



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](mailto:lbsys@polyu.edu.hk) providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

Pao Yue-kong Library, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

<http://www.lib.polyu.edu.hk>

**DATA-DRIVEN ANALYTICS OF HUMAN DYNAMICS  
USING PRIVACY-SENSITIVE DATA**

**JIAXING SHEN**

**PhD**

**The Hong Kong Polytechnic University**

**2019**

**The Hong Kong Polytechnic University**

**Department of Computing**

**Data-Driven Analytics of Human  
Dynamics Using Privacy-Sensitive Data**

**Jiaxing SHEN**

*A thesis  
submitted in partial fulfilment of the requirements  
for the degree of*

***Doctor of Philosophy***

January 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 acknowledgement has been made in the text.

\_\_\_\_\_ (Signed)

\_\_\_\_\_ Jiaxing SHEN (Name of student)



# Abstract

Human dynamics is interdisciplinary research which has been extensively investigated in various disciplines from different dimensions. As a result, it leads to somewhat different research focuses like human mobility, international and domestic migration, and population change. In this dissertation, we focus on human dynamics in computer science which refers to *human activities* and *human interactions*.

The rapid development of digital information technologies, like communication technology, sensing technology, and mobile technology, has enabled a mobile and big data era for human dynamics research. These technologies keep track of our lives with digital records of places we go, products we buy, and people we meet. Human dynamics research with data from limited observations or confined experiments has transformed into tons of data records on human communications, interactions, and activities in the naturalistic environment.

In Chapter 2, we study the possibility of user profile inference using privacy-sensitive audio. The contributions are three folds. First, we propose a privacy-sensitive modality for gender identification. The effectiveness and robustness are improved by ensemble feature selection and a two-stage classification. Second, an adaptive correlation-based multichannel VAD algorithm for privacy-sensitive audio is proposed. Last, we bring new insights of gender difference in interruption through analysis of group conversation in natural settings.

In Chapter 3, we utilize the WiFi data to infer relational contextual information. One of our contributions is an effective heuristic that could significantly improve the detection performance of shopping groups. The heuristic indicates APs under which groups appear more frequently and barely separate should have larger weights in measuring customer similarity. The second

contribution is to apply matrix factorization to detect groups without extra clustering processes. Matrix factorization could properly handle data issues in the measured similarity including noise filtering and data completion. Besides, imposing a sparsity constraint to the factorization process could derive the clustering results directly.

In Chapter 4, we explore the relative contextual information based on the WiFi data and study the impact of human presence on wireless coverage. We identified the correlation between wireless coverage and the number of on-site people. Another contribution is the two observations of heuristics which could improve room-level localization. On the one hand, the duration of visit in different shops is different. On the other, different shops have different popularity in attracting customers at different time slots. These two features can be exploited to distinguish locations with similar wireless fingerprints.



# List of Publications

1. **J Shen**, O Lederman, J Cao, S Tang, and A Pentland. “Trying to Be Heard: Gender and Group Dynamics”.<sup>†</sup>
2. **J Shen**, J Cao, O Lederman, S Tang, and A Pentland. “LEO: Automatic Personality Recognition Using Privacy-Sensitive Audio”.<sup>†</sup>
3. **J Shen**, J Cao, and X Liu. “BaG: Behavior-aware Group Detection in Crowded Urban Spaces using WiFi Probes”. Accepted by *The Web Conference (WWW)*, 2019.
4. **J Shen**, O Lederman, J Cao, F Berg, S Tang, and A Pentland. “GINA: Group Gender Identification Using Privacy-Sensitive Audio Data”. Accepted by *IEEE International Conference on Data Mining (ICDM)*, 2018.<sup>‡</sup>
5. **J Shen**, J Cao, X Liu, and S Tang. “SNOW: Detecting Shopping Groups Using WiFi”. Accepted by *IEEE Internet of Things Journal (IoT-J, IF: 7.596)*, 2018.<sup>‡</sup>
6. **J Shen**, Y Lau, and J Cao, “Data-Driven Mall Advertising”, a chapter in *Smart Marketing with the Internet of Things (IGI Global)*, 2018.
7. **J Shen**, J Cao, X Liu, and C Zhang. “DMAD: Data-Driven Measuring of Wi-Fi Access Point Deployment in Urban Spaces”. Accepted by *ACM Transactions on Intelligent Systems and Technology (TIST, 5-year IF: 10.47)*, 2017.<sup>‡</sup>

---

<sup>†</sup>In submission

<sup>‡</sup>Presented in this thesis

8. **J Shen**, J Cao, X Liu, J Wen, and Y Chen. “Feature-Based Room-Level Localization of Unmodified Smartphones”. In *Smart City 360°*, 2016.
9. E Wen, J Cao, **J Shen**, and X Liu. “Fraus: Launching Cost-efficient and Scalable Mobile Click Fraud Has Never Been So Easy”. In *IEEE Conference on Communications and Network Security (CNS)*, 2018.
10. X Liu, J Wen, S Tang, J Cao, and **J Shen**. “City-Hunter: Hunting Smartphones in Urban Areas”. In *IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2017.
11. Y Sahni, J Cao, **J Shen**, “Challenges and Opportunities in Designing Smart Spaces”, a chapter in *Internet of Everything* (Springer), Kuan-Ching Li, Beniamino DiMartino, Laurence Yang and Antonio Esposito (Eds.), 2017.
12. Y Chen, M Guo, **J Shen**, and J Cao. “A graph-based method for indoor subarea localization with zero-configuration”. In (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld) [**Best paper**], 2016.
13. Y Chen, M Guo, **J Shen**, and J Cao. “GraphLoc: a graph-based method for indoor subarea localization with zero-configuration”. In *Personal and Ubiquitous Computing*, 2016.

# Acknowledgements

First of all, I would like to pay my highest level of gratitude to my supervisor Prof. Jiannong Cao for his continuous efforts, great patience, and insights comments. Besides, Prof. Alex ‘Sandy’ Pentland, a wise and kind professor from MIT Media Lab, also helped me a lot during my visit to his lab. Both professors taught me so much about research and life. Without their guidance, I would never make it!

During my PhD, if I made any tiny achievement, that must owe to my advisors and co-authors. I really appreciate the efforts and suggestions from Dr. Xuefeng Liu, Mr. Jiaqi Wen, Dr. Shaojie Tang, Mr. Yuvraj Sahni, Dr. Chisheng Zhang, Dr. Yuanyi Chen, and Mr. Oren Lederman. I also appreciate the assistance from Dr. Yuhong Feng and her students for helping the experiment.

Special thanks to my office colleagues, Dr. Yuqi Wang, Mr. Zimu Zheng, Mr. Yanwen Wang, Dr. Bo Tang, Mr. Yuan Huang, Mr. Zhe Li, Dr. Andy He, and Dr. Yu Li for their companion, support, and tolerance. We used to have a lot of happy hours. Thank you, guys!

Over this time, I also got the opportunity to work and learn from many brilliant researchers. They are Dr. Guobin Shen, Dr. Yuanqing Zheng, Dr. Xiulong Liu, Dr. Wengen Li, Dr. Lei Yang, Dr. Junbin Liang, Mr. Rui Liu, Dr. Fuliang Li, Dr. Zongjian He, Dr. Guangqin Liang, Dr. Gang Yao, Dr. Hongliang Lu, Dr. Weiqin Liu, Dr. Xuanjia Qiu, Dr. Tao Li, Ms. Wanyu Lin, Dr. Linchuan Xu, Dr. Chao Ma, and Mr. Yaguang Huangfu.

Many promising young researchers from our group gave me plenty of support. I appreciated it! They are Mr. Yu Yang, Mr. Shan Jiang, Ms. Yanni Yang, Mr. Hanqing Wu, Mr. Dan Li, Mr. Ruosong Yang, Mr. Zhuo Li, Ms. Jia Wang, Mr. Zhiyuan Wen, Mr. Jinlin Chen, Mr. Chuntao Ding, and many others.

I made a lot of friends in PolyU. Appreciate them for sharing my happiness and sorrows. I used to play badminton with Dr. Tiangang Wang, Dr. Zhibo Wang, and Dr. Minglei Li. Miss that time so much! I also need to thank Dr. Yadie Yang, Dr. Wenjian Xu, Mr. Quanyu Dai, Mr. Lei Han, Dr. Yunfei Long, Dr. Yutian Tang, Dr. Shang Gao, Dr. Zhe Peng, Dr. Zhaoyan Shen, Ms. Jin Xiao, Ms. Xinye Lu, Ms. Chengyao Chen, Dr. Zi Yang, and many others. I also appreciate the warm assistance from staff of our department. They are Ms. Carmen Au, Ms. Christy Au, Ms. Jolie Chick, Ms. Anna Cheng, and many others.

During my one-year visit to MIT, I made new friends there and they always helped me with pleasure. They are Ms. Yan Leng, Mr. Yuan Yuan, Mr. Yafei Zhang, Mr. Chuang Deng, Mr. Longxu Yan, Ms. Yuanye Deng, Ms. Tianni, and many others.

For some special friends who might never read this, thank you! I would never forget those moments we used to have.

At last, I would like to express my sincere appreciation to my parents and other family members. I love you so much!

# Table of contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background . . . . .	1
1.1.1	Human Dynamics Analytics . . . . .	1
1.1.2	Privacy-Sensitive Data . . . . .	6
1.2	Research Focus . . . . .	8
1.2.1	Research Challenges & Methodologies . . . . .	9
1.2.2	Research Framework . . . . .	10
1.3	Literature Review . . . . .	12
1.3.1	Taxonomy of Modeling Emphases . . . . .	12
1.3.2	Taxonomy of Input Modalities . . . . .	13
1.4	Thesis Organization . . . . .	16
<b>2</b>	<b>GINA: Group Gender Identification Using Privacy-Sensitive Audio Data</b>	<b>17</b>
2.1	Introduction . . . . .	18
2.2	System Overview . . . . .	20
2.3	System Design . . . . .	21
2.3.1	Privacy-Sensitive Data Collection . . . . .	22
2.3.2	Voice Activity Detection . . . . .	22
2.3.3	Conversational Feature Extraction . . . . .	25
2.3.4	Group Gender Identification . . . . .	30
2.4	Experimental Evaluation . . . . .	32
2.4.1	Settings . . . . .	32
2.4.2	Evaluation . . . . .	33
2.5	Related Work . . . . .	36
2.5.1	Gender detection . . . . .	37
2.5.2	Gender differences and interruption . . . . .	38
2.6	Conclusion . . . . .	39
<b>3</b>	<b>SNOW: Detecting Shopping Groups Using WiFi</b>	<b>41</b>
3.1	Introduction . . . . .	41
3.2	System Design . . . . .	45

3.2.1	Data Collection . . . . .	45
3.2.2	Similarity Measurement . . . . .	47
3.2.3	AP Significance Estimation . . . . .	51
3.2.4	Group Detection . . . . .	52
3.3	Experimental Evaluation . . . . .	54
3.3.1	Settings . . . . .	54
3.3.2	Evaluation . . . . .	56
3.4	Related Work . . . . .	60
3.5	Discussion . . . . .	61
3.6	Conclusion . . . . .	62
<b>4</b>	<b>DMAD: Data-Driven Measuring of Wi-Fi Access Point Deployment</b>	<b>63</b>
4.1	Introduction . . . . .	64
4.2	Overview . . . . .	66
4.2.1	Preliminaries . . . . .	67
4.2.2	Framework of DMAD . . . . .	73
4.3	System Design . . . . .	74
4.3.1	Data Collection . . . . .	74
4.3.2	Device Classification . . . . .	81
4.3.3	Area Localization and Density Calculation . . . . .	82
4.3.4	Dead Spots Estimation . . . . .	87
4.4	Experiments and Results . . . . .	88
4.4.1	Setup . . . . .	89
4.4.2	Evaluation . . . . .	90
4.5	Related Work . . . . .	97
4.6	Conclusion . . . . .	101
<b>5</b>	<b>Conclusion &amp; Future Work</b>	<b>103</b>
	<b>References</b>	<b>105</b>

# List of Figures

1.1	Illustration of human dynamics with related areas and confused notions. . . . .	2
1.2	Research framework. . . . .	11
1.3	The three work in the 3D design space. . . . .	12
2.1	Overview of GINA. . . . .	21
2.2	An example result of multichannel VAD on a meeting with four participants between 18:12:50 and 18:13:50. . . . .	25
2.3	Illustration of conversational features. Underlined bold text represent turn-taking features, the other bold text represent interruption features. . . . .	25
2.4	Effectiveness analysis of turn-taking features. (a) ~ (d) PDFs of different features; (e) ROC curves of features. . . . .	26
2.5	Analysis of who interrupts who with PDFs of four-class interruption and results of Mann-Whitney U test for different types of interruption. (a) Type I interruption; (b) Type II interruption; (c) Type I and Type II interruption. . . . .	27
2.6	Analysis of inters under three types of interruption. . . . .	28
2.7	Analysis of intees under three types of interruption. . . . .	29
2.8	Feature importance of all the features in a Random Forest consisting of 100 trees. . . . .	29
2.9	An illustration of ensemble feature selection and the two-stage classification in an iteration of cross-validation. . . . .	30
2.10	Stacked histogram of number of meetings and meeting duration for all study groups. . . . .	33
2.11	Illustration of baseline approaches. . . . .	33
2.12	The impact of different $\theta$ . . . . .	35
2.13	Performance of gender composition detection with different models. . . . .	35
2.14	Comparison of performance using different classification models. (a) Precision; (b) Recall; (c) F1-score. . . . .	36

3.1	Answers to the online survey problem: “How often will you get separated with your companion(s) in these regions?” . . . . .	42
3.2	A toy example for illustrating the main idea of SNOW. . . . .	43
3.3	An overview of SNOW. . . . .	45
3.4	The three-step preprocessing of the WiFi data. . . . .	47
3.5	Distance distributions of groups and non-groups using RSS. . . . .	48
3.6	Distance distributions of groups and non-groups using RSS trend. . . . .	48
3.7	ROC curves of using RSS and RSS trend. . . . .	48
3.8	RSS of group members using different smartphones. In both cases, group members stick together all the time, but there exist gaps in their RSS signals. . . . .	50
3.9	A simple example for calculating AP weights. . . . .	52
3.10	An illustration of partitioning customers into different crowds with the temporal constraint. . . . .	53
3.11	Detailed information of the collected data in 3 weeks. . . . .	55
3.12	An illustration of baseline approaches. . . . .	56
3.13	Distribution and CDF of customers’ dwell time in the mall. . . . .	57
3.14	Impact of the amount of Bluetooth data in estimating AP weights. . . . .	59
3.15	Performance under different AP density on semi-labeled dataset. . . . .	59
4.1	A simple illustration of the impact of human beings on wireless coverage. (a) Ideal coverage of APs; (b) Real coverage of APs in the presence of walking people. The shadow areas in (b) are potential dead spots caused by human beings. . . . .	65
4.2	Example of using Wi-Fi data to represent wireless coverage status. The unit of $T_c$ and $T_t$ are both minute. Coverage ratio $\rho = T_c/T_t$ . . . . .	68
4.3	Illustration of using connectivity matrix $M$ to calculate coverage ratio vector $\Omega$ . $and()$ does the AND-operation of the column vector. . . . .	69
4.4	Typical examples of different coverage statuses at different locations from 12:00 ~ 13:00. The black bar of an AP indicates the AP can “hear”* from the device on that location. Intuitively, the order of coverage status is: Case I > Case II > Case III. . . . .	70
4.5	Illustration of translating a connectivity matrix into probability of dead spots. . . . .	71
4.6	Correlation analysis of the average <i>ratio of change</i> and the average <i>number of people</i> from the data collected in a shopping mall in Shenzhen for 46 days. . . . .	73



4.7	The framework of DMAD. It consists of four components, data collection, device classification, area localization and density calculation, and dead spots estimation. . . . .	74
4.8	AP deployment and expected coverage area on the ground floor of the mall. . . . .	75
4.9	Grid partition on expected coverage area of the ground floor. . .	75
4.10	A simple illustration of the Wi-Fi data collection. . . . .	75
4.11	Flow chart of collecting Wi-Fi data in APs. . . . .	76
4.12	Histogram of duration time of around 100 customers from three different shops. (a) A fast food restaurant; (b) A traditional Chinese restaurant; (c) A woman accessories shop. . . . .	79
4.13	Distribution of total number and surveyed number of shops. . .	80
4.14	Regression analysis between some predictor variables and mean of the duration distribution. . . . .	81
4.15	Regression analysis between some predictor variables and standard deviation of the duration distribution. . . . .	81
4.16	Decision tree for classifying static and mobile devices. . . . .	81
4.17	Raw data of a static device on 5 June 2015. . . . .	82
4.18	Impact of different $a$ and $b$ on the percentages of static and mobile devices from $D_w$ . . . . .	91
4.19	Precision and recall of static device classification and mobile device classification. . . . .	91
4.20	Precision and recall of static device classification with different $c$ . . .	91
4.21	Average number of static and mobile devices of each day in a week. . . . .	92
4.22	Average number of people appear in different hours of a day. . .	92
4.23	Comparison of CDFs of total duration time of non-static devices in each day. . . . .	92
4.24	Accuracy of floor localization and grid localization for different methods. . . . .	93
4.25	The impact of different $k$ on the accuracy. . . . .	93
4.26	Heat map of human density on the ground floor in different time slots. (a) 9:00; (b) 12:00; (c) 18:00. . . . .	94
4.27	Distribution of fingerprints and devices used in collecting fingerprints. We use devices from three manufacturers to collect fingerprints for area localization. . . . .	95
4.28	Confusion matrix of localization accuracy using different kinds of devices for training and testing. . . . .	95

4.29	Precision, recall, and F-score of dead spots estimation under different $e$ . . . . .	96
4.30	CDF of $P_{DS}$ of grids from different floors. . . . .	96
4.31	Heat map of severity of grids on the ground floor in different time slots. (a) 9:00; (b) 12:00; (c) 18:00. . . . .	97
4.32	F-score of linear and exponential decay functions under different $e$ . . . . .	97
4.33	Pareto chart of additional dead spots. The sorted grids are equally separated into 10 groups. . . . .	97

# List of Tables

3.1	Performance comparison on both datasets. . . . .	57
4.1	Notions used in this chapter. . . . .	67
4.2	A fraction of raw Wi-Fi data. MAC has been hashed. . . . .	76
4.3	A fraction of unlabeled shop data. Some of fields like, floor, location, and average spend is not shown in the table. . . . .	80
4.4	Details of the testing data. . . . .	90
4.5	Confusion matrix of device classification. . . . .	90



# Introduction

This chapter consists of four main sections: Background, Research Focus, Literature Review, and Thesis organization. Firstly, the background information including human dynamics and privacy-sensitive data is introduced in Background. Then, in Research Focus, I elaborate on research challenges, general methodologies, and the research framework. After that, two taxonomies of existing works are presented in Literature Review. Lastly, Thesis Organization reveals the overall structure of this dissertation.

## 1.1 Background

This section is to provide readers with essential knowledge of the motivation and background of this dissertation. Two key concepts, including human dynamics analytics and privacy-sensitive data, are highlighted.

### 1.1.1 Human Dynamics Analytics

To understand human dynamics, we firstly give the basic definition, then show its development of different times, and lastly, provide some emerging trends of human dynamics research from the perspective of computer science.

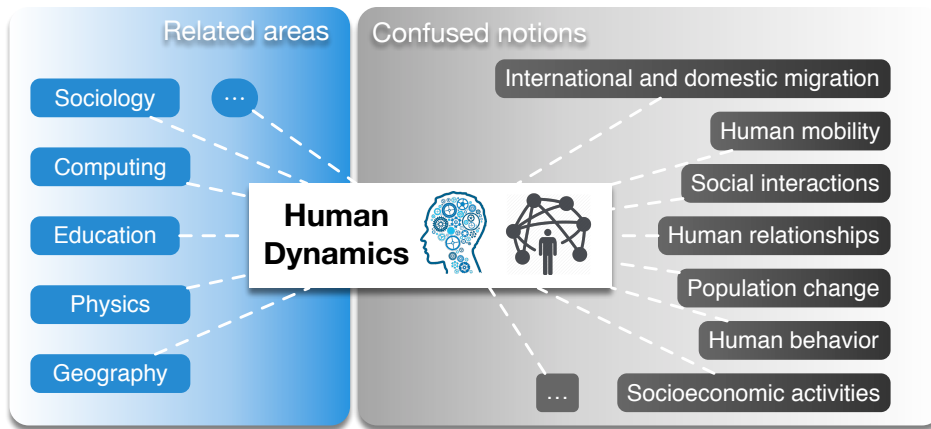
#### **Definitions of human dynamics**

Human dynamics is interdisciplinary research which roots in human history [135]. It is defined in Wikipedia\* as “a branch of complex systems research in statistical physics whose main goal is to understand human behavior”.

With the flourish of modern technologies, human dynamics has now been widely used and extensively investigated in various disciplines from different dimensions [136]. As illustrated in Figure 1.1, many fields like geography, psychology, and sociology define human dynamics from different perspectives which leads to somewhat different research focuses. For example, in physics,

---

\*[https://en.wikipedia.org/wiki/Human\\_dynamics](https://en.wikipedia.org/wiki/Human_dynamics)



**Fig. 1.1:** Illustration of human dynamics with related areas and confused notions.

the focus of human dynamics is the force of change and the dynamic spatiotemporal relationships of the observed objects at both micro (individual) and macro (group) perspectives [136]. The focuses in computational social science are investigating social and behavioral relationships and interactions through social simulation, modeling, network analysis, and media analysis [79]. Computer scientists are interested in developing new spatial analytics, data mining, and machine learning approaches to understand the disaggregated data of human activities from web services, social media, and various open data sources [175].

Without a commonly accepted definition, human dynamics has been confused with many topics like social interactions, human mobility, human relationships, socioeconomic activities, population change, international and domestic migration, and other variants pertinent to human activities which are shown in Figure 1.1. Human dynamics in this thesis is defined as combination of human activities and interactions according [175]. Human activities are various things we do in our daily lives like talking, walking and shopping, while human interactions emphasis more about the interplay within an activity. For example, we may talk other people, which is a human-human interaction. We may also talk to a social robot, which is human-computer interaction on the other hand.

Studying human dynamics could help us gain a comprehensive understanding of human societies including human activities like the evolving patterns over time and space and human interactions like how social influence propagate, and how people collaborate with each other. Besides, it also enables various applications ranging from business intelligence to public health. In busi-

ness intelligence, we could predict organization sustainability via inspecting gender inequality; improve service quality by examining facility utilization; enhance marketing campaigns by looking into the composition of target audiences. In public health, we could predict the spread of infectious disease via analyzing social network; treat mental disorders through detecting early symptoms.

## **Development of human dynamics**

The term of human dynamics is borrowed from physics, which investigates dynamics by studying objects' motions and movements using mathematic equations and physical laws. As one of the early proposers of "human dynamics", Finch emphasized the value of regional geography and indicated that "forces of human dynamics" are not amenable to direct observation [38]. At that time (the 1930s), it was quite challenging to observe the forces of human dynamics, let alone human dynamics. This was mostly due to the deficiency of tools or technologies to collect the observational data, especially at a large scale. Traditional data collection methods like interview and survey methods are labor-intensive and time-consuming in collecting and recording human activities and interactions. As a result, it is prohibited to apply theoretical frameworks to examining human dynamics at a community or society level.

With the rapid development of digital information technologies, including information and communication technology, sensing technology, location-aware technology, and mobile technology, tremendous convenience has been brought into urban lives. More importantly, it has dramatically changed the patterns of human activities and interactions [135]. Although the fundamental human needs remain almost the same as before, the ways we fulfill these activities have changed significantly owing to these modern technologies. For instance, online social networks and mobile social network have reshaped the way people connect with friends through numerous services and information available on the Internet. Laptops and smartphones are other examples which changed the way we interact with the world. With an Internet connection and appropriate devices, it is now feasible to conduct various office tasks anywhere. When looking for a dining place, the recommendation made by strangers from smartphone apps could often be a superb choice.

Modern technologies have not only introduced changes to human dynamics but also enabled our capability of collecting the detailed data about human dynamics. As indicated by Pentland, when we enjoy the convenience brought by those technologies we also leave behind many virtual “bread crumbs”—digital records of places we go, products we buy, and people we meet [112]. These bread crumbs could tell a more comprehensive and accurate story of our daily lives than we could in the interview or survey. We might carefully update Facebook status and deliver tweets according to some standards of the day. On the contrary, digital bread crumbs could record our behavior that actually happened in an authentic way. In general, these digital records successfully transformed human dynamics research with data from limited observations or confined experiments into tons of data records on human communications, interactions, and activities in the naturalistic environment. A mobile and big data era for human dynamics research has arrived ever since [136].

Although the dilemma of lacking data has been alleviated by the mobile and big data age, our knowledge about the implications of human dynamics to the communities is still too shallow to answer important questions in human dynamics. To make smart decisions for a better future of our communities and societies, more efforts are required to gain insights into human dynamics.

### **Emerging trends of human dynamics**

This subsection aims to list emerging questions of importance in human dynamics according to the literature.

1. **User privacy erosion:** Due to the prevalence of IoT devices and social media services, privacy concerns and information disclosure risks are becoming a major concern for both researchers and the public. A common privacy concern is the leakage of user locations when using “check-in” functions in social media to reveal their physical locations [86]. Another type of privacy concern relies in collecting spontaneous data in a naturalistic environment since it requires recording video or audio of people in unconstrained and unpredictable situations, both public and private. There is little control over who or what might be recorded. Private content and uninvolved parties could be recorded without their consent [168]. To alleviate the privacy issues, methods like obfuscation [144] might be helpful in hiding sensitive locations in



Big Data. Besides, privacy-sensitive data modality should be considered with a priority in studying human dynamics.

2. Incomplete user profiles: Conventional ways of data collection like interview and survey could build a comprehensive user profile including gender, age, occupation, and other demographics. However, many datasets especially big data sources collected by IoT devices do not contain such detailed demographic information. Without knowing demographics, human dynamic research based on big data and social media data might be biased. For example, as reported that the actual users of social media services are mostly from the young generation. Around 75% of Twitter users are in the age range of 15 ~ 25<sup>†</sup>. Therefore, data and messages collected from social media only represent a small fraction of the whole population [150]. Further research work is needed to infer demographics effectively and reliably in the mobile and big data era.
  
3. Missing contextual information: Every single human activity takes place within a context. The focus of human dynamics research is not only just about human but also the environment and the situation they interact with. The environment always plays an essential role in understanding human dynamics since it could influence and reshape human behaviors. Based on the enumeration of context by Shaw and Sui [135], there are at least three context spaces: relative space, relational space, and mental space. First, relative context space contains the basic spatiotemporal information of an individual. Second, relational context space is about the various relationships among different entities. Third, mental context space includes emotion, perception, motivation, etc. Unfortunately, many data sources, especially those collected from natural settings, lack such context information which hinders a better understanding of human dynamics. Recent research interests in places and semantics are good examples of deriving meanings behind human dynamics based on the context.

---

<sup>†</sup>[www.beevolve.com](http://www.beevolve.com)

## 1.1.2 Privacy-Sensitive Data

With the development of human dynamics, privacy is becoming a sensitive yet important aspect. We first introduce what is privacy-sensitive data. Then the motivation of using such data modalities are described. Lastly, we provide some example privacy-sensitive data modalities used in this dissertation.

### Definition of privacy-sensitive data

Privacy possesses different meanings in different situations. For an individual, the privacy concern usually arises from sensitive personally identifiable information (PII)<sup>‡</sup> which might result in crimes like identity theft if disclosed. PII refers to data that could be traced back to an individual and lead to harm to that person if disclosed. Typical PII include biometric data, personally identifiable financial information (PIFI), medical information and unique identifiers such as passport and social security number. For an organization, privacy refers to any information that poses a risk to the organization if discovered by the public or a competitor. Some of the typical information are acquisition plans, financial data, trade secrets, supplier and customer information.

European Community regulates personal data [36]

‘personal data’ shall mean any information relating to an identified or identifiable natural person (‘data subject’); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity;

and specifies that data can be

kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Member States shall lay down appropriate safeguards for personal data stored for longer periods for historical, statistical or scientific use.

---

<sup>‡</sup><https://whatis.techtarget.com/definition/sensitive-informations>

The key thing is “identifiable”. As long as the data cannot be traced to an individual, the regulations do not apply. Therefore, *privacy-sensitive data* (or privacy-preserving data) in this thesis is defined as any information that satisfies the following two requirements [29].

1. The data could not be used to trace back to an individual or an organization directly (Untraceability).
2. The data do not contain any private content that requires the consent of the involved parties (Privacy).

However, absolute privacy cannot be guaranteed. Even though the data might be processed with proper anonymization, researchers found that anonymity cannot be promised. For example, MIT researchers studied credit card records of 3 months from 1.1 million people and found that 4 spatiotemporal points are enough to uniquely re-identify 90% of individuals [33]. Besides, knowing the price of the transaction or the gender of the individual could further increase the risk of reidentification.

### **Motivation of using privacy-sensitive data**

As introduced in Section 1.1.1, privacy concern is increasingly serious, which might be the greatest barrier to the development of human dynamics in the mobile and big data era. There exist at least two reasons motivate us to use privacy-sensitive data.

An obvious advantage is to alleviate the severity of privacy erosion, especially for large-scale data collection. An example scenario is the shopping mall where video surveillance is usually used for customer monitoring. But it might be inappropriate to use the video data for customer analysis. The capture video data has a high risk of privacy erosion since certain customers could be recognized, which violates the requirement of anonymity. In this case, WiFi data might be a better choice since it is more privacy-sensitive. The detailed explanation is described in “examples of privacy-sensitive data modalities”.

Another advantage is to raise the possibilities of studying human dynamics in many private spaces. Previously, many experiments of human dynamics are conducted in the controlled lab environment, which might not capture

the latent factors in the naturalistic environment. To capture the real-world human activities, privacy is a big concern. With privacy-sensitive data, study human dynamics in natural setting gradually come into reality.

## **Examples of privacy-sensitive data modalities**

Although privacy cannot be guaranteed, its concern could be greatly reduced by privacy-sensitive data. In the mobile and big data age, various data modalities have been explored and some of them are naturally more privacy-sensitive. For example, the WiFi data collected from smartphones are more sensitive than video data in terms of privacy. Unlike individuals could be recognized directly in video data, people are identified by virtual proxies, MAC addresses of smartphones, in WiFi data. Since there is a gap between the virtual proxy and the real identity, WiFi data is more privacy-preserving than traditional video data.

We could also achieve privacy-sensitive modalities by appropriate processing of traditional modalities. Take audio for example, assuming access to raw audio is impractical for most real-world situations and impedes collecting truly natural data [168]. An alternative is to collect privacy-sensitive audio [81]. The microphone signal is sampled at 700 Hz and generates an average amplitude reading every 50 milliseconds to ensure raw audio is not recorded nor can it be reconstructed.

## **1.2 Research Focus**

The main theme of the thesis is data-driven analytics of human dynamics based on privacy-sensitive data. In this section, we will introduce the main research challenges, general methodologies and a research framework covering the works presented in this thesis.

## 1.2.1 Research Challenges & Methodologies

### **C1: Low quality of privacy-sensitive data**

As introduced in Section 1.1.2, privacy-sensitive audio data has a much lower sampling rate than normal raw audio and it is further processed with a mean filter. Actually, most privacy-sensitive data could alleviate the privacy concern since they are significantly less informative. Even for WiFi data, it is naturally sparse and noisy as the timing of sending wireless packets is opportunistic and wireless signal is vulnerable. These indicate privacy-sensitive data are usually low-quality in terms of data granularity and data purity.

One of the challenging issues caused by low data resolution is the difficulty to extract adequate and effective features for high-level applications. For example, it is barely possible to extract popular voice features like pitch and first formant from privacy-sensitive audio data. Other issues like data missing and noisy readings also pose serious challenges. Take the WiFi data for instance, without appropriate processing, it is difficult to calculate a similarity between pairwise smartphone users.

To address the aforementioned challenging issues, we propose the following approaches as a general guidance.

- Integrate knowledge from other domains and devise new features. An example is presented in Chapter 2 where we extracted conversational features rather than voice features to identify gender. As indicated in sociology literature [122, 181, 50, 104], the way people take turns and interrupt each other could also reveal their gender information.
- Fuse data from multiple sources since additional information are usually beneficial. As illustrated in Chapter 4, we use a probabilistic approach to locate customers based on the WiFi data. However, due to the sparsity of the data, it is difficult to achieve satisfactory performance. Therefore, we fuse the PoI data and derive a more accurate prior probability.

### **C2: Dynamics of human behavior**

Human behavior is dynamic in nature. We all know that different people may behave differently. Even for the same person, his or her behaviors

keep changing. People may behave differently in different situations due to environmental, personal, and behavioral factors. A typical example is the shopping behavior. When shopping with companions, some people choose to walk together through the whole process, while other people might get separated from their companions from time to time. Take the conversation behavior of a group of people as another example. People usually interrupt each other during a conversation. However, researchers find that interruption is more evenly distributed in same-gendered group conversations [103]. Besides, customers' indoor mobility patterns also reveal dynamic characteristic. Group customers have different PoI preferences from individual customers.

The variation in human behavior poses a serious challenge to the effectiveness and robustness of system performance, sometimes even results in a new research problem. For instance, in gender identification, the dynamics of conversational behavior significantly influence the performance since the extracted conversational features have large variations and are thus difficult to capture group dynamics. For social group detection, as group customers happen to separate sometimes, it is difficult to address shopping group detection with conventional detection methods since their assumption is that a social group do not separate. This makes shopping group detection a new research problem.

To address the dynamics of human behavior, an effective way is to infer the contextual information first. As described in Introduction, every single human activity takes place within a context, the environment and the situation people interact with. This contextual information could affect human behavior and could thus partially explain the dynamics of human behavior. An example is shown in Chapter 2, we infer the gender composition as an extra input for gender identification since the composition plays a latent role in people's turn-taking behaviors and interruption patterns.

