frames is specified. The frames are also recognized as Wi-Fi packets which can be categorized into three types based on the communication purposes: management frames, control frames, and data frames. Each packet consists of essential details for the communication. One of the most significant components of the data is the MAC-ID of the transmitting device. A MAC-ID is considered as a unique identifier of an individual Wi-Fi device. To be more specific, a 48-bit MAC-ID is assigned to a network interface controller (NIC) of an individual device. In a 48-bit MAC-ID, the first three octets are the organizationally unique identifier (OUI) of the NIC's manufacturer. The OUIs are normally regulated by the IEEE to ensure their global uniqueness. Then, the last three octets are assigned by the manufacturer based on the uniqueness constraint. Figure A.1 demonstrates an example of a 48-bit MAC-ID.

¹ Institute of Electrical and Electronics Engineers

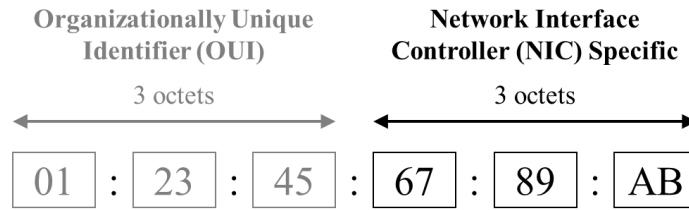


Figure A.1: A 48-bit MAC address

A Wi-Fi device requires a connection to be established for communicating with other devices. The connection can be based on the ad hoc mode or the infrastructure mode. In the ad hoc mode, Wi-Fi devices are connected in the form of a peer-to-peer network which allows communication between devices within each other's range. A Wi-Fi connection in the infrastructure mode is more general due to its capability of supporting more connections. In addition, Wi-Fi devices can be connected to the Internet apart from communicating with each other. To do so, an access point (AP) linking the device to the Internet is necessary. The Wi-Fi devices need to connect with an AP within their coverage area. Figure A.2 illustrates the two modes of Wi-Fi connection.

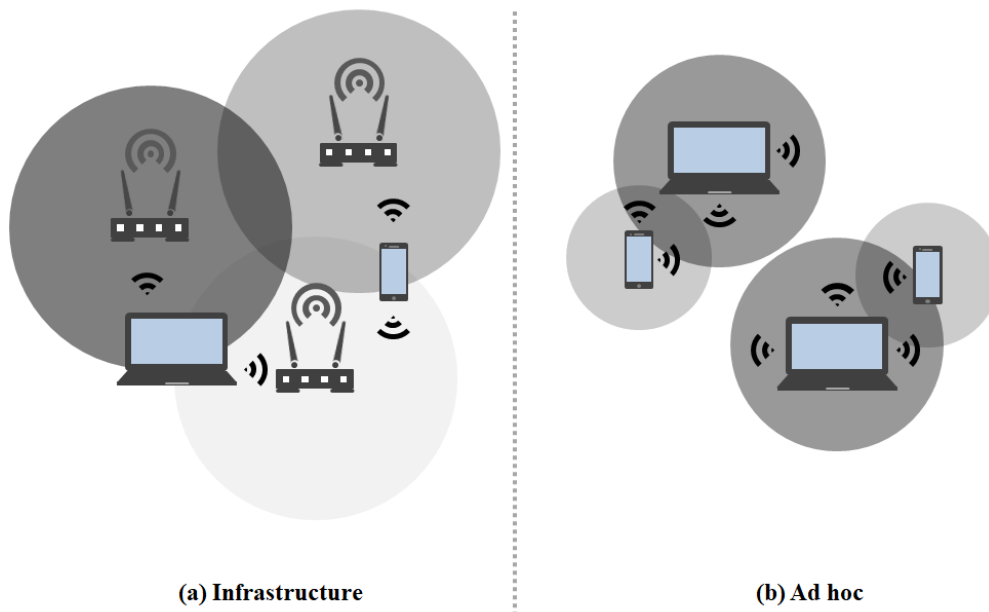


Figure A.2: Wi-Fi connection in (a) an infrastructure mode, and (b) an ad hoc mode

To connect to a wireless network, a Wi-Fi device first needs to search for the nearby available networks. Two types of service discovery are defined in the IEEE 802.11 standard: passive scanning and active scanning. During passive scanning, a Wi-Fi device listens to each Wi-Fi channel for the beacon frames broadcast by the nearby APs. Here, an AP is responsible for broadcasting beacon (BEA) packets at regular intervals in order to declare its presence

and provide the information of the wireless network. The BEAs are broadcast in approximately 100 milliseconds by default (Velayos and Karlsson, 2003).

During active scanning, a Wi-Fi device broadcasts probe request (PRQ) packets to the individual channels and waits for probe response (PRP) packets from the nearby APs. A PRQ may be broadcast for requesting the information from a specific AP by identifying a specific Service Set Identifier (SSID) in the PRQ packet or may be broadcast to all APs using the broadcast SSID. Basically, most Wi-Fi devices perform active scanning since the devices can quickly receive the response from APs. In contrast, the devices need to wait for a periodic BEA in passive scanning. Moreover, the devices could miss the BEA if the dwell time on a channel is insufficient. Figure A.3 shows the mechanisms of the two scanning modes.

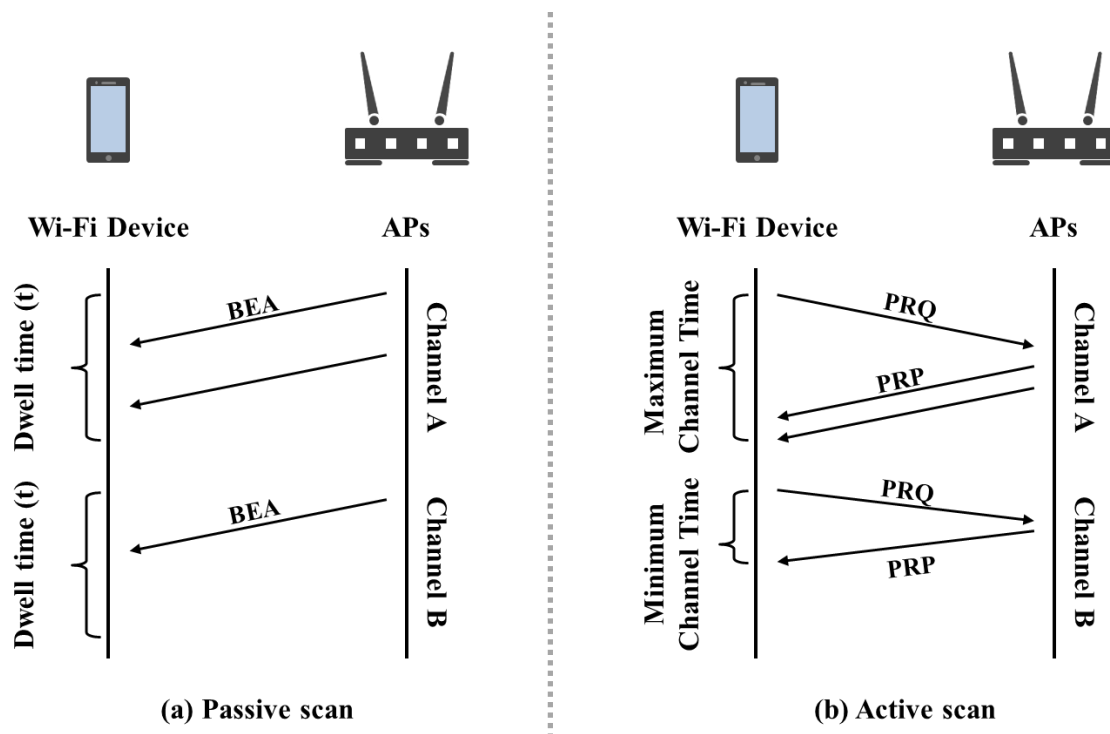


Figure A.3: The Wi-Fi discovery phase for (a) passive scanning and (b) active scanning

Once a connection is established for a Wi-Fi device, the device is enabled to communicate with other devices in the wireless network. Data frames can be used for delivering the communication details afterward. Regardless of the AP association status, it can be noted that a Wi-Fi device regularly transmits a number of packets for maintaining a connection to a wireless network and for exchanging information. In a secure connection, the packets are encrypted and only authorized Wi-Fi devices can access the information. However, it is necessary for the basic information in the packets to be broadcast. For example, the transmitter's MAC-ID is crucial since the receiver must be able to identify the source of a received packet. Therefore, the basic information can be overheard by any Wi-Fi devices in

the range. The underlying mechanism for Wi-Fi communication provides the opportunity for exploiting Wi-Fi data in order to extract useful information for human mobility analysis.

A.2 Wi-Fi device discovery

A Wi-Fi scanner is responsible for collecting passive Wi-Fi data. The scanner can be an AP which is specially configured for collecting Wi-Fi data, or a dedicated Wi-Fi monitoring device which has been developed for monitoring purposes. The scanner is responsible for listening to the Wi-Fi communication in nearby Wi-Fi networks. Here, the scanner acts as a common Wi-Fi device which is able to receive Wi-Fi packets from other devices in the operating range. Since the scanner is not an authorized receiver, only non-encrypted information in Wi-Fi packets can be accessed e.g. the MAC-ID. However, the basic information is adequate for identifying individual Wi-Fi devices. Therefore, Wi-Fi devices can be observed passively by a Wi-Fi scanner without any interference in Wi-Fi communication. It is worth noting that accessing an encrypted message is privacy intrusive. A Wi-Fi scanner must not include any decryption capability.

In the related literature, a group of researchers developed a Wi-Fi scanner for active monitoring (Musa and Eriksson, 2012). The scanner was modified to behave as an emulated AP. Hence, the scanner had the capability to transmit Wi-Fi packets to other Wi-Fi devices within range. The transmitted packets were designed to stimulate the data transmission of Wi-Fi devices. Their proposed method increased the number of detected devices as well as the number of packets from each device. Nevertheless, the benefits of active monitoring could be arguable in practice.

Passive Wi-Fi monitoring can be implemented in two ways. Firstly, the scanner can operate in a mobile sensing mode. A portable scanner can be carried by people or placed in a vehicle as an in-vehicle sensing device. The mobile sensing mode poses extra variation in mobility analysis since the environments in the monitoring range will keep changing while the carrier is moving. In most cases, the Wi-Fi data captured from such conditions are practical only when topological information of the surrounding Wi-Fi networks is available (Bulut et al., 2014). The mobile sensing mode is suitable for studies that focus on observing the mobility and/or the surroundings of a particular target (Zhang et al., 2018).

In contrast, a Wi-Fi scanner can operate in a fixed-location sensing mode. By installing a Wi-Fi scanner at a certain location, the scanner can observe the changes to the Wi-Fi devices in a particular study area. The fixed-location sensing mode was more attractive in the previous studies since the Wi-Fi data could be processed to derive insightful analysis of people

mobility over the observed area. Due to the limitations of the scanner's detection range, Wi-Fi device tracking based on fixed-location sensing has been suggested for indoor mobility analysis, particularly in small-scale study areas such as a hospital or a campus. The study area can be extended by increasing the locations which Wi-Fi scanners are installed to cover a larger space.

Regardless of the operating modes, there are two significant parameters of a Wi-Fi scanner that need to be configured: the detection range and the monitoring cycle.

A.2.1 Detection range

The detection range can be inconsistent among different Wi-Fi scanners. The range is dependent on a number of factors:

- The specification of the Wi-Fi scanner such as transmitter power, antenna type, and antenna gain (Abedi et al., 2015);
- The specification of a detected Wi-Fi device e.g. transmitter power (Yoo et al., 2014); and
- Interference of environmental and/or physical obstacles in wireless communication (Abedi et al., 2014).

Therefore, the very first step is to justify the scanner detection range. A regular method for estimating the detection range is to conduct experiments using various models of Wi-Fi devices (e.g. smartphones). The scanner is capable of detecting Wi-Fi packets while the devices are in the detection range. By gradually changing the devices' position to be further away from the scanner, the detection range can be determined at the distance at which most of the devices are no longer detected. It is worth noting that not all Wi-Fi devices may have the same detection range. Consequently, the detection range can be assumed from the results of a majority of devices. An example of detection range estimation from an empirical study will be discussed in Appendix B.3.

A.2.2 Monitoring cycle

Wi-Fi packets can be broadcast to all channels in the radio frequency, or only transmitted to a specific channel. The transmission is reliant on the objective of an individual Wi-Fi packet. For example, the PRQ packet is usually broadcast to all channels in order to search for available wireless networks, whereas the data frame is generally sent to an authorized receiver in a specific channel. In order to capture the channel-specific packets, a Wi-Fi scanner needs to listen for the packets on the same channel. Without prior knowledge of the

Wi-Fi traffic on each channel, the scanner can perform channel hopping which loops over the available channels so as to capture most of the Wi-Fi packets. This implies that the scanner cannot capture every Wi-Fi packet in the detection range since the packets in other channels will be omitted while the monitoring is performed on another specific channel.

The monitoring cycle can be configured by specifying the operational parameters. First, the number of channels to be observed within a time unit can be determined. Second, the scanner can be configured to capture the packets from some specific channels with significant traffic. The Wi-Fi traffic on each channel can be evaluated empirically. Freudiger (2015) conducted a set of experiments to investigate the effects of the monitoring cycle on the number of detected Wi-Fi packets. To observe the number of captured packets, the Wi-Fi scanners in the experiments were operated with different configurations both in the channel-specific mode and in the channel hopping mode.

Appendix B

Experiments on Wi-Fi data characteristics

B.1 Equipment

- **Wi-Fi scanners**

Portable Wi-Fi scanners were used for collecting the Wi-Fi data as shown in Figure B.1. Each scanner comprised of three parts. First, a small single-board, low-cost computer called a Raspberry Pi was developed to perform the role of a scanner. Second, a portable charger was used as a power source for the computer. Third, with a plug-in WLAN module, the computer was then able to perform Wi-Fi monitoring on wireless networks. Here a module called Wi-Pi was plugged into the computer as an extension. The module supports IEEE 802.11 b/g/n standards and the maximum transmission power is -20 dBm.

The Raspberry Pi is operated on the Linux Operating System. To develop the passive Wi-Fi monitoring ability, an open-source wireless network monitoring tool called Kismet was integrated into the computer. Kismet has configuration parameters for Wi-Fi monitoring. Wi-Fi traffic can be monitored in a channel-specific mode to monitor user-specific channels, or a channel-hopping mode can be used to monitor the entire frequency band. Furthermore, the monitoring can be configured for each mode. The monitoring channels can be specified in the channel-specific mode, while the channel hop velocity can be identified in the channel-hopping mode. In the designed experiments, the Wi-Fi scanners were operated in a channel-hopping mode. The captured Wi-Fi data were recorded in log files. The files can be opened using a network analyzer such as Wireshark.

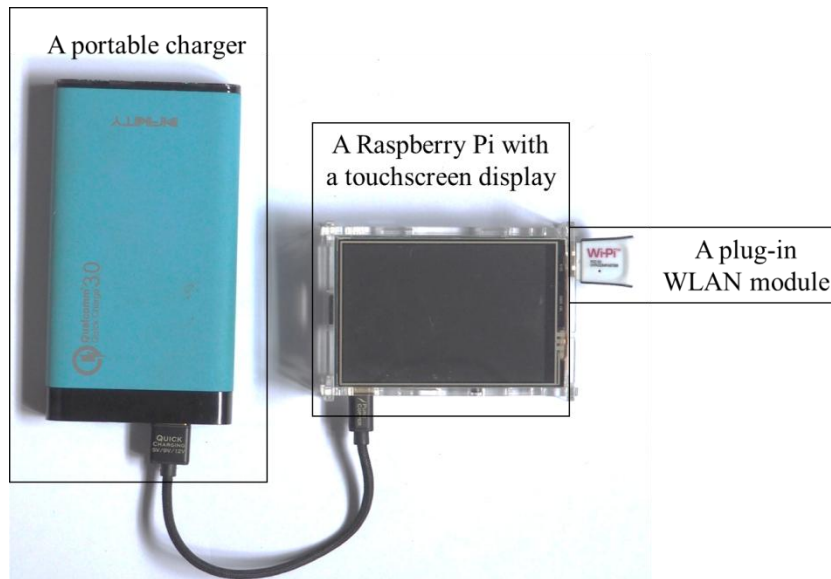


Figure B.1: A portable Wi-Fi scanner used in this study

- **Wi-Fi device**

A number of Wi-Fi devices with known MAC-IDs are required for observing the Wi-Fi data from individual devices. Several Wi-Fi devices were used based on the objectives of each experiment. For example, various device models were considered for observing the differences in Wi-Fi data characteristics among the devices. The models were selected from the smartphone market share to be representative of the majority of Wi-Fi devices. Moreover, the state of the devices was set up during the experiments since the data characteristics of an individual device may be different in various device states.

- **General equipment setup**

The equipment setup for the designed experiments is illustrated in Figure B.2. A number of Wi-Fi scanners (WM) were placed at a location, and the Wi-Fi devices (MD) were placed away from the scanner location at a distance, d at the same ground level. The Wi-Fi devices were arranged in the same orientation.

B.2 Detection events

The characteristics of detection events are essential for mobility analysis. In particular, the event period strongly affects the reliability of time information such as activity duration. An experiment was designed for evaluating the effects of some factors on the quantity of detection events. In the following sub-sections, the designed experiments are described with their objectives followed by the experimental results and discussion.

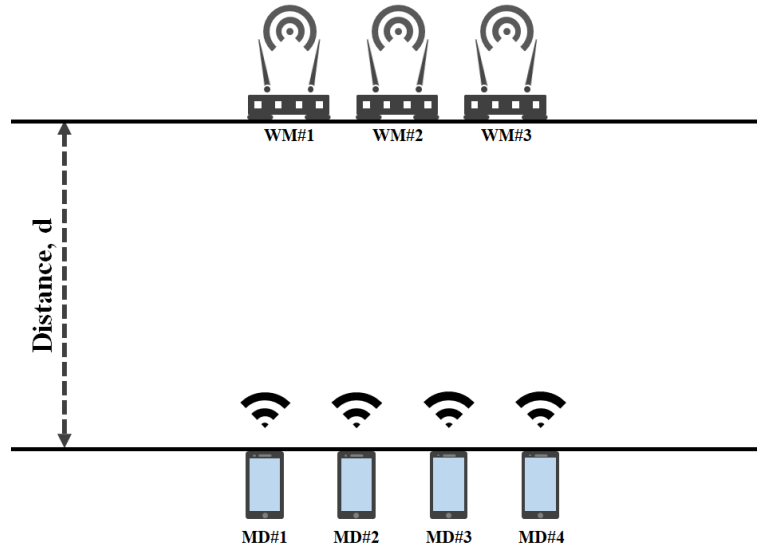


Figure B.2: General equipment setup for the designed experiments

B.2.1 Experiment#1: observing detection events and event periods

- *Objectives:*
 - To study the effects of (i) Wi-Fi device models and (ii) devices' states on the quantity of detection events.
- *Equipment and setup:*
 - Wi-Fi scanner: a Wi-Fi scanner operating in a channel-hopping mode which hops a single channel per second².
 - Wi-Fi devices: three Wi-Fi devices including 2 smartphones and 1 tablet
 - iPhone SE (iOS 10.3.3),
 - Samsung Galaxy Note 5 (Android version 6.0.1), and
 - Samsung Galaxy Tab A (Android version 5.0.2).

The Wi-Fi devices were not associated with an AP and the devices' screens remained on at the screen showing the available Wi-Fi networks.
 - Equipment setup: The equipment was placed on an even floor. The distance, d , between Wi-Fi devices and the scanner was 10 cm.
- *Location:* a private residence.
- *Duration:* 4 hours 5 minutes
- *Description:* The Wi-Fi devices were configured in four various states during the experiment in order to observe the Wi-Fi data during the various states. Each device remained in each state for one hour. The four states are summarized in Table B.1 based on the access point association and the device state.

² The channel hop velocity can be configured to hop 1-10 channels per second. For instance, if the velocity is 5 channels per second, each channel will be monitored for 1/5 of a second.

Table B.1: Summary of Wi-Fi device states during the experiment

Scenario	AP association	Device state
1	×	Stand by (on the screen showing Wi-Fi networks)
2	×	Idle (screen off)
3	✓*	Stand by (on the home screen)
4	✓*	Idle (screen off)

*All devices were associated with the same AP.