## 1.2.2 Research Framework

As shown in Figure 1.2, the research framework consists of 6 layers. On the bottom two layers, we collect PS data from different human activities and interactions including talking, walking and shopping. In certain scenarios, we

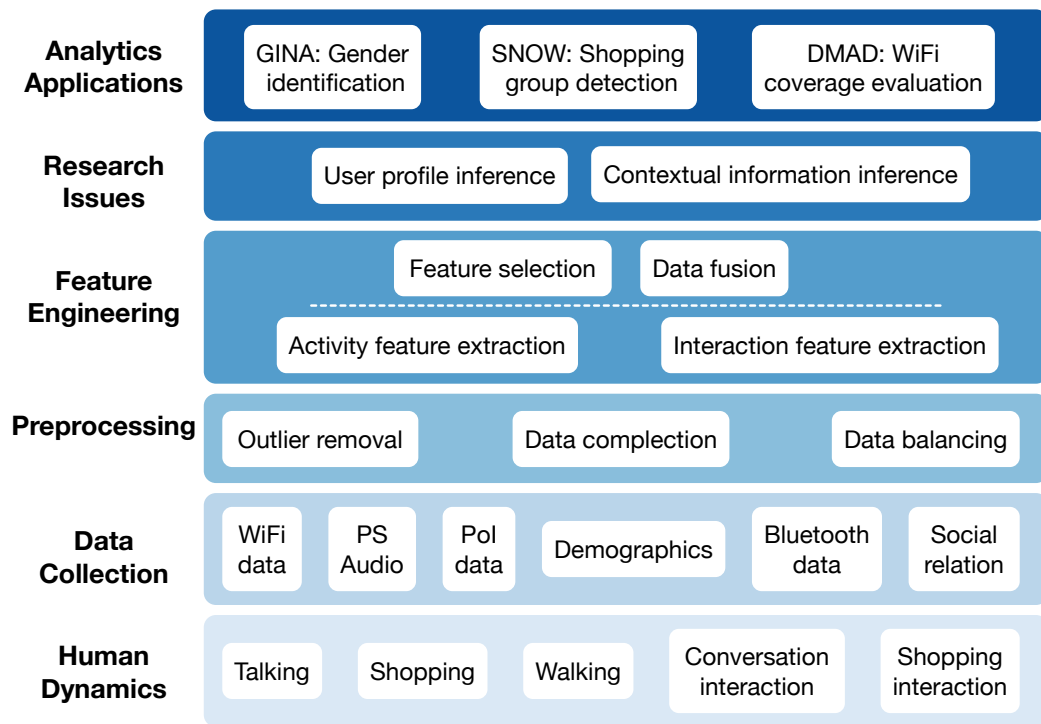
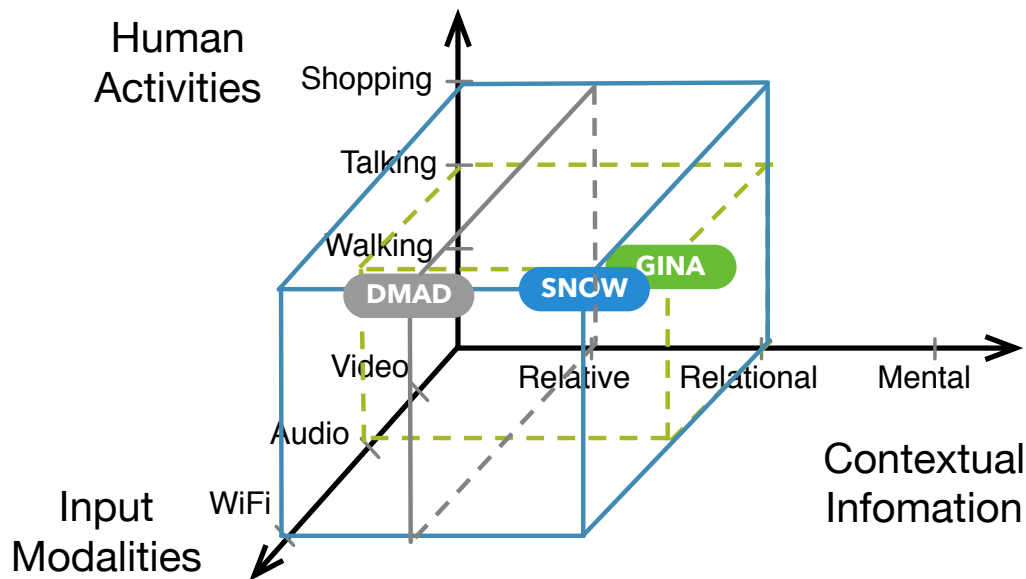


Fig. 1.2: Research framework.

also need to fuse other information like PoI and Blue-tooth data to improve system performance.

To clean the raw data and make preparation for extracting features, we need to address some common data issues including outliers, data missing, and data imbalance in Preprocessing Layer. For example, to address data missing and noisy readings in the WiFi data, we have to remove outliers and apply certain data completion techniques like matrix factorization. More details could be found in the following three chapters. Feature Engineering Layer is to extract effective features from the processed raw data. In some scenarios, we need to rely on the related domain knowledge to devise features. Take the voice based gender identification as an example, voice features are mostly used. But we also found that communication styles could also reflect gender difference from sociology literature and devise conversational features. In Research Issue Layer, we mainly address user profile inference and contextual information inference as both issues are increasingly popular in studying human dynamics. Lastly, in Analytics Applications Layer, we focus on different tasks like gender identification, group detection, and wireless coverage estimation.



**Fig. 1.3:** The three work in the 3D design space.

Figure 1.3 illustrates the location of the three works in the same 3D design space. The three dimensions are human activities (e.g., shopping and talking), contextual information (relative context, relational context, and mental context), and input modalities (e.g., WiFi and PS audio). For example, GINA use PS audio data to study talking behavior in the relational space.

## 1.3 Literature Review

Due to the immense efforts on human dynamics, there are various taxonomies of literature. In this thesis, we review existing works from the perspectives of modeling emphases and input modalities, respectively. The taxonomy of modeling emphases, the dimension of time and space, is a direct and classical way of literature review. Besides, to illustrate the significance and necessity of privacy-sensitive data in studying human dynamics, we also show a taxonomy of different data modalities.

### 1.3.1 Taxonomy of Modeling Emphases

A direct and natural way of examining the development of human dynamics is based on the modeling emphases: time or space. Time information could be effortlessly recorded owing to the system clock of every digital device. One of



the most notable work is done by Barabási. In his work [9], he found that the distribution of human activities is not random over time. The occurrences of human activities follow the Pareto distribution which is found in many areas ranging from communication to entertainment. More specifically, human activities occur rapidly in bursts followed by extended periods of inactivity. According to Barabási, Pareto statistics reflect some fundamental and generic features of human dynamics among various human activity patterns [175].

After that, the research focus of various human dynamics studies became modeling the timing, frequency, waiting time of human activities and interactions. For example, Zhou et al revealed the origin of power-laws in the ratings of movies and presented a systematic exploration of the time intervals between two consecutive ratings of movies [180]. Gonçalves and Ramasco analyzed web logs generated by university students [47] and Oliveira and Vazquez focused on the inter-event time of interactions [111].

With the rapid development of social networks, the ability to get mobile users' physical location information drives human dynamics development from the temporal dimension to the spatial dimension. As one of the most accessible social media data, Twitter data have simulated various researchers to explore the space dimension in human dynamics. For example, Vosoughi et al studied the spread of true and fake news online in a network space [155]. Others also examined semantics meanings of individual visits [6], international and domestic migration patterns [176]. Many researchers also utilize the detailed phone call records to examine human dynamics, like the daily rhythms of city life [3], work-home commuting patterns [76], and distribution of human convergence or divergence [37].

While research of human dynamics in the time dimension reveals the law of burst and heavy-tailed distributions of human activities, research in the space dimension remains exploratory, descriptive, or forensic [175].

### 1.3.2 Taxonomy of Input Modalities

During the development of human dynamics, different data modalities have been explored including traditional modalities like records from survey, video, and audio as well as emerging ones like WiFi data, phone call records, and social media data.

Different modalities usually have unique advantages in certain scenarios. With increasing awareness of privacy concerns, privacy-sensitive modalities should be preferred when multiple modalities are available. Take crowd density estimation as an example which aims to infer the quantity of people in a given area. Both video [117] and WiFi data [162] are feasible options, the latter should be of higher priority. As mentioned WiFi is more privacy-sensitive than video since user identity could be directly revealed in the video surveillance.

## **Video**

Video is among the most popular modalities used for human dynamics due to its prevalence. In areas like biometric research, researchers are interested in understanding and interpreting human behavior in complex environments using video [150]. The most basic steps to understand human dynamics is the ability of tracking individuals in video sequences. Besides, video surveillance is also used for large-scale crowd analysis [124].

For gender identification, vision-based approaches exploit information from the face and whole body (either from a still image or gait sequence) to recognize human gender. It is usually based on appearance differences like face and body, and behavior differences like gait [108].

For group detection, vision-based approaches regard group detection as a task of clustering a set of users' trajectories into disjoint subsets [45, 142].

## **Audio**

Audio is another common modality for understanding human dynamics, focusing on the perspective of communication and interaction. Researchers could examine emotion and information flow within an organization could be captured through communication [168, 81].

For gender identification, voice-based methods rely on discriminative features extracted from human voices. The intuition is that different genders have different acoustic characteristics due to physiological differences (like glottis, vocal tract thickness) [4] and phonetic differences [141]. The most frequently used features are pitch [60] and first formant [119], which are closely related

to voice sources and vocal tract, respectively. Generally, the pitch and the formant frequencies of females are higher than that of males.

## WiFi

WiFi data is becoming increasingly popular due to the penetration of smart devices like smartphones. Media Access Control (MAC) address of each WiFi-enabled device is a proxy of each smartphone user. Since the WiFi data could be collected in a passive and non-intrusive way, many researchers utilized WiFi data to study human dynamics in many real-life scenarios [71, 56].

For group detection, WiFi probe is increasingly popular. The probe contains significant information like timestamp, smartphone MAC address, RSSI, and Service Set Identifier (SSID), which enables a wide range of applications like passive tracking [35, 139], crowd counting[128, 169], and facility utilization analysis [114]. Compared to other approaches, probe-based approaches do not require high deployment cost or user intervention. SSID and RSSI are two frequently used information to detect groups. Cunche et. al. [31, 10, 24] link different smartphones through SSID similarity.

## Other modalities

As mentioned in Section 1.3.1, social media data and phone call records provided great opportunities for human dynamic research in the space dimension, like international and domestic migration patterns [176], distribution of human convergence or divergence [37], and the daily rhythms of city life [3].

For group detection, sensor-based approaches use wearable devices or install apps on smartphones to collect users' behavioral data. Groups are detected through correlation analysis of multiple sensor data. For instance, MIT researchers use specially designed wearable devices called "Sociometric Badges" [109, 110] to measure group behavior through face-to-face interaction and physical proximity. Some research works [70, 84, 131] combine several sensor modalities (WiFi, accelerometer, compass, etc.) to measures users' similarity.

## 1.4 Thesis Organization

The rest of the dissertation is organized as follows. The next following three chapters contain independent research works focusing on human dynamic using privacy-sensitive data. For each work, it has a general outline like introduction, system overview, system design, experimental evaluation, related works, and conclusion of that work.

In Chapter 2, we introduce a data mining system (GINA) that could identify the gender information of a group of people using their privacy-sensitive audio data. Chapter 3 presents a shopping group detection system (SNOA) which are based on WiFi data. Chapter 4 describes a data-driven approach for evaluating the quality of wireless networks. The last chapter concludes the whole dissertation.

# GINA: Group Gender Identification Using Privacy-Sensitive Audio Data

Group gender is essential in understanding social interaction and group dynamics. With the increasing privacy concerns of studying face-to-face communication in natural settings, many participants are not open to raw audio recording. Existing voice-based gender identification methods rely on acoustic characteristics caused by physiological differences and phonetic differences. However, these methods might become ineffective with privacy-sensitive audio for two main reasons. First, compared to raw audio, privacy-sensitive audio contains significantly fewer acoustic features. Moreover, natural settings generate various uncertainties in the audio data. In this chapter, we make the first attempt to identify group gender using privacy-sensitive audio. Instead of extracting acoustic features from privacy-sensitive audio, we focus on conversational features including turn-taking behaviors and interruption patterns. However, conversational behaviors are unstable in gender identification as human behaviors are affected by many factors like emotion and environment. We utilize ensemble feature selection and a two-stage classification to improve the effectiveness and robustness of our approach. Ensemble feature selection could reduce the risk of choosing an unstable subset of features by aggregating the outputs of multiple feature selectors. In the first stage, we infer the gender composition (mixed-gender or same-gender) of a group which is used as an additional input feature for identifying group gender in the second stage. The estimated gender composition significantly improves the performance as it could partially account for the dynamics in conversational behaviors. According to the experimental evaluation of 100 people in 273 meetings, the proposed method outperforms baseline approaches and achieves an F1-score of 0.77 using linear SVM.

## 2.1 Introduction

Group gender plays an essential role in understanding social interaction and group dynamics [149, 116]. It is also the foundation of promising research like gender inequality [39] and gender difference [179]. With the prevalence of studying spontaneous face-to-face communication in natural settings [140, 14, 147], it becomes unprecedentedly important to identify group gender through privacy-sensitive audio data. Because face-to-face conversation is a dominant and the richest communication modality available to humans [12, 167]. Such communication could capture real emotions and represent true information flow within an organization [168, 81].

Gender identification using privacy-sensitive data is based on ethical and practical needs. Collecting truly spontaneous conversation requires recording people in unconstrained and unpredictable situations, both public and private. There is little control over who or what might be recorded. Private content and uninvolved parties could be recorded without their consent - a scenario that, if raw audio is involved, is always unethical and sometimes illegal. Therefore, assuming access to raw audio is impractical for most real-world situations and impedes collecting truly natural data [168]. An alternative is to collect privacy-sensitive audio [81]. The microphone signal is sampled at 700 Hz and generates an average amplitude reading every 50 milliseconds to ensure raw audio is not recorded nor can it be reconstructed.

Existing voice-based gender identification methods rely on distinctive acoustic characteristics caused by physiological differences (like glottis, vocal tract thickness) and phonetic differences [141, 4]. Those features are extracted from raw audio. Various identification systems have been proposed with different acoustic features and classification models [60, 119, 1, 75, 4]. The most frequently used features are pitch [60] and first formant [119] which are related to voice sources and vocal tract, respectively.

Despite extensive efforts on voice-based methods, existing solutions might become ineffective with privacy-sensitive audio for two main reasons. First, compared to raw audio, privacy-sensitive audio is too coarse-grained and it is extremely hard to extract valuable acoustic features from it. Moreover, due to natural settings, privacy-sensitive audio contains various uncertainties like background noises. These uncertainties pose serious challenges for existing

methods. For example, estimating fundamental frequency with different levels of noises is difficult [4].

In this chapter, we aim to achieve group gender identification using privacy-sensitive audio (GINA). Instead of extracting acoustic features from privacy-sensitive audio, we focus on conversational behaviors. The rationale is that conversational behaviors could reflect gender difference. Many sociology studies have reported explicit relationships between gender and conversational behaviors including turn-taking behaviors and interruption patterns [122, 181, 50, 104]. Take the length of speaking turns as an example, women have shorter speaking turns [123]. Also, men are more likely to interrupt women than the opposite [178]. Different from previous studies whose data are collected in laboratories, we conduct extensive experiments using data collected in natural settings and observe similar patterns. For example, we find that the average turn length of women (2.6 seconds) is shorter than that of men (3.2 seconds). Besides, contrary to most existing findings on interruption, we find that women interrupt men more often than vice versa.

The vision of GINA, however, entails two significant challenges when applied to real conditions. 1) *Transforming privacy-sensitive audio into voice activities encounters problems including low-resolution audio and unexpected dynamics of spontaneous conversation.* On one hand, the low-resolution audio hinders extracting acoustic features. This makes existing approaches, like multi-class classification, ineffective. On the other, spontaneous conversation in natural settings contains various uncertainties. For example, unpredictable noise and people movement could affect the robustness of existing methods. 2) *Although conversational behaviors reflect gender difference to some extent, their instability reduces the robustness and effectiveness of gender identification.* People's conversational behaviors are affected by many factors including internal factors (like emotions) and external factors (like gender composition of the meeting [103, 171]). For example, people behave differently when in mixed-gender and same-gender groups [103]. This results in unstableness and even inconsistency of conversational behaviors and thus affects the performance of gender identification.

To address the first challenge, we propose a correlation-based multichannel voice activity detection (VAD) algorithm. Traditional approaches try to separate voice signals from other people (crosstalk) because crosstalk imposes negative effects on voice applications. However, we observe that crosstalk

is beneficial as it generates correlation in privacy-sensitive audio. Based on the observation, we could identify moments when only one person speaks. Then we extract their speaking features to detect voice activities adaptively. For the second challenge, we have made two efforts. To reduce the variance of the performance, we adopt ensemble feature selection which reduces the variance of F-score by over 10%. It is often reported that several different feature subsets may yield equally optimal results, and ensemble feature selection may reduce the risk of choosing an unstable subset [126, 130]. To improve the general identification performance, we propose a two-stage classification method. In the first stage, we predict one of the external factors (gender composition) as an additional input feature for gender identification in the second stage. This approach could improve F-score by over 10% because gender composition could partially explain the dynamics of conversational behaviors.

According to our experimental evaluation of 100 people in 273 meetings, with a total length of 438 hours, GINA improves the performance of baseline approaches by 8.5% on average. GINA could achieve an F1-score of 0.77 using linear SVM. The contribution of this chapter is summarized as follows.

- We propose a privacy-sensitive modality (conversational behaviors) for gender identification. The performance is improved by ensemble feature selection and a two-stage classification method.
- An adaptive correlation-based multichannel VAD algorithm for privacy-sensitive audio is proposed.
- We analyze group conversation in natural settings and bring new insights of gender difference in interruption.

The remainder of this chapter is organized as follows. An overview is introduced in Section 2.2. We elaborate on design details of the proposed system in Section 2.3. Section 2.4 illustrates the experimental evaluation of the data collected in real-life scenarios. Related works are introduced in Section 2.5, and we conclude this work in the last section.

## 2.2 System Overview

In this section, we give an overview of GINA. As illustrated in Figure 2.1, the proposed system consists of four main components, including Privacy-



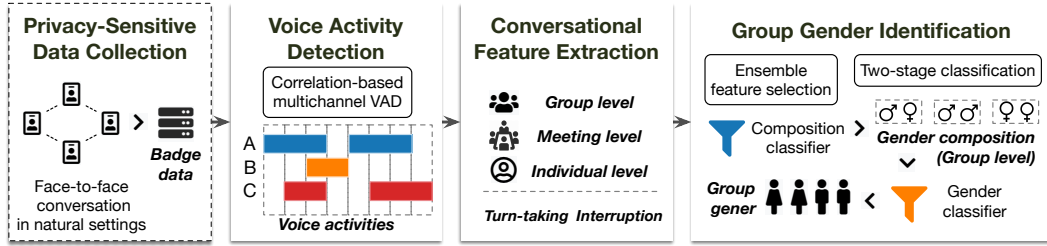


Fig. 2.1: Overview of GINA.

Sensitive Data Collection, Voice Activity Detection, Conversational Feature Extraction, and Group Gender Identification.

GINA is motivated by the ethical and legal issues arising from studying spontaneous face-to-face conversation. To this end, we exploit electronic badges [81] to collect privacy-sensitive audio data in *Privacy-Sensitive Data Collection*. We briefly introduce this component as it is not our main contribution. More details could be found in [81]. After collecting the badge data, it is processed with the devised multichannel VAD algorithm in *Voice Activity Detection*. This step mainly transforms privacy-sensitive audio data into voice activities or conversational behaviors. In *Conversational Feature Extraction*, we extract two kinds of features, namely turn-taking features and interruption features, for group gender identification. These features could be further divided into individual level, meeting level and group level. We also demonstrate the effectiveness analysis of those features and new insights of gender difference in interruption patterns. Lastly, we introduce the proposed two-stage classification method in *Group Gender Identification*. It is related to two classifiers: composition classifier and gender classifier. In the composition classifier, we predict the latent information of gender composition as an additional group level feature. Because people’s conversational behaviors vary in groups with different gender composition (mixed gender and same gender). Then we apply ensemble feature selection to three different levels of features to select stable feature subsets. Finally, we exploit the gender classifier to identify group gender based on the selected features.

## 2.3 System Design

### 2.3.1 Privacy-Sensitive Data Collection

As indicated in [168], it is a large problem to assume access to raw audio recordings in collecting spontaneous face-to-face conversational data. Therefore, we adopt a platform that uses a privacy-sensitive data collection style [81]. The platform exploits electronic badges [80] which embed multiple sensors like RFID, Bluetooth, and microphone to monitor face-to-face interaction of badge wearers.

The badge samples the microphone signal at 700 Hz and creates an average amplitude reading every 50 milliseconds. The averaged amplitude generally reflects the fluctuation of badge wearers' volume. In one second, every badge generates 20 volume data points. We call the timespan of one second as a *frame*. The collected badge data is privacy-sensitive as no raw audio is recorded and the audio cannot be re-generated from the stored samples.

### 2.3.2 Voice Activity Detection

Multichannel voice activity detection (VAD) is to detect whether a user in a channel speaks or not. Privacy-sensitive though the badge data is, it brings new challenges in VAD due to the low resolution of the badge data and unpredictable dynamics of spontaneous conversation in natural settings.

One type of traditional VAD is based on multi-class classification. Related features are extracted from raw audio first and then classification models like Hidden Markov Model [164] or Gaussian Mixture Model [113] are utilized to detect voice activities. However, most of the features could not be extracted from the privacy-sensitive audio data. Besides, it might be difficult to adapt to scenarios without training data.

Another type of methods regards VAD as blind source separation and solves it using independent component analysis (ICA) [99]. However, ICA assumes stationary mixing of the signal, i.e., requires participants to remain fixed at locations. This constraint is hard to satisfy in natural settings as participants would walk around and show some demos during the meeting. Apart from this, it is also difficult to find thresholds to separate speech and noise on the de-mixed signals, which are not resilient to different environments.

Traditional approaches try to separate voice signals from other people (crosstalk) because crosstalk imposes negative effects on voice applications. However, we find that crosstalk is beneficial as it generates correlation in privacy-sensitive audio. When only one badge wearer speaks, other people's badge signals are highly correlated with the speaker's badge signal due to crosstalk. Voice signal from different people could be regarded as independent random variables. Without the effect of crosstalk, the correlation of voice signals from two speakers should obey a zero mean normal distribution. Given a set of participants  $\mathbf{P}$  within a meeting, the badge data  $\mathbf{S}_i$  of wearer  $i$  in a frame could be represented as:

$$\mathbf{S}_i = \underbrace{\mathbf{V}_i}_{\text{Local speech}} + \underbrace{\sum_{j \in \mathbf{P}} \phi_{ij} \cdot \mathbf{V}_j}_{\text{Crosstalk}} + \underbrace{\rho_d + \rho_e}_{\text{Noise}}, j \neq i$$

where  $\mathbf{V}_i$  is the voice signal from the wearer in the same frame,  $\phi_{ij}$  is a attenuation factor of voice over the distance between wear  $i$  and  $j$ ,  $\rho_d$  and  $\rho_e$  are device and environmental noise respectively. The badge signal is a mixture of *local speech* (voice from the badge wearer), *crosstalk* (voice from other participants), and noise (device and environmental noise). When only participant  $i$  speaks during frame  $k$ , the badge signal of  $\mathbf{S}_i(k)$  and  $\mathbf{S}_j(k)$  could be reduced to Equation 2.1. It is clear that approximated  $\mathbf{S}_i(k)$  and approximated  $\mathbf{S}_j(k)$  are linearly correlated.

$$\begin{cases} \mathbf{S}_i(k) = \mathbf{V}_i(k) + \rho \approx \mathbf{V}_i(k) \\ \mathbf{S}_j(k) = \phi_{ij} \cdot \mathbf{V}_i(k) + \rho \approx \phi_{ij} \cdot \mathbf{V}_i(k) \end{cases} \quad (2.1)$$

Based on the observation, we propose a correlation-based multichannel VAD algorithm as shown in Algorithm 1\*. The algorithm takes the badge data  $\mathcal{F}_b$  from a whole meeting as input and derives voice activities  $\mathcal{F}_r$  for all participants. It consists of three main steps. We extract a set of all frames by the union of each participant's frames (line 2). The first step (line 3 ~ 6) is to find a subset of frames  $\mathcal{F}_g$  that only one wearer speaks or only one local speech exists (denote as *genuine speak* information). The selection criteria are two-fold. First, the person  $p$  must has the highest mean volume to make sure his badge signal is not caused by crosstalk. Second, other people's badge signal are all highly correlated with the person  $p$  which ensures  $p$  is the only

\*The code: <https://github.com/HumanDynamics/openbadge-analysis>

---

**Algorithm 1: Correlation-based multichannel VAD.**

---

**Input** :  $\mathbf{P}$ : a set of participants in a meeting;  
 $\mathcal{F}_b$ : a directory of badge data for all participants;  
**Output**:  $\mathcal{F}_r$ : a directory of voice activities for all participants

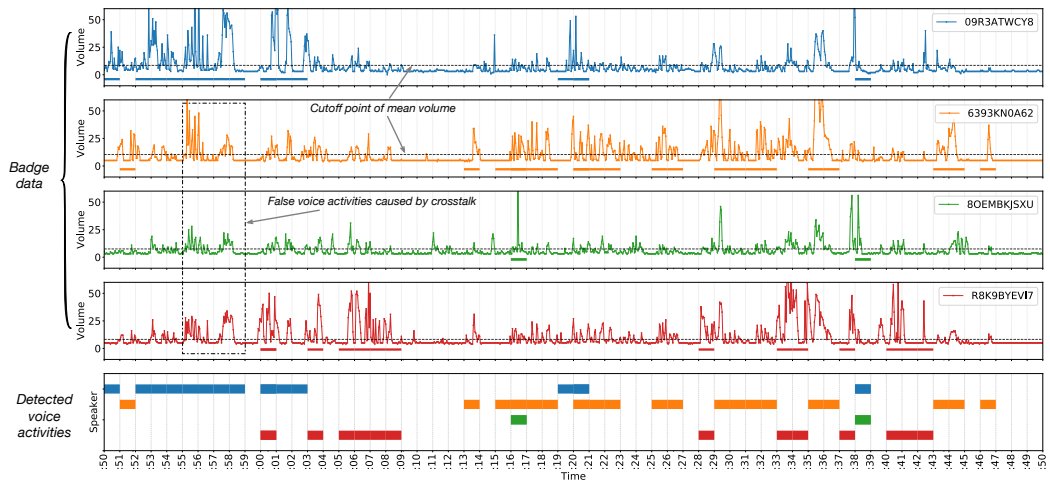
- 1 Initialize empty directories:  $\mathcal{F}_g, \mathcal{F}_a, \mathcal{F}_r$ ;
- 2  $\mathbf{F} \leftarrow \bigcup_l \mathcal{F}_b(l)$ (frame); //  $\mathbf{F}$  is a set of all frames in the meeting  
/\* Step 1: Detect genuine speak information \*/
- 3 **foreach** frame  $k \in \mathbf{F}$  **do**
- 4      $p \leftarrow \operatorname{argmax}(\operatorname{mean}(\mathbf{S}_i(k))), i, p \in \mathbf{P}$ ;
- 5     **if**  $\forall j \in \mathbf{P}, \operatorname{corr}(p, j) > \theta$  **then**
- 6         Add frame  $k$  to  $\mathcal{F}_g(j)$ ;
- 7     /\* Step 2: Detect all speak information \*/
- 8      $\mathcal{C} \leftarrow \operatorname{get-clf-rules}(\mathcal{F}_g)$ ; // Find classification rule for each person
- 9     **foreach** frame  $k \in \mathbf{F}$  **do**
- 10         **foreach**  $p \in \mathbf{P}$  **do**
- 11             **if**  $\operatorname{mean}(\mathbf{S}_p) \geq \mathcal{C}(p, \text{'mean'})$  **or**  $\operatorname{std}(\mathbf{S}_p) \geq \mathcal{C}(p, \text{'std'})$  **then**
- 12                 Add frame  $k$  to  $\mathcal{F}_a(j)$ ;
- 13     /\* Step 3: Detect real speak information \*/
- 14      $\mathbf{F}_r(p) = \mathcal{F}_g(j) \cup \mathcal{F}_a(j)$ ;
- 15     **foreach** frame  $k \in \bigcup_l \mathcal{F}_a(l)$  **do**
- 16         **if**  $\forall i, j (j \neq i) \in \mathbf{P}, \operatorname{corr}(\mathbf{S}_i(k), \mathbf{S}_j(k)) > \theta$  **then**
- 17              $p \leftarrow \operatorname{argmin}(\operatorname{mean}(\mathbf{S}_i(k)), \operatorname{mean}(\mathbf{S}_j(k))), p \in \mathbf{P}$  ;
- 18             Remove frame  $k$  from  $\mathcal{F}_r(p)$ ;
- 19     Function  $\operatorname{get-clf-rules}(\mathcal{F}_g)$ ;
- 20     **Input** :  $\mathcal{F}_g$ : A directory of frames when only one person speaks
- 21     **Output**:  $\mathcal{C}$  A directory of classification rules for each person
- 22     **foreach**  $p \in \mathbf{P}$  **do**
- 23          $\mathbf{D}_t(p, \text{'mean'}) \leftarrow$  distribution of mean volume in a frame when  $p$  talks;
- 24          $\mathbf{D}_s(p, \text{'mean'}) \leftarrow$  distribution of mean volume when  $p$  remains silent;
- 25          $\mathbf{D}_t(p, \text{'std'}), \mathbf{D}_s(p, \text{'std'}) \leftarrow$  distributions of standard deviation of volume;
- 26          $\mathcal{C}(p, \text{'mean'}) \leftarrow$  intersection of  $\mathbf{D}_t(p, \text{'mean'})$  and  $\mathbf{D}_s(p, \text{'mean'})$   $\mathcal{C}(p, \text{'std'}) \leftarrow$  intersection of  $\mathbf{D}_t(p, \text{'std'})$  and  $\mathbf{D}_s(p, \text{'std'})$
- 27     **return**  $\mathcal{C}$

---

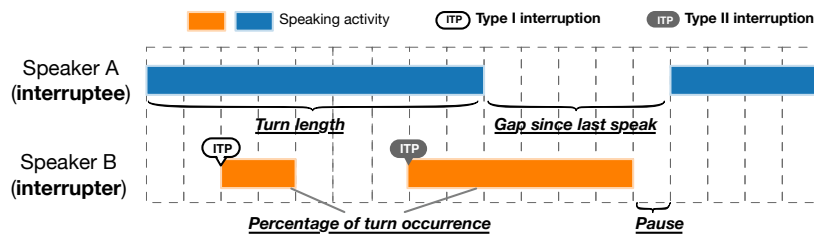
speaker. Parameter  $\theta$  is a threshold of correlation to detect crosstalk. We further discuss this parameter in Experimental Evaluation.

The second step (line 7 ~ 11) detects all frames that a person is likely to speak by applying classification rules learned from  $\mathcal{F}_g$  (all speak information). Given  $\mathcal{F}_g$ , we could identify frames of two situations for a person: talking and silence. Through comparison of both situations, we could identify cutoff points of the statistical features (mean and variance) of the volume.

Since the detected voice activities could be caused by crosstalk, the last step (line 12 ~ 16) is to remove such false activities (real speak information). The voice signal of two speakers are expected to be random independent variables, so do their badge data. For pairwise wearers, if their badge signals



**Fig. 2.2:** An example result of multichannel VAD on a meeting with four participants between 18:12:50 and 18:13:50.



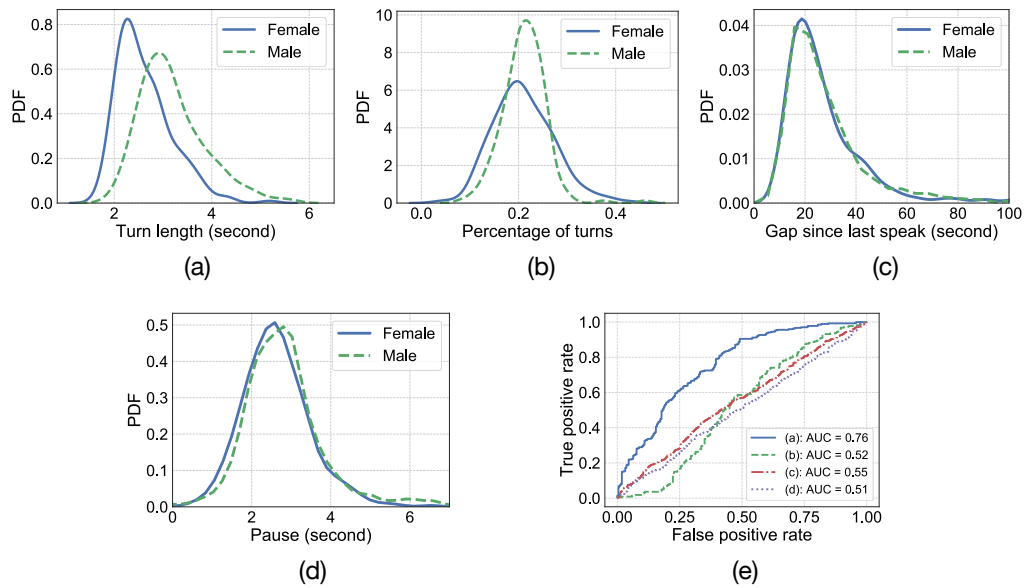
**Fig. 2.3:** Illustration of conversational features. Underlined bold text represent turn-taking features, the other bold text represent interruption features.

are strongly correlated (correlation  $\geq \theta$ ), we remove the frame for the wearer who has the weaker volume as it might be caused by crosstalk.

An example result of multichannel VAD is illustrated in Figure 2.2. The first four sub-figures reveal the badge data collected from four participants. It is clear that participants' badge signals in the box exceed their cutoff point of mean volume. However, these false activities are just caused by crosstalk from the blue participant. The last sub-figure illustrates the detected voice activities for all participants.

### 2.3.3 Conversational Feature Extraction

After Voice Activity Detection, privacy-sensitive audio data is transformed into voice activities. From the detected voice activities, we define and extract two kinds of conversational features, turn-taking features and interruption features, which are shown in Figure 2.3.



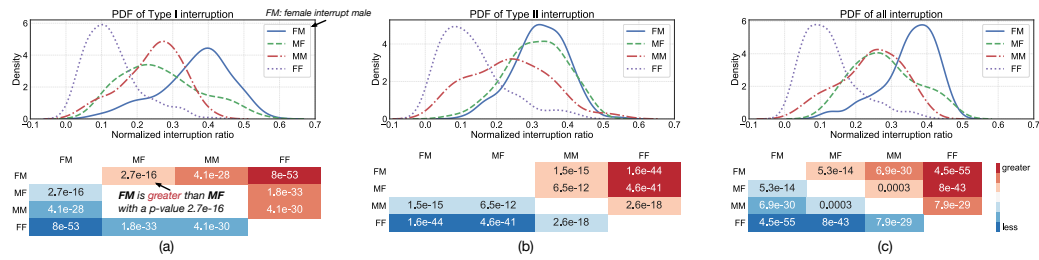
**Fig. 2.4:** Effectiveness analysis of turn-taking features. (a) ~ (d) PDFs of different features; (e) ROC curves of features.