Experimental results

Table B.2 shows the number of detection events during each scenario. It can be observed that the number of detection events was different for the various device models, as well as for the different states of an individual device. Firstly, all devices broadcast numerous PRQ packets when the devices were not connected to an AP (Scenario 1 and 2) since the devices attempted to search for a Wi-Fi connection during the various states. However, the device state had a considerable effect on the number of detection events which were significantly reduced in Scenario 2 when the screens were turned off. This could imply that the AP searching of the devices decreased during their idle states.

Table B.2: Number of detection events during each device state

Scenario	No. of detection events		
	iPhone	Samsung	Tablet
1	506	681	708
2	60	21	22
3	4 (no data frame)	11 (9 data frames)	N/A ³
4	10 (1 data frame)	7 (1 data frame)	N/A

Next, the number of detection events was found to be different among the Wi-Fi devices with an AP association regardless of the device state. Although the effect of the device state on the unassociated Wi-Fi devices (Scenarios 1 and 2) is conclusive, the effects were unobvious for the associated devices (Scenarios 3 and 4). Hence, the detection events were further investigated to reach a reasonable conclusion. For the iPhone, most of the detection events in Scenarios 3 and 4 were PRQ packets. There was only one data packet from ten detection events in Scenario 4. More data frames were captured from the Samsung smartphone. There were nine data frames in Scenario 3 when the device was in home screen mode, and a data frame was found in Scenario 4. Nonetheless, there were no detection events from the Samsung tablet during the AP association.

³ Not available

Two major conclusions can be drawn from the investigation. First, PRQs were still occasionally broadcast during the AP association states. Basically, the objective of these PRQs is to maintain the Wi-Fi connection. However, the mechanism might be different for the Samsung tablet with an AP association since there were no observed PRQs. More experiments are required to reach a solid conclusion for the case of tablets. Second, the number of detection events during the AP association was more strongly affected by the data frames than by the device state. An investigation into the installed smartphone applications was conducted. It was found that the applications in the iPhone were provided by the manufacturer with no other additional installations, while social media applications (e.g. Facebook) were installed in the Samsung. Since the Samsung device had more active applications in the background, it resulted in a higher chance of data frame transmission upon the application requests. To this end, it can be concluded that the number of detection events during the AP association states is affected by the transmission of data frames. The detection events could be significantly increased in cases where the device is being used by the user.

In addition to the detection events, the statistics of the event periods can be observed. As detection events can be observed as a burst of a few seconds, it has been suggested in previous related literature that a burst usually lasts for 3 seconds and a low-pass filter (LPF) should be applied to reduce potential bias in statistical analysis (Hu et al., 2015). Herein, for each MAC-ID, a detection event is considered if the MAC-ID has not been detected within the previous 3-second time window. Otherwise, the detection event will be omitted. Table B.3 summarizes the average and standard deviation of event periods. The results were calculated separately to compare the statistical values from the detection events with and without applying the LPF.

Table B.3: Summary of event periods

Scenario	Dataset	Event period (s)					
		iPhone		Samsung		Tablet	
		Average	S.D.	Average	S.D.	Average	S.D.
1	All	7	9	5	6	5	7
	LPF	14	7	12	4	13	5
2	All	51	25	165	283	161	221
	LPF	53	23	550	205	376	167
3	All	762	1,320	326	224	N/A	N/A
	LPF	2,287	N/A	362	167	N/A	N/A
4	All	712	1,171	621	892	N/A	N/A
	LPF	1,282	1,210	1,242	744	N/A	N/A

The results show that the values of the average event period were significantly lower without the LPF since multiple detection events can be detected simultaneously in a second. In such cases, the event period between the consecutive detection events is zero. As a result, the average event period can be biased. In the example of the iPhone in Scenario 1, it is more

correct to assume that a detection event will be captured every 14 seconds on average rather than every 7 seconds. Moreover, using the LPF dataset is more practical for calculating the standard deviation which can represent the distribution of event periods. Figure B.3 demonstrates the comparison of average event periods from individual devices in each scenario. The LPF was applied for calculating the average values.

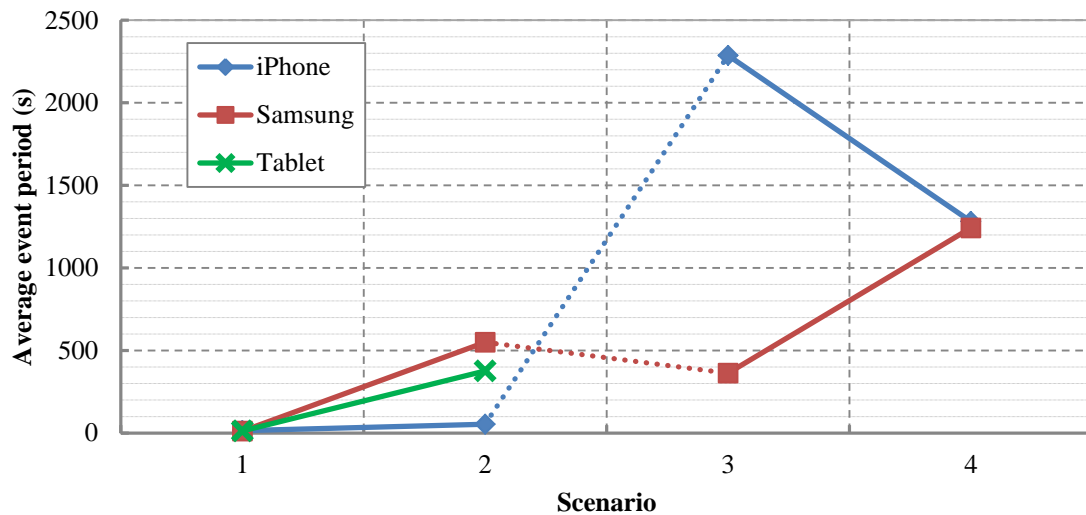


Figure B.3: Average event periods with an LPF

The trend of event periods is consistent with the number of detection events. In general, the period was shorter when more detection events were available. An obvious trend can be observed when the devices were in non-association states. The average event periods were considerably short for all devices in Scenario 1, while event periods were larger for Scenario 2. During the association states, the event periods were more varied. In some cases, only a few detection events were available if there was no data transmission between the Wi-Fi device and the associated access point, resulting in the event period becoming longer. For instance, four detection events were captured from the iPhone in an hour during Scenario 3. However, only two valid observations remained after performing the LPF. As a result, the event period was about 38 minutes during the one hour.

To sum up, the quantity of detection events is dependent on a number of factors. The experimental results show that the model and the state of Wi-Fi devices can be contributing factors. First, the number of detection events is varied in different Wi-Fi devices in the same state. The mechanisms for broadcasting PRQs may not be consistent for all versions of mobile operating systems (i.e. iOS and Android). In particular, the duration of performing an active scan can be different. Second, the state of the Wi-Fi devices strongly affects the number of detection events, especially in the AP association states. Third, an LPF is essential for analyzing statistical information from detection events. Finally, the average event period is assumed to be inversely proportional to the number of detection events. In other words, the

time period between consecutive events tends to be shorter when there is a higher frequency of detection events.

B.3 Received Signal Strength Indicator

RSSI is one part of the broadcast data in a Wi-Fi packet. The RSSI value indicates the intensity of the received radio signal. Owing to the potential of using RSSI for locating Wi-Fi devices, the information has been exploited especially for indoor localization. Basically, the position of a Wi-Fi device can be estimated based on the location of available Wi-Fi scanners (or APs). Various positioning techniques have been developed to improve the accuracy of the estimated positions. Two techniques are widely implemented, namely trilateration-based and fingerprint-based techniques. Both techniques require multiple fixed Wi-Fi scanners with the geographical information of the scanners' locations, as well as prior knowledge on the RSSI at various positions. The trilateration-based methods rely on radio wave propagation models, while the fingerprint-based methods estimate the device position using the RSSI fingerprint database. The positioning accuracy is dependent on the number of Wi-Fi scanners and prior RSSI measurements.

To understand the characteristics of RSSI data, this section aims to evaluate the reliability of RSSI data based on a single Wi-Fi scanner. Two experiments were designed in order to study the variation of RSSI in various Wi-Fi devices, as well as the effects of physical obstacles in RSSI data.

B.3.1 Experiment#2: observing RSSI

- *Objectives:*
 - To study the effects of the distance between Wi-Fi devices and a Wi-Fi scanner on missed detection probabilities.
- *Equipment and setup:*
 - Wi-Fi scanner: a Wi-Fi scanner operating in a channel-hopping mode which hops a single channel per second.
 - Wi-Fi devices: four Wi-Fi devices including 3 smartphones and 1 tablet
 - iPhone SE (iOS 10.3.3),
 - Samsung Galaxy Note 3 (Android version 4.4.2),
 - Samsung Galaxy Tab A (Android version 5.0.2), and
 - Lenovo A7000 (Android version 5.0.2).

The Wi-Fi devices were set to the same orientation and state. There was no AP association and the devices' screens remained on at the screen showing the available Wi-Fi networks.

- Equipment setup: The equipment was placed on an even floor. The distance, d , between the Wi-Fi devices and the scanner was changed to 15 different distances i.e. 1m, 2m, 3m, 4m, 5m, 6m, 7m, 8m, 9m, 10m, 15m, 20m, 30m, 40m, and 50m.
- Location: an open space on Block Z building, The Hong Kong Polytechnic University.
- Duration: 1 hour 30 minutes
- Description: The Wi-Fi devices were placed at fifteen different distances during the experiment in order to observe the Wi-Fi data detected from various distances. The devices remained at each distance for 5 minutes.

Experimental results

RSSI observations from the four Wi-Fi devices are illustrated in Figures B.4-B.7. The box-and-whisker plots show the summary of RSSI observations including the minimum, the three quartiles, and the maximum, which were measured at individual distances. Also, the mean and standard deviation (S.D.) are shown in Tables B.4-B.7.

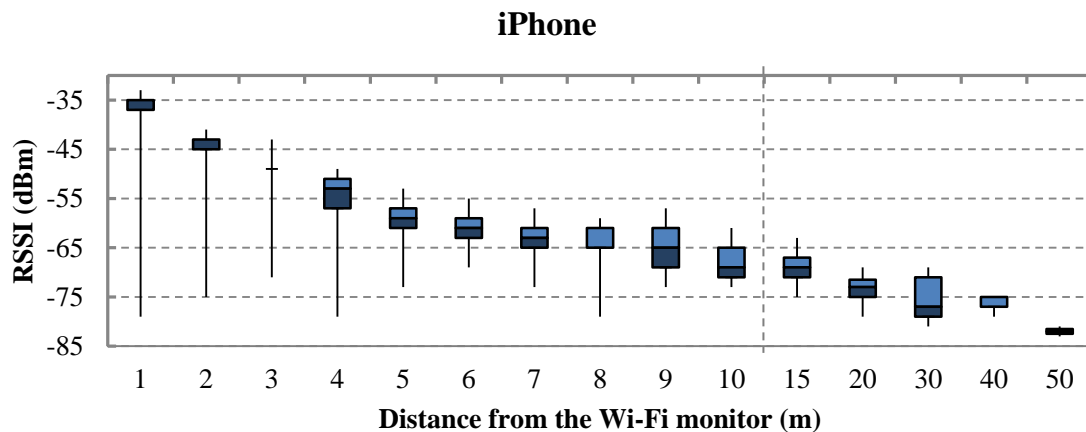


Figure B.4: The RSSI plot (an iPhone)

Table B.4: The mean and S.D. of RSSI (an iPhone)

Distance (m)	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50
Mean (dBm)	-37	-46	-50	-55	-59	-62	-63	-64	-65	-68	-69	-73	-75	-77	-82
S.D. (dBm)	8	7	3	5	4	4	4	4	4	3	3	2	4	1	1

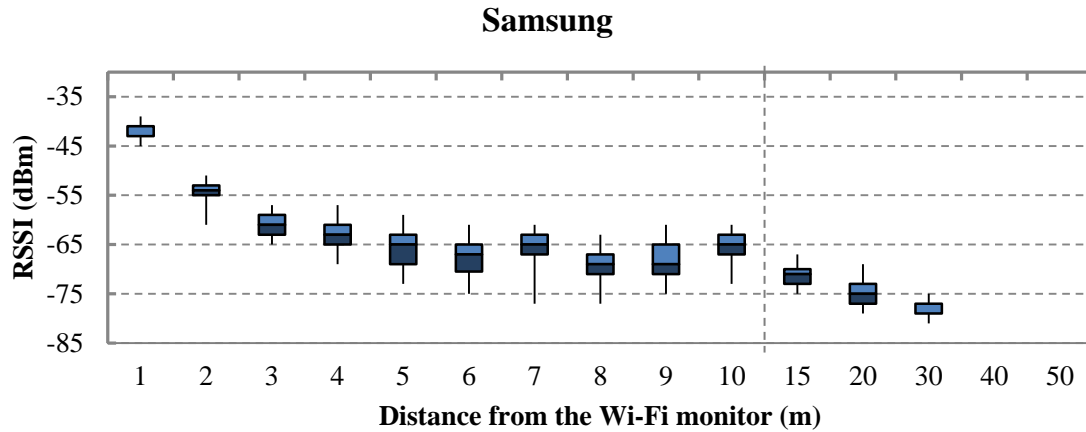


Figure B.5: The RSSI plot (a Samsung smartphone)

Table B.5: The mean and S.D. of RSSI (a Samsung smartphone)

Distance (m)	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50
Mean (dBm)	-42	-54	-61	-63	-66	-68	-66	-69	-68	-66	-71	-75	-78	-	-
S.D. (dBm)	1	2	2	3	4	3	4	4	4	3	2	3	2	-	-

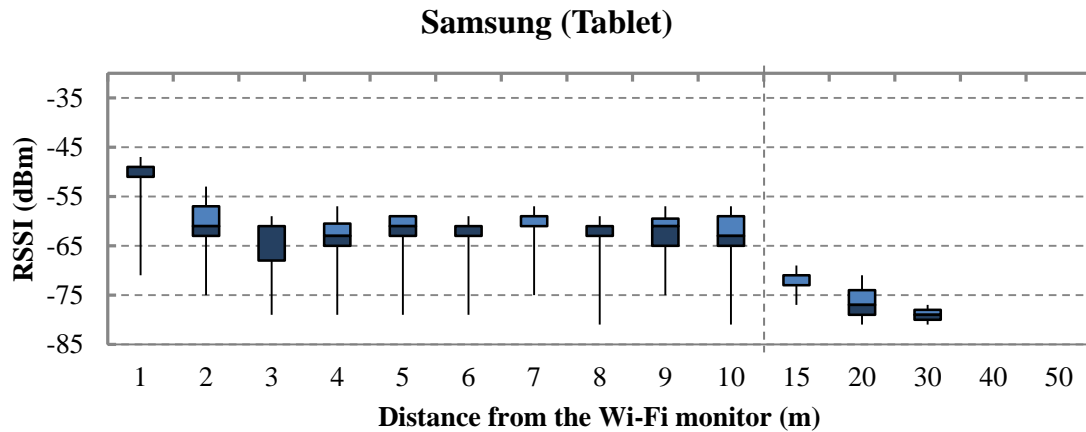


Figure B.6: The RSSI plot (a Samsung tablet)

Table B.6: The mean and S.D. of RSSI (a Samsung tablet)

Distance (m)	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50
Mean (dBm)	-51	-61	-65	-63	-63	-62	-62	-64	-63	-63	-73	-76	-79	-	-
S.D. (dBm)	6	6	7	5	6	3	4	5	4	5	2	4	3	-	-

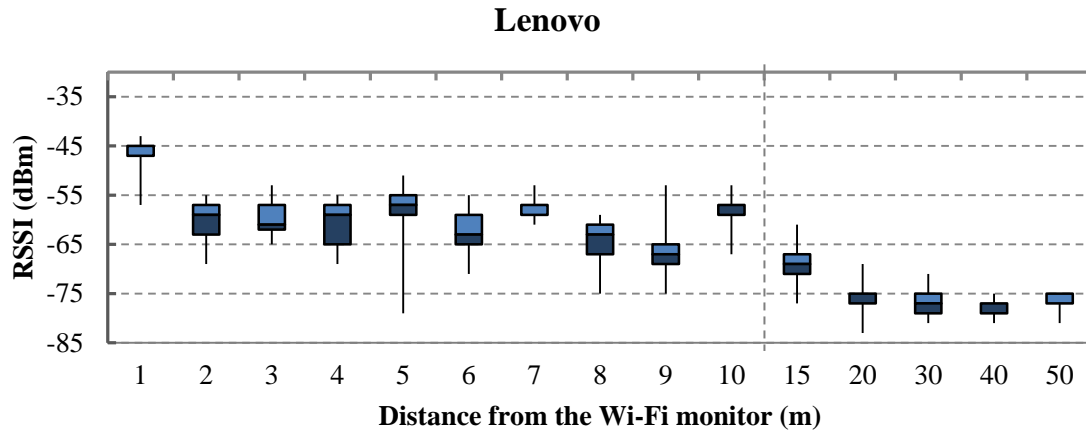


Figure B.7: The RSSI plot (a Lenovo smartphone)

Table B.7: The mean and S.D. of RSSI (a Lenovo smartphone)

Distance (m)	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50
Mean (dBm)	-47	-61	-60	-60	-58	-62	-58	-64	-68	-58	-69	-76	-77	-78	-76
S.D. (dBm)	3	4	3	4	3	4	2	4	3	2	3	3	3	2	2

According to the RSSI plots of the four different Wi-Fi devices, a number of observations should be highlighted as follows.

- First, a decreasing trend of RSSI can be observed from all Wi-Fi devices when the distance between the Wi-Fi devices and the Wi-Fi scanner was increased. In particular, the trends are most obvious when the distances were greater than 10 meters.
- Second, there was a variation of RSSI values at each individual distance. In Table B.4, it can be seen that the standard deviation was able to reach 8 dBm. The degree of variation at the same distance was different for each Wi-Fi device. For example, the standard deviations of RSSI at the 1-meter distance varied from 1 to 8 dBm for the four devices. In addition, it can be noticed from Figure B.4 that a low RSSI can be observed (-79 dBm), even though the Wi-Fi device was very close to the scanner (1 meter).
- Third, the mean RSSI of the four devices were different at the same distance. Moreover, some devices were no longer detected at long distances (i.e. the Samsung devices at 40 meters). It is noteworthy that the minimum RSSI observation in the experiment was -83 dBm. Therefore, it can be assumed that the Samsung devices

were broadcasting Wi-Fi packets but the RSSI was too weak to detect them. This method can be used for estimating the scanner detection range.

It can be concluded that the relationship between RSSI and the distance between a Wi-Fi device and a Wi-Fi scanner is non-linear. The statistics of RSSI varied among different Wi-Fi devices.

B.3.2 Experiment#3: observing RSSI with physical obstacles

This experiment was extended from Experiment#2. In particular, a backpack was used as a physical obstacle and the Wi-Fi devices were placed in the backpack.

- *Objectives:*
 - To study the effects of (i) Wi-Fi device models and (ii) physical obstacles on the quality of RSSI data.
- *Equipment and setup:*
 - Equipment setup: The distance, d , between the Wi-Fi devices and the scanner was changed to three different distances: 1m, 5m, and 10m. In addition, all Wi-Fi devices were put into a backpack which was carried by a volunteer with the same orientation as in Experiment#2.
- *Duration:* 16 minutes
- *Description:* The Wi-Fi devices were placed in a backpack at the three distances during the experiment. The devices remained at each distance for 5 minutes. The RSSI observations in this experiment will be compared with the observations derived from the same distances in Experiment#2.

Experimental results

Figures B.8-B.11 show the comparison of RSSI observations from the four Wi-Fi devices. Each figure presents a summary of the RSSI observed from each of three distances: 1, 5, and 10 meters. In addition, the RSSI observations from the cases with physical obstacles are plotted against those from cases without obstacles at each of the distances. Here, the cases with physical obstacles are marked with an asterisk (*) in the plots. A summary of mean and standard deviations is shown in Tables B.8-B.11.