## Turn-taking features

Turn-taking features include turn length (how long a person’s turn lasts), the percentage of turn occurrence (how frequently a person speaks), pause between consecutive turns, and gap since last speak as indicated in the literature [122].

Through analysis of the data collected from MIT Sloan Fellows program (See Section 2.4), we find that some of these features might not be effective. Figure 2.4(a) ~ (d) depict the probability density functions (PDFs) of four different features. As shown in Figure 2.4(a), females have shorter turn length than males. According to Figure 2.4(b), females have larger turn-taking variations. Besides, there seem no significant gender difference in gap since last speak and turn pauses as indicated by Figure 2.4(c) and (d).

To compare the effectiveness of those turn-taking features in gender identification, we exploit Receiver Operating Characteristic (ROC) curve, which is usually used to illustrate the diagnostic ability of a binary classifier system as its discrimination threshold varies. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. As shown in Figure 2.4(e), it is clear that the effectiveness of turn length is much better than the others.



**Fig. 2.5:** Analysis of who interrupts who with PDFs of four-class interruption and results of Mann-Whitney U test for different types of interruption. (a) Type I interruption; (b) Type II interruption; (c) Type I and Type II interruption.

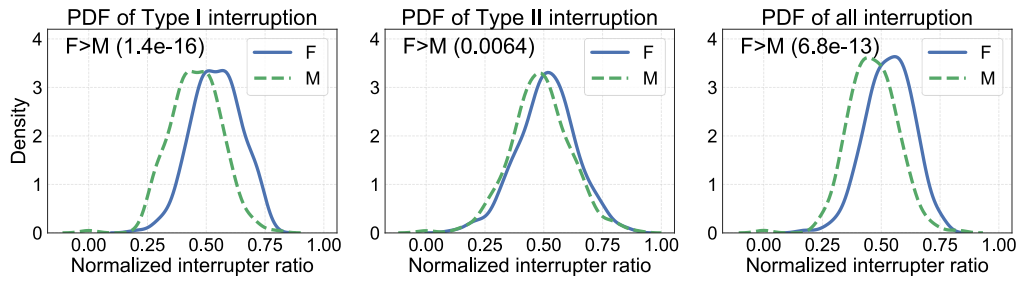
## Interruption features

According to literature, interruption consists of cooperative and disruptive interruption which could reflect gender difference [146, 178]. Cooperative interruption is usually words of agreement and support or anticipation of how other people's sentences and thoughts would end. Disruptive interruption, on the other hand, is described as having a tendency to switch the topic or take the floor. The detailed description of interruption and gender difference is stated in Related Work (Section 2.5.2).

However, cooperative and disruptive interruption might be too complex and difficult to detect without context information. In Figure 2.3, we define two roles in interruption. An interrupter is a person who starts his turn before others' turns finish while an interruptee is a person that is interrupted. Besides, we also define two types of interruption. Type I interruption is more likely to be a mixture of unsuccessful interruption and cooperative interruption, while Type II interruption is mostly successful interruption.

After analyzing the collected data, we find that generally women interrupt men more frequently which is contrary to the most existing findings in sociology studies [181, 178]. The analysis of interruption consists of three parts, who interrupts who, interrupter, and interruptee.

*Who interrupts who:* There are four classes of interruption, namely FM (female interrupt male), MF, MM, and FF, in a mixed-gender group meeting.



**Fig. 2.6:** Analysis of inters under three types of interruption.

Given the fact that the numbers of both genders are different, we calculate interruption ratios as shown in the matrix.

<b>Interruption ratios</b>					
FF	FM	=	$\frac{I_{FF}}{I_F \cdot N_F}$	$\frac{I_{FM}}{I_F \cdot N_M}$	$I_{FF}$ : Number of FF interruption
MF	MM		$\frac{I_{MF}}{I_M \cdot N_F}$	$\frac{I_{MM}}{I_M \cdot N_M}$	$I_F$ : Number interruption started by females $N_F$ : Number of females in group

The normalized interruption ratio is a normalization of each ratio over their total sum. As shown in Figure 2.5, we plot PDFs of four classes of interruption in three different situations. To show the relation of pairwise classes of interruption, we resort to Mann-Whitney U test which is a nonparametric test. The null hypothesis of the test is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample. We derive interesting results that in different situations, the relations of four-class interruption are also different. For all interruption, the relationship of four-class interruption is  $FM > MF > MM > FF$ . For Type I interruption, the relationship mostly holds except there is no significant difference between MF and MM. The PDFs of Type II interruption indicate that there is no significant difference in Type II interruption between female interrupt male and male interrupt female.

*Interrupter:* The role of gender as interrupters is analyzed in Figure 2.6. We show PDFs of male and female interrupters under three different types of interruption. The normalized interrupter ratio is simply calculated using the percentage of male or female interrupter over all interrupters. We could find that females are more likely to initiate interruptions especially Type I interruption. This is reasonable since a significant part of Type I interruption is cooperative interruption which is favored by women.



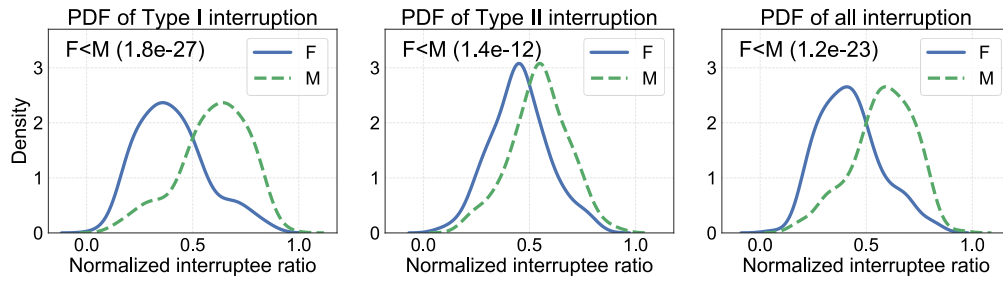


Fig. 2.7: Analysis of intees under three types of interruption.

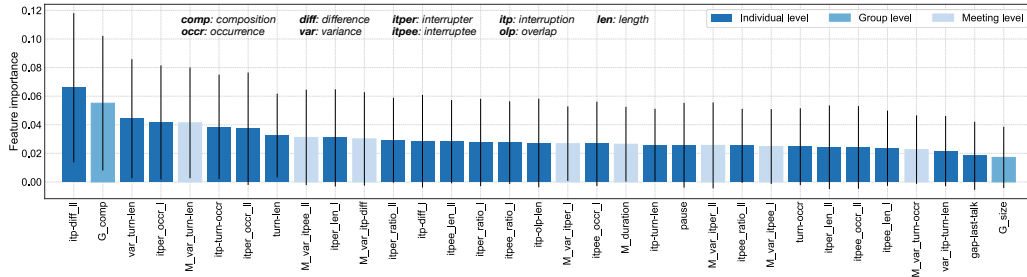
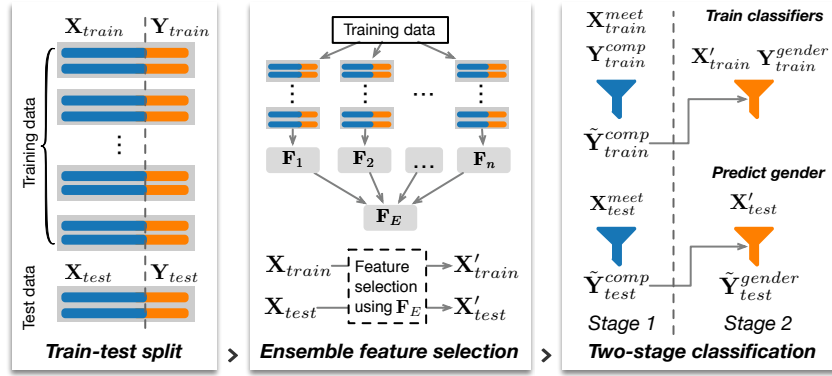


Fig. 2.8: Feature importance of all the features in a Random Forest consisting of 100 trees.

*Interruptee*: Similar to the analysis of interrupters, we also analyze interruptees. The results in Figure 2.7 indicate males are far more likely to be interrupted in different types of interruption.

Turn-taking behaviors and interruption patterns could both reflect gender difference. Therefore we devise three levels of features based on turn-taking and interruption. Figure 2.8 includes different levels of features we use. Features start with an ‘M’ is a meeting level feature, ‘G’ indicates group level features, while the rest are individual level features. For example, feature *itper\_len\_I* means the average length of Type I interruption when a participant acts as an interrupter. Feature *itpee\_occ* means the occurrence of interruption when a participant acts as an interruptee. Feature *itp-diff* is the difference between *itper\_occ* and *itpee\_occ*.

We also show the importance of those features in Figure 2.8. A Random Forest of 100 trees is used to evaluate their importance on an artificial classification task. Each bar represents the importance of a certain feature, along with its inter-tree variability. We could notice two things. First, it is nontrivial to select a subset of features that are very informative. Second, almost all the features have a large deviation in different trees. This also reflects the instability of conversational behaviors.



**Fig. 2.9:** An illustration of ensemble feature selection and the two-stage classification in an iteration of cross-validation.

### 2.3.4 Group Gender Identification

The last step is to predict group gender based on the extracted features. Specifically, it consists of the following 2 steps: ensemble feature selection and two-stage classification which are illustrated in Figure 2.9. First of all, in an iteration of  $k$ -fold cross-validation, we choose  $(k - 1)$  folds as training data and the rest fold as test data. The input data ( $\mathbf{X}$ ) consists of three counterparts, individual level feature, meeting level feature, and group level features:  $\mathbf{X} = \{\mathbf{X}^{idl}, \mathbf{X}^{meet}, \mathbf{X}^{group}\}$ . The label ( $\mathbf{Y}$ ) consists of two parts, composition and gender:  $\mathbf{Y} = \{\mathbf{Y}^{comp}, \mathbf{Y}^{gender}\}$ . Each fold contains the data from one or more groups. Second, we further separate the training data into  $n$  folds for training ensemble feature selector  $\mathbf{F}_E$ . The selector  $\mathbf{F}_E$  is applied to select a subset of features for both training ( $\mathbf{X}'_{train}$ ) and test ( $\mathbf{X}'_{test}$ ) data respectively. Lastly, the training data is used to train two classifiers (composition classifier and gender classifier). During the testing, the estimated composition ( $\tilde{\mathbf{Y}}_{test}^{comp}$ ) and selected input data ( $\mathbf{X}'_{test}$ ) are fed into the gender classifier to infer genders ( $\tilde{\mathbf{Y}}_{test}^{gender}$ ).

#### Ensemble feature selection

As introduced in Introduction, although conversational behaviors could reflect gender difference, such behaviors are unstable sometimes inconsistent. The potential reasons for those changes rely on the complex nature of human dynamics. Many factors could affect people's conversational behaviors including internal factors like emotions and external factors like gender composition of a meeting [103].

To improve the performance of using conversational behaviors, feature selection is essential. The objectives of feature selection are usually three-fold: improving the prediction performance, providing faster and more cost-effective predictors, and facilitating a better understanding of the underlying process. Furthermore, to handle the instability of conversational behaviors, we adopt ensemble feature selection (EFS). The idea of ensemble feature selection resembles ensemble learning. It is often reported that several different feature subsets may yield equally optimal results in large feature or small sample size domains. EFS could reduce the risk of choosing an unstable subset [49]. Besides, the representational power of a particular feature selector might constrain its search space such that optimal subsets cannot be reached. With EFS this problem could be alleviated by aggregating the outputs of several feature selectors [49].

Among several ways of ensemble, we adopt homogeneous ensemble [130]. It is not only easy to implement, but also more fair to evaluate its effectiveness with the standalone feature selector. Homogeneous ensemble applies the same feature selection method to different training data. As illustrated in Figure 2.9, we separate the training data into  $n$  folds and apply  $n$  feature selectors of the ensemble  $\mathbf{E} = \{\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_n\}$  to each  $(n - 1)$  folds of training data. Each selector  $\mathbf{F}_i$  outputs a weight vector ( $\mathbf{f}_i$ ) of all features with  $\mathbf{f}_i^j$  representing the weight of the  $j$ -th feature. To derive a general weight vector  $\mathbf{f}_E$  from all weight vectors, we use an average as shown in Equation 2.2. Lastly, a subset of features is selected with the mean feature weight of  $\mathbf{f}_E$  as a threshold.

$$\mathbf{f}_E^j = \frac{1}{n} \cdot \sum_{i=1}^n \mathbf{f}_i^j \quad (2.2)$$

## Two-stage classification

We find gender composition, one of the external factors, could be inferred accurately using meeting level features. Therefore, we propose a two-stage classification method as shown in Figure 2.9. In the first stage, we infer the latent information of gender composition and treat it as an additional input feature for group gender identification in the second stage. In both stages, we choose popular classification models like linear SVM and Random Forest.

In the first stage, we leverage meeting level features of each group to predict its gender composition. Each participant in the meeting has two roles,

interrupting others (as interrupter) and being interrupted by others (as interruptee). The variance of the difference between interrupter and interruptee in a meeting ( $M\_var\_itp\_diff$ ) is a good indicator of gender composition. Same-gendered groups tend to have smaller variance. Because interruption is reported more evenly distributed in same-gendered groups [103]. In the second stage, we combine the selected features and the inferred gender composition as input to predict gender for the whole group.

## 2.4 Experimental Evaluation

### 2.4.1 Settings

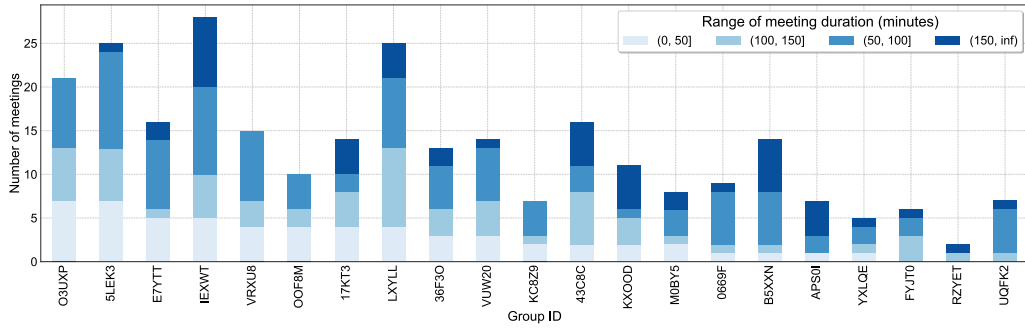
#### Setup

The privacy-sensitive audio data is collected from spontaneous face-to-face meetings of MIT Sloan Fellows class of 2016/17 for about 4 weeks. 100 out of the 110 students participated in the study, including 31 females and 69 males. They came from 35 different countries and had an average age of  $37.41 \pm 4.45$  years (mean  $\pm$  standard deviation) as well as an average work experience of  $13.78 \pm 4.24$  years. All participants gave written informed consent about their participation in the study.

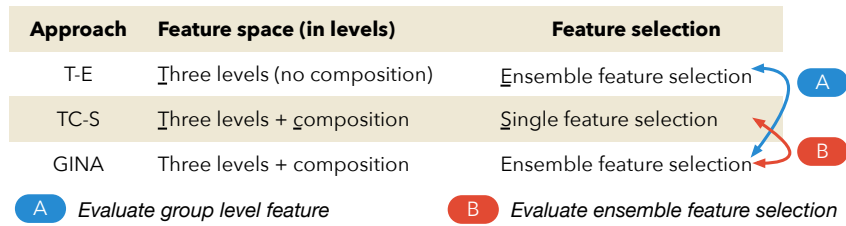
Great importance is attached to group collaboration in the MIT Sloan Fellows program. Therefore, Sloan Fellows are assigned to study groups of four or five students before the program starts. The guideline of the group assignment ensures if it is a mixed gender group there are at least 2 students of the same gender. There are 21 study groups including 5 same-gender groups and 15 mixed-gender groups. These groups are consistent over the whole program, and the students within these groups regularly meet to study and work on the courses together. They are free in how often and how long they meet.

#### Dataset

During the experiment, we collect 273 effective meetings with a total length of 438.25 hours from 21 groups. We show the number of meetings and their duration for each group in Figure 2.10. On average, each group had 13



**Fig. 2.10:** Stacked histogram of number of meetings and meeting duration for all study groups.



**Fig. 2.11:** Illustration of baseline approaches.

meetings, but still, some groups had no more than 5 meetings. Besides, over half of those meetings last for more than 100 minutes.

## 2.4.2 Evaluation

### Baseline approaches

To evaluate the effectiveness of ensemble feature selection and gender composition, we propose two other approaches as baselines. The detailed configuration of the approaches are illustrated in Figure 2.11.

Feature selection techniques can be divided into three categories based on how they interact with the classifier. Filter methods directly operate on the dataset by providing a feature weighting, ranking or subset as output. The advantage of being fast and independent of the classification model but at the cost of inferior results. Wrapper methods perform a search in the space of feature subsets, guided by the outcome of the model (like classification performance on cross-validation of the training set). Their results are reported better than filter methods, but at the cost of an increased computational cost. Lastly, embedded methods use internal information of the classification model to perform feature selection (e.g., use of the weight

vector in support vector machines). They often provide a good trade-off between performance and computational cost [125]. Therefore, a decision tree based embedded feature selection method is used.

## Evaluation metrics

Gender identification is essentially a binary classification problem. We use metrics based on precision, recall, and F1-score to evaluate the performance of the proposed system. When the target label is male (i.e.,  $X$  is set to male), precision, recall and F1-score for male is calculated as follows.

$$\left\{ \begin{array}{l} \text{precision}(p) = \frac{\text{tp}}{\text{tp}+\text{fp}} \\ \text{recall}(r) = \frac{\text{tp}}{\text{tp}+\text{fn}} \\ \text{F1-score} = 2 \cdot \frac{p \cdot r}{p+r} \end{array} \right.$$

		Truth		$X$	$\tilde{X}$	
		$X$	tp	fp	$X$ Target label {female, male}	
Prediction	$X$	tp	fp			
	$\tilde{X}$	fn	tn	$\tilde{X}$	Non-target label	

Considering the imbalance in numbers of females and males, we use a weighted version of those metrics. The weighted F1-score is calculated with Equation 2.3 where  $S_F$  is the support of female or the number of true female instances and  $F1_F$  is the F1-score for females. The weighted precision and weighted recall are derived in a similar way.

$$F1 = \frac{S_F}{S_F + S_M} \cdot F1_F + \frac{S_M}{S_F + S_M} \cdot F1_M \quad (2.3)$$

## Parameter selection

Parameter  $\theta$  in Voice Activity Detection (Section 2.3.2) is a threshold for detecting crosstalk. Different values of  $\theta$  lead to different genuine speak information ( $\mathcal{F}_g$ , in Section 2.3.2).

Generally, large  $\theta$  could derive better accuracy because the frames selected as genuine speak ( $\mathcal{F}_g$ ) becomes more strict. However, it will also lead to large deviation as the number of frames in  $\mathcal{F}_g$  decreases. On the contrary, small  $\theta$  will result in more false detections of genuine speak and thus reduce the accuracy but the number of frames are more than adequate. Given two distributions  $D_t(p, 'mean')$  (distribution of mean volume when  $p$  talks) and  $D_s(p, 'mean')$  ( $p$  remains silent), the larger their distance measured in KL divergence the better. As shown in Figure 2.12, with the increase of  $\theta$ ,

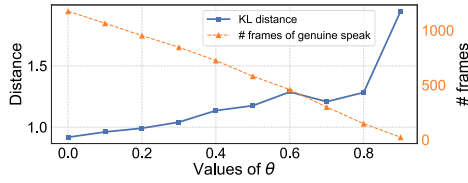


Fig. 2.12: The impact of different  $\theta$ .

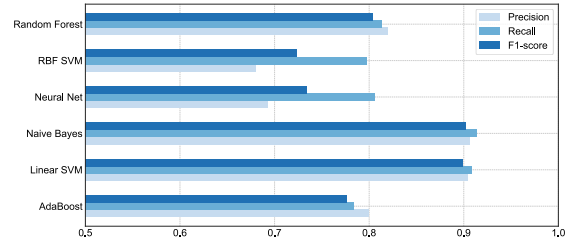


Fig. 2.13: Performance of gender composition detection with different models.

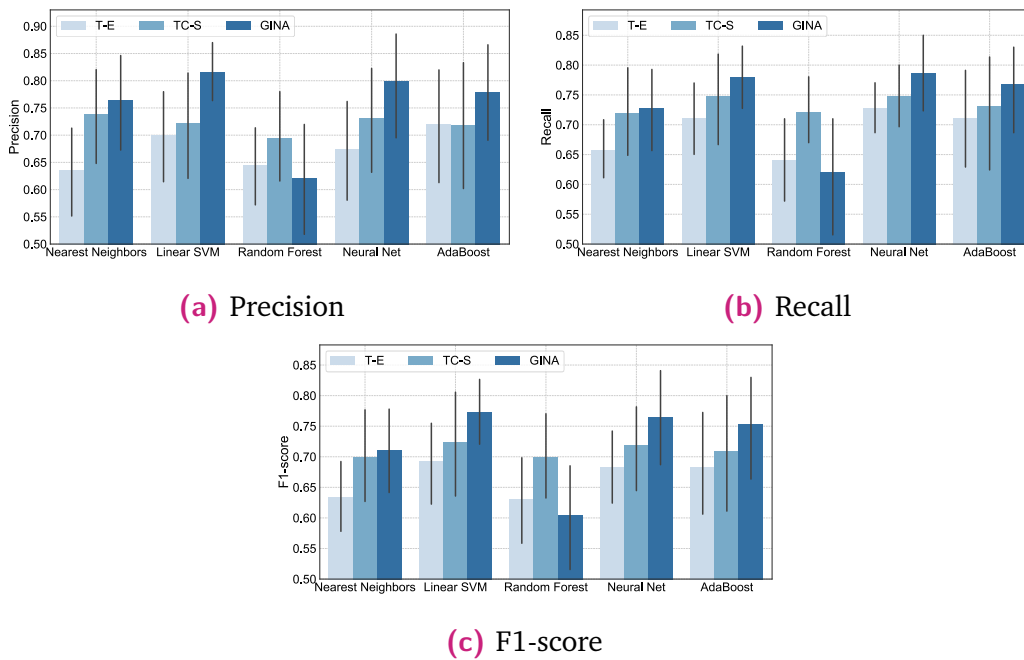
the mean distance also increases with while the number of genuine speak frames decreases. This is a trade-off between accuracy and deviation, we experimentally set  $\theta = 0.5$  in our scenario.

### Performance of gender composition detection

We evaluate the performance of gender composition detection with 10-fold cross-validation. Because the number of groups is small, we repeat the cross-validation process for 5 times and show the average performance in Figure 2.13. Naive bayes and linear SVM outperform other models and achieve a weighted F1-score around 0.9. This indicates the meeting level features we extract have the potential to capture gender composition effectively. Because same-gendered groups and mixed-gendered groups have distinct meeting behaviors. Same-gendered groups have evenly distributed interruption patterns. The gap between a person being an interrupter and an interruptee are close to each other in the same-gendered groups. While in mixed-gendered groups, women tend to have large gap while men are likely to have small gaps. This is reflected in the analysis on *who interrupts who*. Therefore the variance of gaps is larger in the mixed-gendered groups.

### Performance of group gender identification

We evaluate the performance of baseline approaches on selected classification models including Nearest Neighbor, Linear SVM, Random Forest, Neural Network and AdaBoost. The parameter settings for all models are consistent with different baselines. As shown in Figure 2.14, for most of the models, the order of performance is GINA > TC-S > T-E. On average, GINA outperforms T-E and TC-S by 11.62% and 5.37% in F1-score respectively except on Random Forest. This indicates that the inferred gender composition and



**Fig. 2.14:** Comparison of performance using different classification models. (a) Precision; (b) Recall; (c) F1-score.

ensemble feature selection are effective in improving the performance of gender identification.

Not only the performance, but ensemble feature selection could also reduce the variance of performance. Without Random Forest, on average GINA reduces the variance of Precision and Recall by 17.28% and 7.15%. As explained, ensemble feature selection could reduce the risk of choosing an unstable subset of features by aggregating the outputs of several feature selectors.

As shown in Figure 2.8, the feature of gender composition is the second most important feature. On average, this additional feature could improve the Precision and Recall by 15.99% and 9.15%. Gender composition could partially account for the instability of conversational behaviors and thus increase the interpretability of conversational features.

## 2.5 Related Work



## 2.5.1 Gender detection

Gender identification has been studied for decades in different areas. Various modalities like vision, online behaviors and voice have been utilized for this purpose. Different application scenarios have varying preferences of modalities. For example, vision-based methods are the first choice in systems where user cooperation is not required, like surveillance systems. In speech recognition, voice-based approaches are preferred.

Vision-based approaches exploit information from the face and whole body (either from a still image or gait sequence) to recognize human gender. It is usually based on appearance differences like face and body, and behavior differences like gait. More details on the utilized techniques and challenging issues could be found in the survey [108].

Vision, voice as well as handwriting are traditional modalities for gender identification. With the development of digital advertising, users' online behaviors like video viewing behaviors [177] and web browsing behaviors [58] are used for gender identification recently. This type of approaches is based on preference differences and behaviors differences.

Among all different modalities, voice is the most related to conversational behaviors. Voice-based methods rely on discriminative features extracted from human voices. The intuition is that different genders have different acoustic characteristics due to physiological differences (like glottis, vocal tract thickness) [4] and phonetic differences [141]. Various identification systems with different classification models and different types of features have been reported in the literature [60, 119, 1, 75, 4]. The most frequently used features are pitch [60] and first formant [119], which are closely related to voice sources and vocal tract, respectively. Generally, the pitch and the formant frequencies of females are higher than that of males. Moreover, as pointed out in [4], other traditional acoustic features like linear predictive coefficients (LPC), linear predictive cepstral coefficients (LPCC), Mel-frequency cepstral coefficients (MFCC), perceptual linear predictive coefficients (PLP), and relative spectral PLP coefficients (RASTA-PLP) are used in the literature for gender identification.

The majority of aforementioned acoustic features depend on accurate estimation of the fundamental frequency which itself is a challenging task.

Therefore, Alhussein et. al. propose a new single-value feature in the form of area under the modified voice contour (MVC) in [4]. The proposed feature is independent of fundamental frequency and is proved promising compared to existing features.

Besides, there is a trend of combining multiple features for gender identification in recent work [85, 1]. For example, Abouelenien et. al. extract features from five different modalities, including acoustic, linguistic, visual, thermal, and physiological, to identify gender [1].

## 2.5.2 Gender differences and interruption

The occurrence of overlap and interruption have been found closely related to gender in many sociology studies [181, 178]. The classic study by Zimmerman and West found that in same-sex conversations, interruptions were rare and appeared to be evenly distributed between speakers, whereas in cross-sex conversations, almost all the interruptions were initiated by male speakers [181]. A well-adopted explanation is males tend to show dominance by interrupting females. Many other works have found similarly that men interrupt more than women.

However, a few studies have different findings. For example, Hannah et. al. found no significant difference between interruption and gender [50]. Murray and Covelli even had a contrary discovery that women interrupt more than men [104]. One potential reason for the diverse findings is multiple conceptual and operational definitions of interruptions used in the literature [178]. Interruption is a complex interactional phenomenon with rich meanings, diverse functions, and various structural features [178]. There exist two different types of interruption, cooperative and disruptive, in literature [146, 178]. Cooperative interruption is usually words of agreement and support or anticipation of how other people's sentences and thoughts would end. This type of interruption is reported characteristic of women's style of speech [122] that might have a potentially positive influence on the interpersonal relationship between speakers. Disruptive interruption, on the other hand, is described as tending to switch the topic or take the floor. This type of interruption is attributed to men's style that might have the potential to bear negatively on the interpersonal relationship between speakers.

## 2.6 Conclusion

In this chapter, we propose a data mining system (GINA) to identify group gender through privacy-sensitive audio data. Our contribution are three-fold. First, we propose a privacy-sensitive modality for gender identification. The effectiveness and robustness are improved by ensemble feature selection and a two-stage classification. Second, an adaptive correlation-based multichannel VAD algorithm for privacy-sensitive audio is proposed. Last, we bring new insights of gender difference in interruption through analysis of group conversation in natural settings. According to experimental evaluation, GINA could effectively identify group gender with an F1-score 0.77 using Linear SVM.

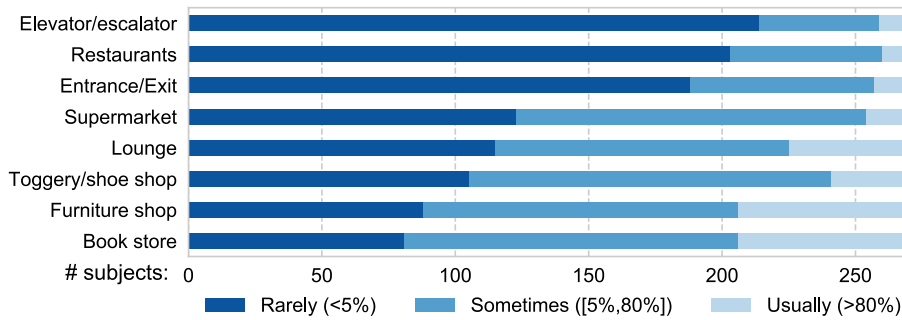


# SNOW: Detecting Shopping Groups Using WiFi

Detecting shopping groups is gaining popularity as it enables various applications ranging from marketing to advertising. Existing methods exploit WiFi probe requests to detect shopping groups by identifying co-located customers. However, the probe request is prone to suffer from device heterogeneity which might pose a severe data sparseness problem. More importantly, we find that a certain amount of shopping groups would separate sometimes which makes traditional methods unreliable. In this chapter, we propose a shopping group detection system using WiFi (SNOW). Instead of collecting probe requests, SNOW utilizes the WiFi data from smartphones associated with the deployed access points (APs). We could thus obtain data from different devices and even ensure a data granularity of seconds using Arping. Besides, we exploit an effective heuristic extracted from two observations of shopping group dynamics to improve the detection performance. First, the probability of group separation differs in diverse areas. Second, the proportion of group participation and individual engagement differs in different activities of the mall. Therefore, APs under which shopping groups appear more frequently and barely separate should contribute more in measuring customer similarity. Lastly, we represent the measured similarity into a matrix format and apply matrix factorization with a sparsity constraint to derive grouping results directly. According to our experiments in a large shopping mall, SNOW improves the detection performance of baseline approaches by 13.2% on average.

## 3.1 Introduction

Detecting shopping groups is not only the foundation of many areas but also an enabler of various applications ranging from marketing to advertising [56]. The insights of shopping groups can help retailers to provide a context-specific incentive to potential customers on the one hand and add more intelligence to their business analytics on the other [131].

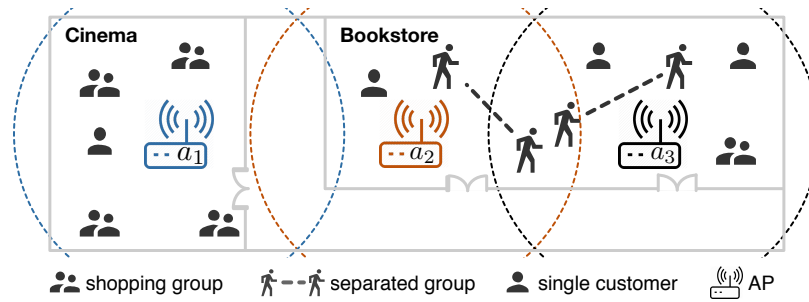


**Fig. 3.1:** Answers to the online survey problem: “How often will you get separated with your companion(s) in these regions?”

A *group* refers to people with similar properties or behaviors like network association histories [10] and mobility patterns [71, 56]. *Detecting shopping groups* is defined as a task to cluster a set of customers into disjoint subsets.

Some existing works detect co-located people as groups with *probe requests* (or probes) [71, 56]. Those probes are broadcast by smartphones to seek information about nearby access points (APs). Received Signal Strength Indicator (RSSI) contained in probes could be used to represent smartphone users’ mobility information. Compared to other approaches, the probe method requires neither high deployment cost (e.g., deploy cameras [45, 142]) nor user intervention (e.g., carry wearable devices or install mobile applications [70, 131, 109]).

However, the probe method might have two difficulties in detecting shopping groups. The first problem results from the probe request itself. Pervasive as the probe is, it suffers from multiple issues like MAC randomization [42], meaningless devices [138], and especially device heterogeneity [42]. The timing of sending probes are mainly determined by user-device interaction and the internal mechanism of the device. Therefore, different devices might generate data with various granularities which makes it difficult to measure customer similarity [42]. The second problem arises from the shopping group. Existing group detection methods assume group members always stick together while shopping groups might sometimes get separated. According to our online survey of 268 subjects, most group customers are often separated with their companions in the mall, especially in bookstores. The detailed answers to the survey problem are shown in Figure 3.1. This fact requires group detection methods can not only distinguish strangers who are close to each but also identify groups that might disperse.



**Fig. 3.2:** A toy example for illustrating the main idea of SNOW.

We ask the following question: *can we reliably detect shopping groups using WiFi?* In this chapter, we provide an affirmative answer by proposing **SNOW**. Instead of sniffing controversial probe requests, we collect the *WiFi data* from customers who associate with the deployed APs. WiFi data refers to the information contributed by any captured wireless traffic. According to a survey [115], over 75% people use public WiFi, which indicates the WiFi data is also pervasive enough for many application scenarios. Due to extra wireless packets, WiFi data could derive more continuous information for different devices. As WiFi data comes from only connected customers, extra efforts to handle MAC randomization and remove meaningless devices could be exempted. To handle the dynamics of shopping groups, we exploit an effective heuristic derived from two key observations. *Observation I*: the probability of group separation differs in diverse areas, which is also reflected in the survey results in Figure 3.1. We think this happens might due to different interests of group members. *Observation II*: the proportion of group participation and individual engagement differs in different activities of the mall. It is reported when considering whether to engage in hedonic and public activities like going to a movie alone, individual consumers anticipate negative inferences from others about their social connectedness that reduce their interests of engaging in such activities [120].

We show a toy example in Figure 3.2 to highlight the main idea of SNOW. We could see that less group separation occurs in the cinema (*Observation I*), indicating there would be less false negative detections (groups are detected as strangers). Besides, the ratio of engaged groups over individuals is higher in the cinema (*Observation II*), showing the probability of false positive detection (strangers are detected as groups) would be smaller. Therefore, customer similarity measured in the cinema is more important than that of the bookstore. Accordingly, the AP deployed in the cinema ( $a_1$ ) should bear higher importance than other APs.

The vision of SNOW, however, entails significant challenges when applied to real conditions. First, it is difficult to compare customers' WiFi data directly since the data from different devices are usually defined on different time instants with different lengths. Besides, other issues like packet loss and not sending any packets make it even difficult to measure the customer similarity. Second, the measured similarity could be incomplete and noisy which might lead to false detections like detecting strangers as a group and vice versa.