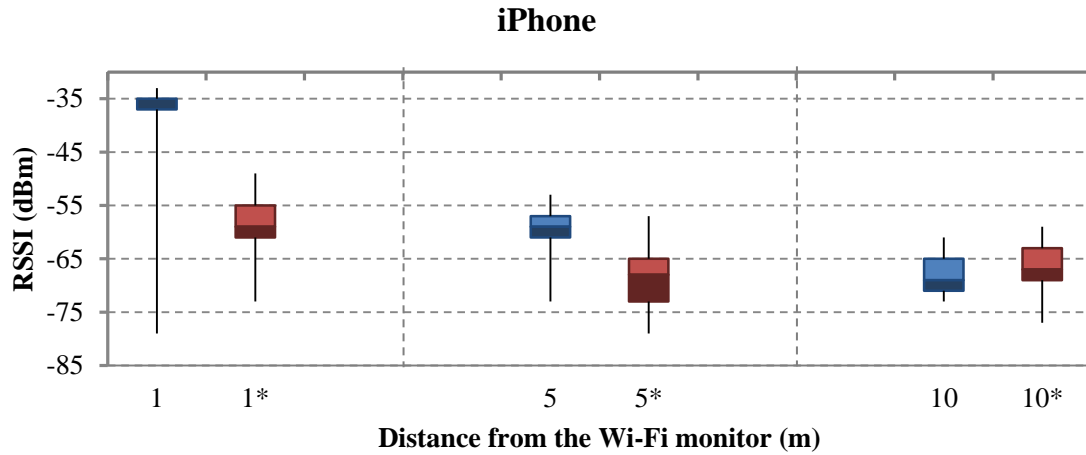


Figure B.8: The RSSI plots for the cases with and without obstacles (an iPhone)

Table B.8: The mean and S.D. of the RSSI for the cases with and without obstacles (an iPhone)

Distance (m)	1	1*	5	5*	10	10*
Mean (dBm)	-37	-59	-59	-69	-68	-67
S.D. (dBm)	8	5	4	5	3	4

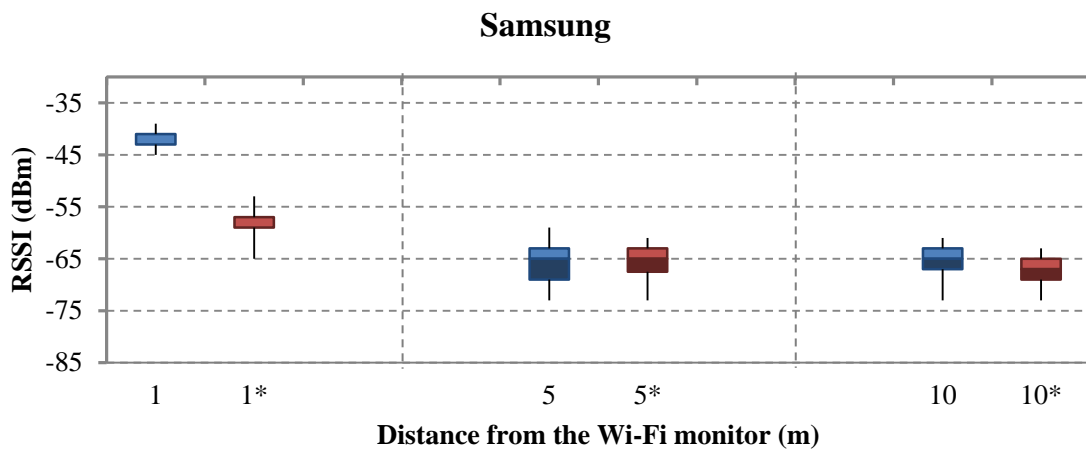


Figure B.9: The RSSI plots for the cases with and without obstacles (a Samsung smartphone)

Table B.9: The mean and S.D. of the RSSI for the cases with and without obstacles (a Samsung smartphone)

Distance (m)	1	1*	5	5*	10	10*
Mean (dBm)	-42	-58	-66	-66	-66	-67
S.D. (dBm)	1	4	3	3	3	3

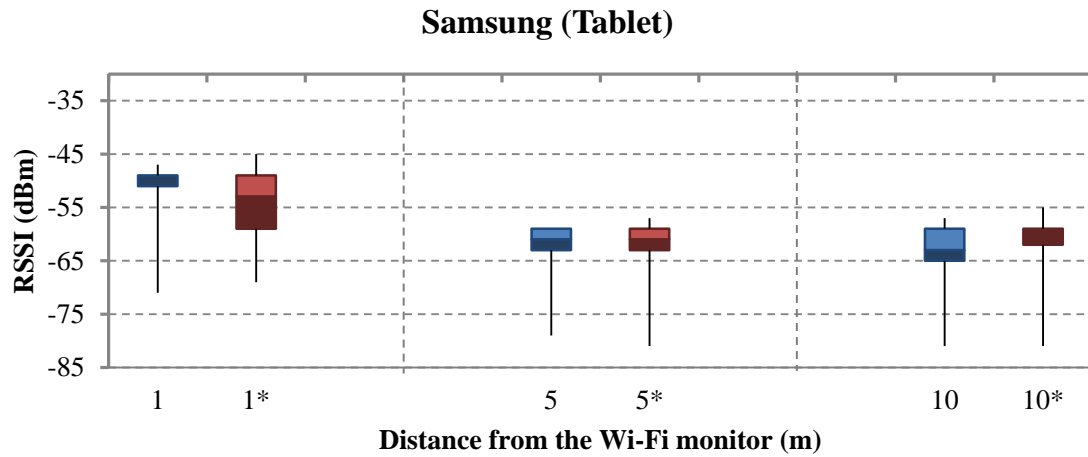


Figure B.10: The RSSI plots for the cases with and without obstacles (a Samsung tablet)

Table B.10: The mean and S.D. of the RSSI for the cases with and without obstacles (a Samsung tablet)

Distance (m)	1	1*	5	5*	10	10*
Mean (dBm)	-51	-54	-63	-63	-63	-62
S.D. (dBm)	6	6	6	7	5	6

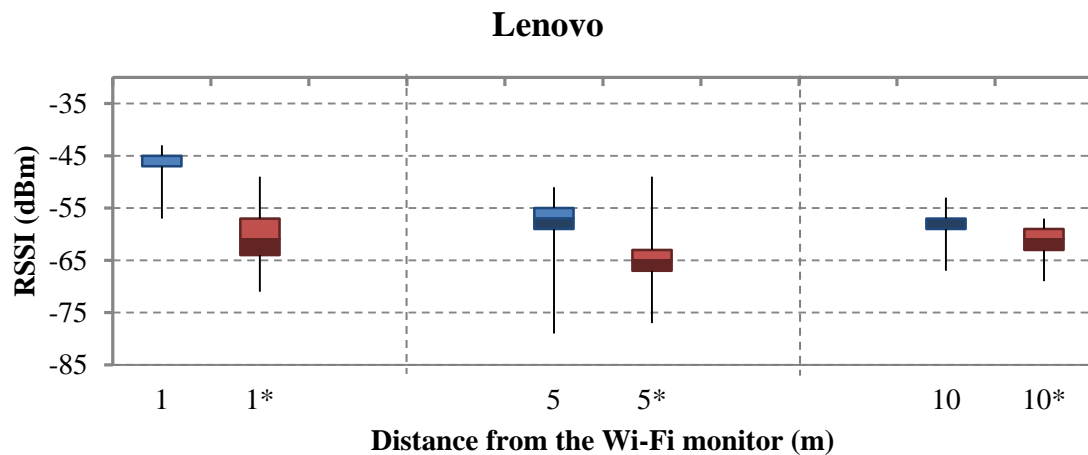


Figure B.11: The RSSI plots for the cases with and without obstacles (a Lenovo smartphone)

Table B.11: The mean and S.D. of the RSSI for the cases with and without obstacles (a Lenovo smartphone)

Distance (m)	1	1*	5	5*	10	10*
Mean (dBm)	-47	-60	-58	-65	-58	-61
S.D. (dBm)	3	5	3	4	2	3

Since most RSSI observations tend to be lower in the cases with physical obstacles, it can be inferred that the loss of RSSI can occur when the radio signal is required to pass through physical objects. The signal attenuation is noticeable for the shorter distance between the Wi-Fi scanner and the Wi-Fi devices i.e. 1 meter. The degree of attenuation decreased for the longer distances and finally comparable in some cases (e.g. the iPhone device at a 10-meter distance, the Samsung devices at 5-meter and 10-meter distances).

B.4 Missed detection

B.4.1 Experiment#4: observing missed detection from various Wi-Fi devices

This experiment was extended from Experiment#1. Therefore, the Wi-Fi scanner and Wi-Fi devices used in the experiments were the same, as were the locations. The only differences were in the experimental design as described in the following.

- *Objectives:*
 - To study the effects of Wi-Fi device models on the missed detection probability.
- *Duration:* 1 hour

Experimental results

The missed detection probability of a Wi-Fi device can be evaluated using the ratio of the captured detection events to the total Wi-Fi packets broadcast by the device. Firstly, the number of detection events captured from each controlled Wi-Fi device was counted. Next, the total number of PRQ packets can be estimated using the difference in the sequence number (SN) between the first detection event and the last detection event. Table B.12 shows the evaluation results from the three controlled devices. It is noteworthy that it was not possible to estimate the total number of PRQs from the iPhone due to the iPhone Operating System (iOS). The SN of consecutive PRQs from the iPhone is in a non-sequential order. It can be assumed that iOS may provide additional security for the user's privacy. The implemented security of the iOS (i.e. MAC randomization) is described in Section 3.5.4.

Table B.12: Evaluation of missed detection probability

Wi-Fi Device	Detection Event	Total PRQ	Missed Detection (%)
iPhone	506	N/A	N/A
Samsung smartphone	681	9,330	92.70
Samsung tablet	708	15,820	95.52

The results show that the probability of missed detection is significant even in the most active and stable stage of Wi-Fi devices in the controlled experiment. It can be noticed that the number of PRQs broadcast varies in different devices. However, the results also indicate that the missed detection probability does not always increase for the devices which broadcast more PRQs (e.g. the Samsung tablet). It can be assumed that the active scanning cycle of the Samsung smartphone had more encounters with the monitoring cycle of the Wi-Fi scanner.