To address the first challenge, we propose a three-step data preprocessing including time interpolation, noise filtering, and non-effective instant removal. After the preprocessing, issues like packet loss and not sending any packets could be appropriately addressed. Then we could measure pairwise customer similarity based on the WiFi data. For the second challenge, we propose to represent customer similarity into a matrix format and apply matrix factorization (MF) to derive the grouping results. The advantages are two-fold. First, MF is a popular approach for noise filtering and data completion by decomposing the input matrix into several factor matrices. Second, MF with sparsity constraint is an alternative for clustering. Therefore, we could directly derive the grouping results without extra clustering processes by imposing the sparsity constraint to the decomposition.

According to our experimental evaluation in a large mall, SNOW could achieve robust and reliable performance in detecting shopping groups. Compared to baseline approaches, SNOW improves the performance of detection by 13.8% and 12.6% on labeled and semi-labeled datasets, respectively.

The contributions of this work are summarized as follows.

- We extract an effective heuristic from observations of shopping group dynamics that could significantly improve the detection performance.
- We propose a general three-step preprocessing method for processing the WiFi data.
- We evaluate the proposed system using data collected in a large shopping mall for three weeks.

The remainder of this chapter is organized as follows. We present design details of the proposed system in Section 3.2. Section 3.3 demonstrates results of the experimental evaluation. Related work is introduced in Section



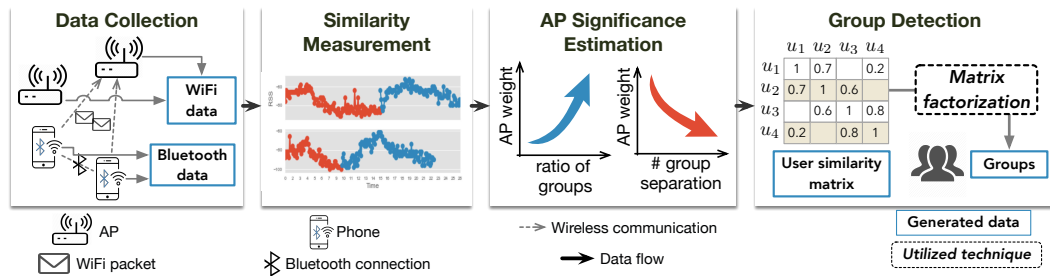


Fig. 3.3: An overview of SNOW.

3.4. We provide further discussion on issues that might be unclear in Section 3.5. Finally, we conclude this chapter in the last section.

## 3.2 System Design

In this section, we elaborate on design details of SNOW. Figure 3.3 shows an overview of the proposed system which consists of four main components. First, Data Collection is to collect the WiFi data from customers who associate with the deployed APs. We also collect the Bluetooth data from volunteers to estimate the significance of different APs. Note the Bluetooth data is not required when the system is in service. Second, we measure pairwise customer similarity based on their WiFi data in Similarity Measurement. Third, AP Significance Estimation exploits both WiFi and Bluetooth data to evaluate AP weights. Last, the customer similarity is refined combining both the AP weights and the WiFi data. We further represent the refined similarities in a matrix and apply matrix factorization to detect shopping groups.

### 3.2.1 Data Collection

#### WiFi data

The WiFi data refers to the information contributed by any wireless traffic of connected customers. It would not violate customers' privacy since only non-sensitive information in the packet header is used. Compared to probe requests, the WiFi data has two advantages. First, smartphones use the MAC randomization mechanism to protect user privacy nowadays [42]. Probe requests might also come from passers-by outside the mall. Therefore, it usually requires extra processes for probe methods to handle MAC random-

ization and filter out passers-by. While these efforts could be exempt for the WiFi data as only consumers would take the initiative to connect to the deployed APs in the mall. Second, the timing of sending probes are mainly determined by user-device interaction and the internal mechanism of devices. Generally, Android devices send more probes than iOS devices and devices with old operating systems send more probes.[42]. Therefore, probe requests might generate sparse data with different data granularity. For the WiFi data, however, it is always available for both iOS and Android devices once connected to the AP.

We exploit off-the-shelf WiFi APs to collect and store the WiFi data and upload them to a server daily. Each AP works under OpenWrt (a GNU/Linux distribution for embedded devices) and uses IW (a tool for managing wireless configuration) to collect WiFi data from connected devices. An example of using IW is “*iw interface station dump*” (*interface* is the wireless interface of the AP). We extract two fields from the output of IW. The first field is *signal* which indicates RSSI. The second field *inactive time* refers to an interval since receiving the last packet. We execute the command every second to extract both fields and record the signal information when it updates.

However, when the smartphone is not used, the granularity of WiFi data could be unsatisfying. To overcome this issue, we use Arping to force the connected device to send packets more frequently. Arping is a tool for discovering a MAC address given an IP address. IP addresses of the connected devices could be found in ARP table of DHCP lease. This act might cause more energy consumption and we further discuss this issue in Section 3.5. According to our experiments under laboratory environment, Arping could on average boost the data granularity by 43% for devices with different systems and ensure a data granularity of seconds.

## **Bluetooth data**

We also collect the Bluetooth data from volunteers’ smartphones for significance estimation of different APs. When the system is applied in service, this data is not required for customers.

We develop an Android application to scan and record its associated network information and nearby Bluetooth devices every 30 seconds. The interval is determined empirically as it takes several seconds to complete the scan

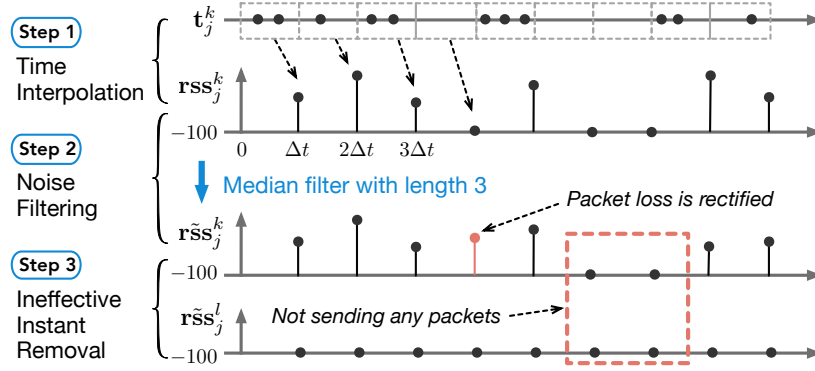


Fig. 3.4: The three-step preprocessing of the WiFi data.

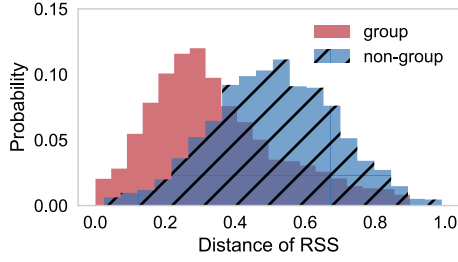
process. The app collects data entries in the format of  $[time, local\_address, associated\_MAC, scan\_info]$ . The parameter  $local\_address$  is the Bluetooth address of the device,  $associated\_MAC$  refers to the MAC address of the connected AP, and  $scan\_info$  is a list of scanned device addresses and corresponding RSSI. Given a group of two users, if their Bluetooth RSS is smaller than a threshold for a certain period of time, then it is regarded as a group separation. The settings of the RSS threshold and the period might vary in different scenarios. According to our experiments, the performance peaks when the RSS threshold is set to  $-90$  and the period is set to 2 minutes.

### 3.2.2 Similarity Measurement

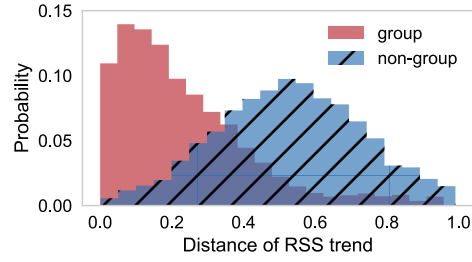
#### Data preprocessing

The WiFi data from different devices usually have different lengths. They might be defined on different time instants. Also, conditions like packet loss and not sending any packets should be properly handled. To address those issues, we propose a three-step data preprocessing: time interpolation, noise filtering, and non-effective instant removal. An illustration of the preprocessing procedure is shown in Figure 3.4.

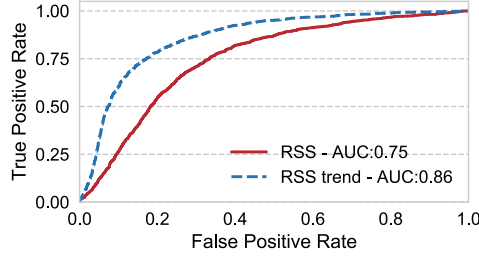
In step 1, we align the WiFi data by generating RSS vectors on an equally spaced time instants. As shown in Figure 3.4,  $\mathbf{P}_j^k$  represents all packets received by AP  $k$  from smartphone  $j$ . We generate a RSS vector out of  $\mathbf{P}_j^k$  on the unified time instants  $\mathbf{T}_u = [0, \Delta t, \dots, n\Delta t]$ . The RSS vector of packets received by AP  $k$  from smartphone  $j$  is denoted as  $\text{rss}_j^k = [\text{rss}_j^k(0), \text{rss}_j^k(1), \dots, \text{rss}_j^k(p)]$ , where  $\text{rss}_j^k(p)$  represents the median RSS value of packets during the period of time  $[p\Delta t, (p+1)\Delta t]$ . If no packets are re-



**Fig. 3.5:** Distance distributions of groups and non-groups using RSS.



**Fig. 3.6:** Distance distributions of groups and non-groups using RSS trend.



**Fig. 3.7:** ROC curves of using RSS and RSS trend.

ceived, we take the typical lowest RSS value ( $-100$  db) as a replacement. We call a RSS value a valid RSS if it is unequal to  $-100$ .

In step 2, we filter out two types of noises with median filter for each RSS vector. The filtered RSS vector is represented as  $\tilde{r}ss_j^k$  for the given input  $rss_j^k$ . As shown in Figure 3.4,  $rss_j^k$  might have intermittent  $-100$  which are mainly caused by packet loss. These  $-100$  among RSS are one of the noises. The other type of noise is the isolated RSS that appears among a long sequence of  $-100$  which might be caused by device noise or multipath effect.

The main idea of the median filter is to run through the signal entry by entry, replacing each entry with the median of neighboring entries. For a sequence of  $x$ , a median filter with length  $2n + 1$  will generate a sequence  $y$  which is defined as:

$$y(i) = \text{median}([x(i - n), \dots, x(i), \dots, x(i + n)]) \quad (3.1)$$

The length of the median filter is determined by the maximum number ( $n$ ) of packet loss that is allowed during the period of  $(2n + 1)\Delta t$ . For example, we set  $2n + 1 = 3$  in Figure 3.4. Given a certain time instant  $p\Delta t$ , such a median filter could generate a valid  $\tilde{r}ss_j^k(p)$  if the number of instants with packet loss is no more than 1 during  $[(p - 1)\Delta t, (p + 1)\Delta t]$ .

In step 3, we remove ineffective time instants when a smartphone does not send any packet. We represent the effective time instants of smartphone  $j$  as  $\mathbf{T}_j$  which is defined in Equation 3.2. Parameter  $\mathbf{K}$  is a set of all APs.

$$\mathbf{T}_j = \{t \mid \exists k \in \mathbf{K}, \text{r}\tilde{\text{ss}}_j^k(t) \neq -100, t \in \mathbf{T}_u\} \quad (3.2)$$

Effective times refer to time instants when a smartphone send packet(s) or have a valid RSS. When  $\text{r}\tilde{\text{ss}}_j^k(p) = -100$ , it indicates that AP  $k$  does not receive any packets from the smartphone  $j$  during  $[p\Delta t, (p+1)\Delta t]$ . If all APs do not receive packets from the devices in a certain instant, it is believed that the smartphone does not send any packets.

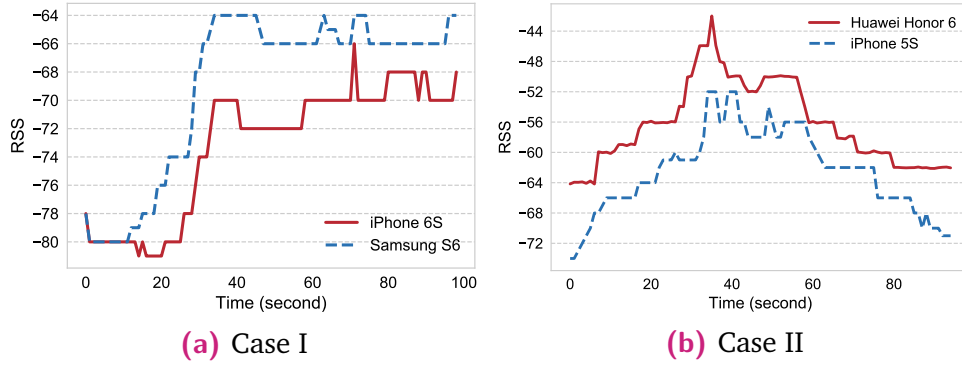
### Similarity measurement with RSS

According to [152], the RSS difference of customers  $i$  and  $j$  at AP  $k$  could be measured with Equation 3.3.  $\mathbf{T}_{i \cap j} = \mathbf{T}_i \cap \mathbf{T}_j$  represents the intersection of effective time instants from customers  $i$  and  $j$ .

$$\mathbf{d}_k(i, j) = \frac{1}{|\mathbf{T}_{i \cap j}|} \sqrt{\sum_{t \in \mathbf{T}_{i \cap j}} [\text{r}\tilde{\text{ss}}_i^k(t) - \text{r}\tilde{\text{ss}}_j^k(t)]^2} \quad (3.3)$$

We could derive an overall RSS distance  $\mathbf{d}(i, j)$  by averaging  $\mathbf{d}_k(i, j)$  over all APs. However, we should note that the RSS distance at a certain AP would change over time. Instead of calculating a single statistic of  $\mathbf{d}_k(i, j)$  for the whole testing period, a better way is to look at the distribution consisting of multiple  $\mathbf{d}_k(i, j)$  from different time slots and different APs. Intuitively, if  $i$  and  $j$  are in the same group, the center of the distribution should be close to 0.

For pairwise customers, we could obtain two distributions  $\mathcal{D}_g$  and  $\mathcal{D}_{\bar{g}}$  of RSS difference from shopping groups and non-groups respectively. Whether RSS is an appropriate feature could be verified with the WiFi data collected from 104 volunteer shopping groups by comparing  $\mathcal{D}_g$  and  $\mathcal{D}_{\bar{g}}$ . As an illustration, Figure 3.5 shows both distributions from our experiments described in Section 3.3. In particular, the unit time instant  $\Delta t = 1s$ , and the length of the median filter is set to 11. Both parameters are determined empirically and experimentally. Using the testing data involving 104 groups in three weeks, we have generated a pool containing about 9,675 data samples from group pairs and another pool containing about 32,410 samples from non-group pairs.



**Fig. 3.8:** RSS of group members using different smartphones. In both cases, group members stick together all the time, but there exist gaps in their RSS signals.

Although  $\mathcal{D}_g$  and  $\mathcal{D}_{\bar{g}}$  are obviously different, the level of difference is still not large enough to derive satisfying group detection performance. Since there is a large overlap between the two distributions, it is difficult to find a threshold to differentiate both distributions. After an investigation, we realize that one of the potential reasons for the large overlap is device heterogeneity. Even though group members stick together all the time, the difference between their smartphones could lead to a large gap. Figure 3.8 shows two typical examples. In both cases, group members with different smartphones walk closely with each other, but there still exist certain gaps in their RSS signals which might be caused by hardware difference.

### Similarity measurement with RSS trend

We find through experiments that when two customers are walking together closely, the change of their RSS reveals quite similar patterns, which could be exploited for group detection. From Figure 3.8 we also notice that despite the gap in group members' RSS signals, their general trends are similar. This property has been observed in [137] and [139] that although RSS is very unstable, the trend of RSS is relatively stable. RSS values increase or decrease when approaching or leaving an AP. This has been utilized for indoor localization in [139, 82].

Following the procedures as described in the previous subsection, we could find out the filtered common RSS vectors  $\tilde{\mathbf{r}}_i^k(\mathbf{T}_{i \cap j})$  and  $\tilde{\mathbf{r}}_j^k(\mathbf{T}_{i \cap j})$  for smartphone  $i$  and  $j$ . Let  $X = \tilde{\mathbf{r}}_i^k(\mathbf{T}_{i \cap j})$  and  $Y = \tilde{\mathbf{r}}_j^k(\mathbf{T}_{i \cap j})$ , we calculate the distance of RSS trend with

$$\mathbf{d}'_k(i, j) = 1 - \rho(X, Y) = 1 - \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}, \quad (3.4)$$

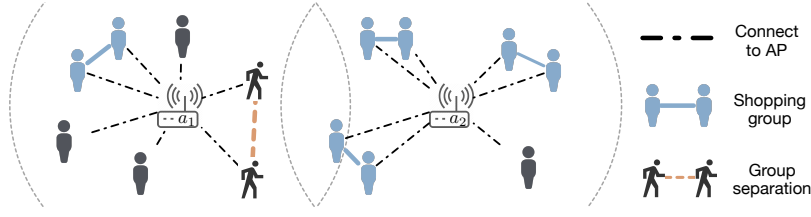
where  $\rho(X, Y)$  is the Pearson correlation coefficient of sequence  $X$  and  $Y$ ,  $\sigma_X$  is the standard deviation of  $X$ ,  $cov$  is the covariance defined as  $cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$ , and  $\mu_X$  is the mean of  $X$ . Here Pearson correlation is chosen for its better simplicity and efficiency compared with other measurements like DTW (Dynamic Time Warping).

Figure 3.6 shows distributions from group pairs and non-group pairs, we can find the overlap is smaller than using RSS values. To compare features in Figure 3.5 and Figure 3.6 more objectively, we plot the receiver operating characteristic (ROC) curve of both methods. ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. True positive occurs when we correctly detect a customer pair as a group. False positive occurs when two strangers are improperly detected as a group. From Figure 3.7, we could see the performance of using RSS trend is better than RSS.

### 3.2.3 AP Significance Estimation

Although we could detect groups by identifying customers with similar RSS trend, it is not good enough as shopping groups might naturally separate sometimes. However, we have two observations that indicate different APs should have different weights in measuring customer similarity. First, groups are more likely to separate in certain areas like bookstores and supermarkets as group members might have different interests. APs in those areas should have smaller weight due to frequent separation. Second, the ratio of customer groups over individuals is higher in public entertainment areas. A study of customer behaviors indicates individual customers are less interested in public entertainment activities since they anticipate negative inferences from others about their social connectedness. Therefore, we propose a probabilistic representation of different AP weights.

To calculate the probability, we combine the WiFi data and the Bluetooth data collected from volunteers. The following information could be extracted from the combined data: the number of individual customers connected to AP  $k$  ( $N_i^k$ ); the number of group customers in AP  $k$  ( $N_g^k$ ); the number of shopping groups in AP  $k$  ( $M_g^k$ ); and the number of group separation in AP



**Fig. 3.9:** A simple example for calculating AP weights.

$k$  ( $M_s^k$ ). Figure 3.9 illustrates an example with 2 APs and some individual customers and group customers. For AP  $a_1$ ,  $N_i^1 = 3$ ,  $N_g^1 = 4$ ,  $M_g^1 = 2$ , and  $M_s^1 = 1$ .

We calculate a posterior probability ( $P(G\bar{D}|A)$ ) as the AP weight in Equation 3.5.  $A = \{a_1, \dots, a_n\}$  is a variable indicating the target AP, event  $D$  represents groups appear in the AP, and event  $\bar{D}$  means groups do not disperse within the AP coverage. Therefore,  $P(G\bar{D}|A)$  refers to the probability that groups appear and do not separate within the coverage of a certain AP.

$$P(G\bar{D}|A) = P(G|A) \cdot P(\bar{D}|A) = \frac{N_g}{N_g + N_i} \cdot \left(1 - \frac{M_s}{M_g}\right) \quad (3.5)$$

In the example of Figure 3.9,  $w_1 = P(G\bar{D}|a_1) = \frac{4}{4+3} \cdot \left(1 - \frac{1}{2}\right) = 0.29$ ,  $w_2 = P(G\bar{D}|a_2) = \frac{6}{6+1} \cdot \left(1 - \frac{0}{3}\right) = 0.86$ . It is clear that  $a_2$  has a larger weight than  $a_1$  and thus is more important in measuring customer similarity.

### 3.2.4 Group Detection

#### Customer similarity matrix

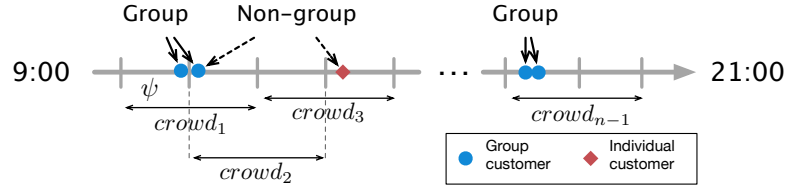
Given the weights of different APs, we could refine the similarity with:

$$\mathbf{Sim}(i, j) = \frac{\sum_{k=1}^K w_k \cdot \mathbf{d}'_k(i, j)}{\sum_{k=1}^K w_k}, \quad (3.6)$$

where  $\mathbf{Sim}(i, j) \in [0, 1]$  is the refined similarity between customers  $i$  and  $j$ ,  $\mathbf{d}'_k(i, j)$  is the similarity under AP  $k$  measured with RSS trend, and  $w_k$  is the weight of AP  $k$ .

Instead of detecting shopping groups out of all customers in a whole day, we first utilize the temporal constraint of groups to separate customers into different crowds and then identify groups out of each crowd. We equally





**Fig. 3.10:** An illustration of partitioning customers into different crowds with the temporal constraint.

partition the business hours of the mall into non-overlapping fragments using a threshold  $\psi$  which is determined by the customers' dwell time. The partition process is depicted in Figure 3.10. For each adjacent segment, we measure customers' similarity and construct a similarity matrix. The idea is quite straightforward. If two customers have a large gap in the time domain, they are more likely to be strangers.

### Group detection with matrix factorization

Group detection is essentially a hard clustering problem which means each user can only belong to a cluster or not. Existing works mostly apply graph clustering methods like Markov clustering to detect groups. Here we resort to matrix factorization for the following two reasons. First, the constructed matrix has some noises, like strangers being regarded as groups and vice versa. Matrix factorization can help in reducing these noises and preserving the latent group information. Second, matrix factorization can directly derive clustering results by imposing a sparseness constraint, which is similar to K-means but the performance is much better.

Given a similarity matrix  $\mathcal{A} \in \mathbb{R}^{m \times m}$  and an integer  $k < m$ , matrix factorization aims to find two factors  $\mathcal{W} \in \mathbb{R}^{m \times k}$  and  $\mathcal{H} \in \mathbb{R}^{m \times k}$  such that  $\mathcal{A} \approx \mathcal{W}\mathcal{H}^T$ . The solutions can be found by solving the optimization problem with nonnegative and sparseness constraints:

$$\min_{\mathcal{W}, \mathcal{H}} \frac{1}{2} \left[ \|\mathcal{A} - \mathcal{W}\mathcal{H}^T\|_F^2 + \eta \|\mathcal{W}\|_F^2 + \beta \sum_{i=1}^m \|\mathcal{H}(i, :)\|_1^2 \right] \quad (3.7)$$

*s.t.*  $\mathcal{W}, \mathcal{H} \geq 0$ ,

where  $\|\cdot\|_F$  means Frobenius Norm which has a Gaussian noise interpretation and the objective function can be easily transformed into a matrix trace version,  $\mathcal{H}(i, :)$  is the  $i$ -th row vector of  $\mathcal{H}$ . Parameter  $\eta > 0$  controls the size of the elements of  $\mathcal{W}$ . It is usually determined by the largest element

of input matrices [69]. Parameter  $\beta > 0$  balances the trade-off between the accuracy of approximation and the sparseness of  $\mathcal{H}$ . A larger value of  $\beta$  implies stronger sparseness while smaller values of  $\beta$  can achieve better accuracy of approximation. The imposed nonnegative constraint is due to physical meanings (similarity of pairwise users) of entries in the original matrix. Positive factors facilitate direct physical connections. The sparseness on the  $\mathcal{H}$  factor could directly derive the clustering results with  $\ell_1$ -norm regularization.

Although Equation 3.7 is a non-convex problem, it is convex separately in each factor, i.e., finding the optimal factor  $\mathcal{W}$  corresponding to fixed factors  $\mathcal{H}$  reduces to a convex optimization problem. Algorithms based on alternating nonnegative least squares (ANLS) are often used for sparse nonnegative matrix factorization. More details of solving the optimization problem and determining the appropriate value of  $k$  can be found in [69].
















## 3.3 Experimental Evaluation

### 3.3.1 Settings

#### Setup

We conduct experiments in a large shopping mall with 4 floors covering an area of  $4890m^2$ . This mall is at the bottom of an office building which is adjacent to a subway station. Most of the shops in the mall are related to food like restaurants and bakeries. There are originally 20 APs installed in the mall for customers to access the Internet. We use those APs to collect the WiFi data. The configurations of the AP are as follows: AR9341 (WLAN chip), 64M (RAM), and 8M (Flash Memory).

Within three weeks, we conduct 34 experiments at different times of a day. For each experiment, we recruit 2  $\sim$  4 volunteer groups with each group containing 2  $\sim$  4 customers and record their MAC addresses and grouping information. The majority of experiments last less than 3 hours. To ensure authenticity, volunteers are only told to keep the smartphone WiFi function enabled without knowing the purpose of experiments.

	Week 1	Week 2	Week 3	
Sources			 	 Volunteer group  Single volunteer customer
Data	 	 		 } Unlabeled customers  }
Purpose	AP significance estimation		Evaluation	 } Data  }

**Fig. 3.11:** Detailed information of the collected data in 3 weeks.

## Dataset

During 3 weeks, we collect the WiFi data from volunteer customers and other customers. Detailed information about the collected data is illustrated in Figure 3.11. For the first two weeks, we record volunteer customers' WiFi data and Bluetooth data to estimate the different significance of the deployed APs. For the last week, we record the WiFi data from volunteer and non-volunteer customers appeared in the mall to evaluate group detection performance.

The WiFi data comes from 104 volunteer shopping groups during 34 experiments, including 258 group pairs (positive data samples) and 864 non-group pairs (negative samples). For example, given an experiment with two shopping groups, each group has 3 customers, then we have  $2 \times C_3^2 = 6$  positive samples and  $C_3^1 \times C_3^1 = 9$  negative samples. We call the dataset above *labeled dataset* which contains only volunteers' data. In other words, we know the relation of all pairwise customers in the dataset. We also have a *semi-labeled dataset* that includes both volunteers and other customers which we do not know their grouping information. But one thing for sure is that volunteers and other customers must be strangers. Therefore, the semi-labeled dataset has much more negative samples than labeled dataset.

The Bluetooth data are from 58 volunteers during the first two weeks. Combined with the WiFi data, we find that customers groups are more likely to separate in places like restrooms (27.7%) and cosmetics shops (21.5%). One potential reason is that the mall has limited types of shops and most of the shops are related to food. As reported by our survey results, customers are not frequently get separated in restaurants.

Code	Similarity Measurement	Clustering Approach
AG	RSS trend + AP significance	Graph clustering
NM	RSS trend (No AP significance)	Matrix factorization
SNOW	RSS trend + AP significance	Matrix factorization

A Evaluate matrix factorization      B Evaluate AP significance

**Fig. 3.12:** An illustration of baseline approaches.

## 3.3.2 Evaluation

### Baseline approaches

To detect shopping groups using WiFi data, similarity measuring and group clustering are two essential steps. We have different baselines for evaluating different steps. Since it is demonstrated that RSS trend is better than RSS, all baseline approaches are based on RSS trend.

As illustrated in Figure 3.12, to evaluate the effectiveness of AP significance and matrix factorization, we need two baseline approaches apart from SNOW. The first baseline is called AG which measures customer similarity with estimating AP significance and then constructs a user graph with each node representing a customer and each edge representing the similarity between pairwise customers. Then AG detects groups with the help of Markov Cluster algorithm (MCL) [56, 131]. MCL works well when the cluster size is small and it does not require the number of clusters as an input. The second baseline is NM which measures customer similarity without AP significance and then detects groups using matrix factorization.

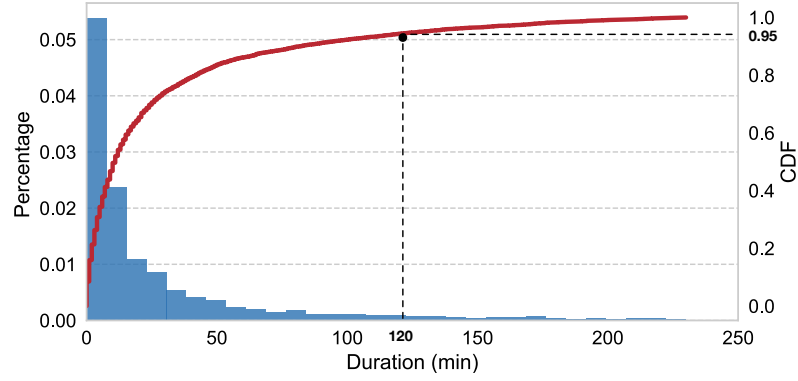
### Evaluation metric

As pointed out in [142], there is no consensus on which metrics should be used to evaluate groups detection. Here we use *precision* and *recall* to measure the performance of group detection, which are defined as:

$$\begin{cases} \text{precision} = \frac{tp}{tp+fp} \\ \text{recall} = \frac{tp}{tp+fn} \end{cases}$$

		Truth	
		$g$	$\tilde{g}$
Detection	$g$	tp	fp
	$\tilde{g}$	fn	tn

$g$ : Group  
 $\tilde{g}$ : Non-group



**Fig. 3.13:** Distribution and CDF of customers' dwell time in the mall.

**Table 3.1:** Performance comparison on both datasets.

	Labeled Dataset			Semi-labeled Dataset		
	P <sup>1</sup>	R <sup>2</sup>	F <sup>3</sup>	P	R	F
AG	0.863	0.841	0.852	0.782	0.811	0.796
NM	0.750	0.787	0.768	0.691	0.734	0.712
SNOW	0.912	0.927	<b>0.919</b>	0.833	0.860	<b>0.846</b>

<sup>1</sup> Precision    <sup>2</sup> Recall    <sup>3</sup> F-score

As shown in the confusion matrix,  $tp$  is the number of cases that positive samples being detected as groups. We also use a combined metric *F-score* defined in Equation 3.8 to represent the general performance.

$$F\text{-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (3.8)$$

### Parameter selection

We have two important parameters to determine for SNOW. First, parameter  $\psi$  represents the maximum duration time of customers. As shown in Figure 3.13, over 90% customers stay in the mall for less than 2 hours. Therefore, we simply set  $\psi = 120$  (minutes). Second, parameter  $\beta$  balances the tradeoff between accuracy of approximation and sparseness. Even though the performance is not that sensitive to  $\beta$ , too big  $\beta$  is undesirable since that might lead to worse approximation. Therefore, we set  $\beta = 0.3$  for all methods.

## Performance evaluation

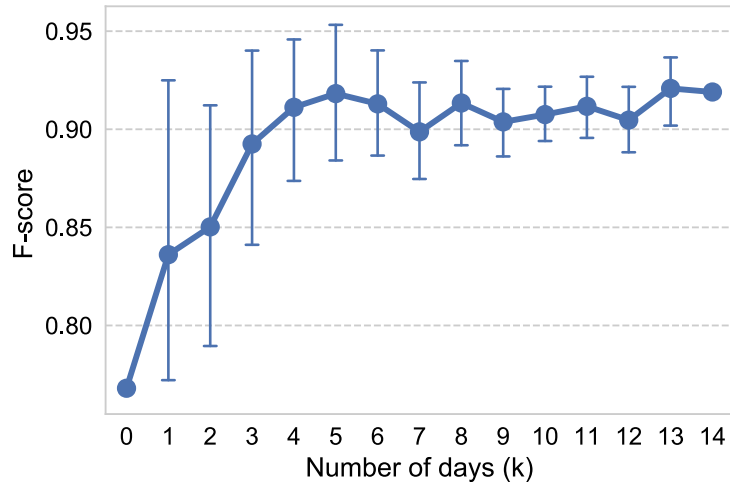
We evaluate the performance of SNOW and baseline approaches. As shown in Table 3.1, SNOW outperforms baseline approaches on both datasets by 6.3% ~ 19.7%. The performance of all three methods on semi-labeled data are slightly worse than that of labeled dataset. This effect is reasonable and could be explained by the following reasons. First, there are much more customers in semi-labeled datasets which bring in more noise for the clustering process. Second, the number of negative samples increased sharply that may cause more potential false positive detections and decrease precision.

To evaluate the effectiveness of matrix factorization, we could compare SNOW and AG. On average, SNOW outperforms AG by 7.1% in F-score which means matrix factorization could achieve better results than graph clustering. As explained, one potential reason is that matrix factorization could reduce the effect of false positive and false negative detections to some extent. Matrix factorization is known for removing noises and preserving latent group information.

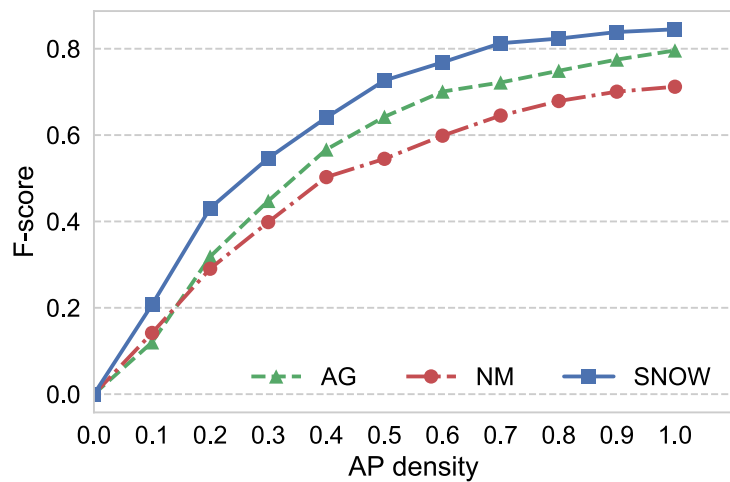
As one of the core components, AP significance is supposed to capture the dynamics of shopping groups. We could evaluate the effectiveness of AP significance by comparing SNOW and NM. From Table 3.1, on average the F-score of SNOW is 19.3% better than that of NM. This indicates that estimating AP significance could greatly improve the performance of detecting shopping groups. Since a certain number of shopping groups may actually separate from time to time, their similarity in the RSS space can be affected. Straightforward as this method is, it reflects the pattern of most shopping group activities and successfully refines customers' similarity in the signal space.

## The impact of Bluetooth data

The Bluetooth data are critical in AP Significance Estimation (Section 3.2.3). Intuitively, more adequate Bluetooth data can achieve more accurate approximation to group separation in real situations. Figure 3.14 shows the detailed performance of SNOW on the label dataset using different numbers of the Bluetooth data. When the number is  $k$ , we randomly choose  $k$  days' Bluetooth data to calculate the AP significance and refine customer similarity. When  $k$  is small, the final performance may not be good enough and there exists a



**Fig. 3.14:** Impact of the amount of Bluetooth data in estimating AP weights.



**Fig. 3.15:** Performance under different AP density on semi-labeled dataset.

large deviation. With the increase of  $k$ , the performance gradually increases and becomes more stable (the deviation gets smaller). In our scenario, we could see that using no less than one week Bluetooth data achieves relatively stable performance.

### The impact of AP density

Since different scenarios may have different AP deployments and AP densities, we evaluate the performance of different methods with various AP densities. To derive different AP density, we adopt a sampling method over the semi-labeled dataset. For example, to evaluate the performance under 0.8 AP density, we randomly choose 16 out of 20 APs and use the WiFi data of chosen APs for all users. We average the results for 100 times to eliminate the impact of randomness. As illustrated in Figure 3.15, the performance drops as AP

density decreases. One of the potential reasons is information loss. However, we can also find that SNOW still outperforms baseline approaches under different AP density.

## 3.4 Related Work

With the development of IoT [95, 94, 45, 70, 56], there exist various group detection systems. However, none of them are particularly designed for shopping groups. These methods detect groups mainly by separating strangers who are close to each other but overlook the fact that shopping groups might separate sometimes. Under this situation, existing methods might generate many false negative detections and thus degrade the usability of the system. Literature methods can be classified as vision-based approaches, sensor-based approaches, and probe-based approaches according to different means.

*Vision-based approaches* regard group detection as a task of clustering a set of users' trajectories into disjoint subsets [45, 142]. However, this kind of methods have some apparent limitations. First of all, the most significant problem is privacy erosion. Besides, video surveillance suffers from environmental issues such as non-line-of-sight, and low brightness.

*Sensor-based approaches* use wearable devices or install apps on smartphones to collect users' behavioral data. Groups are detected through correlation analysis of multiple sensor data. For instance, MIT researchers use specially designed wearable devices called "Sociometric Badges" [109, 110] to measure group behavior through face-to-face interaction and physical proximity. Some research works [70, 84, 131] combine several sensor modalities (WiFi, accelerometer, compass, etc.) to measure users' similarity. However, these methods might be difficult to collect data on a scale, as they require user intervention which would be cumbersome in some scenarios. Besides, engaging multiple sensors drains smartphone battery more quickly.

*Probe-based approaches* utilize the information contained in probe requests to detect groups. The probe contains significant information like timestamp, smartphone MAC address, RSSI, and Service Set Identifier (SSID), which enables a wide range of applications like passive tracking [35, 139], crowd counting [128, 169], and facility utilization analysis [114]. Compared to other approaches, probe-based approaches do not require high deployment



cost or user intervention. SSID and RSSI are two frequently used information to detect groups. Cunche et. al. [31, 10, 24] link different smartphones through SSID similarity. However, 80% of the devices reply with empty SSID list [59], approaches that rely on SSID may not work well anymore. Then researchers' focus transfer to RSSI which indicate users' mobility. Kjærgaard et. al. [71] extract spatial features, signal-strength features, and pseudo-spatial features from signal strength to detect social groups which they call pedestrian flocks. It is found that the performance of spatial features is unreliable since mapping RSSI into locations is not accurate enough. Besides, the mapping process itself is usually time-consuming and labor-intensive. To avoid the cumbersome mapping process, directly measure the similarity of RSSI fingerprints to detect co-located mobile users. These methods get rid of absolute locations, thus eliminate labor-intensive calibration and protect users' privacy. SocialProbe [56] considers the hardware diversity and uses the normalized RSSI vector to achieve co-location detection. However, the timing of sending probes are mainly determined by user-device interaction and internal mechanism of the device. Different devices might generate various data granularity which makes it hard to compare their similarity [42].

## 3.5 Discussion

In this section, we provide further discussions to clarify potential issues that might be confusing and unclear. Issues to be discussed including AP deployment, the energy issue, and the generality of the system.

For AP deployment, we do not have any special requirements since we do not care about where exactly customers visit. Although our observations are related to different areas in the mall, we do not need to locate the customers. Because we assume that when customers are in different areas they would connect to different APs. This assumption stands in most of the practical scenarios because the coverage radius of common APs is tens of meters. Besides, AP deployments in shopping malls are usually conducted by experts which would try to use as less APs as possible while ensuring the network quality.

For the energy issue, it is no doubt that SNOW will increase energy consumption since we exploit Arping to lure smartphones to send more packets.

However, we also need to note two points. First, the amount of extra energy consumed is very limited due to the low frequency of sending Arping packets and their small packet size [143]. Second, we notify customers of the potential energy consumption in the agreement when they initially connect to the deployed APs. If they really care about the extra energy to be consumed they would disconnect from those APs by themselves.

Lastly, the proposed system could be generalized to different shopping malls since we do not rely on specific scenarios or any device configurations. As discussed above, we do not have special requirements for the AP deployment. Our observations of group dynamics are also independent of places. They are based on the online survey with over 250 subjects (Figure 3.1) and the research result of consumer behaviors [120] from Americans, Chinese, and Indian respondents.

## 3.6 Conclusion

In this chapter, we propose a practical shopping group detection system (SNOW) using WiFi. One of our contributions is an effective heuristic that could significantly improve the detection performance of shopping groups. The heuristic indicates APs under which groups appear more frequently and barely separate should have larger weights in measuring customer similarity. Our second contribution is to apply matrix factorization to detect groups without extra clustering processes. Matrix factorization could properly handle data issues in the measured similarity including noise filtering and data completion. Besides, imposing a sparsity constraint to the factorization process could derive the clustering results directly. Finally, we conduct extensive experiments in a large shopping mall to validate the performance of SNOW. Experimental results indicate SNOW can detect over 90% groups with a precision of 91.2%.

# DMAD: Data-Driven Measuring of Wi-Fi Access Point Deployment 4

Wireless networks offer many advantages over wired local area networks such as scalability and mobility. Strategically deployed wireless networks can achieve multiple objectives like traffic offloading, network coverage and indoor localization. To this end, various mathematical models and optimization algorithms have been proposed to find optimal deployments of access points (APs).

However, wireless signals can be blocked by human body, especially in crowded urban spaces. As a result, the real coverage of an on-site AP deployment may shrink to some degree and lead to unexpected dead spots (areas without wireless coverage). Dead spots are undesirable, since they degrade the user experience in network service continuity on one hand, and on the other hand paralyze some applications and services like tracking and monitoring when users are in these areas. Nevertheless, it is nontrivial for existing methods to analyze the impact of human beings on wireless coverage. Site surveys are too time-consuming and labor-intensive to conduct. It is also infeasible for simulation methods to predict the number of on-site people.

In this chapter, we propose DMAD, a Data-driven Measuring of Wi-Fi Access point Deployment, which not only estimates potential dead spots of an on-site AP deployment but also quantifies their severity, using simple Wi-Fi data collected from the on-site deployment and shop profiles from the Internet. DMAD firstly classifies static devices and mobile devices with a decision-tree classifier. Then it locates mobile devices to grid-level locations based on shop popularities, wireless signal, and visit duration. Lastly, DMAD estimates the probability of dead spots for each grid during different time slots and derives their severity considering the probability and the number of potential users.

The analysis of Wi-Fi data from static devices indicates that the Pearson Correlation Coefficient of wireless coverage status and the number of on-site people is over 0.7, which confirms that human beings may have a significant

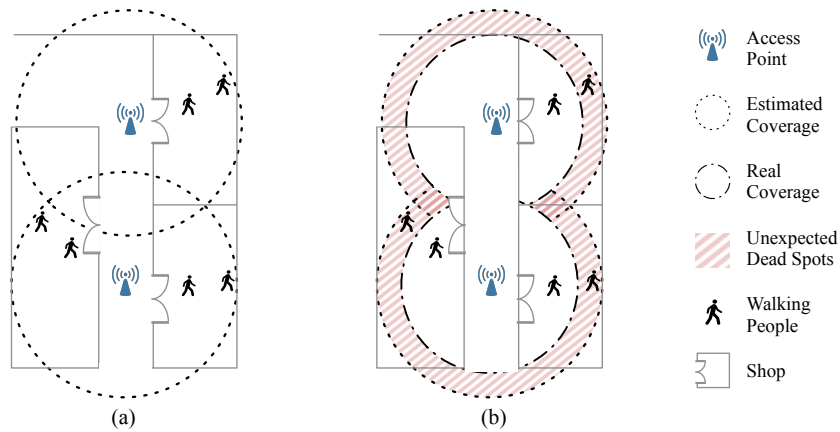
impact on wireless coverage. We also conduct extensive experiments in a large shopping mall in Shenzhen. The evaluation results demonstrate that DMAD can find around 70% of dead spots with a precision of over 70%.

## 4.1 Introduction

Wireless networks are remarkably important in modern societies, not only just for wireless communication, but also as a key enabler of numerous novel applications. One of well-known examples is wireless based tracking and monitoring systems [105, 139, 7, 161]. Besides, wireless networks offer many advantages over wired local area networks such as scalability and mobility. It is convenient to access network resources from any locations within the coverage of wireless networks. It can also be set up easily in a quick and expandable way. Lastly, wireless networks are cost-effective since wiring costs are eliminated or reduced.

As one of the predominant problems, wireless networks layout problem or access point (AP) deployment problem has been extensively studied over decades [2], since strategically deployed wireless networks can achieve multiple objectives like maximizing ratios of traffic offloading [18] and wireless coverage [23], and improving indoor localization accuracy [23]. Existing solutions can be classified as site surveys and simulation approaches. For site surveys, engineers with electronic monitoring equipment such as spectrum analyzers walk throughout the facility to measure wireless signal quality. Based on the information, engineers attempt to identify potential locations for APs that would minimize the disruption of service [121]. However, site surveys require specialized equipment and extensive manpower which is quite expensive, especially for large areas. Therefore, many simulation approaches [88, 101, 163] are proposed by modeling the AP deployment problem as an optimization problem. Those simulation methods are mostly built on the basis of propagation loss models which characterize how wireless signal attenuate over distances and different obstacles. Simulation methods usually consist of two stages, an iterative stage to calculate the minimum number of APs and an optimization stage to find out optimal locations of those APs towards one or more objectives.

However, Wi-Fi signals can be blocked by human body [132], especially in crowded urban spaces. As a result, it could result in unexpected dead spots



**Fig. 4.1:** A simple illustration of the impact of human beings on wireless coverage. (a) Ideal coverage of APs; (b) Real coverage of APs in the presence of walking people. The shadow areas in (b) are potential dead spots caused by human beings.

(areas without wireless transmission coverage) as illustrated in Figure 4.1 (b), where walking people cause real coverage of the on-site AP deployment to shrink to some extent. These dead spots are undesirable, since they degrade the user experience in network service continuity on one hand, and on the other hand paralyze some applications and services like tracking and monitoring when users are in these areas. Nevertheless, it is nontrivial for existing methods to analyze the impact of human beings on wireless coverage. It is too time-consuming and labor-intensive to measure wireless coverage status for a long time using site survey methods. Also, site surveys may disrupt the ongoing activities (like shopping activities) in the facility. For simulation methods, it is infeasible to consider the impact since the number of people cannot be determined. Moreover, neither of site surveys nor simulation approaches is able to evaluate the severity of different dead spots in a quantitative way.

As explained above, wireless networks can suffer from unpredictable influences of changeable interactions between multiple devices, specific hardware, and human activities. These influences might further lead to a difference between real-world functioning and design-time functioning [74]. Recently, the data-driven design of intelligent wireless networks is gaining popularity due to its capability to better understand the behavior of complex systems that cannot be easily modeled or simulated. Data science or “data-driven research” is a research approach that uses real-life data to gain insights about the behavior of systems. It enables the analysis of various systems to assess whether they work according to the intended design and as seen in simulations.

In this chapter, we propose **DMAD**, a **Data-driven Measuring of Wi-Fi Access point Deployment** to estimate dead spots and quantify their severity based on simple Wi-Fi data collected from the on-site AP deployment and shop data from the Internet. DMAD firstly classifies static devices and mobile devices with a decision-tree classifier. Then it locates mobile devices to shop-level locations on the basis of two observations of heuristics. 1) We find that the visit duration in different shops is different, for example, people stay longer in restaurants than in clothing shops. 2) Different shops have different popularity in attracting customers, thus the probability of people appearing in a shop should closely relate to the popularity. These two observations could help to improve the accuracy of shop-level localization. Lastly, for each area, we estimate the probability of a dead spot in different time slots and derive their severity combining the probability and the number of people. Since if a dead spot appears in an area with more potential users, its severity should be higher.

The contributions of this work are summarized as follows.

- To the best of our knowledge, we are the first to propose the AP deployment measuring problem (ADM).
- We also propose a data-driven approach (DMAD) to solve the ADM problem, which can identify around 70% of dead spots with a precision of over 70%.
- The performance of DMAD is carefully evaluated using data collected from a real AP deployment of a large shopping in Shenzhen.

The remainder of the chapter is organized as follows. In the next section, we summarize the related work. In section 4.2, we give an overview, including the preliminaries, the feasibility of using Wi-Fi data to study wireless coverage, the impact of people on wireless coverage, and the framework of DMAD. Section 4.3.1 to Section 4.3.4 elaborate on each component of DMAD. In section 4.4, we present the detailed evaluation of each component and followed by a conclusion.

## 4.2 Overview

In this section, we give an overview of DMAD by introducing the basics of Wi-Fi AP deployment measuring problem, studying the feasibility of using

**Table 4.1:** Notions used in this chapter.

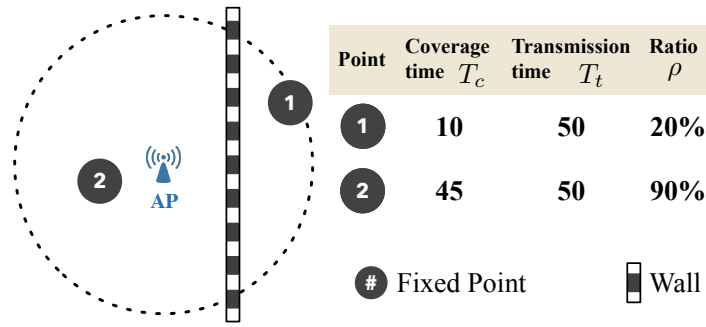
Symbol	Explanation
$\mathcal{A}$	A set of APs, $\mathcal{A} = \{a_1, a_2, \dots\}$
$\mathcal{S}$	A set of shops, $\mathcal{S} = \{s_1, s_2, \dots\}$
$\mathcal{D}$	A set of smart devices, $\mathcal{D} = \{d_1, d_2, \dots\}$
$\mathcal{G}$	A set of non-overlapping grids, $\mathcal{G} = \{g_1, g_2, \dots\}$
$\mathcal{T}$	A set of time slots, $\mathcal{T} = \{t_1, t_2, \dots\}$ , $t_i$ is a period of time
$M_j$	Connectivity matrix of $d_j$ , $M_j = [V_j(1) \quad V_j(2) \quad \dots]$
$V_j(i)$	Connectivity vector of $d_j$ at time $i$ , $V_j(i) = [v_{i1} \quad v_{i2} \quad \dots \quad v_{i \mathcal{A} }]^T$
$v_{ij}$	Binary variable, $v_{ij} \leftarrow 1$ if $a_j$ hears from the device at time $i$
$\rho$	Coverage ratio of an AP, $\rho = T_c/T_t$
$T_c$	Coverage time, how long an AP can hear from a device
$T_t$	Transmission time, how long a device sends packets
$\zeta$	Ratio of change, $\zeta \in [-1, 1]$
$\Omega_j(i)$	Coverage ratio vector of $d_j$ during $t_i$ , $\Omega_j(i) = [\rho_1 \quad \dots \quad \rho_{ \mathcal{A} }]^T$
$\mathbf{D}_w$	Unlabeled Wi-Fi data, $\mathbf{D}_w = \{\mathcal{E}_1, \mathcal{E}_2, \dots\}$ , $\mathcal{E}_1 = (a_i, d_j, t_{start}^k, t_{end}^k)$
$\mathbf{D}_w^*$	Labeled Wi-Fi data, with the label of grid information
$\mathbf{D}_s$	Unlabeled Shop data, $\mathbf{D}_s = \{\mathcal{I}_1, \mathcal{I}_2, \dots\}$ , $\mathcal{I}_1$ is a set of attributes
$\mathbf{D}_s^*$	Labeled Shop data, labels are # of people and their duration time
$\hat{\mathcal{L}}_j(i)$	Estimated location of device $d_j$ at time $i$
$R$	$R = (n_{ij})$ , $n_{ij}$ is the number of people in shop $s_j$ during $t_i$
$H$	$H = (\eta_{ij})$ , $\eta_{ij}$ is number of people in $g_j$ during $t_i$
$\mathcal{S}_j$	A set of shops that are located in grid $g_j$
$\bar{\mathcal{N}}_j$	A set of grids that are neighboring to grid $g_j$
$\mathbf{V}_j$	A set of connectivity vectors collected in grid $g_j$

Wi-Fi data to measure wireless coverage and dead spots, and investigating the impact of people on wireless coverage.

## 4.2.1 Preliminaries

DMAD is a data-driven approach to measuring wireless coverage of a given AP deployment. First, we collect Wi-Fi data from deployed APs and shop data from the Internet. Then we conduct a comprehensive analysis on the data to estimate dead spots and quantify their severity. Some of the notions used in this chapter are listed in Table 4.1.

In this subsection, we first investigate the feasibility of using Wi-Fi to measure wireless coverage and dead spots. Then we discuss the latent issues of coverage ratio and dead spots. Lastly, we study the impact of human presence on wireless coverage.



**Fig. 4.2:** Example of using Wi-Fi data to represent wireless coverage status. The unit of  $T_c$  and  $T_t$  are both minute. Coverage ratio  $\rho = T_c/T_t$ .

## Feasibility Study

Before estimating dead spots, we study the feasibility of using Wi-Fi data to measure AP coverage status and how to represent dead spots.

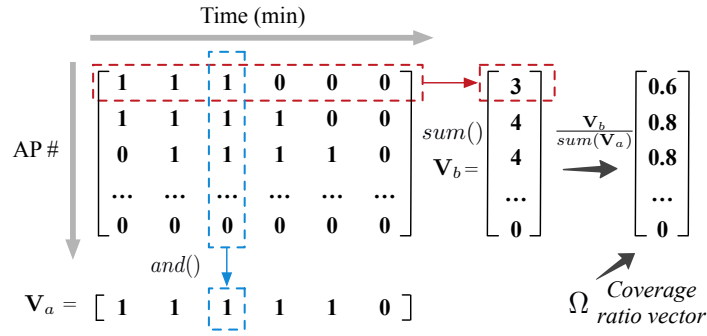
Figure 4.2 depicts a simple scenario with one AP and two fixed points. There are two smartphones at both points respectively, keeping broadcasting Wi-Fi packets. Given transmission time  $T_t$  and coverage time  $T_c$  of both devices, how to measure the wireless coverage status at both points?

The transmission time represents the total amount of time of sending packets on the smartphone, while the coverage time means the total amount of time of receiving packets from a smartphone on the AP side. As smartphones would send many packets within one minute, we count  $T_t$  and  $T_c$  in a granularity of one minute, which means if the smartphone send any packet(s) in the duration of one minute we increase  $T_t$  by 1. Due to packet loss and physical constraint (like distance), we have  $T_t \geq T_c$ .

Ideally, if the coverage status is good enough,  $T_c$  would approximate to  $T_t$ . Otherwise,  $T_c$  would be much smaller than  $T_t$ . Based on this intuition, we use a coverage ratio  $\rho = T_c/T_t$  to represent the wireless coverage status on a fixed point. In the example of Figure 4.2,  $\rho_1 = 10/50 = 0.2$ ,  $\rho_2 = 45/50 = 0.9$ ,  $\rho_k$  represents the coverage status on the  $k$ -th point. The larger the ratio is, the better the coverage is.

Then what is relationship between coverage status and dead spots under this simple scenario? Simply speaking, a point with good coverage status (i.e., large coverage ratio) is impossible to be a dead spot. Instead, those with





**Fig. 4.3:** Illustration of using connectivity matrix  $M$  to calculate coverage ratio vector  $\Omega$ .  $and()$  does the AND-operation of the column vector.

terrible coverage status are more likely to be dead spots. So we propose a probabilistic representation for dead spots based on coverage ratio as illustrated in Equation 4.1.  $P_{DS}(p_i)$  is the probability that point  $p_i$  is a dead spot.

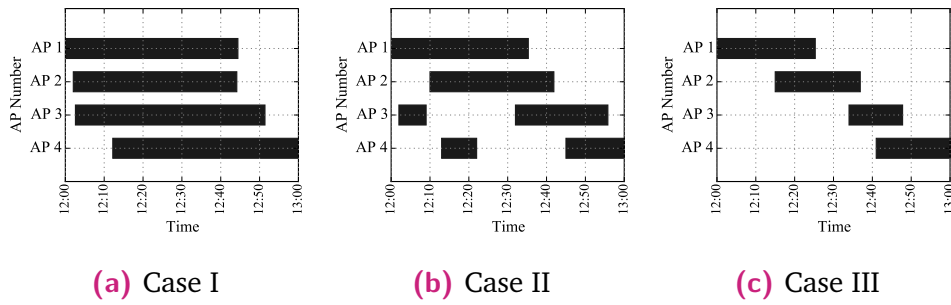
$$P_{DS}(p_i) = 1 - \rho_i \quad (4.1)$$

Since wireless coverage status would change over time, using deterministic representation of  $P_{DS}$  might be error-prone, it is better to utilize such a probabilistic representation. Under this definition,  $P_{DS}(p_1) = 0.8$  and  $P_{DS}(p_2) = 0.1$ , which means  $p_1$  is more likely to be a dead spot.

However, the example in Figure 4.2 just illustrates the simplest case. In real scenarios, we have three imperative issues. Firstly, we can never know the exact transmission time  $T_t$  of a smartphone by passively sniffing its Wi-Fi data. Secondly, a location could be covered by multiple APs with the same SSID. Lastly, the relation between coverage status and dead spots is much more complex than that of the simplest example.

For the first issue, we assume that if the smartphone sends any packets, at least one AP would receive the packet; otherwise, the smartphone is not sending any packets. Based on this assumption, we can calculate an approximation  $\hat{T}_t$  of  $T_t$ .

For the second issue, when a location is covered by multiple APs, we can transform the Wi-Fi data from user's device  $d_j$  into a connectivity matrix  $M_j$  using Algorithm 2 described in Section 4.3.1. Figure 4.3 gives an example connectivity matrix and shows how to derive the coverage status from the matrix. Each row vector  $V_j(i)$  in  $M_j$  is called a connectivity vector. It shows the connectivity information of the device with all APs at a specific time  $i$ . For



**Fig. 4.4:** Typical examples of different coverage statuses at different locations from 12:00 ~ 13:00. The black bar of an AP indicates the AP can “hear”\* from the device on that location. Intuitively, the order of coverage status is: Case I > Case II > Case III.

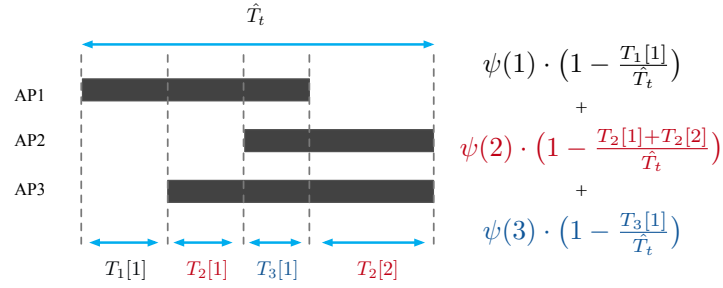
example,  $V_j(i) = [1 \ 1 \ 0 \ 0]^T$  means the device  $d_j$  is within the coverage of AP  $a_1$  and  $a_2$  at time  $i$ . As shown in Figure 4.3,  $\hat{T}_t$  can be calculated by summing up the vector  $V_a$ . Then for each AP, we can calculate its coverage time, and then derive its coverage ratio.  $\Omega = [\rho_1 \ \rho_2 \ \dots]^T$  is a coverage ratio vector containing coverage ratios of all APs.

For the last issue, our basic idea is still that a point with terrible coverage is more likely to be a dead spot, but it requires more meticulous design.

We show three typical coverage status in Figure 4.4. Intuitively, the order of coverage status should be Case I > Case II > Case III. Since in Case I, the coverage ratios of 4 APs are very large; while in Case III, the ratios are all quite small. This ranking can be explained from another perspective, all large ratios indicate the location is covered by multiple APs for most of the time, thus the probability of dead spots is significantly smaller than that of all small ratios.

Based on the observation above, we devise Equation 4.2 to map a connectivity matrix  $M_j$  into the probability of dead spots. Actually,  $1 - \text{sum}(T_i)/\hat{T}_t$  is the coverage ratio when the location or area is covered by  $i$  APs.  $T_i$  is an array of coverage time when covered by  $i$  APs.  $\text{sum}(T_i)$  sums up  $T_i$ . Since  $\hat{T}_t$  is just an approximation of the real transmission time as mentioned in the earlier part of this section, we use a decay function  $\psi(i)$  to represent the initial probability of dead spots when covered by  $i$  APs. Figure 4.5 illustrates the

\*“hear” means the AP receives any packet(s) from the device



**Fig. 4.5:** Illustration of translating a connectivity matrix into probability of dead spots.

idea of Equation 4.2 by showing an example of applying the equation.  $T_2[1]$  means the first coverage time when the location is covered by two APs.

$$P_{DS}(M_j) = \sum_{i=1}^{|\mathcal{A}|} \left( \psi(i) \cdot \left(1 - \frac{\text{sum}(T_i)}{\hat{T}_t}\right) \right) \quad (4.2)$$

### More Discussion on Coverage Ratio and Dead Spots

Here we discuss two issues to clarify both concepts and eliminate potential misunderstandings of coverage ratio and dead spots.

The first issue is about measuring coverage status using data collected on the AP side. If an AP can hear from a device, it is very likely that the device can also hear from the AP. However, even though a device can hear from an AP, the AP sometimes cannot hear back from the device. Since the transmit power of an AP is usually larger than that of a mobile device. This indicates that using data collected on APs and data on mobile devices represent different coverage. The coverage from the AP side is a proper subset of the coverage from the device side. DMAD focuses on the former coverage which is more meaningful. If the AP cannot hear from the device, a range of services and applications residing on the AP side like passive tracking [105] cannot work. Worse still, devices cannot access the Internet.

The second issue is about situations where DMAD cannot work. DMAD does not estimate dead spots by directly checking whether there is wireless coverage or not which is the main idea of site survey. Instead, it estimates the probability of dead spots in a given area (around  $20m \times 20m$ ) for a period of time based on the coverage status. The coverage status cannot be calculated without coverage time  $T_c$  and estimated transmission time  $\hat{T}_t$ . Both  $T_c$  and  $\hat{T}_t$

are derived from the packets heard on the AP side. Therefore, if a device has already been on a dead spot, or the device does not send any packets, DMAD cannot work properly.

However, in real scenarios, both situations are very rare. Firstly, DMAD only estimates the probability of dead spots in expected coverage area, which are supposed to have wireless coverage in normal circumstances. Dead spots in expected coverage area are caused by human body and change with on-site people and cannot cover a large area. Therefore it is quite rare that devices are within dead spots all the time. Secondly, smartphones are keeping broadcasting packets even not in use, which is explained in Section 4.3.1.

### Impact of People on Wireless Coverage

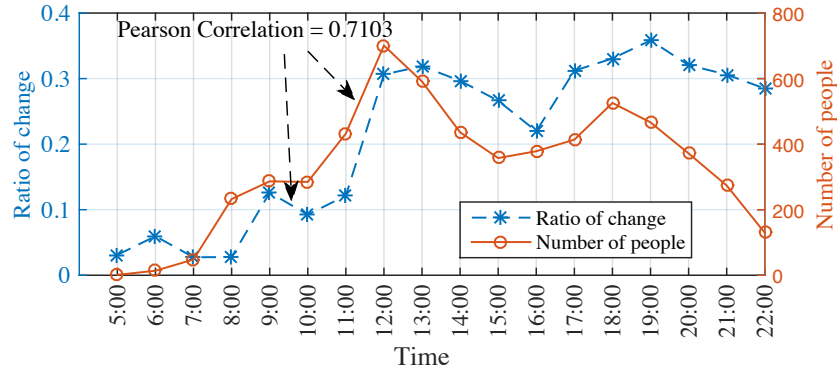
We manage to find some static devices which are fixed in locations, such as desktops, smart TVs, and IP cameras, in a large shopping mall in Shenzhen. How to find those devices is well-explained in Section 4.3.2. For each static device, we transform its Wi-Fi data into 24 connectivity matrices, with each matrix represents the connectivity information for one hour. Then we calculate the coverage ratio vector from connectivity matrix following the procedure in Figure 4.3.

Usually, early in the morning, there are no people except few on-duty securities in a shopping mall. Therefore, coverage ratio vector  $\Omega(s)$  of that period of time reflects the wireless coverage without people. While coverage ratio vector  $\Omega(d)$  during the time of 5:00 ~ 22:00 indicates the wireless coverage in the presence of human.

We use Equation 4.3 ~ 4.4 to measure the change from  $\Omega(s)$  to  $\Omega(d)$ . The output of the function is a *ratio of change*  $\zeta \in [-1, 1]$ . The larger  $\zeta$  is, the poorer the coverage status is compared to  $\Omega(s)$ .

$$\zeta = \left( \frac{\Omega(s)}{\text{sum}(\Omega(s))} \right)^T \cdot \frac{\Omega(s) - \Omega(d)}{\text{sum}(\Omega(s) - \Omega(d))} \quad (4.3)$$

$$\text{sum}(\Omega(s)) = \sum_{i=1}^{|\mathcal{A}|} \rho_i \quad (4.4)$$



**Fig. 4.6:** Correlation analysis of the average ratio of change and the average number of people from the data collected in a shopping mall in Shenzhen for 46 days.

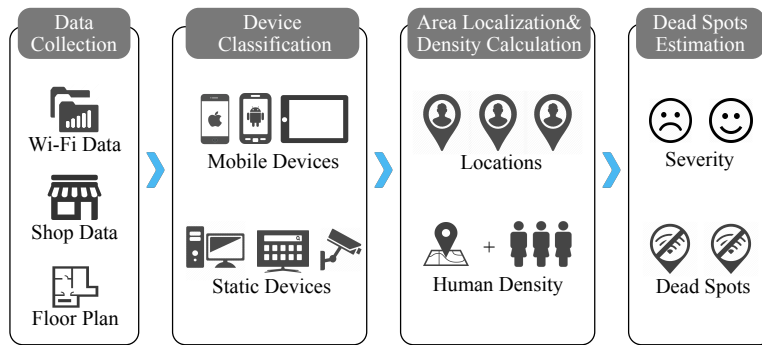
In Equation 4.3,  $\left(\frac{\Omega(s)}{\text{sum}(\Omega(s))}\right)^T$  is the transpose of normalized  $\Omega(s)$ . Those APs with large coverage ratios play dominating roles in coverage status, therefore their changes should have a larger weight than that of APs with small ratios.

To calculate the ratio of change, we set 3:00 ~ 4:00 as  $\Omega(s)$  and each hour in 5:00 ~ 22:00 as  $\Omega(d)$ . Then based on the data collected from the shopping mall in 46 days, we derive the average ratio of change for each hour.

Compared to static devices, it is easier to find mobile devices which are carried by people in the mall. The basic idea is to use the unique MAC address of mobile device to represent a mobile user. Detailed process can be found in Section 4.3.2. So we can also calculate the average number of people in different hours from the accumulated data. The results of correlation analysis of the ratio of change and the number of people are shown in Figure 4.6. We can see that the Pearson Correlation Coefficient of them is over 0.7, which indicates people might have a non-negligible impact on wireless coverage and the impact increases with the number of on-site people.

## 4.2.2 Framework of DMAD

DMAD has four components as depicted in Figure 4.7. The first component is “Data collection”, which is to collect desired input data, including Wi-Fi data, shop data, and the floor plan. Then the data are used for “Device classification”, which sorts out static devices. After that, we use “Area localization and density calculation” to estimate mobile devices’ shop-level locations and



**Fig. 4.7:** The framework of DMAD. It consists of four components, data collection, device classification, area localization and density calculation, and dead spots estimation.

analyze human density in different areas and time slots respectively. Lastly, “Dead spots estimation” estimates the probability of dead spots and their severity.

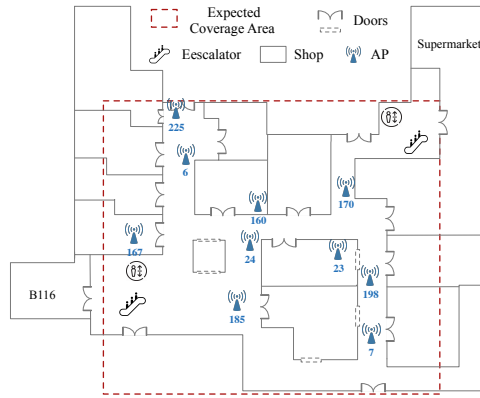
## 4.3 System Design

In this section, we elaborate on the design details of DMAD. It consists of four main components, namely Data collection, Device classification, Area localization and density calculation, and Dead spots estimation.

### 4.3.1 Data Collection

Data collection is the first component of DMAD, it serves as data input for the whole system. We collect two sources of data, Wi-Fi data from deployed APs, and shop data from the Internet. The purposes of collecting Wi-Fi data is to determine grid locations, measure wireless coverage, and estimate dead spots of an on-site AP deployment. While shop data can help to improve the accuracy of area localization. In this section, we show that both of Wi-Fi data and shop data can be readily collected by introducing details of the data collection processes.

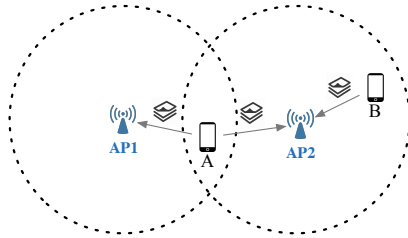
Wi-Fi data consists of two parts, a large amount of unlabeled Wi-Fi data  $D_w$  collected from mobile users inside the mall, and a small amount of labeled Wi-Fi data  $D_w^*$  from volunteers.