The detection events from an individual Wi-Fi device can be used for further investigation. Table B.13 presents an example of consecutive detection events from the Samsung smartphone. Here the SN of each detection event is included for further discussion. The SN is general information in a Wi-Fi packet identifying the sequence of packets broadcast from a Wi-Fi device. The purpose of SN is to avoid duplicated transmission in the Wi-Fi communication protocol.

Table B.13: An example of detection events from a Wi-Fi device

Detection Event No.	Time (s)	Channel	SN
1	1	1	529
2	1	1	530
3	11	7	585
4	21	9	633
5	21	9	634
7	31	11	676
8	31	11	681
9	31	11	682
10	41	13	729
10	41	13	730

It can be observed that the ten consecutive detection events from the Wi-Fi device did not have successive packet sequence numbers. Only partial PRQs were captured by the Wi-Fi scanner. Two assumptions can be drawn from the information presented in Table B.13. First, in the channel-hopping mode, missed detection can occur when the monitoring is performed in a specific channel but the Wi-Fi packets are broadcast in other channels. Suppose that the Wi-Fi device has an active scanning cycle which broadcasts a number of PRQs in the individual channels sequentially, and the Wi-Fi scanner has a monitoring cycle for the channel-hopping mode e.g. one channel per second. Hence, the PRQs will be captured only if the active scanning cycle and the monitoring cycle are encountered and performed on the same channel. As the detection events are mostly available in every 10-second period, it can be assumed that the two cycles are encountered in every 10-second period when a Wi-Fi device is in a stable state.

The other evidence is that the SN of the detection events captured in the same second is generally continuous. Such detection events are the PRQs which were captured from the same channel since the monitoring was configured to perform for one second on each channel. On the other hand, the SN is skipped for the detection events with distinct detection times. For example, the SN of detection event number 3 is 585, which is skipped from the SN

530 of detection event number 2. This implies that the Wi-Fi device broadcast the PRQs with SN 531-587 during the period from 1st second to the 11th second. However, the active scanning cycle and the monitoring cycle were not encountered, resulting in the missed detection of such PRQs.

The second assumption is that the captured Wi-Fi packets may not be the total amount of all packets broadcast on the channel. The detection events during the 31st second could be an example. According to the SN of the detection events, it can be assumed that at least four PRQs with SN 677-680 were missing. Missed detection in the channel is possible since numerous Wi-Fi packets can be broadcast on the same channel by various Wi-Fi devices and the Wi-Fi scanner may be incapable of capturing all the packets within the monitoring cycle.

B.4.2 Experiment#5: observing missed detection from various distance

The Wi-Fi data captured from Experiment#2 are used for studying the variation of RSSI data. The main objective of this experiment is to study the effects of (i) Wi-Fi device models and (ii) the distance between Wi-Fi devices and a Wi-Fi scanner on the quality of RSSI data.

Experimental results

Figure B.12 demonstrates the number of detection events from the four Wi-Fi devices in Experiment#2 based on the distance from the Wi-Fi scanner.

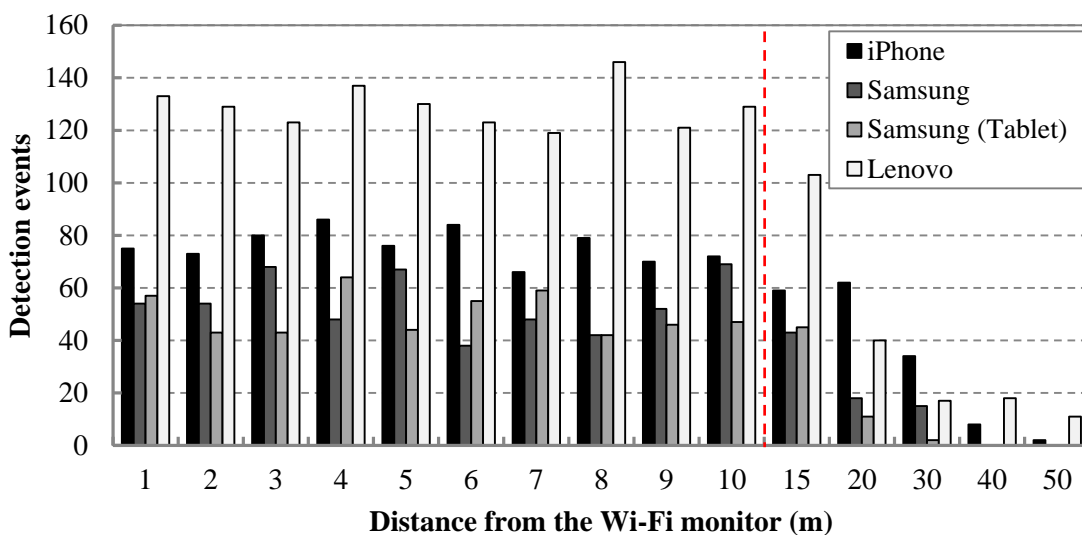


Figure B.12: Comparison of detection events based on the distance of Wi-Fi devices from a Wi-Fi scanner

The results show that the number of detection events is affected by the model of the device as well as by the distance between the Wi-Fi device and the Wi-Fi scanner. Assuming that an individual device broadcast an equal number of PRQs during its time at each distance, the missed detection probability increased noticeably when the distance was greater than 20 meters. This means that the sample size for summarizing mobility information tends to be reduced for greater distances.

To this end, the effects of a Wi-Fi device's conditions on missed detection have been discussed including the device's model, its state, and its distance from the Wi-Fi scanner. In this experiment, missed detection is further investigated using several Wi-Fi scanners which are configured with different channel hopping velocities. Figure B.13 shows the concept of Wi-Fi monitoring in a channel hopping mode using two different velocities.

Suppose that the monitoring is performed on each channel for an equal period of time, T . It will take 6.5 seconds for the Wi-Fi scanners using the two channels per second (ch/s) velocity to complete a monitoring cycle, and 13 seconds for the scanners using the 1 ch/s velocity. It can be observed that the 2 ch/s velocity will spend an equal amount of time on each channel as the 1 ch/s velocity within 2 monitoring cycles. Hence, an initial assumption may be that the channel hopping velocity does not significantly affect missed detection. However, missed detection still depends on the encounter probability between the active scanning cycle and the monitoring cycle. The following experiments were designed to study the effect of different channel hopping velocities on missed detection.

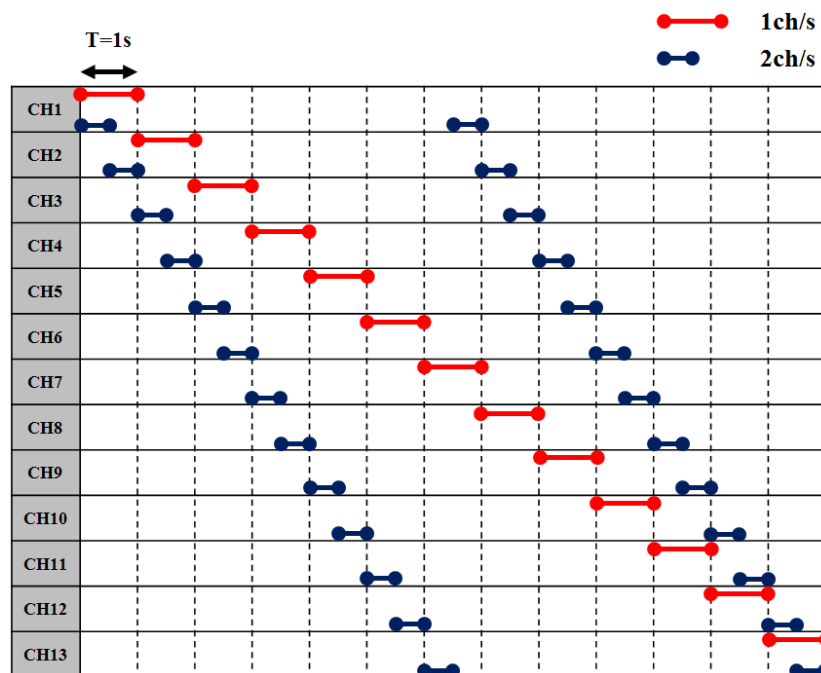


Figure B.13: Wi-Fi monitoring using different channel-hopping velocities

B.4.3 Experiment#6: observing missed detection based on different Wi-Fi scanner setups

This experiment was extended from Experiment#2. Therefore, the Wi-Fi scanner and Wi-Fi devices used in the experiments were the same, as were the locations. Two additional Wi-Fi scanners were included in this experiment. The differences in the experimental design are described as follows.

- *Objectives:*
 - To study the effects of channel hopping velocity on the quality of Wi-Fi data.
- *Equipment and setup:*
 - Wi-Fi scanner: three Wi-Fi scanners operating in the channel-hopping mode with different channel hopping velocities including 1, 5, and 10 ch/s. The three velocities were selected based on the range of velocities which are available for the developed Wi-Fi scanners. The minimum and maximum velocities are 1 ch/s and 10 ch/s, respectively.
- *Duration:* 1 hour
- *Description:* This experiment was conducted at the same time as Experiment#2. Two more Wi-Fi scanners were included to collect Wi-Fi data for 1 hour when the devices were placed at 1-10 meters from the scanner.

Experimental results

In this experiment, the missed detection was initially evaluated in terms of monitoring capacity. Table B.14 shows the monitoring capacity of the three Wi-Fi scanners which were configured with different channel hopping velocities. The capacity was measured based on the number of detection events during the experiment. Then the capacity was generalized using the average number of detection events per second. In Table B.14, the overall monitoring capacity was evaluated using every captured detection event. Moreover, the number of detection events from the controlled Wi-Fi devices per second was calculated in order to study the effect of channel hopping velocity on missed detection.

Table B.14: Wi-Fi monitoring capacity

Channel hopping velocity (channels/s)	Monitoring capacity (events/s)				
	Overall	Samsung	iPhone	Tablet	Lenovo
1	23	0.18	0.25	0.17	0.42
5	21	0.15	0.23	0.14	0.34
10	15	0.14	0.20	0.14	0.33

Although the effect of channel hopping velocity on missed detection was assumed to be insignificant, the results show that the monitoring capacity was enhanced when the monitoring was performed on lesser channels in a second. A rational explanation for this could be based on the encounter probability between the monitoring cycle of a Wi-Fi scanner and the active scanning cycle of individual Wi-Fi devices. According to the experimental results, it can be assumed that a higher channel hopping velocity has more chance of missed detection even though the three scanners spent an approximately equal amount of time on each channel in the long run (i.e. in the 1-hour experiment).