**Fig. 4.8:** AP deployment and expected coverage area on the ground floor of the mall.



**Fig. 4.9:** Grid partition on expected coverage area of the ground floor.



AP	Device	Start time	End time
AP1	A	10:00	10:20
AP2	A	9:50	10:10
AP2	B	10:12	10:32

AP   
 Smartphone   
 Wi-Fi packets

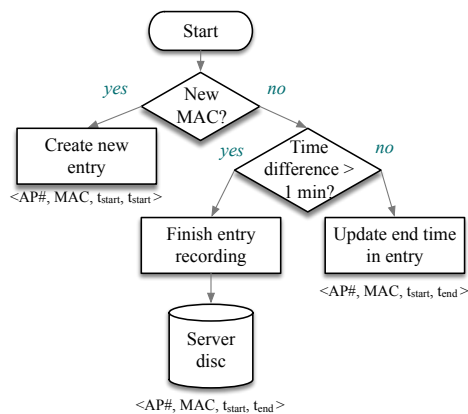
**Fig. 4.10:** A simple illustration of the Wi-Fi data collection.

### Unlabeled Wi-Fi data $D_w$

$D_w$  is collected from a large shopping mall in Shenzhen, where we have previously installed 48 APs among 5 floors. The original purpose of the Wi-Fi network is to provide Internet access for customers in common areas, but we also find that it can be utilized for other applications or services, like indoor localization [139]. Here we study the problem of measuring Wi-Fi AP deployment in expected coverage areas based on the accumulated Wi-Fi data (46 days in total, starting from 1 May 2015). Figure 4.8 shows the AP installation and expected coverage area on the ground floor.

The unlabeled Wi-Fi data is passively collected from users' smartphones. As smartphones keep broadcasting Wi-Fi packets [42] which can be sniffed by off-the-shelf APs. Even users are not using Wi-Fi services, smartphones send out packets (e.g., probe requests) intermittently [105]. Figure 4.10 illustrates a simple scenario and lists some descriptive data records.

Each AP works under OpenWrt<sup>†</sup> system, with a monitor mode virtual network interface<sup>‡</sup> enabled. We run Tcpcdump (a utility for capturing network traffic) to sniff nearby wireless traffic. More specifically, we use each AP to collect tuples in the format of  $\langle AP\#, MAC, t_{start}, t_{end} \rangle$  and once the entry is finished, the AP uploads it to the server and then deletes the local entry file. Detailed process is illustrated in the flowchart of Figure 4.11.



**Fig. 4.11:** Flow chart of collecting Wi-Fi data in APs.

**Table 4.2:** A fraction of raw Wi-Fi data. MAC has been hashed.

AP #	MAC	$T_{start}$	$T_{end}$
47	3891527	1431652262	1431652271
12	160458	1431721805	1431721810
6	1200164	1431800823	1431800828
30	528517	1431879976	1431880176
6	1585005	1431951873	1431952173
14	398316	1431968982	1431968987
2	685499	1432033589	1432034997
4	102681	1432114009	1432114014
...	...	...	...
22	1114093	1432160871	1432160896
25	1832169	1432514915	1432515031
46	4234664	1432302241	1432302400
8	493476	1432324267	1432324290
22	100731	1432386985	1432387036

As can be seen from the flowchart, collecting the Wi-Fi data does not require analyzing each packet and extracting the information like received signal strength indicator. Instead, we only record the connectivity information, i.e., whether the smartphone is under the coverage of an AP. The advantages are twofold, on the one hand, the connectivity information is easy to collect, it does not add too much burden to those APs. On the other hand, it saves much space compared to storing information from every packet, which could be incredibly huge in volume (several Giga bytes from all APs for only one day).

Table 4.2 shows a small fraction of the raw Wi-Fi data. It has four fields,  $AP\#$  shows the id of the AP that hears from the device;  $MAC$  is the hashed MAC address of the device;  $T_{start}$  is a timestamp that the device is heard for the first time; and lastly  $T_{end}$  is a timestamp that the device is last heard by the AP.

<sup>†</sup>OpenWrt (<https://openwrt.org/>) is a highly extensible GNU/Linux distribution for embedded devices (typically wireless routers).  
<sup>‡</sup>virtual network interface, <https://wiki.openwrt.org/doc/networking/network.interfaces>



Then the raw Wi-Fi data is transformed into connectivity matrix using Algorithm 2. An example of connectivity matrix can be found in Figure 4.3.

---

**Algorithm 2:** Transform raw Wi-Fi data into connectivity matrices

---

**Data:** Raw Wi-Fi data from all APs in a day

**Result:**  $K$  connectivity matrices:  $\{M_1, \dots, M_K\}$

```

1  $users \leftarrow$  Group the raw data by the field of MAC address;
2 for  $user_i \in users$  do
3    $entries_i \leftarrow$   $user_i$ 's raw Wi-Fi data ;
4   Create a zero ( $|\mathcal{A}| \times 1440$ ) matrix  $M_i = (m_{ij})$ ;
5   for  $entry \in entries_i$  do
6      $entry \leftarrow (a_j, d_i, t_s, t_e)$  ;
7     transform  $t_s, t_e$  into the order of the matrix  $s, e$ ;
8     for  $k \in [s, e]$  do
9        $m_{jk} \leftarrow 1$ 

```

---

### Collecting labeled Wi-Fi data $D_w^*$

The label of  $D_w^*$  is the grid information which is manually separated. We separate the expected coverage area of the mall into 60 grids, Figure 4.9 illustrates the grid partition of the ground floor. To collect the data, we engage over 20 volunteers in a week with different smartphones including popular iOS and Android devices. The purpose of  $D_w^*$  is for area localization (in Section 4.3.3), which estimates people's area locations based on their Wi-Fi data.

Volunteers are required to collect some "wireless fingerprints" in specific grids following the procedure below. Firstly get to the grid, turn on the Wi-Fi function and record the start time, then they walk around within the grid, after visiting all feasible locations of the grid, record the end time and turn Wi-Fi off. It usually takes 5 to 10 minutes to finish a collection process.

Since the presence of people can block wireless signal and it will have a negative impact on localization performance, we separate the daytime into several time slots  $\mathcal{T} = \{t_1, t_2, \dots\}$  and collected fingerprints for each time slot. In this way, we collect over 1,500 Wi-Fi data entries for  $D_w^*$ .

Although DMAD requires such a labeling process, it takes much less effort compared to site survey. As described in Section 4.4.1, even simplified site survey usually takes 500 ~ 700 seconds to check whether dead spots exist in a grid. While the labeling work takes shorter time (300 ~ 600 seconds). More importantly, dead spots are related to human activities, the detection results may become invalid over time. To detect dead spots next time, site survey needs to start from scratch, while DMAD does not bother to do that, since it merely requires a one-time investment.

Shop data also consists of two parts, unlabeled shop data  $D_s$  from the Internet, and some labeled shop data  $D_s^*$  collected by volunteers.

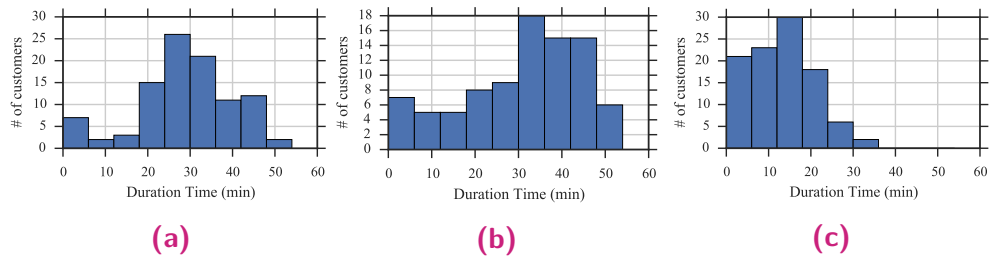
### Collecting labeled shop data $D_s^*$

The labels of shop data are the number of people and their visit duration in different shops and time slots respectively. The number of people is used to calculate a prior probability that people appear in a shop. While visit duration is another kind of “fingerprints”, we observe that the time spends in visiting different shop is also different, and we take it as another feature to distinguish users’ area locations.

We collect shop data in different time slots in a day since shops have different popularities during different time slots. For example, restaurants gain more customers during dinner time than clothing shops. To collect the ground truth about the number of people in a shop, we send volunteers to different shops to count the number of customers at different time slots. It usually takes 1 or 2 minutes for volunteers to finish the data collection task. The ground truth is represented in a matrix  $R$  in Equation 4.5, where  $n_{ij}$  is the number of people in shop  $s_j$  during time slot  $t_i$ .

$$R = \begin{bmatrix} n_{11} & n_{12} & n_{13} & \dots & n_{1|S|} \\ n_{21} & n_{22} & n_{23} & \dots & n_{2|S|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_{|\mathcal{T}|1} & n_{|\mathcal{T}|2} & n_{|\mathcal{T}|3} & \dots & n_{|\mathcal{T}||S|} \end{bmatrix} \quad (4.5)$$

For visit duration, we use a distribution to represent the duration time in a shop, as it differs from person to person even for the same shop. To collect the data, we ask volunteers to stay near the entrance or the exit of a shop and



**Fig. 4.12:** Histogram of duration time of around 100 customers from three different shops. (a) A fast food restaurant; (b) A traditional Chinese restaurant; (c) A woman accessories shop.

record the visit duration of customers. Figure 4.12 shows the duration time of three different shops with around 100 samples. Generally, the distribution can be approximated using a normal distribution.

However, it is too labor-intensive and time-consuming to collect the distribution of duration time for all shops. We believe that the duration time of a shop is closely related to its type and user ratings. For example, people usually stay in restaurants for around 20 minutes. Also, if the restaurant has a pleasant environment and satisfactory services, customers may choose to stay longer. These observations can be quickly verified from the comparison of the three shops in Figure 4.12. So we just collect duration time in some typical shops of each category and crawl all shop profiles from the Internet. Then we utilize machine learning techniques to predict the distribution of unlabeled shops. Detailed explanation can be found in Section 9.

### Collecting unlabeled shop data $D_s$

We collect shop profiles in that mall from Dianping<sup>§</sup> and AutoNavi<sup>¶</sup>. For each shop in that mall, we crawl its type (like clothing shop, restaurant), location, the number of positive comments (comments with more than 3 stars), and user ratings about products, environment, and services between 1 May 2015 and 15 June 2015. We exploit Scrapy<sup>||</sup> to crawl the desired data and save them to a local file. A fraction of collected data is shown in Table 4.3.

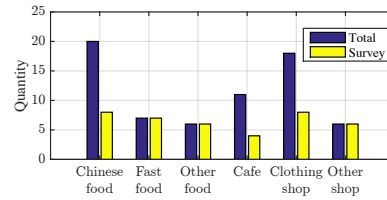
<sup>§</sup>Dianping (<https://www.dianping.com/>), a popular Chinese group buying website for locally found consumer products and retail services.

<sup>¶</sup>AutoNavi (<http://www.gaode.com/>), a well-known map website in China.

<sup>||</sup>Scrapy (<https://scrapy.org/>), an open source and collaborative framework for extracting data from websites.

**Table 4.3:** A fraction of unlabeled shop data. Some of fields like, floor, location, and average spend is not shown in the table.

Name	Type	Likes	Product	Env	Service
Cafe de Coral	Restaurant	41	7.6	7.8	7.6
King of Pastry	Restaurant	3	6.8	6.8	6.9
Benbo	Clothing	2	7.1	7.1	7.1
Shiny Nail	Make-up	32	7.9	8.2	8.3
Costa coffee	Cafe	4	6.8	7.1	7
belle	Clothing	3	6.8	6.8	6.8
Muji	Clothing	8	7.3	7.3	7.3



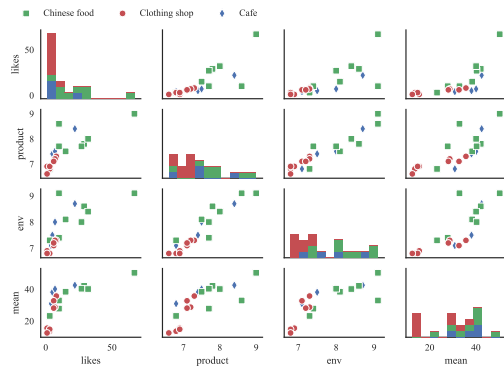
**Fig. 4.13:** Distribution of total number and surveyed number of shops.

There are 68 shops of interest in that mall, and we classify those shops into 6 categories. The total number and the number of surveyed shops are shown in Figure 4.13. Among the 6 categories, Chinese food restaurants, clothing shops, and cafes are top 3 categories in terms of total number and we collect duration time from some of these categories. For other categories, we just collect data from all shops.

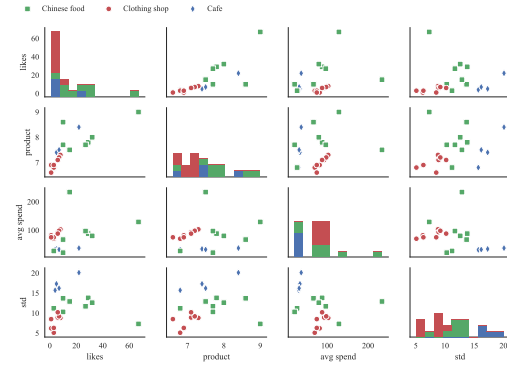
To predict mean and standard variance of the distribution is a regression problem. The predictor variables are shop type, location, average spend, and user ratings (include service, product, environment, number of positive comments). The response variables are mean and standard deviation (std) of the visit duration distribution. Both response variables are independent, so we can simply use two regression models to regress them.

We conduct regression analysis and show the relation between some predictor variables and both response variables respectively in Figure 4.14 ~ 4.15. For mean, we can see that, there exists a strong linear-log relation [13] between predictors and the response. So we use ordinary least square to estimate the unknown parameters. The regression results indicate that R-squared of the model is 0.810. We also utilize 5-fold cross validation to evaluate the accuracy of the regression model. The root mean squared error is 4.611.

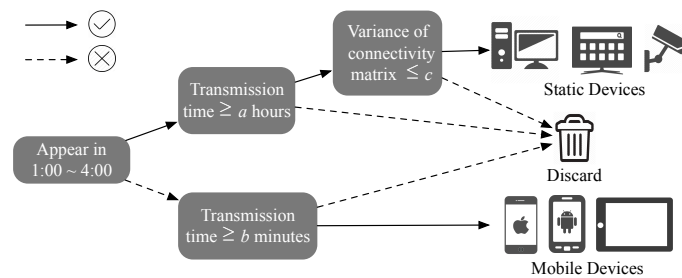
For std, there does not exist obvious relation from the perspective of all data, but the data shows strong cohesion within the same kind of shops. So for each category we use a simple linear regression model to regress the response. We use 2-fold cross validation to evaluate those linear regression models. The average root mean squared error is 3.623.



**Fig. 4.14:** Regression analysis between some predictor variables and mean of the duration distribution.



**Fig. 4.15:** Regression analysis between some predictor variables and standard deviation of the duration distribution.



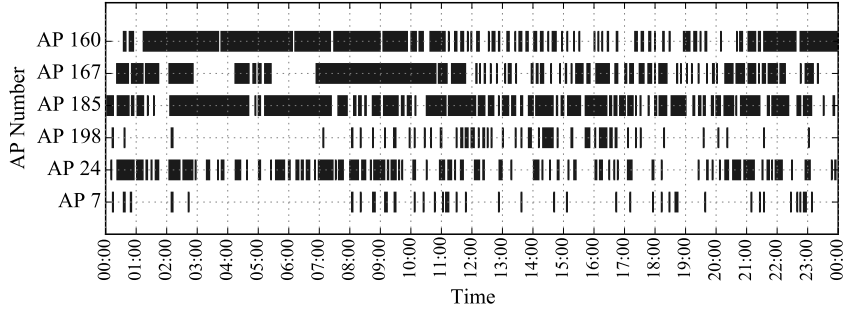
**Fig. 4.16:** Decision tree for classifying static and mobile devices.

### 4.3.2 Device Classification

Device classification classifies devices as static devices and mobile devices. Static devices are defined as devices fixed in locations like desktops and IP cameras, while mobile devices could easily change their locations with the help of people, such as smartphones and tablets.

Static devices can be utilized to study the impact of people on wireless coverage status of those fixed locations. The results are demonstrated in Section 4.2.1. For mobile devices, their Wi-Fi data can be exploited to infer people’s locations and mobility patterns.

We propose a decision-tree classifier to classify devices, as illustrated in Figure 4.16. First of all, the most distinguishing feature between mobile and static devices is that static devices still work early in the morning, here we choose 1:00 ~ 4:00 AM. Figure 4.17 visualizes the Wi-Fi data of a static devices on 5 June 2015.



**Fig. 4.17:** Raw data of a static device on 5 June 2015.

Then for mobile devices, we filter out those devices which may come from passers-by using a threshold of  $b$  minutes. For static devices, we use a threshold  $a$  to filter out devices with short transmission time. Besides, we check the mobility to remove static devices whose locations changed over time. The mobility can also help to remove mobile devices that are left by some shop owners unintentionally.

To check the mobility, we calculate a variance  $\gamma$  of a connectivity matrix, if the variance exceeds a threshold, it should not be a static device. Connectivity matrix  $M_j$  is shown in Equation 4.6.  $\gamma$  is calculated using Equation 4.7 where  $var(X)$  calculates the statistical variance.

$$M_j = [V_j(1) \quad V_j(2) \quad \cdots \quad V_j(k)] = \begin{bmatrix} v_{11} & v_{21} & v_{31} & \cdots & v_{k1} \\ v_{12} & v_{22} & v_{32} & \cdots & v_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ v_{1|\mathcal{A}|} & v_{2|\mathcal{A}|} & v_{3|\mathcal{A}|} & \cdots & v_{k|\mathcal{A}|} \end{bmatrix} \quad (4.6)$$

$$\gamma = \frac{\sum_{i=1}^{|\mathcal{A}|} var([v_{1i} \quad v_{2i} \quad \cdots \quad v_{ki}])}{|\mathcal{A}|}, \gamma \in [0, 0.25] \quad (4.7)$$

### 4.3.3 Area Localization and Density Calculation

Area localization and density calculation are the core component in DMAD. In this section, we elaborate on our proposed solutions and demonstrate that although the connectivity information is coarse-grained for fine-grained localization, it is still feasible to locate users to grid-level locations.

We firstly separate the floor plan into 60 non-overlapping areas (or grids, denoted as  $\mathcal{G} = \{g_1, g_2, \cdots\}$ ) manually. Most of the grids contain one or

more shops, few of them contain only common areas. Then, based on users' Wi-Fi data, we are able to derive their grid locations. We also have two observations of heuristics that can be utilized to improve the accuracy of area localization. First of all, different shops attract different numbers of people. Besides, the visit duration in various types of shops is different.

### Area localization

Wi-Fi based indoor localization has been extensively studied in the past decades [105, 139, 7, 161]. The output of those systems can be classified into geometric locations (represented in coordinates) and semantic locations. Our problem belongs to the latter category and we find two existing methods that can be used to solve this problem.

The first method is centroid method [17], the main idea of which is quite simple. Given a connectivity vector  $V_j(i) = [\sigma_1 \ \sigma_2 \ \dots \ \sigma_{|\mathcal{A}|}]^T$  the estimated location  $\hat{\mathcal{L}}_j(i)$  can be calculated using Equation 4.8 ~ 4.10, where  $\Phi = \begin{bmatrix} x_1 & x_2 & \dots & x_{|\mathcal{A}|} \\ y_1 & y_2 & \dots & y_{|\mathcal{A}|} \end{bmatrix}$  is the coordinate vector of all APs. Based on  $(\hat{x}, \hat{y})$ , the grid location can be determined with ease. However, this method works well only if the density of APs is high enough [159], it may work poorly in our scenario due to low AP density and multiple floors.

$$\hat{\mathcal{L}}_j(i) = \begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix} = \frac{1}{\text{sum}(V_j(i))} \Phi \cdot V_j(i) \quad (4.8)$$

$$\hat{x} = \frac{1}{\sum_{k=1}^{|\mathcal{A}|} \sigma_k} \sum_{t=1}^{|\mathcal{A}|} x_t \cdot \sigma_t \quad (4.9)$$

$$\hat{y} = \frac{1}{\sum_{k=1}^{|\mathcal{A}|} \sigma_k} \sum_{t=1}^{|\mathcal{A}|} y_t \cdot \sigma_t \quad (4.10)$$

Another method is fingerprinting method [7, 165, 173, 139] which consists of training phase and testing phase. The training phase is to construct a fingerprint database which requires a simple site survey to collect the connectivity information of APs in all grids. In the testing phase, given a measured connectivity vector, we compare the vector with that of all grids in the database and use the best match as the estimated user location.

However, the RF signal is vulnerable to environmental disturbances and varies over time, which degrades the performances of deterministic fingerprinting approaches. Some researchers proposed probabilistic fingerprinting method [173] which is based on statistical inference between the reported signal information and stored fingerprints. Specifically, given a measured connectivity vector  $V_m$ , the objective is to find a grid  $g$  ( $g \in \mathcal{G}$ ) which maximizes the posterior probability, i.e.,  $\arg \max_g P(g|V_m)$ . Traditional probabilistic fingerprinting calculates  $P(G|V_m)$  ( $G$  is a variable representing all  $g$ ) using Equation 4.11.

$$P(G|V_m) = \frac{P(V_m|G) \cdot P(G)}{P(V_m)} \quad (4.11)$$

Most of the case, previous works regard  $P(G)$  as uniform distribution, i.e.,  $P(G) = 1/|\mathcal{G}|$ . However it is not the case in real scenarios. We observe that various shops have different popularities and the number of customers they attract is thus different. In a similar way, different grids have different attractiveness, therefore  $P(G)$  should be different from grid to grid and time to time. Here we use the number of people on shops to model the popularities of each grid and derive a more practical and accurate estimation of  $P(G)$ .

We also notice that the length of visit to different types of shops is different. Therefore, besides the Wi-Fi signal, we also exploit the visit duration to distinguish different grids. Equation 4.12 shows how we calculate the probability of people in all grids, where  $G$  is a variable for different grids in  $\mathcal{G}$ ,  $T$  is duration time,  $W$  is the measured Wi-Fi data during the period of  $T$ . To calculate  $P(G|WT)$ , we need to know  $P(G)$ ,  $P(T|G)$ , and  $P(W|G)$ , which are described as follows.

$$\begin{aligned} P(G|WT) &= \frac{P(WT|G) \cdot P(G)}{P(WT)} \\ &= \frac{P(W|G) \cdot P(T|G) \cdot P(G)}{P(WT)} \propto P(W|G) \cdot P(T|G) \cdot P(G) \end{aligned} \quad (4.12)$$

$P(G)$  is the probability that people appear in a specific grid. As grids are closely relate to shops, we use the matrix  $R = (n_{ij})$  (in Section 9) to derive the priori probability. The probability people appear in grid  $g_j$  is calculated using Equation 4.13 ~ 4.14.  $P(s_j, t_i)$  is the probability that people appear in shop  $s_j$  during time slot  $t_i$ .  $P(g_j, t_i)$  represents the probability people



appear in  $g_j$  during  $t_i$ .  $\mathbb{S}_j$  is a set of shops that are in the range of  $g_j$ , and  $\mathbb{N}_j$  represents a set of grids that are neighboring to  $g_j$ . If there are no shops in grid  $g_l$ , we use the average probability from all neighboring grids  $\mathbb{N}_l$  of  $g_l$  as alternative.

$$P(s_j, t_i) = \frac{n_{ij}}{\sum_{k=1}^{|\mathbb{S}|} n_{ik}} \quad (4.13)$$

$$P(g_j, t_i) = \begin{cases} \sum_{s_k \in \mathbb{S}_j} P(s_k, t_i), & \text{if } |\mathbb{S}_j| \neq 0 \\ \sum_{g_l \in \mathbb{N}_j} P(g_l, t_i) / |\mathbb{N}_j|, & \text{if } |\mathbb{S}_j| = 0 \end{cases} \quad (4.14)$$

$P(T|G)$  is the probability that how long people will stay in a given grid. Similar to using the number of shops to estimate the number of grids, we calculate the distribution ( $\mu_g$  and  $\sigma_g^2$ ) of duration time for a grid using Equation 4.15 ~ 4.16.  $\mathbb{S}$  is a set of shops that are in the range of grid  $g$ . If  $|\mathbb{S}_j| = 0$ , which means there is no shops in  $g_j$ , the distribution of such grids are collected manually.

$$\mu_g = \frac{1}{|\mathbb{S}|} \cdot \sum_{s_k \in \mathbb{S}} \mu_s, \quad \text{if } |\mathbb{S}| \neq 0 \quad (4.15)$$

$$\sigma_g^2 = \frac{1}{|\mathbb{S}|^2} \cdot \sum_{s_k \in \mathbb{S}} \sigma_s^2, \quad \text{if } |\mathbb{S}| \neq 0 \quad (4.16)$$

We also have two methods to find the duration time of a user in different grids. The most direct way is to exploit traditional area localization methods to map the Wi-Fi data to grid locations, then based on the locations to derive the duration time. The detailed process is illustrated in Algorithm 3.

But this method performs poorly since it relies on existing fingerprinting methods which cannot achieve adequate accuracy. Also, the two parameters are hard to tune.

Another method is to apply subsequence time series clustering techniques. Subsequence clustering is performed on a single time series to group interesting subsequence time series data in the same cluster [21]. There are also several methods to solve the subsequence clustering problem, like hierarchical clustering, partitioning clustering, density based clustering, and etc.

---

**Algorithm 3:** A sliding window approach on  $M_j$ .

---

**Data:** Wi-Fi data of user  $j$ ,  $M_j = [V_j(1) \ \cdots \ V_j(k)]$

**Result:** A set of subsets

- 1 Determine the length  $T_w$  of the sliding window, a threshold  $\lambda_w$ ;
  - 2 **for**  $i \in \text{range}(1, k, T_w)$  **do**
  - 3     **for**  $V_j \in [V_i \ \cdots \ V_{i+T_w-1}]$  **do**
  - 4         Estimate  $\hat{g}_j$  based on  $V_j$ , using  $P(G|V_m)$  ;
  - 5         Calculate the percentage of  $\hat{g}_j$  among all estimated  $\hat{g}$ ;
  - 6         **if** the percentage of  $\hat{g}_j \geq \lambda_w$  **then**
  - 7             The grid of all this window is  $\hat{g}_j$ ;
  - 8 Merge the neighboring windows with same grid information as a subset;
- 

Different methods have different advantages and disadvantages, the detailed explanation can be found in [182].

Here we choose hierarchical clustering, one of the reasons is the generality. Since it does not require any parameters, such as the number of clusters. The procedure of the algorithm is shown in Algorithm 4.

---

**Algorithm 4:** Hierarchical clustering on  $M_j$ .

---

**Data:** Wi-Fi data of user  $j$ ,  $M_j = [V_j(1) \ \cdots \ V_j(K)]$

**Result:** Clusters,  $C$

- 1 Calculate the Euclidean distance matrix  $M_D$  of  $M_j$ ;
  - 2 **while** not every  $V_j(l)$  in clusters **do**
  - 3     Find two  $C_i$  or  $V_j(l)$  with minimum Euclidean distance;
  - 4     Merge the two  $C_i$  or  $V_j(l)$  to produce a new cluster;
  - 5     Update  $M_D$  by calculating distances between new cluster and other clusters;
- 

$W$  in  $P(W|G)$  is a set of connectivity vectors of a device  $d_p$ ,  $W = \{V_p(1), \dots, V_p(K)\}$ , where  $K$  is the size of the cluster. Given a connectivity vector  $V_p(i)$ , the probability that it is within  $g_j$  can be calculated using Equation 4.17.  $\mathbb{V}_j$  is a set of connectivity vectors (also called fingerprints) collected in  $g_j$ ,  $\mathbb{V}_j(l)$  means the  $l$ -th vector.  $\|V_p(i) - \mathbb{V}_j(l)\| / |\mathcal{A}|$  calculates the normalized Euclidean distance between  $V_p(i)$  and the  $l$ -th fingerprints in grid  $g_j$ . We use average probability of all connectivity vectors in  $W$  to represent  $P(W|G)$  in Equation 4.18.

$$P(V_p(i)|G = g_j) = 1 - \frac{1}{|\mathbb{V}_j|} \sum_{l=1}^{|\mathbb{V}_j|} \frac{\|V_p(i) - \mathbb{V}_j(l)\|}{|\mathcal{A}|} \quad (4.17)$$

$$P(W|G = g_j) = \frac{1}{k} \sum_{i=1}^k P(V_i|G = g_j) \quad (4.18)$$

## Density calculation

Density calculation is quite simple compared to area localization. Based on the grid information derived from area localization, for each time slot and each grid, we count the number of people in that grid as density information ( $H$ ), which is represented in Equation 4.19.  $\eta_{ij}$  is the number of people that are within grid  $g_j$  during time slot  $t_i$ .  $\omega_j(i)$  represents the density of  $g_j$  during  $t_i$  and can be calculated using Equation 4.20.  $|\mathcal{T}|$  is the number of time slots,  $|\mathcal{G}|$  is the number of grids.

$$H = \begin{bmatrix} \eta_{11} & \eta_{12} & \eta_{13} & \cdots & \eta_{1|\mathcal{G}|} \\ \eta_{21} & \eta_{22} & \eta_{23} & \cdots & \eta_{2|\mathcal{G}|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \eta_{|\mathcal{T}|1} & \eta_{|\mathcal{T}|2} & \eta_{|\mathcal{T}|3} & \cdots & \eta_{|\mathcal{T}||\mathcal{G}|} \end{bmatrix} \quad (4.19)$$

$$\omega_j(i) = \frac{\eta_{ij}}{\sum_{i=1}^{|\mathcal{T}|} \eta_{ij}} \quad (4.20)$$

### 4.3.4 Dead Spots Estimation

After area localization, each connectivity vector  $V_j(i)$  is associated with a grid information. We separate all connectivity matrices according to time slot and grid. Then for each grid and time slot, there is a set of connectivity matrices. Given the information, a key issue here is how to translate those connectivity matrices into the probability of dead spots, which will be introduced in this section and how to quantify their severity.

#### From connectivity matrices to probability of dead spots

In Section 4.2.1, we have proposed Equation 4.2 to transform a connectivity matrix into probability of dead spots  $P_{DS}(M_j)$ . However, the procedure just converts one connectivity matrix to the probability of dead spots, then how to handle multiple connectivity matrices? We believe that devices with larger

transmission time are more reliable for estimating dead spots. In extreme cases, when the transmission time of a device is very short, its coverage ratio could be highly biased. A reasonable explanation is that larger transmission time corresponds to larger sampling sizes and thus is more reliable.

Given all connectivity matrices  $\{M_1(i), \dots, M_K(i)\}$  during time slot  $t_i$  in grid  $g_j$ , we calculate  $\tau_j(i)$  (the probability of dead spots in  $g_j$  during  $t_i$ ) using Equation 4.21 ~ 4.22, where  $\hat{T}_t(k)$  is the estimated transmission time of  $M_k$ .

$$\tau_j(i) = \sum_{k=1}^K w_k \cdot P_{DS}(M_k) \quad (4.21)$$

$$w_k = \frac{\hat{T}_t(k)}{\sum_{p=1}^K \hat{T}_t(p)} \quad (4.22)$$

### Severity of dead spots

Different dead spots have different severity, if authorities of the facility want to fix some of them, they must want to start with the most critical ones. Obviously, the higher probability of the dead spots, the severer it is. If the probability of two locations is the same, what matters is the number of people. Therefore, the severity of a dead spot is not only related to its possibility but also closely associated with the number of potential users around that dead spot.

Here we combine the probability of dead spots  $\tau_j(i)$  and human density  $\omega_j(i)$ , to derive severity of dead spots.  $\beta$  is the significance factor for human density.

$$\lambda_i^j = \beta \cdot \omega_j(i) + (1 - \beta) \cdot \tau_j(i) \quad (4.23)$$

## 4.4 Experiments and Results

In this section, we firstly introduce the experimental setup and then present the evaluation of each component. Specifically, we carefully study the performance of device classification, area localization, and dead spots estimation. For each of them, we introduce evaluation metrics, baseline approaches, parameter selection, final results, and further discussions if any.

### 4.4.1 Setup

We carry out experiments in a large shopping mall with 5 floors (Ground, and 1st ~ 4th floors) and a total area of over 30,000  $m^2$  in Shenzhen. There are 68 shops and 48 APs in the mall, the floor plan and AP deployment of the ground floor is shown in Figure 4.8 in Section 4.3.1. We manually separate the mall into 60 grids, most of the grids contain at least one shop, few grids contain only common areas. The partition on the ground floor is shown in Figure 4.9. There are few shops, like *B116* on the floor plan, that are not within the expected coverage areas, so we do not take them into consideration. During the period of 46 days, we collect  $|\mathbf{D}_w| = 8,268,462$  Wi-Fi data entries from 726,920 devices.

To evaluate the performances of different components of DMAD, we engage over 20 volunteers to collect testing data for a period of one week. Below shows the tasks that are conducted by volunteers. Table 4.4 lists detailed information of the testing data for different issues.

1. Put some smartphones, including both iOS and Android devices, which keep broadcasting Wi-Fi packets in some predefined locations, like counter desks, and store rooms for a whole day.
2. Do window shopping as usual without preassigned destinations. Record their visiting histories, including the visited grid, start time, and visit duration.
3. Conduct simplified site survey with smartphones. Check is there any dead spots in a specific grid during a specific time slot.

As for the simplified site survey, it is conducted using smartphones rather than spectrum analyzers. To detect whether a grid  $g_j$  has dead spots or not during time slot  $t_i$ , we ask volunteers to go and test every feasible points within  $g_i$ . The granularity of test points is about 4 meters. Generally, there are around 25 points in a grid. For each point, volunteers are required to go there and turn on their Wi-Fi and check the AP list. If the target SSID (“Intown\_Free\_Wi-Fi”) is not in the list, or the network cannot be associated, then that test point is a dead spot. Usually, it takes 20 ~ 30 seconds to finish testing one point.

**Table 4.4:** Details of the testing data.

Issue	Data from task	Data format	# of data
Device classification	I, II, III	<MAC, mobile/static>	249
Area localization	II	<MAC, visit record>	456
Dead spots estimation	III	< $g_j, t_i$ , Boolean-DS >	5127

**Table 4.5:** Confusion matrix of device classification.

		Predicted condition		
		Static	Mobile	Others
True condition	Static	$N_{11}$	$N_{12}$	$N_{13}$
	Mobile	$N_{21}$	$N_{22}$	$N_{23}$
	Others	$N_{31}$	$N_{32}$	$N_{33}$

## 4.4.2 Evaluation

In this subsection, we evaluate the performance of different system components, including device classification, area localization, and dead spots estimation.