A major impact of missed detection is the accuracy of the occupancy time of Wi-Fi devices. Suppose that the occupancy time of a device is the duration of time that the device is in the detection area of a Wi-Fi scanner, and the presence time of a device is the duration of time from the first detection timestamp to the last one. In a specific time window, the presence time is usually shorter than the actual occupancy duration due to missed detection. The effects of missed detection can be further investigated from the mean absolute percentage error (MAPE) of the controlled devices' occupancy time. Since the controlled devices were physically in the detection area for the entire 1-hour experiment, the error of device occupancy time during a time window is the difference between the presence time and the specific time window. Then, the MAPE is calculated from the ratio of the error to the time window so as to describe the deviation of presence time from the occupancy time. The MAPE can be calculated as follows:

$$MAPE = \frac{1}{N} \times \frac{(\textit{occupancy time} - \textit{presence time})}{\textit{occupancy time}} \quad (\text{B.1})$$

where N is the number of available samples for calculating the MAPE.

Figure B.14 shows the MAPE of occupancy time. The samples from the Samsung smartphone were used for illustrating the error. The MAPE was calculated from the three Wi-Fi scanners, using various time windows in order to compare the effects of missed detection based on the duration of occupancy time. In addition, a re-sampling method was applied for the more reliable MAPE. For each time window, a new sample from individual devices was taken into account in every 30-second period except for the 1-hour time window.

The results show that both channel hopping velocity and the time window affect the accuracy of occupancy time. First, the error is decreased when the time windows are longer since the occupancy time error remained at a corresponding degree while the time duration is increased. Second, the error is increased with the higher channel hopping velocities. The accuracy results based on channel hopping velocities are also consistent with the trend of Wi-Fi monitoring capacity. Since the lower velocity provides a higher monitoring capacity, the probability of missed detection could be lower and this results in higher accuracy.

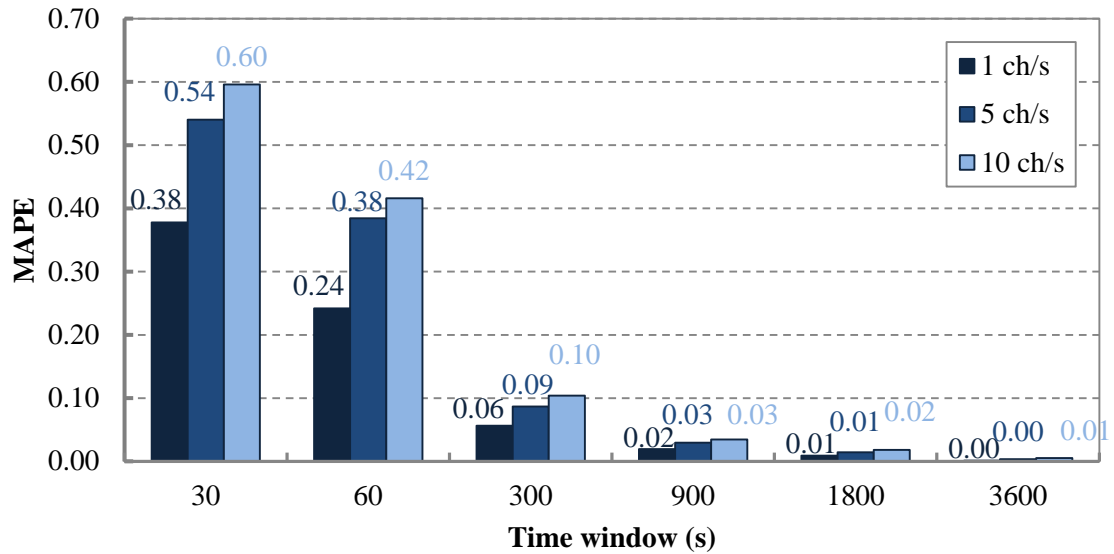


Figure B.14: MAPE of occupancy time

B.4.4 Experiment#7: observing missed detection in a crowded space

- *Objectives:*
 - To study the capability of Wi-Fi scanners to detect Wi-Fi data in a crowded space.
 - To study the effects of channel hopping velocity on the quality of Wi-Fi data in a crowded space.
- *Equipment and setup:*
 - Wi-Fi scanners: three Wi-Fi scanners operating in the channel-hopping mode with different channel hopping velocities including 1, 5, and 10 channel(s) per second.
 - Equipment setup: The equipment was in a box and placed on an even floor. The distance, d , between the Wi-Fi devices and the scanner was 5 cm.
- *Location:* an open space at Victoria Harbour, East Tsim Sha Tsui, Kowloon, Hong Kong
- *Duration:* 1 hour starting from 11.15 p.m. during an event for New Year count down and fireworks, 2017. It is noteworthy that there were temporary closures of the road and pedestrian walkways connecting to Victoria Harbour. The closures started from 11 p.m. resulting in low pedestrian flows within the location until the end of the event at 12.15 a.m.
- *Description:* The equipment was maintained in the same state during the experiment in order to observe the Wi-Fi data detected in the crowded space. The data will be compared with the results from Experiment#6.

Experimental results

In the same way as for Experiment#6, Table B.15 shows the monitoring capacity of the three Wi-Fi scanners. Both the overall capacity and the capacity of the controlled devices were evaluated.

Table B.15: Wi-Fi monitoring capacity in a crowded space

Channel hopping velocity (channels/s)	Monitoring capacity (events/s)				
	Overall	Samsung	iPhone	Tablet	Lenovo
1	32	0.15	N/A	N/A	0.39
5	27	0.13	N/A	N/A	0.32
10	N/A	N/A	N/A	N/A	N/A

It can be noted that some results are not available in Table B.15 due to technical problems during the experiment. First, the Wi-Fi scanner using the 10 channels/s velocity was unable to capture Wi-Fi traffic. Second, the Samsung tablet was not in a stable state during the experiment. The experimental results from the device are therefore invalid and should be excluded. In addition to the technical issues, the iPhone device was unable to be identified in crowded environments. Due to the security setting of iOS, the MAC-ID among the broadcast PRQs was not the actual MAC-ID of the iPhone device. In the previous experiments, the MAC of the iPhone could be identified using RSSI information and/or the device's presence time. However, this experiment was conducted in a crowded space. In this case, these methods cannot be used since there were several MAC-IDs with similar RSSI and presence time.

The monitoring capacity trend from the two Wi-Fi scanners is consistent with the results in Experiment#6. Furthermore, the results from the two experiments in Tables B.14 and B.15 can be compared in order to study the impact of Wi-Fi traffic volume on the monitoring capacity. The results indicate that the overall capacity improved in Experiment#7 which was conducted in the midst of a high load of Wi-Fi traffic during a New Year event. With the same configuration of Wi-Fi scanners in the two experiments, the results imply that the monitoring capacity in Experiment#6 did not reach the maximum capacity of the scanners. In contrast, the improved monitoring capacity in Experiment#7 can be assumed to be close to maximum capacity based on the Wi-Fi traffic in Victoria Harbour from thousands of people attending the countdown event.

Although the overall capacity of the Wi-Fi scanners improved in Experiment#7, the results indicate that the scanner captured fewer detection events from the controlled devices. An explanation for this could be that the excessive volume of Wi-Fi traffic increased the missed detection probability of an individual Wi-Fi device. There is a higher chance that the Wi-Fi data from an individual device will be included in the set of captured detection events when the same Wi-Fi scanner performs in lighter Wi-Fi traffic. Since missed detection is affected by the volume of Wi-Fi traffic, the error in the devices' occupancy time was also

investigated. Figure B.15 shows the MAPE of the occupancy time of the Samsung smartphone.

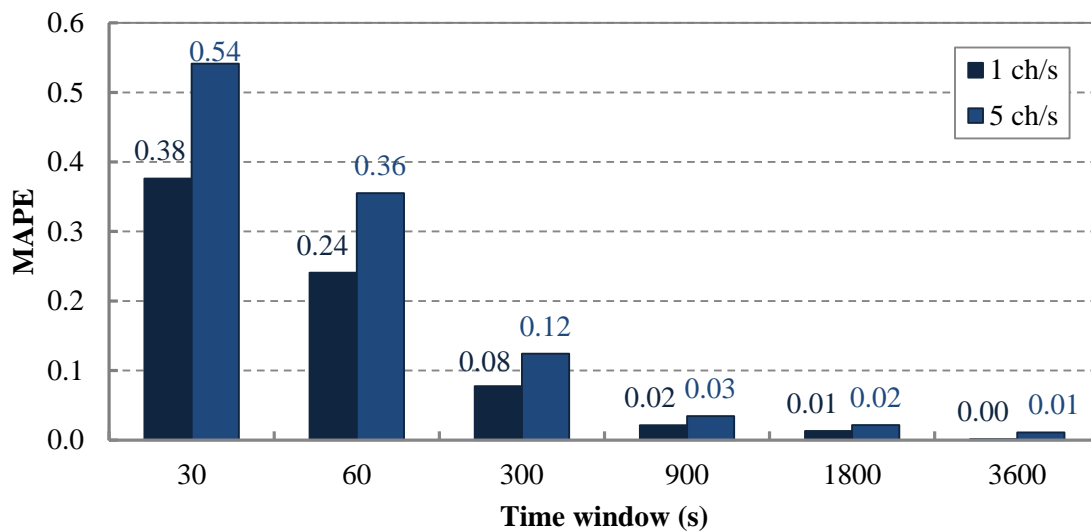


Figure B.15: MAPE of occupancy time in a crowded space

Although the monitoring capacity of the Samsung smartphone decreased in the crowded space, the error in the device's occupancy time is still comparable to the MAPE results from Experiment#6. The main reason for this could be that the increased number of detection events within the small time window was insignificant. For example, the Wi-Fi scanner using the 1 channel/s velocity detected 1.8 more detection events than Experiment#6 in a 60-second time window. In this case, the additional detection events may not strongly affect the MAPE of the occupancy time especially when the detection events were captured in the same timestamp along with other detection events. For the longer time windows, the MAPE is usually inconsiderable due to the ratio of the small occupancy time error to the large time window.

To highlight the necessity of using the channel hopping mode for Wi-Fi monitoring, two datasets including the detection events from both Experiment no.6 and no.7 were further processed. The data were captured by the Wi-Fi scanner using the 1 channel/s velocity. Figures B.16 and B.17 demonstrate the proportion of the detection events of the two datasets on each Wi-Fi channel.

It can be seen that most of the detection events from Experiment#6 were captured from three major channels: 1, 6, and 11. Since the experiment was conducted on a university campus, it should be highlighted that the three channels were configured for providing in-campus Wi-Fi services. Hence, the results may imply that most detection events were captured from the nearby access points in the campus and the Wi-Fi devices which were associated to the access points.

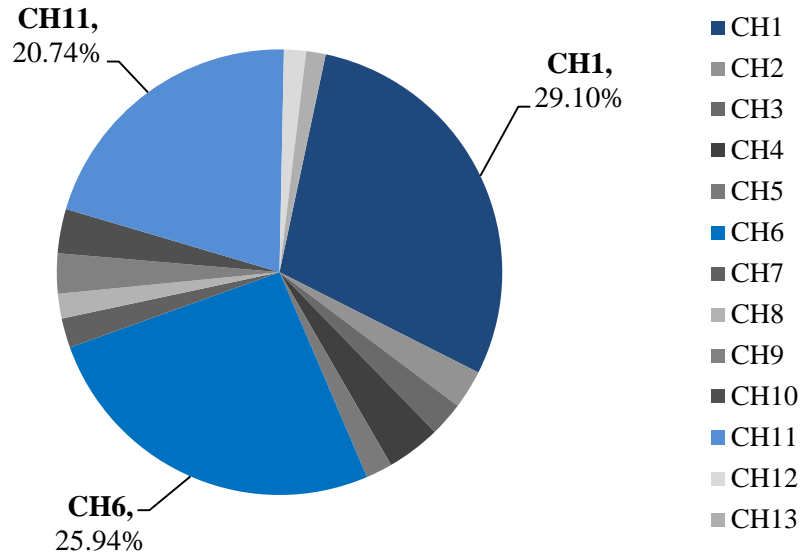


Figure B.16: The proportion of detection events from Experiment#6 in each Wi-Fi channels

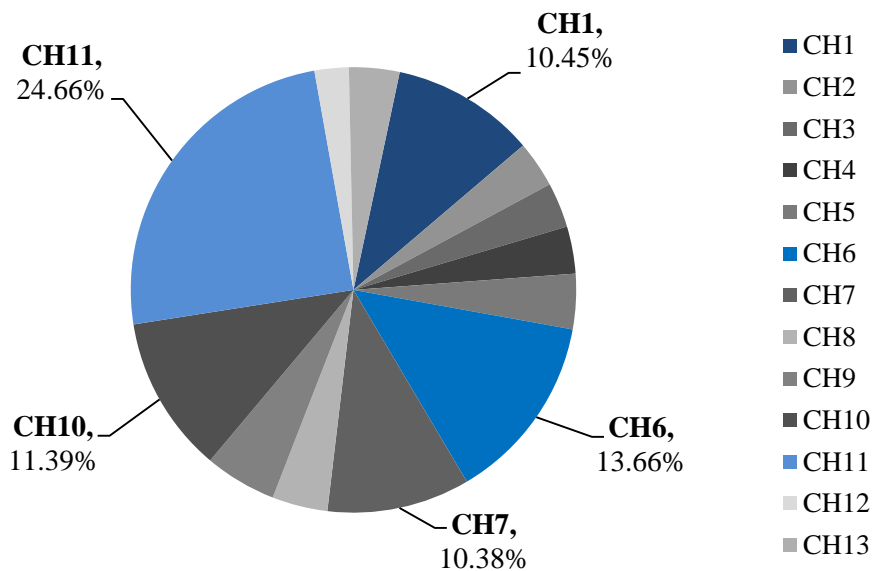


Figure B.17: The proportion of detection events from Experiment#7 in each Wi-Fi channels

The proportion is different for the other dataset. The three channels still hold the majority of the detection events, even though the other channels also had greater proportions, particularly channels 7 and 10. Since the experiment was conducted in an open space, the information about the channels of nearby Wi-Fi access points is inconclusive. The majority of detection events might have been detected from personal Wi-Fi devices rather than from the nearby access points. The two different proportions of detection events demonstrate that Wi-Fi traffic in a particular location can vary. Therefore, the configuration of the Wi-Fi monitoring mode is important for maximizing the detection probability of the target devices based on the

objectives of the mobility analysis. To study the usage of in-campus Wi-Fi services, a set of Wi-Fi scanners can be operated in a channel-specific mode which aims to observe the specific channels of the Wi-Fi services.

Appendix C

Examples of raw data

C.1 Passive Wi-Fi data

Table C.1: Examples of raw Wi-Fi data

Timestamp	Source MAC-ID	Destination MAC-ID	RSSI (dBm)	Info.
16:54:27	Luxul_2e:03:d2	Broadcast	-75	Beacon frame, SN=962, FN=0, Flags=....., BI=100, SSID=Mirpuri 2.4G
16:54:27	SamsungE_51:73:64	Broadcast	-63	Probe Request, SN=3034, FN=0, Flags=....., SSID=Broadcast
16:54:28	ArubaNet_5d:5a:25	Broadcast	-71	Beacon frame, SN=3221, FN=0, Flags=....., BI=100, SSID=USLS
16:54:28	LenovoMo_c8:34:db	Broadcast	-57	Probe Request, SN=574, FN=0, Flags=....., SSID=Broadcast
16:54:28	TendaTec_c1:4e:5c	Broadcast	-35	Probe Request, SN=2128, FN=0, Flags=....., SSID=Broadcast
16:54:28	ArubaNet_74:eb:a5	Broadcast	-69	Beacon frame, SN=2740, FN=0, Flags=....., BI=100, SSID=USLS
16:54:28	HuaweiTe_56:18:b9	Broadcast	-77	Probe Request, SN=3507, FN=0, Flags=....., SSID=Broadcast

C.2 Observed passenger waiting times

Table C.2: Examples of observed passenger waiting times

No.	Passenger arrival time	Bus departure time	Boarded bus line
1	17:33:20	17:38:01	112
2	17:35:50	17:43:33	171
3	17:38:15	17:43:33	171
4	17:43:12	17:46:04	102
5	17:32:30	17:48:16	117
6	17:32:13	17:53:05	118
7	17:48:12	17:56:15	104
8	17:52:17	17:56:15	104
9	17:54:21	17:56:15	104
10	17:50:58	17:56:28	171
11	17:49:37	17:56:28	171
12	17:54:35	17:56:39	118
13	17:56:43	17:57:38	112
14	17:57:18	17:57:38	112
15	17:48:19	17:57:54	102
16	17:52:38	17:58:08	112
17	18:02:58	18:04:03	118
18	17:58:42	18:06:05	117
19	17:58:54	18:07:56	102
20	18:04:52	18:07:56	102
21	18:07:11	18:07:56	102
22	18:04:59	18:08:28	118
23	18:08:35	18:11:51	118
24	18:10:43	18:11:51	118
25	18:12:35	18:16:31	102
26	18:13:06	18:16:31	102
27	18:17:42	18:21:53	112
28	18:19:38	18:21:53	112
29	18:07:11	18:22:08	104
30	18:22:47	18:23:34	112
31	18:22:26	18:24:20	171
32	18:31:45	18:41:10	171
33	18:40:40	18:41:10	171
34	18:33:41	18:41:37	102
35	18:36:45	18:41:37	102

C.3 GPS bus data

Date: 2/7/2019
 Starting time: 7:49
 Bus line: 24
 Bus license no: 50143

Table C.3: Examples of raw GPS data

Timestamp	Latitude	Longitude	Speed
2019-07-02 07:45:07	13.84283	100.49353	25
2019-07-02 07:45:22	13.84272	100.49306	16
2019-07-02 07:45:37	13.84261	100.49264	0
2019-07-02 07:45:52	13.84256	100.49244	15
2019-07-02 07:46:07	13.84247	100.49214	9
2019-07-02 07:46:22	13.84236	100.49178	9
2019-07-02 07:46:37	13.84222	100.49167	0
2019-07-02 07:46:52	13.84225	100.49167	1
2019-07-02 07:47:07	13.84228	100.49156	8
2019-07-02 07:47:22	13.84244	100.49164	4
2019-07-02 07:47:37	13.84247	100.49181	0
2019-07-02 07:47:52	13.84253	100.49203	4
2019-07-02 07:48:07	13.84256	100.49214	1
2019-07-02 07:48:22	13.84264	100.49256	15
2019-07-02 07:48:37	13.84283	100.49342	18
2019-07-02 07:48:52	13.84286	100.49353	0
2019-07-02 07:49:07	13.84289	100.49367	3
2019-07-02 07:49:22	13.84292	100.49375	4
2019-07-02 07:49:37	13.84294	100.49389	1
2019-07-02 07:49:52	13.84294	100.49392	1
2019-07-02 07:50:07	13.84297	100.49400	2
2019-07-02 07:50:22	13.84300	100.49422	0
2019-07-02 07:50:37	13.84300	100.49422	0
2019-07-02 07:50:52	13.84306	100.49444	9
2019-07-02 07:51:07	13.84314	100.49481	15
2019-07-02 07:51:22	13.84336	100.49592	40
2019-07-02 07:51:37	13.84358	100.49711	30
2019-07-02 07:51:52	13.84381	100.49867	43
2019-07-02 07:52:07	13.84389	100.49947	23
2019-07-02 07:52:22	13.84392	100.50083	35
2019-07-02 07:52:37	13.84397	100.50192	17
2019-07-02 07:52:52	13.84394	100.50228	23
2019-07-02 07:53:07	13.84392	100.50336	10

C.4 Passenger boarding and alighting at bus stops

Date: 2/7/2019

Starting time: 7:49

Bus line: 24

Bus license no: 50143

Table C.4: Examples of passenger boarding and alighting data

Bus stop no.	Boarding passengers	Alighting passengers	On-board passengers
11	10	1	25
12	8	3	30
14	3	1	32
15	1	0	33
16	2	5	30
18	1	5	26
19	5	4	27
20	1	6	22
21	2	0	24
22	5	4	25
24	0	2	23
25	4	10	17
26	0	1	16
27	5	3	18
28	1	3	16