### Performance of device classification

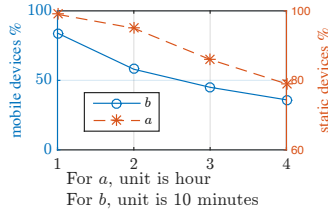
Device classification classifies a smartphone as a static device or a mobile device using a decision tree classifier. Here we study the evaluation metric, discuss the selection of system parameters, and show the final results.

*Evaluation metric* : Since this is a classification problem, we use *precision* and *recall* to evaluate the performance, where  $precision_i = N_{ii} / \sum_j N_{ji}$ ,  $recall_i = N_{ii} / \sum_j N_{ij}$ , and  $N$  is a confusion matrix as explained in Table 4.5.

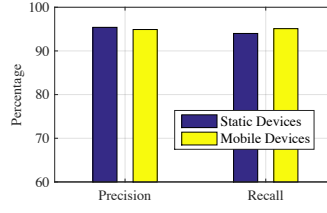
*Parameter selection*: In this component, we have three parameters, a threshold of transmission time  $a$  for static devices, a threshold of transmission time for mobile devices  $b$ , and a threshold for variance of connectivity matrix  $c$ .

The precision and recall of device classification are not sensitive to parameters  $a$  and  $b$ . But different values of  $a$  and  $b$  can affect of the number of static and mobile devices that we can derive from  $D_w$ . Figure 4.18 shows the percentages of static and mobile devices under different value of  $a$  and  $b$ .

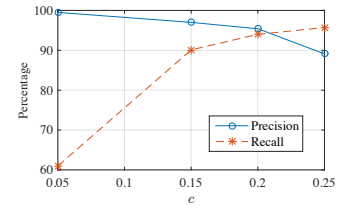
We set  $a = 2$ , which means static device should send packets for at least 2 hours. Since if  $a$  is too small, we cannot calculate the change of coverage ratio. When  $a$  is too large, we may miss many static devices.



**Fig. 4.18:** Impact of different  $a$  and  $b$  on the percentages of static and mobile devices from  $D_w$ .



**Fig. 4.19:** Precision and recall of static device classification and mobile device classification.



**Fig. 4.20:** Precision and recall of static device classification with different  $c$ .

For mobile devices, we set  $b = 10$ . If  $b$  is too large, it may miss a large number of mobile users. On the contrary, if  $b$  is too small, those devices with small transmission time may have a side effect on DMAD, as their data may be collected from passers-by which could be highly biased.

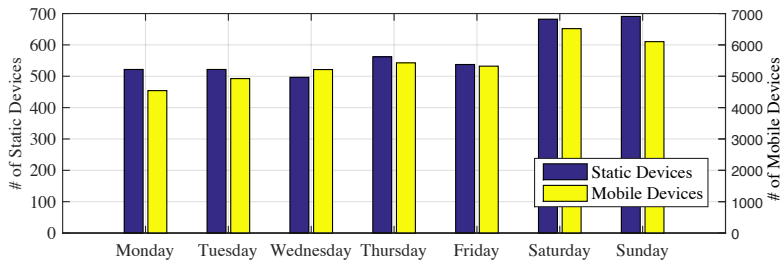
For  $c$ , we set it to 0.2 which is derived from  $D_w^*$ . Since if  $c$  is too large it cannot restrict the mobility of static devices, while  $c$  is too small, it cannot tolerate errors.

*Results :* Figure 4.19 shows the precision and recall of device classification for both static devices and mobile devices. Also, we study the impact of different  $c$  on static devices as illustrated in Figure 4.20. The precision slightly reduces when  $c$  increase.

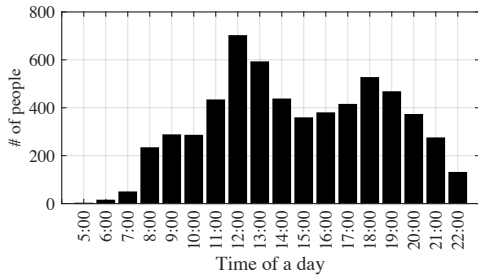
The results of device classification over  $D_w$  indicate that among 726,920 devices, the majority of them (83.1%) are from passers-by of the mall, while 14.98% of them are mobile devices and only (1.92%) of them are static devices.

Figure 4.21 shows the average number of static and mobile devices of each day in a week. Interestingly, most the of days, the number of mobile devices is around ten times larger than that of static devices. During weekends, both static devices and mobile devices increase since more people go shopping in holidays. Also from Figure 4.22 we can see that during dinner time (12:00 and 18:00) the number of people peaks.

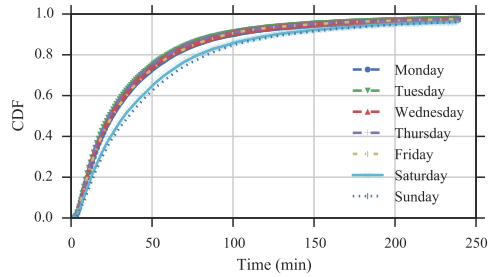
We also analyze the duration time of non-static devices of each day in a week and the results are shown in Figure 4.23. From the results, we can find that the majority (80%) of people stay in the shopping mall for less than 1



**Fig. 4.21:** Average number of static and mobile devices of each day in a week.



**Fig. 4.22:** Average number of people appear in different hours of a day.



**Fig. 4.23:** Comparison of CDFs of total duration time of non-static devices in each day.

hour during weekdays. While during weekends, people stay there for longer time.

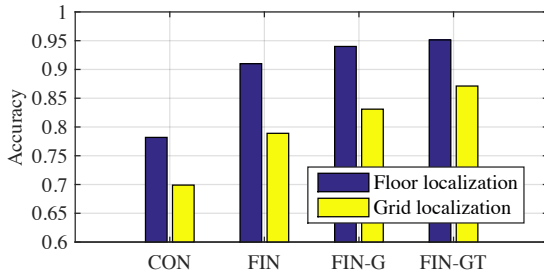
### Performance of area localization

Area localization is to determine users' grid locations according to their connectivity information. We look into evaluation metric and baseline approaches of area localization as well as the performance of grid localization and floor localization. Floor localization is to determine users' floor information, which is more coarse-grained than grid information.

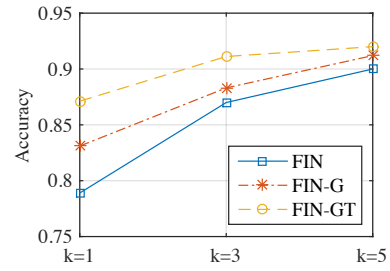
*Evaluation metric* : Area localization is essentially a classification problem, each grid can be regarded as a class. So we use  $accuracy = N_c/N_t$  to measure the performance of floor localization and area localization.  $N_c$  is the number of correctly estimated test cases, while  $N_t$  is the total number of test cases.

To have a comprehensive understanding of different methods, we also evaluate the  $accuracy = N_c^k/N_t$  of top- $k$  results for some methods, where  $N_c^k$  is the number of test cases that the top  $k$  estimated results cover the true results.





**Fig. 4.24:** Accuracy of floor localization and grid localization for different methods.



**Fig. 4.25:** The impact of different  $k$  on the accuracy.

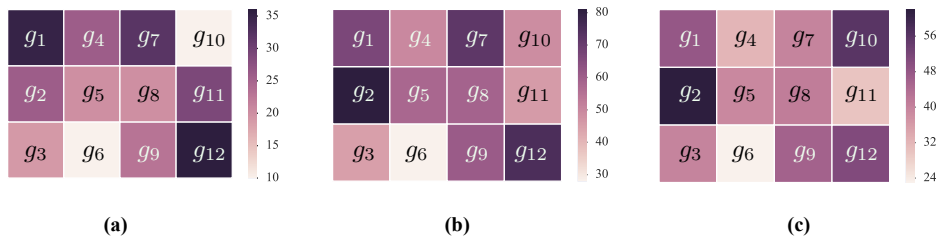
*Baselines:* The baseline approaches for area localization are centroid method and fingerprinting method.

Centroid method is denoted as “CEN”, for this method we need to transform the estimated location  $\hat{\mathcal{L}}$  to estimated grid  $\hat{g}$  by returning the grid which  $\hat{\mathcal{L}}$  belongs to. It also happens when  $\hat{\mathcal{L}}$  are calculated from multiple APs from different floors. In this case, we determine the floor information by using the closest floor that  $\hat{\mathcal{L}}$  is close to.

Another baseline series is probabilistic fingerprinting methods. We denote the traditional method without  $P(G)$  and  $P(T|G)$  as “FIN”. “FIN-G” is a method considering non-uniformed  $P(G)$ , and “FIN-GT” is our proposed method in MDAD which considers both non-uniformed  $P(G)$  (shop popularity) and  $P(T|G)$  (visit duration).

*Results :* The evaluation results of localization are shown in Figure 4.24 and Figure 4.25, which indicate our proposed approaches (“FIN-G” and “FIN-GT”) outperform centroid method and conventional fingerprinting method by over 10%. The potential reasons are that for centroid method, it works well when the AP deployment density is high, but the requirement can hardly be satisfied in real scenarios. Also for conventional fingerprinting methods, due to similar fingerprints in different grids and vulnerability of wireless signal, coarse-grained wireless fingerprints alone cannot achieve high localization accuracy.

“FIN-G” and “FIN-GT” utilize a more realistic priori probability of people appearing in different grids. Besides, “FIN-GT” exploits an additional feature of visit duration to separate grids with similar wireless fingerprints.



**Fig. 4.26:** Heat map of human density on the ground floor in different time slots. (a) 9:00; (b) 12:00; (c) 18:00.

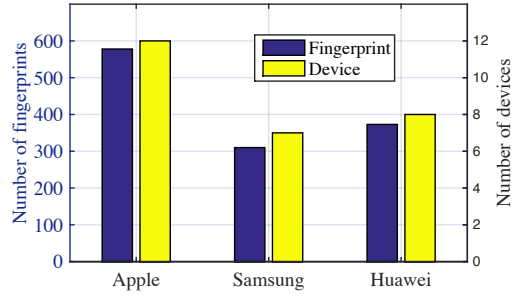
The human density of the ground floor in different time slots in a day is visualized in Figure 4.26. We can see that, at the different time in a day, different grids have varied popularity. For example,  $g_{10}$  is a supermarket, which has more customers in the night than that in the morning. But compared to other grids, some grids like  $g_6$  which is a common area have a small group of people all the time. Generally, we find the following rules from the human density data.

- Grids that are close to entrances or exits are likely to have more people.
- The number of people in grids that contain restaurants peaks at dinner time, i.e., 12:00 and 18:00.

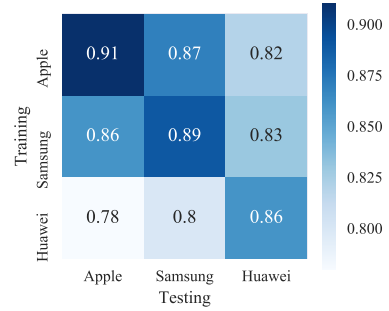
*Further discussion:* Since different mobile devices may have different transmit power, DMAD collects the fingerprints in all grid using devices from different manufacturers. Here we discuss the impact of fingerprints from devices of different manufacturers on the accuracy of area localization. Figure 4.27 shows the distribution of collected fingerprints and the number of devices used to collect fingerprints.

Figure 4.28 shows the localization accuracy of using different kinds of devices for training and testing. As we can see that using the same kind of devices for training and testing can achieve better performance, because different kinds of devices generate different fingerprints. The results are from part of the grids, as we do not have fingerprints of all three devices in all grids.

Among cases where using the same kinds of devices for training and testing, Apple devices outperform other devices. One of the reasons is that we have only iPhone5s and iPhone6 and the fingerprints collected from both models are quite similar. For Huawei, we have 4 models (Mate2, Mate7, P7, and P8). The differences between fingerprints are larger than that of Apple devices.



**Fig. 4.27:** Distribution of fingerprints and devices used in collecting fingerprints. We use devices from three manufacturers to collect fingerprints for area localization.



**Fig. 4.28:** Confusion matrix of localization accuracy using different kinds of devices for training and testing.

This indicates that the more devices used to collect the fingerprints, the higher localization accuracy we can achieve. However, collecting so many fingerprints is too time-consuming to implement. So there is a trade-off between accuracy and simplicity of the system.

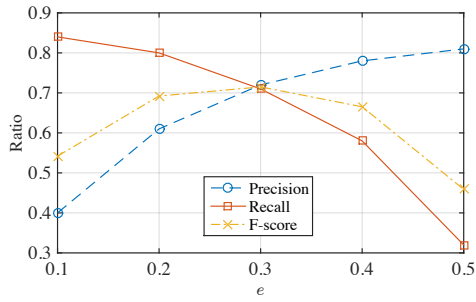
### Performance of dead spots estimation

Dead spots estimation is to estimate the probability of dead spots at a specific grid during a period of time. We study the evaluation metric, parameter selection, and the final results of this component.

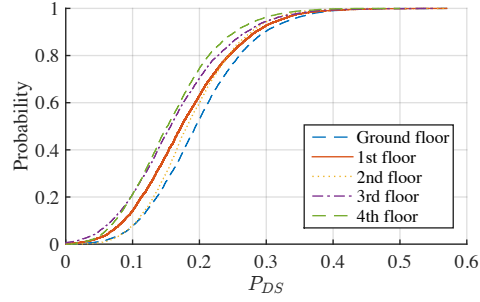
*Evaluation metric* : Estimation of dead spots is a binary classification problem, so we use *precision*, *recall*, and  $F_{score}$  to evaluate its performance.  $precision = tp/(tp + fp)$ ,  $recall = tp/(tp + fn)$ , and  $F_{score} = 2 \cdot precision \cdot recall / (precision + recall)$ .  $tp$  are cases that dead spots are predicted as dead spots;  $tn$  are cases that there are no dead spots and predicted as no dead spots;  $fp$  are cases that are predicted to be dead spots, but in real there is none; and  $fn$  are cases there are dead spots but predicted as no dead spots.

*Parameter selection* : We have two parameters in this component,  $\psi(n)$  is a decay function, and  $\beta$  is the significance factor for human density.  $\psi(n)$  models the probability of dead spot when the location is covered by  $n$  APs.

Here we choose an exponential decay function  $\psi(n) = 1/(2^n)$ . Since the best performance of exponential decay function is better than that of a linear decay function  $\psi = 1/(2 * n)$ , as shown in Figure 4.32.



**Fig. 4.29:** Precision, recall, and F-score of dead spots estimation under different  $e$ .



**Fig. 4.30:** CDF of  $P_{DS}$  of grids from different floors.

We set  $\beta = 0.5$ , which means we regard the human density and the probability of dead spots as equally important.  $\beta$  has nothing to do with the accuracy of dead spots estimation, it serves as an importance factor of human density when calculating severity of a dead spot. If the administrator thinks the number of potential users should be the focus, then  $\beta$  can be set to a larger value.

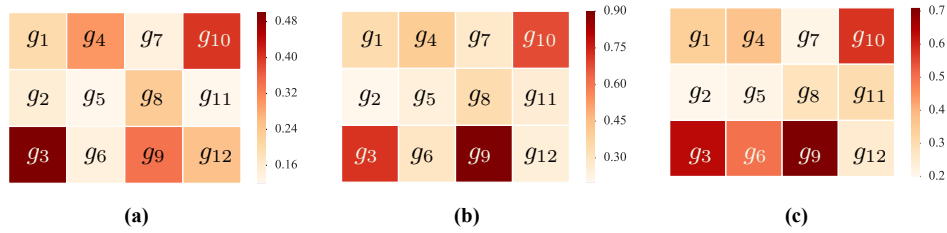
Besides, we also need a threshold  $e$  to determine the existence of dead spots if  $P_{DS} \geq e$ . We set  $e = 0.3$ , since the performance peaks under this value.

*Results :* The results of dead spots estimation are shown in Figure 4.29, which demonstrate that when  $e = 0.3$ , DMAD can identify around 70% of dead spots with a precision over 70%. Also, Figure 4.30 shows the CDF of  $P_{DS}$  of grids in different floors in all time slots. From the results, the lower the floor is, the more dead spots it has. One possible explanation is that more people appear in the lower floors and cause more dead spots.

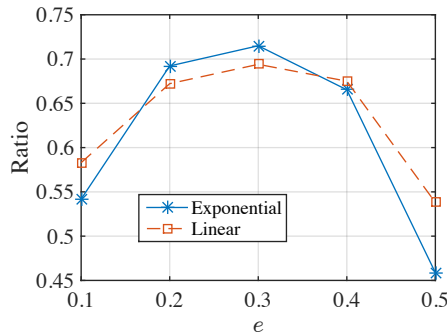
We also derive the normalized severity  $\lambda_i^j / \max(\lambda_i^j)$  of different grids during different time slots. Figure 4.31 shows the severity of grids on the ground floor in different time slots. We can find that some grids like  $g_3$ ,  $g_9$ , and  $g_{10}$ , are more serious over time, which deserve more attention.

#### *Further discussion:*

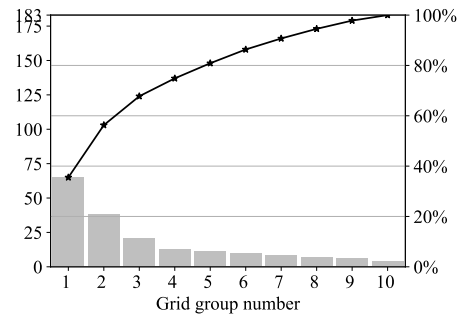
We count the number of dead spots during a whole day among all grids, the average number for weekdays is 972. For weekends, it is reported to have 18.8% more dead spots in average. This result is reasonable, since dead spots are closely related to the people, more people will result in more dead spot. Interestingly, as illustrated in Figure 4.33, we find that the distribution of



**Fig. 4.31:** Heat map of severity of grids on the ground floor in different time slots. (a) 9:00; (b) 12:00; (c) 18:00.



**Fig. 4.32:** F-score of linear and exponential decay functions under different  $e$ .



**Fig. 4.33:** Pareto chart of additional dead spots. The sorted grids are equally separated into 10 groups.

those additional dead spots obeys “70-30 rule”, which means 70% of the additional dead spots are generated by 30% of grids.

## 4.5 Related Work

### AP Deployment Problem

AP deployment problem is to find the minimum number of APs and their optimal locations to achieve one or more objectives. Existing solutions can be classified into two categories, site surveys and simulation approaches. Both methods have their own pros and cons. Site surveys are more accurate and robust, but they require sophisticated electronic monitors and extensive manpower, which is time-consuming and labor-intensive, especially for large areas. Also, site surveys have the potential for disrupting normal operations at the site [2]. Simulations are easy and cheap to conduct, but they cannot precisely characterize the radio frequency (RF) propagation due to the vulnerability of wireless signal and the dynamic environment.

## Site Surveys

Site surveys can provide a solid understanding of the on-site RF behavior, identify any dead spots and reveal areas of channel interference [121]. These surveys are usually conducted by engineers with specialized electronic monitoring equipment such as spectrum analyzers. According to objectives, site surveys can be classified into three categories, predictive modeling surveys, pre-deployment surveys, and post-deployment surveys.

Predictive modeling site surveys use software programs to model the facility and RF environment. Those programs can help to outline the required coverage areas using facility floor plans; estimate RF signal attenuation according to different RF environments; predict the minimum number of APs and their locations. Strictly speaking, predictive modeling surveys belong to simulation methods, as propagation loss models and optimization algorithms are utilized rather than using real equipment to characterize the on-site RF behavior.

Pre-deployment site surveys are often called “AP-on-a-stick” surveys, are performed before setting up a wireless network. In the survey, spectrum analysis is an integral part, which can identify sources of RF interference and dead spots that would cause performance issues. With this survey, a better wireless network design can be achieved by characterizing the RF behavior in the facility, which is uniquely tailored to the physical properties of the environment. It can also be used to verify and adjust a preliminary Wi-Fi network design.

Post-deployment site surveys are performed after the APs have been installed and configured. This type of site surveys reflects the RF signal propagation characteristics of the deployed wireless network. The focus is to validate that the performance of the deployed network matches the original network design.

## Simulation Approaches

Since site surveys are time-consuming and labor-intensive, numerous simulation methods have been proposed to avoid extensive measurements and expensive physical experiments. Simulation methods firstly emulate the RF propagation in the target environment, then model it as a mathemati-

cal problem by using propagation loss models, and finally exploit various optimization algorithms to solve it towards one or more objectives.

Once the propagation loss model is determined, given an AP setting, the signal strength of this AP at any locations on the site can be estimated. Then the problem is to find the minimum number of APs and their optimal locations to satisfy some predefined thresholds like the minimum RSS value. Different simulation methods mainly differ in their propagation loss models, objectives and optimization algorithms.

*Propagation loss models* describe how RF signal attenuate over physical distance and through different obstacles which have been studied extensively over decades [51, 73, 129]. Hashemi conducted a comprehensive survey about mathematical and statistical modeling of individual characteristics of propagation losses in [51]. Other researchers studied propagation loss of signal with multipath characteristics using ray tracing techniques in [73]. While Schoeberl modeled the propagation losses combining ray tracing and Monte Carlo simulation in [129]. All these works indicate that average received signal power decreases logarithmically with distance, which are described in Equation 4.24.

$$PL = PL_0 + 10 \cdot \gamma \cdot \log_{10} \frac{d}{d_0} + \sum_{i=1}^n N_i \cdot L_i \quad (4.24)$$

$PL$  is the total path loss,  $PL_0$  is the path loss at the reference distance  $d_0$ ,  $\gamma$  is the path loss attenuation factor derived from measurements,  $d$  is the length of the path,  $d_0$  is the reference distance,  $N_i$  represents the number of a particular type of obstacles and  $L_i$  represents the loss associated with that type of obstacles.

*Optimization objectives* are usually wireless coverage [72, 88], offloading ratio [18, 68], fingerprint differences [88, 101], and etc. Recently, more works [88, 23] focus on achieving the combination of multiple objectives.

*Optimization methods* like the Nelder-Mead simplex algorithm are adopted to find the optimal AP locations for maximizing the coverage ratio in [40]. In [2], a one-by-one trial method has been proposed to find the minimum number of APs needed to cover a given site. For a given number of APs, the genetic algorithm (GA) optimizer is used to perform AP location optimization. The authors in [153] use a neural network approach to perform propagation

prediction, and adopt an ant colony optimization approach to optimize the AP locations and to maximize the average received power. In [134], the simulated annealing (SA) algorithm is utilized to find the minimum number and optimal transmission power of APs, but the AP locations are not optimized. Wang et al. exploited GA to the placement of APs with heterogeneous costs and capacities in [157, 156]. In [163], Lydon While proposed a multi-objective evolutionary algorithm for three criteria minimized cost, maximized coverage, and minimized service refusal.

## Using Data Science in Wireless Networks

Data science or “data-driven research” is a research approach that uses real-life data to gain insights about the behavior of target systems [74]. It enables the analysis of various systems in order to assess whether they function according to the intended design and as seen in simulations.

Wireless networks can exhibit unpredictable interactions between algorithms from multiple protocol layers, interactions between multiple devices, and hardware specific influences [74]. These interactions may further result in a difference between real-world functioning and design-time functioning. Data science methods can be utilized to detect the actual behavior and hopefully provide insights to improve the system performance.

Numerous research areas like large-scale social networks, advanced business and healthcare processes, have successfully adopted data-driven approaches to analyze networked interactions. In traditional wireless research, it often starts with theoretical models to devise solutions which are then evaluated using a simulator or experimental setup [74]. In contrast to traditional approaches, research works such as [30, 92, 93] use a data-driven approach, starting from large, real-life wireless datasets to extract knowledge about wireless systems.

For example, in [30], the authors proposed a data-driven solution for fingerprinting wireless devices that can help existing network access control systems to enhance network security by allowing access only for certain devices or device types (devices that have the same hardware configuration). Different from traditional security mechanisms that rely on device authentication based on public key cryptography and digital certificates, which could be simply transferred to another device. The proposed data-driven approach



relies on distinguishing devices by looking into the statistical distribution of inter-arrival times between packets generated by the same device and a particular application. The authors formulated this as a classification problem, proved their hypothesis from two testbeds, and finally solved the problem with an artificial neural network model.

## 4.6 Conclusion

In this chapter, we propose DMAD, a data-driven measuring of Wi-Fi AP deployment to estimate dead spots and quantify their severity using both Wi-Fi data and shop data.

Based on the collected data, we firstly classify static devices and mobile devices using a decision-tree classifier. The most distinguishing feature between them is whether they work early in the morning.

Then we locate these devices to shop-level locations based on two observations of heuristics. On the one hand, the duration of visit in different shops is different, for example, people stay longer in restaurants than that of clothing shops. On the other, different shops have different popularity in attracting customers at different time slots, for example, restaurants attract more people during lunch time than clothing shops. These two features can be exploited to distinguish locations with similar wireless fingerprints.

Lastly, for each location, we estimate the probability of dead spots in different time slots and derive their severity combining the dead spots probability and human density. Since if a dead spot appears in a place with a lot of potential users, this dead spot must be severer.

We carefully study the performance of different components of DMAD using real data collected from a large shopping mall. The evaluation results demonstrate that DMAD can identify around 70% of dead spots with a precision over 70%.



## Conclusion & Future Work

In this dissertation, we address the emerging issues in understanding human dynamics using privacy-sensitive data modalities. The first issue is user privacy erosion due to the prevalence of IoT devices and social media services. Human dynamics research in the uncontrolled setting usually involve collecting spontaneous data in a naturalistic environment. Private content and uninvolved parties could be recorded without their consent. The second issue is incomplete user profile. Unlike traditional data collection methods like interview and survey, many IoT collected data sources lack detailed demographic information. Without knowing this information, the results of human dynamic research could be biased. The third issue is missing contextual information. The focus of human dynamics research is not only just about human but also the environment and the situation they interact with. The environment plays an essential role in understanding human dynamics since it could influence and reshape human behaviors.

For the first issue, we propose to use privacy-sensitive data modalities for human dynamics research. For the remaining issues, we also show that it is possible to infer user profiles and contextual information using privacy-sensitive data in the presented three works. However, we entail two grand research challenges when applying privacy-sensitive data for these purposes. First, low quality of privacy-sensitive data brings difficulty in extracting adequate and effective features for high-level applications. Second, dynamics of human behavior poses serious challenges to the effectiveness and robustness of system performance, sometimes even results in a new research problem.

To address the first challenge, the general methodologies are integrating knowledge from other domains, devising new features, and fusing data from multiple sources. As illustrated in Chapter 4, we use a probabilistic approach to locate customers based on the WiFi data. However, due to the sparsity of the data, it is difficult to achieve satisfying performance. Therefore, we fuse the PoI data and derive a more accurate prior probability. Another example is presented in Chapter 2 where we extracted conversational features rather than voice features to identify gender since it is indicated in sociology literature that the way people take turns and interrupt each other could

also reveal their gender information. To address the second challenge, an effective way is to infer the contextual information first. An example is shown in Chapter 2, we infer the gender composition as an extra input for gender identification since the composition plays a latent role in people's turn-taking behaviors and interruption patterns.

As for future work, we mainly have three directions. First, we will investigate more types of human activities like online activities and will try to fuse online and offline behaviors. Second, more attention will be paid to cross-modality research like the combination of multiple modalities as the capability of collecting more types of data keeps increasing. Last, we will conduct more research work on the mental context like sensing and analytics of personalities and emotions. These high-level dynamics play significant roles in understanding human behaviors and interactions.

# References

- [1] Mohamed Abouelenien, Verónica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. “Multimodal gender detection”. In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM. 2017, pp. 302–311 (cit. on pp. 18, 37, 38).
- [2] Martin D Adickes, Richard E Billo, Bryan A Norman, et al. “Optimization of indoor wireless communication network layouts”. In: *IIE Transactions* 34.9 (2002), pp. 823–836 (cit. on pp. 64, 97, 99).
- [3] Rein Ahas, Anto Aasa, Siiri Silm, and Margus Tiru. “Daily rhythms of suburban commuters’ movements in the Tallinn metropolitan area: Case study with mobile positioning data”. In: *Transportation Research Part C: Emerging Technologies* 18.1 (2010), pp. 45–54 (cit. on pp. 13, 15).
- [4] Musaed Alhussein, Zulfiqar Ali, Muhammad Imran, and Wadood Abdul. “Automatic gender detection based on characteristics of vocal folds for mobile healthcare system”. In: *Mobile Information Systems 2016* (2016) (cit. on pp. 14, 18, 19, 37, 38).
- [5] Sally Andrews, David A Ellis, Heather Shaw, and Lukasz Piwek. “Beyond self-report: tools to compare estimated and real-world smartphone use”. In: *PloS one* 10.10 (2015), e0139004.
- [6] Gennady L Andrienko, Natalia V Andrienko, Georg Fuchs, et al. “Extracting Semantics of Individual Places from Movement Data by Analyzing Temporal Patterns of Visits.” In: *COMP@ SIGSPATIAL*. 2013, pp. 9–15 (cit. on p. 13).
- [7] Paramvir Bahl and Venkata N Padmanabhan. “RADAR: An in-building RF-based user location and tracking system”. In: *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*. Vol. 2. Ieee. 2000, pp. 775–784 (cit. on pp. 64, 83).
- [8] Nilanjan Banerjee, Sharad Agarwal, Paramvir Bahl, et al. “Virtual compass: relative positioning to sense mobile social interactions”. In: *Pervasive computing*. Springer, 2010, pp. 1–21.

- [9]Albert-Laszlo Barabasi. “The origin of bursts and heavy tails in human dynamics”. In: *Nature* 435.7039 (2005), p. 207 (cit. on p. 13).
- [10]Marco V Barbera, Alessandro Epasto, Alessandro Mei, Vasile C Perta, and Julinda Stefa. “Signals from the crowd: uncovering social relationships through smartphone probes”. In: *Proceedings of the 2013 conference on Internet measurement conference*. ACM. 2013, pp. 265–276 (cit. on pp. 15, 42, 61).
- [11]Christian Bauckhage. “k-Means Clustering Is Matrix Factorization”. In: *arXiv preprint arXiv:1512.07548* (2015).
- [12]Nancy K Baym, Yan Bing Zhang, and Mei-Chen Lin. “Social interactions across media: Interpersonal communication on the internet, telephone and face-to-face”. In: *New Media & Society* 6.3 (2004), pp. 299–318 (cit. on p. 18).
- [13]Kenneth Benoit. “Linear regression models with logarithmic transformations”. In: *London School of Economics, London* (2011) (cit. on p. 80).
- [14]Rachel Bernstein. “Communication: spontaneous scientists”. In: *Nature* 505.7481 (2014), pp. 121–123 (cit. on p. 18).
- [15]Rod Bond. “Group size and conformity”. In: *Group processes & intergroup relations* 8.4 (2005), pp. 331–354.
- [16]Chloë Brown, Christos Efstratiou, Ilias Leontiadis, et al. “The architecture of innovation: Tracking face-to-face interactions with ubicomp technologies”. In: *Proceedings of ACM Ubicomp, 2014*.
- [17]Nirupama Bulusu, John Heidemann, and Deborah Estrin. “Adaptive beacon placement”. In: *Distributed Computing Systems, 2001. 21st International Conference on*. IEEE. 2001, pp. 489–498 (cit. on p. 83).
- [18]Eyuphan Bulut and Boleslaw K Szymanski. “WiFi access point deployment for efficient mobile data offloading”. In: *ACM SIGMOBILE Mobile Computing and Communications Review* 17.1 (2013), pp. 71–78 (cit. on pp. 64, 99).
- [19]Li-wei Chan, Ji-rung Chiang, Yi-chao Chen, et al. “Collaborative localization: Enhancing wifi-based position estimation with neighborhood links in clusters”. In: *Pervasive Computing*. Springer, 2006, pp. 50–66.
- [20]Donghui Chen and Robert J Plemmons. “Nonnegativity constraints in numerical analysis”. In: *The birth of numerical analysis* 10 (2009), pp. 109–140.
- [21]Jason R Chen. “Making subsequence time series clustering meaningful”. In: *Fifth IEEE International Conference on Data Mining (ICDM’05)*. IEEE. 2005, 8–pp (cit. on p. 85).
- [22]Lei Chen, M Tamer Özsu, and Vincent Oria. “Robust and fast similarity search for moving object trajectories”. In: *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM. 2005, pp. 491–502.

- [23] Qiuyun Chen, Bang Wang, Xianjun Deng, Yijun Mo, and Laurence T Yang. “Placement of access points for indoor wireless coverage and fingerprint-based localization”. In: *High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC\_EUC), 2013 IEEE 10th International Conference on*. IEEE. 2013, pp. 2253–2257 (cit. on pp. 64, 99).
- [24] Ningning Cheng, Prasant Mohapatra, Mathieu Cunche, et al. “Inferring user relationship from hidden information in wlans”. In: *Proceedings of MILCOM, 2012* (cit. on pp. 15, 61).
- [25] Krishna Chintalapudi, Anand Padmanabha Iyer, and Venkata N Padmanabhan. “Indoor localization without the pain”. In: *Proceedings of ACM MobiCom, 2010*.
- [26] Yohan Chon, Suyeon Kim, Seungwoo Lee, et al. “Sensing WiFi packets in the air: practicality and implications in urban mobility monitoring”. In: *Proceedings of ACM Ubicomp, 2014*.
- [27] Shammur Absar Chowdhury, Morena Danieli, and Giuseppe Riccardi. “Annotating and categorizing competition in overlap speech”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE. 2015, pp. 5316–5320.
- [28] Robert B Cialdini and Noah J Goldstein. “Social influence: Compliance and conformity”. In: *Annu. Rev. Psychol.* 55 (2004), pp. 591–621.
- [29] Chris Clifton, Murat Kantarcioglu, and Jaideep Vaidya. “Defining privacy for data mining”. In: *National science foundation workshop on next generation data mining*. Vol. 1. 26. Citeseer. 2002, p. 1 (cit. on p. 7).
- [30] Manuel Crotti, Maurizio Dusi, Francesco Gringoli, and Luca Salgarelli. “Traffic classification through simple statistical fingerprinting”. In: *ACM SIGCOMM Computer Communication Review* 37.1 (2007), pp. 5–16 (cit. on p. 100).
- [31] Mathieu Cunche, Mohamed Ali Kaafar, and Roksana Boreli. “I know who you will meet this evening! linking wireless devices using wi-fi probe requests”. In: *Proceedings of WoWMoM, 2012* (cit. on pp. 15, 61).
- [32] Marzieh Dashti, Mohd Amiruddin Abd Rahman, Hamed Mahmoudi, and Holger Claussen. “Detecting co-located mobile users”. In: *Proceedings of IEEE ICC, 2015*.
- [33] Yves-Alexandre De Montjoye, Laura Radaelli, Vivek Kumar Singh, et al. “Unique in the shopping mall: On the reidentifiability of credit card metadata”. In: *Science* 347.6221 (2015), pp. 536–539 (cit. on p. 7).
- [34] D Shakina Deiv, Mahua Bhattacharya, et al. “Automatic gender identification for hindi speech recognition”. In: *International Journal of Computer Applications, New York* 31.5 (2011), pp. 1–8.

- [35]Adriano Di Luzio, Alessandro Mei, and Julinda Stefa. “Mind your probes: De-anonymization of large crowds through smartphone WiFi probe requests”. In: *Proceedings of IEEE INFOCOM, 2016* (cit. on pp. 15, 60).
- [36]EU Directive. “95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data”. In: *Official Journal of the EC* 23.6 (1995) (cit. on p. 6).
- [37]Zhixiang Fang, Xiping Yang, Yang Xu, Shih-Lung Shaw, and Ling Yin. “Spatiotemporal model for assessing the stability of urban human convergence and divergence patterns”. In: *International Journal of Geographical Information Science* 31.11 (2017), pp. 2119–2141 (cit. on pp. 13, 15).
- [38]Vernor C Finch. “Geographical science and social philosophy”. In: *Annals of the Association of American Geographers* 29.1 (1939), pp. 1–28 (cit. on p. 3).
- [39]Heather L Ford, Cameron Brick, Karine Blaufuss, and Petra S Dekens. “Gender inequity in speaking opportunities at the American Geophysical Union Fall Meeting”. In: *Nature communications* 9 (2018) (cit. on p. 18).
- [40]Steven J Fortune, David M Gay, Brian W Kernighan, et al. “WISE design of indoor wireless systems: practical computation and optimization”. In: *IEEE Computational Science & Engineering* 2.1 (1995), pp. 58–68 (cit. on p. 99).
- [41]Dieter Fox. “KLD-sampling: Adaptive particle filters”. In: *Advances in neural information processing systems*. 2001, pp. 713–720.
- [42]Julien Freudiger. “How talkative is your mobile device?: an experimental study of wi-fi probe requests”. In: *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*. ACM. 2015, p. 8 (cit. on pp. 42, 45, 46, 61, 75).
- [43]Santosh Gaikwad, Bharti Gawali, and SC Mehrotra. “Gender identification using SVM with Combination of MFCC”. In: *Advances in Computational Research* 4.1 (2012).
- [44]Yuan Gao, Qingxuan Yang, Guanfeng Li, et al. “XINS: the anatomy of an indoor positioning and navigation architecture”. In: *Proceedings of the 1st international workshop on Mobile location-based service*. ACM. 2011, pp. 41–50.
- [45]Weina Ge, Robert T Collins, and R Barry Ruback. “Vision-based analysis of small groups in pedestrian crowds”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.5 (2012), pp. 1003–1016 (cit. on pp. 14, 42, 60).
- [46]Nicolas Gillis and François Glineur. “A multilevel approach for nonnegative matrix factorization”. In: *Journal of Computational and Applied Mathematics* 236.7 (2012), pp. 1708–1723.



- [47] Bruno Gonçalves and José J Ramasco. “Human dynamics revealed through Web analytics”. In: *Physical Review E* 78.2 (2008), p. 026123 (cit. on p. 13).
- [48] Joachim Gudmundsson and Marc van Kreveld. “Computing longest duration flocks in trajectory data”. In: *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*. ACM. 2006, pp. 35–42.
- [49] Isabelle Guyon and André Elisseeff. “An introduction to variable and feature selection”. In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182 (cit. on p. 31).
- [50] Annette Hannah and Tamar Murachver. “Gender and conversational style as predictors of conversational behavior”. In: *Journal of Language and Social Psychology* 18.2 (1999), pp. 153–174 (cit. on pp. 9, 19, 38).
- [51] Homayoun Hashemi. “The indoor radio propagation channel”. In: *Proceedings of the IEEE* 81.7 (1993), pp. 943–968 (cit. on p. 99).
- [52] Suining He and S-H Gary Chan. “Wi-Fi fingerprint-based indoor positioning: Recent advances and comparisons”. In: (2015).
- [53] Suining He, S-H Gary Chan, Lei Yu, and Ning Liu. “Calibration-free fusion of step counter and wireless fingerprints for indoor localization”. In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM. 2015, pp. 897–908.
- [54] Suining He, S-H Gary Chan, Lei Yu, and Ning Liu. “Fusing noisy fingerprints with distance bounds for indoor localization”. In: *2015 IEEE Conference on Computer Communications (INFOCOM)*. IEEE. 2015, pp. 2506–2514.
- [55] Sebastian Hilsenbeck, Dmytro Bobkov, Georg Schroth, Robert Huitl, and Eckehard Steinbach. “Graph-based data fusion of pedometer and WiFi measurements for mobile indoor positioning”. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM. 2014, pp. 147–158.
- [56] Hande Hong, Chengwen Luo, and Mun Choon Chan. “SocialProbe: Understanding Social Interaction Through Passive WiFi Monitoring”. In: *Proceedings of ACM MobiQuitous, 2016* (cit. on pp. 15, 41, 42, 56, 60, 61).
- [57] AKM Mahtab Hossain, Yunye Jin, Wee-Seng Soh, and Hien Nguyen Van. “SSD: A robust RF location fingerprint addressing mobile devices’ heterogeneity”. In: *IEEE Transactions on Mobile Computing* 12.1 (2013), pp. 65–77.
- [58] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. “Demographic prediction based on user’s browsing behavior”. In: *Proceedings of the 16th international conference on World Wide Web*. ACM. 2007, pp. 151–160 (cit. on p. 37).

- [59]Xueheng Hu, Lixing Song, Dirk Van Bruggen, and Aaron Striegel. “Is there wifi yet?: How aggressive probe requests deteriorate energy and throughput”. In: *Proceedings of ACM IMC, 2015* (cit. on p. 61).
- [60]Yakun Hu, Dapeng Wu, and Antonio Nucci. “Pitch-based gender identification with two-stage classification”. In: *Security and Communication Networks* 5.2 (2012), pp. 211–225 (cit. on pp. 14, 18, 37).
- [61]Hayley Hung and Daniel Gatica-Perez. “Estimating cohesion in small groups using audio-visual nonverbal behavior”. In: *IEEE Transactions on Multimedia* 12.6 (2010), pp. 563–575.
- [62]Shuja Jamil, Sohaib Khan, Anas Basalamah, and Ahmed Lbath. “Classifying smartphone screen ON/OFF state based on wifi probe patterns”. In: *Proceedings of ACM UbiComp, 2016*.
- [63]Kasthuri Jayarajah, Zaman Lantra, and Archan Misra. “Fusing WiFi and Video Sensing for Accurate Group Detection in Indoor Spaces”. In: *Proceedings of the 3rd International on Workshop on Physical Analytics*. ACM. 2016, pp. 49–54.
- [64]Hoyoung Jeung, Heng Tao Shen, and Xiaofang Zhou. “Convoy queries in spatio-temporal databases”. In: *2008 IEEE 24th International Conference on Data Engineering*. IEEE. 2008, pp. 1457–1459.
- [65]Hoyoung Jeung, Man Lung Yiu, Xiaofang Zhou, Christian S Jensen, and Heng Tao Shen. “Discovery of convoys in trajectory databases”. In: *Proceedings of the VLDB Endowment* 1.1 (2008), pp. 1068–1080.
- [66]Junghyun Jun, Yu Gu, Long Cheng, et al. “Social-loc: Improving indoor localization with social sensing”. In: *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. ACM. 2013, p. 14.
- [67]Hyunsoo Kim and Haesun Park. “Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method”. In: *SIAM journal on matrix analysis and applications* 30.2 (2008), pp. 713–730.
- [68]JaYeong Kim, Nah-Oak Song, Byoung Hoon Jung, Hansung Leem, and Dan Keun Sung. “Placement of WiFi access points for efficient WiFi offloading in an overlay network”. In: *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE. 2013, pp. 3066–3070 (cit. on p. 99).
- [69]Jingu Kim and Haesun Park. *Sparse nonnegative matrix factorization for clustering*. Tech. rep. Georgia Institute of Technology, 2008 (cit. on p. 54).
- [70]Mikkel Baun Kjærgaard, Martin Wirz, Daniel Roggen, and Gerhard Tröster. “Detecting pedestrian flocks by fusion of multi-modal sensors in mobile phones”. In: *Proceedings of ACM UbiComp, 2012* (cit. on pp. 15, 42, 60).

- [71] Mikkel Baun Kjærgaard, Martin Wirz, Daniel Roggen, and Gerhard Tröster. “Mobile sensing of pedestrian flocks in indoor environments using wifi signals”. In: *Proceedings of IEEE PerCom, 2012* (cit. on pp. 15, 42, 61).
- [72] Shahnaz Kouhbor, Julien Ugon, A Rubinov, Alex Kruger, and M Mammadov. “Coverage in WLAN with minimum number of access points”. In: *2006 IEEE 63rd Vehicular Technology Conference*. Vol. 3. IEEE. 2006, pp. 1166–1170 (cit. on p. 99).
- [73] Peter Kreuzgruber, Thomas Brundl, Wolfgang Kuran, and Rainer Gahleitner. “Prediction of indoor radio propagation with the ray splitting model including edge diffraction and rough surfaces”. In: *Vehicular Technology Conference, 1994 IEEE 44th*. IEEE. 1994, pp. 878–882 (cit. on p. 99).
- [74] Merima Kulin, Carolina Fortuna, Eli De Poorter, Dirk Deschrijver, and Ingrid Moerman. “Data-Driven Design of Intelligent Wireless Networks: An Overview and Tutorial”. In: *Sensors* 16.6 (2016), p. 790 (cit. on pp. 65, 100).
- [75] Mamta Kumari and Israj Ali. “An efficient algorithm for Gender Detection using voice samples”. In: *Communication, Control and Intelligent Systems (CCIS), 2015*. IEEE. 2015, pp. 221–226 (cit. on pp. 18, 37).
- [76] Kevin S Kung, Kael Greco, Stanislav Sobolevsky, and Carlo Ratti. “Exploring universal patterns in human home-work commuting from mobile phone data”. In: *PloS one* 9.6 (2014), e96180 (cit. on p. 13).
- [77] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, et al. “A survey of mobile phone sensing”. In: *IEEE Communications magazine* 48.9 (2010).
- [78] Patrick Laube and Stephan Imfeld. “Analyzing relative motion within groups of trackable moving point objects”. In: *International Conference on Geographic Information Science*. Springer. 2002, pp. 132–144.
- [79] David Lazer, Alex Sandy Pentland, Lada Adamic, et al. “Life in the network: the coming age of computational social science”. In: *Science (New York, NY)* 323.5915 (2009), p. 721 (cit. on p. 2).
- [80] Oren Lederman, Dan Calacci, Angus MacMullen, et al. “Open badges: A low-cost toolkit for measuring team communication and dynamics”. In: *arXiv preprint arXiv:1710.01842* (2017) (cit. on p. 22).
- [81] Oren Lederman, Akshay Mohan, Dan Calacci, and Alex Sandy Pentland. “Rhythm: A Unified Measurement Platform for Human Organizations”. In: *IEEE MultiMedia* 25.1 (2018), pp. 26–38 (cit. on pp. 8, 14, 18, 21, 22).
- [82] Dik Lun Lee and Qiuxia Chen. “A model-based wifi localization method”. In: *Proceedings of the 2nd international conference on Scalable information systems*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). 2007, p. 40 (cit. on p. 50).

- [83]Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. “Trajectory clustering: a partition-and-group framework”. In: *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. ACM. 2007, pp. 593–604.
- [84]Youngki Lee, Chulhong Min, Chanyou Hwang, et al. “Sociophone: Everyday face-to-face interaction monitoring platform using multi-phone sensor fusion”. In: *Proceedings of ACM MobiSys, 2013* (cit. on pp. 15, 60).
- [85]Sarah Ita Levitan, Taniya Mishra, and Srinivas Bangalore. “Automatic identification of gender from speech”. In: *Proceeding of Speech Prosody*. 2016, pp. 84–88 (cit. on p. 38).
- [86]Muyuan Li, Haojin Zhu, Zhaoyu Gao, et al. “All your location are belong to us: Breaking mobile social networks for automated user location tracking”. In: *Proceedings of the 15th ACM international symposium on Mobile ad hoc networking and computing*. ACM. 2014, pp. 43–52 (cit. on p. 4).
- [87]Zhenhui Li, Bolin Ding, Jiawei Han, and Roland Kays. “Swarm: Mining relaxed temporal moving object clusters”. In: *Proceedings of the VLDB Endowment* 3.1-2 (2010), pp. 723–734.
- [88]Lin Liao, Weifeng Chen, Chuanlin Zhang, et al. “Two birds with one stone: Wireless access point deployment for both coverage and localization”. In: *IEEE Transactions on Vehicular Technology* 60.5 (2011), pp. 2239–2252 (cit. on pp. 64, 99).
- [89]Hongbo Liu, Yu Gan, Jie Yang, et al. “Push the limit of WiFi based localization for smartphones”. In: *Proceedings of the 18th annual international conference on Mobile computing and networking*. ACM. 2012, pp. 305–316.
- [90]Hui Liu, Houshang Darabi, Pat Banerjee, and Jing Liu. “Survey of wireless indoor positioning techniques and systems”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37.6 (2007), pp. 1067–1080.
- [91]Siyuan Liu, Shuhui Wang, Kasthuri Jayarajah, Archan Misra, and Ramayya Krishnan. “TODMIS: Mining communities from trajectories”. In: *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM. 2013, pp. 2109–2118.
- [92]Tao Liu and Alberto E Cerpa. “Foresee (4C): Wireless link prediction using link features”. In: *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on*. IEEE. 2011, pp. 294–305 (cit. on p. 100).
- [93]Tao Liu and Alberto E Cerpa. “Temporal adaptive link quality prediction with online learning”. In: *ACM Transactions on Sensor Networks (TOSN)* 10.3 (2014), p. 46 (cit. on p. 100).

- [94]Xiulong Liu, Keqiu Li, Geyong Min, et al. “Completely pinpointing the missing RFID tags in a time-efficient way”. In: *IEEE Transactions on Computers* 64.1 (2015), pp. 87–96 (cit. on p. 60).
- [95]Xiulong Liu, Keqiu Li, Alex X Liu, et al. “Multi-category RFID estimation”. In: *IEEE/ACM transactions on networking* 25.1 (2017), pp. 264–277 (cit. on p. 60).
- [97]Corrado Loglisci, Donato Malerba, and Apostolos N Papadopoulos. “Mining Trajectory Data for Discovering Communities of Moving Objects.” In: *EDBT/ICDT Workshops*. 2014, pp. 301–308.
- [98]Kathy Macropol. *Clustering on Graphs: The Markov Cluster Algorithm (MCL)*. Tech. rep. Retrieved 06/19/2013 from [http://www.cs.ucsb.edu/~xyan/classes/CS595D-2009winter/MCL\\_Presentation2.pdf](http://www.cs.ucsb.edu/~xyan/classes/CS595D-2009winter/MCL_Presentation2.pdf).
- [99]S Maraboina, Dorothea Kolossa, PK Bora, and Reinhold Orglmeister. “Multi-speaker voice activity detection using ICA and beampattern analysis”. In: *Signal Processing Conference, 2006 14th European*. IEEE. 2006, pp. 1–5 (cit. on p. 22).
- [100]Miller McPherson, Lynn Smith-Lovin, and James M Cook. “Birds of a feather: Homophily in social networks”. In: *Annual review of sociology* 27.1 (2001), pp. 415–444.
- [101]Weixiao Meng, Ying He, Zhian Deng, and Cheng Li. “Optimized access points deployment for WLAN indoor positioning system”. In: *2012 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE. 2012, pp. 2457–2461 (cit. on pp. 64, 99).
- [102]Brendan Morris and Mohan Trivedi. “Learning trajectory patterns by clustering: Experimental studies and comparative evaluation”. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE. 2009, pp. 312–319.
- [103]Anthony Mulac. “Men’s and women’s talk in same-gender and mixed-gender dyads: Power or polemic?” In: *Journal of Language and Social Psychology* 8.3-4 (1989), pp. 249–270 (cit. on pp. 10, 19, 30, 32).
- [104]Stephen O Murray and Lucille H Covelli. “Women and men speaking at the same time”. In: *Journal of Pragmatics* 12.1 (1988), pp. 103–111 (cit. on pp. 9, 19, 38).
- [105]ABM Musa and Jakob Eriksson. “Tracking unmodified smartphones using wi-fi monitors”. In: *Proceedings of the 10th ACM conference on embedded network sensor systems*. ACM. 2012, pp. 281–294 (cit. on pp. 64, 71, 75, 83).
- [106]Kazuaki Nakamura, Tsukasa Ono, and Noboru Babaguchi. “Detection of groups in crowd considering their activity state”. In: *Proceedings of IEEE ICPR, 2016*.

- [107]Rajalakshmi Nandakumar, Krishna Kant Chintalapudi, and Venkata N Padmanabhan. “Centaur: locating devices in an office environment”. In: *Proceedings of the 18th annual international conference on Mobile computing and networking*. ACM. 2012, pp. 281–292.
- [108]Choon Boon Ng, Yong Haur Tay, and Bok Min Goi. “Vision-based human gender recognition: A survey”. In: *arXiv preprint arXiv:1204.1611* (2012) (cit. on pp. 14, 37).
- [109]Daniel Olguin Olguin, Peter A Gloor, and Alex Sandy Pentland. “Capturing individual and group behavior with wearable sensors”. In: *Proceedings of AAAI, 2009* (cit. on pp. 15, 42, 60).
- [110]Daniel Olguín Olguín, Benjamin N Waber, Taemie Kim, et al. “Sensible organizations: Technology and methodology for automatically measuring organizational behavior”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.1 (2009), pp. 43–55 (cit. on pp. 15, 60).
- [111]J Gama Oliveira and A Vazquez. “Impact of interactions on human dynamics”. In: *Physica A: Statistical Mechanics and its Applications* 388.2-3 (2009), pp. 187–192 (cit. on p. 13).
- [112]Alex Sandy Pentland. “The data-driven society”. In: *Scientific American* 309.4 (2013), pp. 78–83 (cit. on p. 4).
- [113]Thilo Pfau, Daniel PW Ellis, and Andreas Stolcke. “Multispeaker speech activity detection for the ICSI meeting recorder”. In: *Automatic Speech Recognition and Understanding, 2001. ASRU’01. IEEE Workshop on*. IEEE. 2001, pp. 107–110 (cit. on p. 22).
- [114]Thor S Prentow, Antonio J Ruiz-Ruiz, Henrik Blunck, Allan Stisen, and Mikkel B Kjærgaard. “Spatio-temporal facility utilization analysis from exhaustive wifi monitoring”. In: *Pervasive and Mobile Computing* 16 (2015), pp. 305–316 (cit. on pp. 15, 60).
- [115]*Public WiFi Usage Survey*. [https://www.idtheftcenter.org/images/surveys\\_studies/PublicWiFiUsageSurvey.pdf](https://www.idtheftcenter.org/images/surveys_studies/PublicWiFiUsageSurvey.pdf). Accessed: 2017-08-24 (cit. on p. 43).
- [116]Priya Raghur and Ana Valenzuela. “Male—Female Dynamics in Groups: A Field Study of The Weakest Link”. In: *Small Group Research* 41.1 (2010), pp. 41–70 (cit. on p. 18).
- [117]Hidayah Rahmalan, Mark S Nixon, and John N Carter. “On crowd density estimation for surveillance”. In: (2006) (cit. on p. 14).
- [118]Anshul Rai, Krishna Kant Chintalapudi, Venkata N Padmanabhan, and Rijurekha Sen. “Zee: zero-effort crowdsourcing for indoor localization”. In: *Proceedings of the 18th annual international conference on Mobile computing and networking*. ACM. 2012, pp. 293–304.

- [119]Kumar Rakesh, Subhangi Dutta, and Kumara Shama. “Gender Recognition using speech processing techniques in LABVIEW”. In: *International Journal of Advances in Engineering & Technology* 1.2 (2011), pp. 51–63 (cit. on pp. 14, 18, 37).
- [120]Rebecca K Ratner and Rebecca W Hamilton. “Inhibited from bowling alone”. In: *Journal of Consumer Research* 42.2 (2015), pp. 266–283 (cit. on pp. 43, 62).
- [122]Deirdre Reznik. “Gender in interruptive turns at talk-in-interaction”. In: *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics* 4.3 (2004) (cit. on pp. 9, 19, 26, 38).
- [123]Cecilia L Ridgeway. *Gender, interaction, and inequality*. Springer, 1992 (cit. on p. 19).
- [124]Mikel Rodriguez, Josef Sivic, Ivan Laptev, and Jean-Yves Audibert. “Data-driven crowd analysis in videos”. In: *ICCV 2011-13th International Conference on Computer Vision*. IEEE. 2011, pp. 1235–1242 (cit. on p. 14).
- [125]Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. “A review of feature selection techniques in bioinformatics”. In: *bioinformatics* 23.19 (2007), pp. 2507–2517 (cit. on p. 34).
- [126]Yvan Saeys, Thomas Abeel, and Yves Van de Peer. “Robust feature selection using ensemble feature selection techniques”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2008, pp. 313–325 (cit. on p. 20).
- [127]Md Sabbir Rahman Sakib, Md Abdul Quyum, Karl Andersson, Kåre Synnes, and Ulf Körner. “Improving Wi-Fi based indoor positioning using particle filter based on signal strength”. In: *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2014 IEEE Ninth International Conference on*. IEEE. 2014, pp. 1–6.
- [128]Lorenz Schauer, Martin Werner, and Philipp Marcus. “Estimating crowd densities and pedestrian flows using wi-fi and bluetooth”. In: *Proceedings of EAI MOBIQUITOUS, 2014* (cit. on pp. 15, 60).
- [129]T Schoberl. “Combined Monte Carlo simulation and ray tracing method of indoor radio propagation channel”. In: *Microwave Symposium Digest, 1995., IEEE MTT-S International*. IEEE. 1995, pp. 1379–1382 (cit. on p. 99).
- [130]Borja Seijo-Pardo, Iago Porto-Díaz, Verónica Bolón-Canedo, and Amparo Alonso-Betanzos. “Ensemble feature selection: homogeneous and heterogeneous approaches”. In: *Knowledge-Based Systems* 118 (2017), pp. 124–139 (cit. on pp. 20, 31).

- [131]Rijurekha Sen, Youngki Lee, Kasthuri Jayarajah, Archan Misra, and Rajesh Krishna Balan. “Grumon: Fast and accurate group monitoring for heterogeneous urban spaces”. In: *Proceedings of ACM SenSys, 2014* (cit. on pp. 15, 41, 42, 56, 60).
- [132]Souvik Sen, Romit Roy Choudhury, and Srihari Nelakuditi. “SpinLoc: Spin once to know your location”. In: *Proceedings of Workshop on ACM MobiSys, 2012* (cit. on p. 64).
- [133]Vinay Seshadri, Gergely V Zaruba, and Manfred Huber. “A bayesian sampling approach to in-door localization of wireless devices using received signal strength indication”. In: *Third IEEE International Conference on Pervasive Computing and Communications*. IEEE. 2005, pp. 75–84.
- [134]Chhavi Sharma, Yew Fai Wong, Wee-Seng Soh, and Wai-Choong Wong. “Access point placement for fingerprint-based localization”. In: *Communication Systems (ICCS), 2010 IEEE International Conference on*. IEEE. 2010, pp. 238–243 (cit. on p. 100).
- [135]Shih-Lung Shaw and Daniel Sui. “Introduction: Human Dynamics in Perspective”. In: *Human Dynamics Research in Smart and Connected Communities*. Ed. by Shih-Lung Shaw and Daniel Sui. Cham: Springer International Publishing, 2018, pp. 1–11 (cit. on pp. 1, 3, 5).
- [136]Shih-Lung Shaw, Ming-Hsiang Tsou, and Xinyue Ye. “Human dynamics in the mobile and big data era”. In: *International Journal of Geographical Information Science* 30.9 (2016), pp. 1687–1693 (cit. on pp. 1, 2, 4).
- [137]Guobin Shen, Zhuo Chen, Peichao Zhang, Thomas Moscibroda, and Yongguang Zhang. “Walkie-Markie: indoor pathway mapping made easy”. In: *Proceedings of the 10th USENIX conference on Networked Systems Design and Implementation*. USENIX Association. 2013, pp. 85–98 (cit. on p. 50).
- [138]Jiaxing Shen, Jiannong Cao, Xuefeng Liu, and Chisheng Zhang. “DMAD: Data-Driven Measuring of Wi-Fi Access Point Deployment in Urban Spaces”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 9.1 (2017), p. 11 (cit. on p. 42).
- [139]Jiaxing Shen, Jiannong Cao, Xuefeng Liu, Jiaqi Wen, and Yuanyi Chen. “Feature-Based Room-Level Localization of Unmodified Smartphones”. In: *Smart City 360°*. Springer. 2016, pp. 125–136 (cit. on pp. 15, 50, 60, 64, 75, 83).
- [140]Elizabeth Shriberg. “Spontaneous speech: How people really talk and why engineers should care”. In: *Ninth European Conference on Speech Communication and Technology*. 2005 (cit. on p. 18).
- [141]Adrian P Simpson. “Phonetic differences between male and female speech”. In: *Language and Linguistics Compass* 3.2 (2009), pp. 621–640 (cit. on pp. 14, 18, 37).



- [142] Francesco Solera, Simone Calderara, and Rita Cucchiara. “Socially constrained structural learning for groups detection in crowd”. In: *IEEE transactions on pattern analysis and machine intelligence* 38.5 (2016), pp. 995–1008 (cit. on pp. 14, 42, 56, 60).
- [143] Li Sun, Ramanujan K Sheshadri, Wei Zheng, and Dimitrios Koutsonikolas. “Modeling WiFi active power/energy consumption in smartphones”. In: *Distributed Computing Systems (ICDCS), 2014 IEEE 34th International Conference on*. IEEE. 2014, pp. 41–51 (cit. on p. 62).
- [144] Latanya Sweeney. “k-anonymity: A model for protecting privacy”. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 557–570 (cit. on p. 4).
- [145] Lu-An Tang, Yu Zheng, Jing Yuan, et al. “A framework of traveling companion discovery on trajectory data streams”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 5.1 (2013), p. 3.
- [146] Deborah Tannen and Deborah Tannen. *You just don't understand*. Simon & Schuster Audio, 1991 (cit. on pp. 27, 38).
- [147] Louis Ten Bosch, Nelleke Oostdijk, and Jan Peter De Ruiter. “Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues”. In: *International Conference on Text, Speech and Dialogue*. Springer. 2004, pp. 563–570 (cit. on p. 18).
- [148] Robert Tibshirani. “Regression shrinkage and selection via the lasso: a retrospective”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.3 (2011), pp. 273–282.
- [149] Pamela S Tolbert, Mary E Graham, and Alice O Andrews. “Group gender composition and work group relations: Theories, evidence, and issues”. In: (1999) (cit. on p. 18).
- [150] Ming-Hsiang Tsou. “Research challenges and opportunities in mapping social media and Big Data”. In: *Cartography and Geographic Information Science* 42.sup1 (2015), pp. 70–74 (cit. on pp. 5, 14).
- [151] Stijn Marinus Van Dongen. “Graph clustering by flow simulation”. In: (2001).
- [152] Geert Vanderhulst, Afra Mashhadi, Marzieh Dashti, and Fahim Kawsar. “Detecting human encounters from wifi radio signals”. In: *Proceedings of ACM MUM, 2015* (cit. on p. 49).
- [153] Ivan Vilovic, Niksa Burum, and Zvonimir Sipus. “Ant colony approach in optimization of base station position”. In: *2009 3rd European Conference on Antennas and Propagation*. IEEE. 2009, pp. 2882–2886 (cit. on p. 99).
- [154] Michail Vlachos, George Kollios, and Dimitrios Gunopulos. “Discovering similar multidimensional trajectories”. In: *Data Engineering, 2002. Proceedings. 18th International Conference on*. IEEE. 2002, pp. 673–684.

- [155] Soroush Vosoughi, Deb Roy, and Sinan Aral. “The spread of true and false news online”. In: *Science* 359.6380 (2018), pp. 1146–1151 (cit. on p. 13).
- [156] Chen-Shu Wang and Yi-Dung Chen. “Base station deployment with capacity and coverage in WCDMA systems using genetic algorithm at different height”. In: *Genetic and Evolutionary Computing (ICGEC), 2012 Sixth International Conference on*. IEEE. 2012, pp. 546–549 (cit. on p. 100).
- [157] Chen-Shu Wang and Li-Fang Kao. “The optimal deployment of Wi-Fi wireless access points using the genetic algorithm”. In: *Genetic and Evolutionary Computing (ICGEC), 2012 Sixth International Conference on*. IEEE. 2012, pp. 542–545 (cit. on p. 100).
- [158] Jessica JunLin Wang and Sameer Singh. “Video analysis of human dynamics—A survey”. In: *Real-time imaging* 9.5 (2003), pp. 321–346.
- [159] Ji zeng Wang and Hongxu Jin. “Improvement on APIT localization algorithms for wireless sensor networks”. In: *Networks Security, Wireless Communications and Trusted Computing, 2009. NSWCTC’09. International Conference on*. Vol. 1. IEEE. 2009, pp. 719–723 (cit. on p. 83).
- [160] Tian Wang, Weijia Jia, Guoliang Xing, and Minming Li. “Exploiting statistical mobility models for efficient Wi-Fi deployment”. In: *IEEE Transactions on Vehicular Technology* 62.1 (2013), pp. 360–373.
- [161] Yan Wang, Jie Yang, Yingying Chen, et al. “Tracking human queues using single-point signal monitoring”. In: *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. ACM. 2014, pp. 42–54 (cit. on pp. 64, 83).
- [162] Jens Weppner and Paul Lukowicz. “Bluetooth based collaborative crowd density estimation with mobile phones”. In: *Pervasive computing and communications (PerCom), 2013 IEEE international conference on*. IEEE. 2013, pp. 193–200 (cit. on p. 14).
- [163] Lyndon While and Chris McDonald. “Optimising Wi-Fi Installations Using a Multi-Objective Evolutionary Algorithm”. In: *Asia-Pacific Conference on Simulated Evolution and Learning*. Springer. 2014, pp. 747–759 (cit. on pp. 64, 100).
- [164] Stuart N Wrigley, Guy J Brown, Vincent Wan, and Steve Renals. “Speech and crosstalk detection in multichannel audio”. In: *IEEE Transactions on speech and audio processing* 13.1 (2005), pp. 84–91 (cit. on p. 22).
- [165] Chao-Lin Wu, Li-Chen Fu, and Feng-Li Lian. “WLAN location determination in e-home via support vector classification”. In: *Networking, sensing and control, 2004 IEEE international conference on*. Vol. 2. IEEE. 2004, pp. 1026–1031 (cit. on p. 83).

- [166]Ke Wu and Donald G Childers. “Gender recognition from speech. Part I: Coarse analysis”. In: *The journal of the Acoustical society of America* 90.4 (1991), pp. 1828–1840.
- [167]Lynn Wu, Benjamin Waber, Sinan Aral, Erik Brynjolfsson, and Alex Pentland. “Mining face-to-face interaction networks using sociometric badges: Predicting productivity in an it configuration task”. In: (2008) (cit. on p. 18).
- [168]Danny Wyatt, Tanzeem Choudhury, and Henry Kautz. “Capturing spontaneous conversation and social dynamics: A privacy-sensitive data collection effort”. In: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. Vol. 4. IEEE. 2007, pp. IV–213 (cit. on pp. 4, 8, 14, 18, 22).
- [169]Wei Xi, Jizhong Zhao, Xiang-Yang Li, et al. “Electronic frog eye: Counting crowd using wifi”. In: *Proceedings of IEEE INFOCOM, 2014* (cit. on pp. 15, 60).
- [170]Han Xu, Zheng Yang, Zimu Zhou, et al. “Enhancing wifi-based localization with visual clues”. In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM. 2015, pp. 963–974.
- [171]Youqing Xu. *Gender differences in mixed-sex conversations: A study of interruptions*. 2009 (cit. on p. 19).
- [172]Byoung-Kee Yi, HV Jagadish, and Christos Faloutsos. “Efficient retrieval of similar time sequences under time warping”. In: *Data Engineering, 1998. Proceedings., 14th International Conference on*. IEEE. 1998, pp. 201–208.
- [173]Moustafa Youssef and Ashok Agrawala. “The Horus WLAN location determination system”. In: *Proceedings of the 3rd international conference on Mobile systems, applications, and services*. ACM. 2005, pp. 205–218 (cit. on pp. 83, 84).
- [174]Na Yu and Qi Han. “Grace: Recognition of proximity-based intentional groups using collaborative mobile devices”. In: *2014 IEEE 11th International Conference on Mobile Ad Hoc and Sensor Systems*. IEEE. 2014, pp. 10–18.
- [175]May Yuan. “Human dynamics in space and time: A brief history and a view forward”. In: *Transactions in GIS* 22.4 (2018), pp. 900–912 (cit. on pp. 2, 13).
- [176]Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, et al. “Inferring international and internal migration patterns from twitter data”. In: *Proceedings of the 23rd International Conference on World Wide Web*. ACM. 2014, pp. 439–444 (cit. on pp. 13, 15).
- [177]Jie Zhang, Kuang Du, Ruihua Cheng, et al. “Reliable Gender Prediction Based on Users’ Video Viewing Behavior”. In: *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE. 2016, pp. 649–658 (cit. on p. 37).

- [178]Xiaoquan Zhao and Walter Gantz. “Disruptive and cooperative interruptions in prime-time television fiction: The role of gender, status, and topic”. In: *Journal of Communication* 53.2 (2003), pp. 347–362 (cit. on pp. 19, 27, 38).
- [179]Li Zheng, Reipeng Ning, Lin Li, et al. “Gender Differences in Behavioral and Neural Responses to Unfairness Under Social Pressure”. In: *Scientific reports* 7.1 (2017), p. 13498 (cit. on p. 18).
- [180]Tao Zhou, Hoang Anh-Tuan Kiet, Beom Jun Kim, B-H Wang, and Petter Holme. “Role of activity in human dynamics”. In: *EPL (Europhysics Letters)* 82.2 (2008), p. 28002 (cit. on p. 13).
- [181]Don H Zimmermann and Candace West. “Sex roles, interruptions and silences in conversation”. In: *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4* (1996), pp. 211–236 (cit. on pp. 9, 19, 27, 38).
- [182]Seyedjamal Zolhavarieh, Saeed Aghabozorgi, and Ying Wah Teh. “A review of subsequence time series clustering”. In: *The Scientific World Journal* 2014 (2014) (cit. on p. 86).